2016-06-01

Using Perceptually Grounded Semantic Models to Autonomously Convey Meaning Through Visual Art

Derrall L. Heath
Brigham Young University

Using Perceptually Grounded Semantic Models to Autonomously Convey Meaning Through Visual Art

Derrall L. Heath

A dissertation submitted to the faculty of
Brigham Young University
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

Dan Ventura, Chair
Seth Holladay
Mike Jones
Dan Olsen
David Embley

Department of Computer Science

Brigham Young University

June 2016

ABSTRACT

Using Perceptually Grounded Semantic Models to Autonomously Convey Meaning Through Visual Art

Derrall L. Heath
Department of Computer Science, BYU
Doctor of Philosophy

Developing advanced semantic models is important in building computational systems that can not only understand language but also convey ideas and concepts to others. Semantic models can allow a creative image-producing-agent to autonomously produce artifacts that communicate an intended meaning. This notion of communicating meaning through art is often considered a necessary part of eliciting an aesthetic experience in the viewer and can thus enhance the (perceived) creativity of the agent. Computational creativity, a subfield of artificial intelligence, deals with designing computational systems and algorithms that either automatically create original and functional products, or that augment the ability of humans to do so. We present work on DARCI (Digital ARtist Communicating Intention), a system designed to autonomously produce original images that convey meaning. In order for DARCI to automatically express meaning through the art it creates, it must have its own semantic model that is perceptually grounded with visual capabilities.

The work presented here focuses on designing, building, and incorporating advanced semantic and perceptual models into the DARCI system. These semantic models give DARCI a better understanding of the world and enable it to be more autonomous, to better evaluate its own artifacts, and to create artifacts with intention. Through designing, implementing, and studying DARCI, we have developed evaluation methods, models, frameworks, and theories related to the creative process that can be generalized to other domains outside of visual art. Our work on DARCI has even influenced the visual art community through several collaborative efforts, art galleries, and exhibits. We show that the DARCI system is successful at autonomously producing original art that is meaningful to human viewers. We also discuss insights that our efforts have contributed to the field of computational creativity.

Keywords: computational creativity, semantics, perception, visual art, artificial neural networks, deep learning

ACKNOWLEDGMENTS

First of all, I would like to give a big thank you to my advisor, Dr. Dan Ventura. His undaunted optimism and enthusiasm is contagious, and he has been both a mentor and friend. I have appreciated his continual support, guidance, and efforts on my behalf. It's been a blast! I would also like to thank the rest of my committee members for taking time to work with me and provide valuable feedback. Finally, I would like to thank my collaborator and friend, David Norton, whose own research on DARCI helped lay the foundation for my own contributions.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

Computers have become a crucial part of society, from business and industry, to research and government, to our personal lives. With such thorough integration comes the need for more intelligent programs and algorithms that can effectively and efficiently meet the demands of society. Thus, advances in artificial intelligence—the development of computer systems able to perform tasks that normally require human intelligence—should be a high priority. Creativity is an aspect of intelligence that is highly valued because it allows us (humans) to innovate and think outside the box, and it allows us to accomplish tasks perhaps otherwise thought to be impossible. Computational creativity, which is a sub-field of artificial intelligence, is "the philosophy, science, and engineering of computational systems which, by taking on particular responsibilities, exhibit behaviors that unbiased observers would deem to be creative" [28]. Computationally creative systems are applicable in many domains, including: mathematical theories [26], video games [32], art [25, 109], music [33, 83, 115, 175], culinary recipes [116, 160], metaphors [161], narratives [62, 177], poetry [163], humor [13], and many others.

Although there are no universally agreed upon definitions of creativity, most definitions include ideas such as novelty and quality [14]. In the domain of visual art, novelty is often straightforward (the art is new and original). However, what does it mean for a piece of art to have quality? Some would say that quality means art that is simply pleasing to look at. Others may say quality means art that requires a large amount of skill or effort to produce. A very general notion of quality is that art should elicit some type of aesthetic experience in the viewer. Csíkzentmihályi says, "...the aesthetic experience occurs when information coming from the artwork interacts with information

already stored in the viewer's mind..." [36]. This transfer of information implies that quality art communicates some kind of meaning. It implies that the artist created the art with intent and purpose. This idea of intent and purpose suggests that a creative system must be able to appreciate (i.e., evaluate) its own creations [24]. There must be a self-evaluation component that governs and directs the creative process so that when the system is exploring new ideas, it knows which ones are worth pursuing.

If a digital artist intends to create art that conveys an idea, it must first know what that idea is. It must have an understanding of it. It must have its own internal representation of that idea in order to then express that idea through art. In cognitive psychology, the term *semantic memory* refers to the memory of meaning and other concept-based knowledge that allows people to consciously recall general information about the world [152]. An artificial artist would need its own computational model of semantic memory in order to intentionally create art that conveys meaning. Conceptual knowledge is clearly tied to language, where words represent concepts (i.e., they have meaning). The question of what gives words (or concepts) meaning has been studied for years across multiple disciplines. This topic has been approached from a cognitive psychology perspective (already mentioned as semantic memory) [152], from a linguistic perspective (referred to as lexical semantics, or cognitive semantics) [136], and from a computer science perspective (computational semantics) [19].

Cognitive psychologists and linguists typically try to build *semantic models* (computational approximations of some or all aspects of semantics) in order to understand how language and the human mind work. Computer scientists, on the other hand, typically build semantic models in order to solve problems. Computational semantics has its roots in natural language processing and in information retrieval, where the goal is to retrieve documents that match a search query. Developing semantic models help search algorithms do more than just a keyword match. For example, a user could type in the word 'dog' and the system can understand that documents containing words like 'puppy' and 'canine' could also be relevant to the search. In NLP, semantic models are potentially useful for many tasks including word sense disambiguation, semantic parsing, text summarization,

machine translation, topic modeling, text classification, and speech recognition. Other applicable areas include data mining, common sense reasoning, question answering, content-based image/video retrieval, and information extraction.

A computational system that autonomously creates visual art needs a semantic model that allows it to understand language in ways that influence the art it produces. For example, suppose a system is trying to create an image about the concept 'war'. A semantic model will enable the system to understand that 'war' is associated with many other concepts such as 'army', 'battle', 'bloody', 'scary', etc. These other concepts give context to the meaning of 'war' and could be incorporated into the image. Ultimately, however, a semantic model must be grounded in perception in order for the system to understand its own artwork and to truly represent meaning. With visual art, semantic models need to learn associations between words and image qualities. For example, the model could learn that 'happy' images are often bright and colorful, while 'scary' images tend to be dark with jagged lines and edges. The system could then render an image about 'war' in a 'scary' manner (i.e., dark with jagged lines and edges) because 'scary' is associated with 'war'. The final image will in some way artistically convey the meaning of 'war'.

In this dissertation, we argue that the ability to perceive is fundamental to the creative process and enables a system to learn its own semantic model. Even in humans, perception directly influences our ability to think and understand, and the better and more varied our perceptual abilities are, the more we are able to think about, imagine, and ultimately, create [7, 70]. Indeed, every memory and every thought we have is in terms of what we have experienced (perceived) in the past. Most psychologists agree that our perceptions (senses), our conceptual knowledge, and our memories make up our mental model and form the bases of creative tasks like imagination [10, 59]. It has also been suggested that the most creative and influential people are ones that literally perceive (and therefore think) differently [11]. Likewise, a creative computational system must have its own perceptual abilities that allow it to have experiences, to learn, and to ground its semantic model, which can then facilitate and influence the meaningful creation of artifacts.

1.1 The DARCI System

In this dissertation we present a computational system we have developed to explore computational creativity in the domain of visual art. This system, called DARCI (Digital ARtist Communicating Intention), is designed to autonomously create novel and interesting digital images that convey meaning to the viewer. Throughout this dissertation we have developed fundamental (domain independent) theories of how perceptual ability and semantic understanding influence and facilitate the creative process, and we use the DARCI system to experiment with and validate those theories. The DARCI system has been around for several years and previous work has already been done by collaborator David Norton; this dissertation builds off of that prior work. Specifically, the work presented here focuses on designing, building, and incorporating advanced semantic and perceptual models into the DARCI system. These advanced semantic models give DARCI a better understanding of the world and enable it to be more autonomous, to better evaluate its own artifacts, and to create artifacts with intention.

1.2 Background

This dissertation contributes to the area of computational creativity in the visual art domain as well as to computational models of semantics. We first provide background on the computational creativity literature and discuss work related to the visual art domain. We will then provide background for semantic models and talk about work related to semantic models in visual art. Finally, we give background on prior work done on the DARCI system, which will set the stage for the contributions presented in this dissertation.

1.2.1 Computational Creativity and Visual Art

The field of computational creativity begins with the ability to measure and attribute creativity in a computational system, which requires a discussion on what creativity actually means. Although there is no single agreed upon definition of creativity, most definitions involve the ideas of *originality*

and *quality*. Boden's definition of computational creativity is the one most often cited, which argues that a creative system can produce artifacts that are both novel and valuable [14]. In this case novelty means that the artifact produced is new and original, while valuable means the artifact is useful or meaningful in some way. Others have expanded on this notion of creativity including Ritchie's prospective metrics for measuring creativity [138] and Colton's creative tripod [24].

Ritchie discusses creativity in terms of the artifacts generated by a system and attributes creativity to a system if it can produce creative artifacts [138]. In addition to Boden's two criteria, Ritchie adds *typicality*, which is the extent to which an artifact is an example of the class in question. For example, in the visual arts domain, it is the extent to which the artifact is a visual depiction of something. Colton's creative tripod focuses on the creative process rather than the artifacts produced [24]. The notion of quality is characterized in terms of *skill*, or the capacity of a system to produce quality artifacts. The notion of novelty is characterized in terms of *imagination*, or the capacity of the system to produce original and meaningful (non-random) artifacts. Colton then adds an attribute called *appreciation*, or the ability of the system to recognize the quality and novelty of its own artifacts (which emphasizes process over product and the need for self-evaluation).

Many computational systems exist that are designed to produce visual art. In the field of computational creativity, some of the more prominent systems include Harold Cohen's AARON [109] and Simon Colton's Painting Fool [25]. Although these two systems are often cited and talked about in the computational creativity community, both authors have opted to hide the inner workings of the system from the public. For the Painting Fool, this obfuscation is a deliberate attempt to personify the system as part of Colton's experiments with the perception of creativity being a requirement for attribution of creativity. While details governing AARON and The Painting Fool are hidden, many other image generating systems have been openly described in the computational creativity community.

Many such systems center around evolutionary algorithms, due to the innate ability of evolution to yield unpredictable solutions to problems, as discussed by Gero [61]. The use of evolutionary algorithms for generating visual art starts by first instantiating a random population

of possible entities. Each of of these entities is a digital encoding called a genotype, which can be decoded and transformed into an actual image (referred to as the phenotype). Each of the genotypes in the population are usually converted to their respective phenotypes and evaluated by a fitness function. This fitness function determines which genotypes get to pass on their traits to future generations of the evolutionary algorithm. Over multiple iterations (or generations), the population converges toward images that maximize the score from the fitness function. Because art is inherently subjective and difficult to analyze, the fitness function (i.e., evaluation step) is often left to human judgment in most systems that produce evolutionary art. Sims was one of the first to use evolutionary algorithms to produce visual art [148], while a more recent system is Secretan's popular online Picbreeder [145]. In these cases the system generates the art, but it is the human that evaluates and governs the evolutionary process.

Self-evaluation in Artificial Artists

The need to depend on human judgment in the evolutionary process not only is time consuming, but also fails to completely automate the creative process. Ideally, a creative system should be able to govern itself as it explores possible images by evaluating them on its own. There are several evolutionary art systems that incorporate their own fitness functions. Most of these systems extract quantifiable features from the images to evaluate their aesthetics. For example, one system looks at how closely the image color distribution matches the color distribution of high rated Flickr images [128]. Another system uses image compression to measure image complexity [105]. There is also a system that performs a geometric assessment of regions within an image [66].

More advanced methods use a dynamic fitness function that changes based on the evolutionary environment. For example, DiPaola and Gabora designed a system with a fitness function that utilizes two sets of features. The first is how closely the image in question matches a source image and the second is how well the image follows several quantifiable rules of aesthetics. When the evolutionary process becomes stagnant, the algorithm automatically adjusts the weight that it puts on each set of features [46]. The co-evolutionary model designed by Greenfield is another example

of a dynamic fitness function. In his model, the system co-evolves a population of images and a population of image filters together. The fitness functions for each are parameterized by how the two populations interact and thus change and fluctuate as if in an ecological environment [67].

Other systems attempt to directly model a human's judgment of images as the fitness function. Baluja *et al.* first attempted this by using an artificial neural network to model the aesthetic preferences of specific users. The entire pixel space of each image was the input to the neural network and for training data, Beluja used data collected from human evaluated evolutionary art systems [4]. Beluja reported little success because the input space for such a sparse amount of training data was enormous. Machado *et al.* incorporated a dynamic fitness function into their NEvAr system using an artificial neural network model that evaluates the novelty of the system's artifacts [106]. The model is trained to distinguish between famous paintings and images created by NEvAr itself. Instead of using the entire pixel space as done by Baluja, Machado used a finite set of features extracted from each image as inputs to the neural network. If the system recognized an image as its own style, it was rejected. Any new images created by the NEvAr system were then continually fed back into the training data so as to force the system to constantly evolve and change its own style.

None of these systems explicitly try to communicate an intended meaning to the viewer. By contrast, our goal is to take this idea of a dynamic fitness function and tie it to language in a way that allows the system to assess how well an image conveys a concept. To do this, the system needs some notion of semantics.

1.2.2 Semantic Models

The study of computational semantics is a broad field with many applications in areas such as information retrieval, machine translation, topic modeling, question answering, speech recognition, and text summarization. Many of these areas define the notion of semantics in different ways and build models specific to the individual tasks. In this dissertation, we will focus on the more general notion of semantics and more general purpose models.

The question of what gives words meaning has been debated for years; however, it is commonly agreed that a word, at least in part, is given meaning by how the word is associated with and used in conjunction with other words (i.e., its context) [51]. Many computational semantic models consist of building associations between words, essentially forming a large graph that is typically referred to as a semantic network or ontology. In cognitive psychology, meaning is primarily based on memory, or our ability to recall general information about the world (referred to as semantic memory or conceptual knowledge). Over the years, several computational models of semantic memory have been developed based on two primary theoretical frameworks [152]. The first is the hierarchical spreading-activation model of semantic memory described by Collins and Quillian [22]. The second is the prototype model based on the work of Eleanor Rosch [140].

In the hierarchical spreading-activation model, concepts are organized into a semantic network where concepts are connected together either hierarchically through *is-a* relationships (e.g., a robin *is-a* bird and a bird *is-an* animal), or connected through attribute relationships (e.g., a bird *can* fly and a bird *has* feathers). This model provides a parsimonious means of knowledge storage as well as a simple means of generalization. For example, the knowledge that all mammals have hair can be stored by connecting the hair node to the mammal node with a *has* link. Has hair will then generalize to any concept beneath the mammal node connected through the *is-a* relationship. Empirical studies show some limitations in the strict hierarchical structure of the model, and so later versions of the model allowed direct links between any pair of concepts, which used a spreading activation method to search the network for information [23].

In the prototype model, items can be categorized into concepts based on how similar they are to the prototypical example of each concept. For instance, a canary is more similar to a typical bird than it is to a typical dog and can thus be categorized as a bird. In theory, this model seems to match the way humans categorize objects [152]. However, this model does not address the issues of what features should be used to judge similarity and how the exemplars of each category should be chosen.

The biggest limitation of these early models is that they had to be built by hand, which at the time, made them impractical for use in real-world tasks. However, they did establish a pattern for future models. For example, WordNet is a large English language lexical database in which words are grouped together into synonym sets [55]. The noun synonym sets are then linked to other synonym sets hierarchically through *is-a* relationships (called hypernyms, or hyponyms), and include other attribute links (e.g., meronyms and antonyms). WordNet was created by language experts and has been used extensively in many semantic and language processing applications.

Because handmade models are difficult to construct and limited in their scope, other models have made use of crowd-sourcing to build larger semantic networks. Free association norms (FANs) are word associations gathered by asking hundreds of human volunteers to provide the first word that comes to mind when given a cue word [118]. FANs are able to capture many different types of word associations, including word co-ordination (pepper, salt), collocation (trash, can), super-ordination (insect, butterfly), synonymy (starving, hungry), and antonymy (good, bad). Cognitive psychologists use FANs to study memory and language, and they are thought to accurately reflect how people relate concepts and store lexical knowledge [118]. Furthermore, FANs are commonly thought to represent important aspects of meaning [40]. Their use in computational models, however, has been limited to simply being used in baseline metrics of comparison for other models.

ConceptNet is a semantic network created through various methods of crowd-sourcing [101]. Concepts are associated together using any type of link, including *is-a*, *has*, *can*, *used-for*, *has-property*, and so on. ConceptNet has been used successfully in many applications primarily in the domain of common sense reasoning [102]. However, despite the crowd-sourcing, ConceptNet is still limited in its vocabulary, and it does not provide any notion of relationship strength between concepts. Ultimately, the methods discussed so far are limited by having to depend on people to construct word associations. Researchers have been looking for ways to automatically build millions of associations without the need for human involvement.

Vector Space Models (VSMs) are a class of semantic models that use a large corpus to automatically infer semantic information based on patterns of word co-occurrences [159]. They

are also commonly referred to as Corpus-based Semantic Models (CSMs) [6], or Distributional Semantic Models (DSMs) [5]. These models incorporate the idea that similar words will occur in similar contexts and words that are often associated together will often co-occur close together [96]. VSMs have been successfully used on a variety of tasks such as information retrieval [143], multiple choice vocabulary tests [43], multiple choice synonym questions from the TOEFL test [137], and multiple choice analogy questions from the SAT test [158].

One of the most popular VSMs is Latent Semantic Analysis (LSA) [42]. LSA builds either a term $\times$ document or a term $\times$ term matrix from a corpus and then performs Singular Value Decomposition (SVD). SVD reduces the given large sparse matrix to a low-rank approximation of that matrix along with a set of vectors, each representing a word (as well as a set of vectors for each document). These vectors also represent points in "semantic space", and the closer a word's vector is to another in this space, the closer they are in meaning (and the stronger the association between words). For example, the words "dog" and "puppy" should have vectors that are close together in semantic space because the two words are often used in similar situations in a corpus. There are many variations of VSMs, and each is designed to succeed at specific tasks, but all of them use the same basic idea as LSA. More sophisticated models try to include sentence structure to provide additional semantic information [6]. Other models incorporate neural networks that learn the global context of words and can thus represent multiple variations in meaning [81]. The most state-of-the-art VSMs, however, try to build a single semantic model that can generalize and perform well on a variety of semantic tasks [5].

Semantic Models and Visual Art

Semantic computing is a multidisciplinary field that deals with matching the semantics of any kind of content (e.g., images, audio, text, services, hardware, etc) to language for the purposes of retrieving, manipulating, or creating the content [146]. As previously mentioned, our goal is to communicate meaning autonomously through visual art, which requires methods that associate semantics with visual/image information. Image annotation (describing an image with words) and

content-based image retrieval (finding images that match a word) are two major topics in computer vision whose goal is to directly associate images with words automatically. These types of computer vision problems are difficult and extensive research exists that attempts to solve them [103, 169]. These problems, in the general sense, are beyond the scope of this dissertation. We will, however, consider simplified problems in this space that are more directly related to visual art as well as methods for generating images based on natural language.

For example, a system that classifies digital art into five different genres (e.g., expressionism, cubism, impressionism, pop art, and realism) was developed by Zujovic *et al.* [180]. This system extracts both color and gray-scale features from images and then tests classification accuracy with several different machine learning algorithms. Wang Wei-ning *et al.* developed an image retrieval system that uses support vector machines to associate images with a set of 12 emotional word-antonym pairs (e.g., happy-sad, warm-cool, clear-fuzzy, etc) [172]. This system uses a unique set of image features customized for each word pair and is successful at improving retrieval results for the 12 specific emotional word pairs. Both of these systems could potentially be integrated with the image generation systems discussed previously to create images that match one of the genres or convey one of the emotional words.

Bruni *et al.* discovered a way to incorporate images (and other media) into vector space models (VSMs) [17]. They hypothesized that large datasets of labeled images contain semantic information, which can be extracted and combined with corpus data using techniques such as latent semantic analysis. Bruni built a word $\times$ word co-occurrence matrix as usual from the corpus but then added additional columns for visual words from the image dataset. Visual words are extracted from an image (using SIFT), and the labels for that image are the words that co-occur with the visual words from the image. At this point SVD can be applied to the word $\times$ word/visual word matrix as with any VSM. Bruni successfully showed that labeled image datasets can be combined with corpora to automatically build more complete semantic representations.

Semantic models have been incorporated into scene composition systems that automate the tedious task of placing hundreds of objects into a 3D scene. Ken Xu *et al.* created a system that uses

a hand-built semantic database to provide constraints on how objects can be oriented and placed in relation to other objects [176]. Bob Coyne and Richard Sproat developed a system called WordsEye that can take natural language descriptions of 3D scenes as input and automatically render them [34]. For example, the phrase "the black cat is on the chair" will automatically be rendered as a 3D scene with a black cat sitting on a chair. WordsEye incorporates WordNet, various NLP methods, and a finite set of spacial tags (i.e., object specific interpretations of spacial words like "on" or "under") to determine what, where and how to place objects into a scene.

The Story Picture Engine is a system developed by Dhiraj Joshi *et al.* that does automatic text illustration [85]. The system uses information contained in WordNet to determine important keywords from a piece of text. Joshi's system then uses a combination of integrated region mapping for image similarity, discrete state Markov chains, and mutual reinforcement-based ranking in order to select relevant images that best represent the piece of text. Xiaojin Zhu *et al.* created a text-to-picture pictorial communication system that automatically creates 2D collages that graphically convey the gist of the text [179]. The system first analyzes the text using the TextRank algorithm to retrieve keywords that also have high "picturability". The system then uses a clipart database and Google image search to select appropriate images that match the keywords. Finally, the system uses conditional random fields to learn how to spatially arrange the pictures in a layout that best communicates the gist of the text [64].

The last few years have shown a resurgence of deep neural networks (DNNs), especially for computer vision tasks, where they hold current records for several vision benchmarks [52, 153]. Deep learning has the potential to significantly improve visually creative systems as well. A key advantage of DNNs is that they are capable of learning their own image features, while the visual art systems described above all rely on manually engineered features. DNNs have already been used to generate images directly [45, 68, 97]. One particular method, called *gradient ascent* [147], works by essentially using the DNN in reverse. The trained network starts with a random noise image and tries to maximize the activation of the output node corresponding to the desired image class. The network then backpropagates the error into the image itself (keeping the network weights

unchanged) and the image is slightly modified at each iteration to look more and more like the desired class.

Although these systems generate images with the purpose of expressing some meaning, the images produced are not necessarily intended to be considered art, or the systems are merely tools for an animator. In the case of deep learning, it has not yet gained traction in the field of computational creativity. However, in this dissertation we argue for their use in creative systems and incorporate several deep learning models into DARCI.

1.2.3 DARCI Background

The DARCI system was designed to autonomously produce visual art that conveys meaning to the viewer, and the system subscribes to several of the philosophies of creativity described in Section 1.2.1. Specifically, DARCI's autonomous creation process is designed to adhere to Colton's creative tripod [24] and to produce visual art that satisfies Ritchie's essential properties of creativity [138].

The first leg of Colton's tripod, appreciation, is the ability of the system to appreciate its own artifacts (i.e., self-evaluation). Thus an important step was to design a component for DARCI that can evaluate meaning in images to some degree. DARCI's vocabulary was limited to adjectives and thousands of training images were gathered that have been human-labeled with the adjectives that they convey. Multiple neural networks (one for each adjective) were then trained to evaluate how much an image conveys a particular adjective [121]. In order to train these neural networks, relevant image features were discovered to extract from the images. These features had to be general across multiple adjectives and be computationally tractable. Through several iterations of experiments, 51 low-level features were chosen that deal with color, lighting, texture, and basic shapes [127].

With the appreciation component in place, an image rendering component was developed that uses an evolutionary algorithm to take an existing source image and discovers combinations of Photoshop-like features that artistically modifies the image to be more like a particular adjective [123]. This genetic algorithm incorporates Colton's two remaining legs of the tripod, skill

and imagination, by giving DARCI the ability to artistically modify an image (skill), and by allowing DARCI to explore and discover novel ways to modify an image to express an adjective (imagination). The evolutionary algorithm is directed by a fitness function, which is DARCI's appreciation component that evaluates each generation of images to determine which ones best convey a particular adjective. Through several iterations of DARCI, this genetic algorithm has been improved, refined, and evaluated extensively [125, 126].

Additionally, prior work on DARCI has also been focused on evaluation, i.e., how does one tell if DARCI is being successful in creating meaningful art? Thus, much research efforts have focused on designing experiments, human surveys, and metrics to measure DARCI's progress [123, 124, 126]. Prior to this dissertation, DARCI's capabilities consisted of artistically modifying a pre-existing source image to convey a list of adjectives, both provided by the user. In this dissertation, we extend the DARCI system with more advanced semantic and perceptual abilities that allow it to compose its own source image, to discover its own inspiration, and to create more sophisticated art that conveys concepts beyond adjectives.

1.3 Contributions and Summary of Dissertation

There are many computational systems designed to generate visual art in some capacity and several of them have produced beautiful and intricate artwork. However, DARCI is the only system that can, in full autonomy, create visual art that explicitly conveys meaning. It is the only system of its kind with a sophisticated cognitive model that gives it advanced perceptual abilities, gives it broad semantic understanding, and that initiates and drives its creative process. Through designing, implementing, and studying DARCI, we have developed evaluation methods, models, frameworks, and theories related to the creative process that can be generalized to other domains outside of visual art. Our work on DARCI has even influenced the visual art community through several collaborative efforts, art galleries, and exhibits. Here we summarize the chapters that discuss each contribution in detail.

In Chapter 2 we outline this dissertation's first version of the DARCI system in detail[1]. In this iteration, DARCI takes a source image and an adjective as input and artistically modifies the source image to reflect the adjective. We specifically emphasize the ability of the system to modify images in a way that is consistent with the semantic relationships among the various adjectives. For example, rendering an image for 'scary' should look more similar to rendering it for 'creepy' than to rendering it for 'happy' because 'scary' and 'creepy' are more similar in meaning than 'scary' and 'happy'. In this chapter, we also introduce image clustering techniques for evaluating how well DARCI's images reflect the semantic relationships of their intended adjectives, which we use again in Chapter 6. We show that DARCI is successful at generating images that convey their intended meaning in ways consistent with semantic relationships.

In Chapter 3 we begin to develop a more advanced semantic model for DARCI based on word associations[2]. Words are given meaning, in part, by their associations with other words. Thus, by learning how words relate to other words, a system can perform advanced semantic tasks such as answering word analogy questions, matching similar sentences, or playing online word games with humans. We present a novel semantic model that combines word associations automatically learned through written text, with word associations learned directly from humans. We evaluate this model using an online word guessing game, similar to *Catch Phrase*, and show that our model performs better than other pre-existing models.

In Chapter 4 we incorporate the semantic model from Chapter 3 into the DARCI system[3]. This semantic model allows DARCI to generate novel images from scratch (no need for a source image) and to generate images for any concept (not just adjectives). We input into the system any arbitrary concept (e.g., 'war'), and the system composes its own source image by retrieving semantically related words from the semantic model, finding corresponding simple iconic images

[1]Derrall Heath, David Norton, and Dan Ventura, Conveying Semantics Through Visual Metaphor, *ACM Transactions on Intelligent Systems and Technology*, 5(2):31, 2014

[2]Derrall Heath, David Norton, Eric Ringger, and Dan Ventura, Semantic Models as a Combination of Free Association Norms and Corpus-based Correlations, *Proceedings of the 7th IEEE International Conference on Semantic Computing*, pp. 48-55, 2013

[3]Derrall Heath, David Norton, and Dan Ventura, Autonomously Communicating Conceptual Knowledge Through Visual Art, *Proceedings of the 4th International Conference on Computational Creativity*, pp. 97-104, 2013

(representing those related words), then arranging those icons on the image as a collage. The resulting source image is then rendered artistically based on the most semantically related adjective. We use human surveys to evaluate the resulting images and show that DARCI can successfully generate images that convey the meaning of concepts beyond adjectives. Another important contribution is that we show that images that more accurately convey the intended concept are more often liked and considered creative by human viewers. This reinforces the need for meaning and intentionality for attributing creativity to a computational system.

Chapter 5 outlines a computational framework to facilitate imagination using vector space models and associative memory models[4]. The idea is that the meaning of concepts can be partially represented through vector space models (which encode semantic structure in high dimensional vectors). Various perceptual domains (like visual art, music, recipes, etc) can then be associated with these word vectors through associative memory models. Novel artifacts can then be 'imagined' and even generated by taking advantage of the semantic structure encoded in the word vectors. For example, combining the word vectors for 'happy' and 'scary' could produce some novel blending of those two concepts when the associative memory models are run in reverse. We developed this framework based on cognitive theories of imagination in humans and demonstrate an initial prototype.

In Chapter 6 we apply the framework from Chapter 5 to DARCI by incorporating a vector space model and replacing the multiple adjective neural networks with a single model that learns to associate images with entire adjective word vectors[5]. This allows DARCI to take advantage of the semantic structure between words and render images according to adjectives on which it was never explicitly trained. For example, DARCI could be trained on 'scary' and 'dark' images, but not 'creepy' images. DARCI could then "imagine" what a 'creepy' image would look like because 'creepy' is similar in meaning to 'scary' and 'dark'. Even higher level concepts (e.g., 'love',

[4]Derrall Heath, Aaron Dennis, and Dan Ventura, Imagining Imagination: A Computational Framework Using Associative Memory Models and Vector Space Models, *Proceedings of the 6th International Conference on Computational Creativity*, pp. 244–251, 2015

[5]Derrall Heath and Dan Ventura, Creating Images by Learning Image Semantics Using Vector Space Models, *Proceedings of The Thirtieth AAAI Conference on Artificial Intelligence*, 2016

'freedom') can be partially conveyed through the images DARCI renders. We show that DARCI is successfully able to "imagine" the visual qualities of adjectives that it has never previously seen.

In Chapter 7 we draw upon cognitive psychology, neuroscience, and art history to develop a theory regarding the role that perception plays in the creative process[6]. We explore how perception has influenced human thought, imagination, and creativity and discuss how state-of-the-art perceptual models (e.g., deep learning) can be used for imaginative and creative tasks. We demonstrate several methods of generating novel images using deep neural networks and discuss their implications for future creative systems.

In Chapter 8 we incorporate several deep learning models into DARCI to provide the system with better perceptual abilities, which allow DARCI to create more semantically relevant and visually pleasing images[7]. These enhancements also allow DARCI to evaluate and analyze preexisting images in order to find inspiration for its art. We also give DARCI the capability of generating titles and explanations for the art it creates, which enables human viewers to better understand the creative decisions made by DARCI.

In Chapter 9 we outline several opportunities we have had to interact with and influence the art community in conjunction with David Norton's prior work on DARCI. In 2010 we collaborated with Brigham Young University's Visual Arts program, which culminated in an art exhibit, called *Fitness Function*, curated by DARCI in the Harris Fine Arts Center at BYU. In 2011 we participated in the Utah County Art Gallery: Fall Photography and Digital Art Show by having DARCI create two submissions, one of which won 2nd place in the Digital Art category. Later that year, we put on another version of *Fitness Function* in The High Museum of Art in Atlanta, Georgia, as part of the 2011 *ACM Conference on Creativity and Cognition*. This event was covered by Studio 360 on Public Radio International[8]. DARCI also participated in the *GECCO 2013 Evolutionary Art,*

[6]Derrall Heath and Dan Ventura, Before A Computer Can Draw, It Must First Learn To See, *Proceedings of The 7th International Conference on Computational Creativity*, to appear, 2016

[7]Derrall Heath, and Dan Ventura, Autonomously Conveying Inspiration In Visual Art Using Deep Neural Networks, *International Journal of Semantic Computing* , to submit, 2016

[8]http://www.wnyc.org/story/designing-computer-great-taste/

Design, and Creativity Competition held in Amsterdam and collaborated with Colton's The Painting Fool at a public event held in Paris in June 2013.

Chapter 2

Conveying Semantics Through Visual Metaphor[1]

Abstract

In the field of visual art, metaphor is a way to communicate meaning to the viewer. We present a computational system for communicating visual metaphor that can identify adjectives for describing an image based on a low-level visual feature representation of the image. We show that the system can use this visual-linguistic association to render source images that convey the meaning of adjectives in a way consistent with human understanding. Our conclusions are based on a detailed analysis of how the system's artifacts cluster, how these clusters correspond to the semantic relationships of adjectives as documented in WordNet, and how these clusters correspond to human opinion.

[1]Derrall Heath, David Norton, and Dan Ventura, Conveying Semantics Through Visual Metaphor, *ACM Transactions on Intelligent Systems and Technology*, 5(2):31, 2014

2.1 Introduction

Metaphor can be a powerful method for communication. Metaphors are usually associated with linguistics and research has been done in the AI community exploring the interpretation and generation of metaphor within the linguistics domain [164, 165]. However, metaphor also exists in the visual domain. Traditionally, the term *visual metaphor* is used to mean the representation of a target concept by a different source concept [56]. The source replaces the target in the image and can thus convey a new interpretation of the source concept or become a symbol for it. For example, the concept of a dove is often used to represent the target of peace in images. In this example, peace is an abstract concept and a dove is concrete; but, this does not always have to be the case. A visual metaphor can be as simple as using an icon of an envelope to indicate access to email on a computer terminal. Visual metaphor can be a powerful tool for conveying meaning in an image and has seen effective use in art and advertising [16, 87].

We have developed a computer system, called DARCI (Digital ARtist Communicating Intention), that is designed to autonomously create images that convey meaning. The value of visual metaphor in such a system is apparent and is one of the design goals for this system. However, automatic incorporation of visual metaphor in an image is nontrivial, requiring a rather mature understanding of language as well as the ability to visually represent and recognize a word in an image. As a step in this direction, DARCI currently only works with adjectives. Adjectives, more easily than nouns and verbs, can be used to describe an image based on general features such as color distribution, lighting, and repeating patterns. DARCI is presently designed to render an image to communicate a list of specified adjectives using a variety of filters and other image processing techniques. In effect, DARCI can express the meaning of adjectives visually in the images it produces. Since DARCI cannot interpret or render nouns, implementing the traditional view of conveying visual metaphor by replacing a target with a source is severely limited. In a broader sense however, DARCI can represent an adjective, the target, as a unique filter, an abstract source. This sense of visual metaphor does exist in human interaction. For example, the color red is often used to denote danger. For brevity, throughout this paper we will refer to this broader sense of

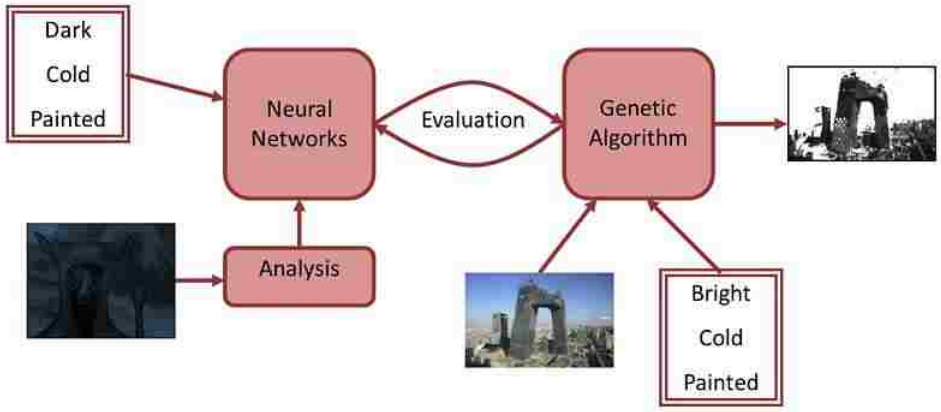


Figure 2.1: Overview of DARCI's artifact creation process.

visual metaphor as simply *visual metaphor*. Though limited, this is an important first step in the direction of conveying fully automated visual metaphor in the narrower, more traditional sense.

DARCI learns the meaning of adjectives, or rather adjective synonym sets—called synsets—by building associations between low-level visual features and the synsets themselves using a series of neural nets trained with human labeled data [121]. The process is augmented by taking advantage of the synset relationships found in WordNet [55]. DARCI renders images by applying various image filters to a source image (not to be confused with the source concept in visual metaphor). The filter settings are learned with an evolutionary mechanism that is governed by the visual-linguistic associations discovered by the neural nets [123]. Rather than the traditional approach to evolutionary art where humans evaluate the products of each generation by hand [139, 145, 148, 156], the fitness function is the output of these neural nets. Figure 2.1 outlines this process of creating artifacts. While there are other approaches using automated and dynamic fitness functions in evolutionary art [4, 46, 66, 67, 105, 128], none explore the communication of meaning as is done here with DARCI.

In previous work we have shown a degree of success in both labeling and generating images with respect to semantic content [121, 123]. However, analyzing the meaning contained in a particular rendering is an inherently subjective task. Thus, we are interested in additional and more

thorough approaches to analyzing the effectiveness of our system. Most of our previous evaluations of DARCI have required the opinion of human volunteers. Unfortunately, using human judges in an evaluation can be costly and highly variable. In this paper we explore an approach to evaluating the system that does not require human interaction. Instead this approach analyzes the way images produced by DARCI cluster and in turn how these cluster relationships correspond with synset relationships found in WordNet. In order to validate our approach, we compare an agglomerative clustering of DARCI's images by their features to how human volunteers' rankings of DARCI's images cluster. Each set of results provides evidence to support our claim that the system is capable of expressing visual metaphor in the artifacts it produces.

We begin this paper by providing a detailed overview of DARCI's algorithms as explained in previous work. We then introduce the methodology for our new evaluation metrics that are the focus of this paper. Next we describe the various experiments we ran with our evaluation methodology in mind. Here we also analyze both how effective our new evaluation metrics are and how effective DARCI is at using visual metaphor to communicate the meaning of adjectives in images. Finally we draw conclusions and suggest future direction for this work.

2.2 System Overview

In previous work we have described and evaluated the various algorithms that operate within DARCI [121, 123]. For convenience, in this section we reproduce those details necessary to appreciate the results of this paper.

2.2.1 Visuo-Linguistic Association

In order for DARCI to make associations between images and their meaning, we present the system with images that are labeled appropriately. For now, we have reduced descriptive labels exclusively to delineated lists of adjectives. Also, since the raw feature space of a typical image is intractable, we have selected 102 real-valued low-level image features to extract from each image [121]. These features were selected based on prior research in the area of image feature

extraction [39, 63, 90, 99, 170, 171] and include general measures of color, light, texture, and shape [121]. Three of these features, for example, are the average RGB (red, green, and blue) values of each pixel in an image. Because the magnitude of these 102 features varies dramatically from feature to feature, we have standardized the image features using a core set of training images (approximately 2000 diverse images). Throughout this paper, when we refer to image features, we are referring to the standardized values.

We use WordNet's [55] database of adjective synsets to give us a large set of descriptive labels. Even though our potential labels are restricted to a single lexical category, the complete set of WordNet adjective synsets can allow for images to be described by their emotional effects, most of their aesthetic qualities, many of their possible associations, and even, to some extent, by their subject.

To collect training data we have created a public website for training DARCI (http://darci.cs.byu.edu). From this website, users are presented with a random image and asked to provide adjectives that describe the image. When users input a word with multiple senses, they are presented with a list of the available senses, along with the WordNet gloss, and asked to select the most appropriate one. Additionally, for each image presented to the user, DARCI lists seven adjectives that it associates with the image. The user is then allowed to flag those labels that are not accurate. This creates strictly negative examples of those synsets, which will be important in the learning process.

Learning image to synset associations is a *multi-label classification* problem [157], meaning each image can be associated with more than one synset. To handle multi-label classification, we use a collection of artificial neural networks (ANNs) that we call appreciation networks. There is an appreciation network for each synset that has a sufficient amount of training data. For the results presented in this paper, that threshold is fifteen positive training instances. As we incrementally accumulate more data, new neural networks can be dynamically added to the collection to accommodate the new synsets. As of writing this paper, there are 211 appreciation networks. This means that DARCI essentially "knows" 211 synsets. The appreciation networks are trained using standard

backpropagation and output a single real value, between 0 and 1, indicating the degree to which a given image can be described by the networks' corresponding synset.

ANNs require a lot of training data to converge. Currently, of the 211 synsets known to DARCI, there are on average just over 33 positive data instances per synset. In order to enhance the amount of positive and negative data used to train the appreciation network, we use antonym relationships found in WordNet in addition to statistical correlations discovered in our data as described in prior research [121].

2.2.2 Image Generation

DARCI uses an evolutionary mechanism to render images so that they visually express the meaning of given synsets. Because of the innate ability of evolution to yield novel solutions to problems, evolutionary methods are frequently used in generative art [61].

Our evolutionary mechanism operates in two modes. The initial mode, which we call *practice mode*, operates by exploring the space of image filters that will render any image according to a single specific synset. For this mode, DARCI creates a separate, persistent gene pool for each synset that the system knows. The second mode, called *commission mode*, operates by exploring the space of image filters that will render a specific image according to a specified list of synsets. For this mode, users prescribe the image and list of synsets that they wish DARCI to render—in other words, they "commission" DARCI. For each commission, DARCI creates a unique gene pool that terminates once the commission is complete. The evolutionary mechanism for both modes functions as follows.

The genotypes that comprise each gene pool are lists of filters (and their accompanying parameters) for processing a source image. The processed image is the phenotype. Many of these filters are similar to those found in Adobe Photoshop and other image editing software. Others come from a series of 1000 filters that Colton et al. discovered using their own evolutionary mechanism [29]. This set of filters, called *Filter Feast*, is divided into categories of aesthetic effect that were discovered by exploring combinations of very basic filters within a tree structure. We

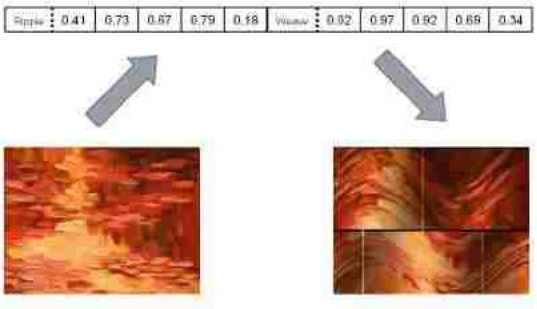


Figure 2.2: Sample genotype (top) applied to a source image (left) resulting in the phenotype (right). The genotype is a list of image filters with parameters. "Ripple" and "Weave" are the names of two (of ninety-two) possible filters. Example image courtesy of William Meire.

have treated *Filter Feast* filters as if each category were a unique filter with a single parameter that specifies the specific filter within the category to use. Figure 2.2 gives an example of a genotype and its phenotype. There are a total of sixty-one traditional filters that we selected for DARCI to use and a total of thirty-one categories of filters from *Filter Feast*, making ninety-two filters available for each genotype. We selected traditional filters that were easily accessible, diverse, fast, and that didn't incorporate alpha values (since our feature extraction techniques cannot yet process alpha values).

Every generation of the evolutionary mechanism, each phenotype is created from the same source image; but, the source image used from generation to generation depends upon which mode the system uses. In commission mode, the source image is the same from generation to generation, while in practice mode the source image for each generation is randomly selected from DARCI's growing image database.

The function used to evaluate the fitness of each phenotype created during the evolutionary process can be expressed by the following equation:

$$\text{Fitness}(f^P) = \lambda_A A(f^P) + \lambda_I I(f^P) \tag{2.1}$$

where f^P is the vector of 102 image features (see Section 2.2.1) for a given phenotype and $A : F^P \rightarrow [0, 1]$ and $I : F^P \rightarrow [0, 1]$ are two functions: appreciation and interest. These functions compute a real-valued score for a given phenotype (here, F^P represents the set of all phenotype feature vectors). $\lambda_A + \lambda_I = 1$, and for now, $\lambda_A = \lambda_I = 0.5$.

The appreciation function A is computed as the weighted sum of the output(s) of the appropriate appreciation network(s), producing a single (normalized) value:

$$A(f^P) = \sum_{w \in C} \alpha_w \text{net}_w(f^P) \tag{2.2}$$

where C is the set of synsets to be portrayed, $\text{net}_w(\cdot)$ is the output of the appreciation network for synset w, and $\alpha_w = 1/|C|$ (though this can, of course, be changed to weight synsets unequally). As indicated, f^P, the feature vector of the phenotype, is the input to each appreciation network.

The interest function I penalizes phenotypes that are either too different from the source image, or are too similar. This can be expressed with the Equations 2.3 - 2.5 as follows:

$$n = \sum_i \sigma(f_i^S, f_i^P) \tag{2.3}$$

$$\sigma(f_i^S, f_i^P) = \begin{cases} 0, & |f_i^S - f_i^P| > 0.3 \\ 1, & |f_i^S - f_i^P| < 0.3 \end{cases} \tag{2.4}$$

where f_i^S represents feature i of the source image and f_i^P represents feature i of the phenotype. The similarity threshold of 0.3 was chosen empirically. Equations 2.3 and 2.4 give a count, n, of how many features between the phenotype and source image are similar. The interest score is calculated using n as follows:

$$I(f^P) = 1 - \begin{cases} \frac{\tau_d - n}{\tau_d}, & n < \tau_d \\ \frac{n - \tau_s}{|f| - \tau_s}, & n > \tau_s \\ 0, & \tau_d \leq n \leq \tau_s \end{cases} \tag{2.5}$$

Number of Sub-Populations	8
Size of Sub-Populations	15
Crossover Rate	0.4
Mutation Rate	0.1
Parameter Mutation Rate	0.9
Migration Rate	0.2
Migration Frequency	0.1
Tournament Selection Rate	0.75
Initial Genotype Length	2 to 4 filters

Table 2.1: Parameters used for the evolutionary mechanism.

where τ_d and τ_s are constants that correspond to the threshold for determining, respectively, when a phenotype is too different from or too similar to the source image. The values $\tau_d = 20$ and $\tau_s = 57$ were used here. $|f|$ is the total number of features analyzed (102). In effect, if there are between 20 and 57 image features that are similar between the source image and the phenotype in question (see Equation 2.3), then the interest score for the phenotype is 1—the maximum score. Otherwise the interest score will degrade according to Equation 2.5.

Our evolutionary mechanism for image generation operates similar to other standard genetic algorithms. As such, there are several more components to the mechanism that require further discussion including the phenotype selection process, crossover, mutation, and migration.

Fitness-based tournament selection determines those genotypes that propagate to the next generation and those genotypes that participate in crossover. One-point "cut and splice" crossover is used to allow for variable length offspring. Crossover is accomplished in two stages: the first occurs at the filter level, so that the two genomes swap an integer number of filters; the second occurs at the parameter level, so that filters on either side of the cut point swap an integer number of parameters. By necessity, parameter list length is preserved for each filter. Table 2.1 shows the parameter settings used.

Mutation rate is the probability that a mutation will occur in each genotype. Parameter mutation rate is the probability that when a mutation occurs, it is a parameter mutation; otherwise, it is a filter mutation. Filter mutation is a wholesale change of a single filter (discrete values),

while parameter mutation is a change in parameter values for a filter (continuous values). When a parameter mutation occurs, anywhere from one to all of the parameters (uniformly chosen) for a single filter in a genotype are changed. The degree of this change, Δl_i, for each parameter, i, is determined by one of the following two equations chosen randomly with equal probability:

$$\Delta l_i = (1 - l_i) \cdot rand\left(0, \frac{(|l| + 1) - |\Delta l|}{|l|}\right) \tag{2.6}$$

$$\Delta l_i = -l_i \cdot rand\left(0, \frac{(|l| + 1) - |\Delta l|}{|l|}\right) \tag{2.7}$$

Here, l_i is the value of parameter i prior to mutation, $|l|$ is the total number of parameters in the mutating filter, $|\Delta l|$ is the number of changing parameters in the mutating filter, and $rand(x, y)$ is a function that uniformly selects a real value between x and y.

Because there are potentially many ideal filter configurations for modeling any given synset, we have implemented sub-populations within each gene pool. This allows the evolutionary mechanism to converge to multiple solutions, all of which could be different and valid. The migration frequency controls the probability that a migration will occur at a given epoch, while the migration rate refers to the percentage of each sub-population that migrates. Migrating genomes are selected uniform randomly, with the exception that the most fit genotype per sub-population is not allowed to migrate. Migration destination is also selected uniform randomly, except that sub-population size balancing is enforced.

Practice gene pools are initialized with random genotypes, while commission gene pools are initialized with the most fit genotypes from the practice gene pools corresponding to the requested synsets. This allows commissions to become more efficient as DARCI practices known synsets. It also provides a mechanism for balancing permanence (artist memory) with growth (artistic progression).

2.3 Evaluation Methodology

In order to evaluate DARCI's artifacts more consistently and without direct human involvement, we use a clustering algorithm to find the inherent groupings of the images. DARCI's artifacts should cluster in ways comparable to relationship clusters found within WordNet.

WordNet is an extensive ontology of the English language consisting of over 117,000 synsets, the basic unit of WordNet, and their relationships. Since individual words can have different meanings, or senses, WordNet is not structured around the words themselves; rather, it is structured around these meanings. A synset is a collection of different words that share the same meaning and can be used interchangeably. For example "bright" and "smart" can mean the same thing and thus together form one synset. Another sense of "bright" means the same thing as a sense of "brilliant" and a sense of "vivid". These form another distinct synset. In WordNet, all synsets are placed into one of four part-of-speech categories: nouns, adjectives, verbs, and adverbs. This paper is primarily concerned with the more than 18,000 adjective synsets.

Many of the relationships documented in WordNet can provide us with a measure of semantic similarity between synsets. For adjectives, these relationships include *adjective clusters*, *satellites*, *antonyms*, and *related concepts*. Most adjective synsets are contained in adjective clusters which consist of two (rarely three) head synsets and their respective satellites. Antonyms are synsets that are opposite in meaning and are commonly associated. The head synsets in an adjective cluster are bound by an antonym relationship making the clusters polar groupings. The satellites of a cluster are synsets similar in meaning to their respective head. Satellites are indirectly antonymous to the opposing head satellites. Related concepts are synsets that do not belong to the same adjective cluster, but share some similarity in meaning. As an example, Figure 2.3 shows the adjective cluster for the antonym pair "peaceful" and "unpeaceful".

Ideally, DARCI's artifacts rendered with synsets that share the same satellite head should tend to cluster more frequently than artifacts rendered with synsets that are only related concepts, which should in turn cluster more frequently than artifacts rendered with more distantly related

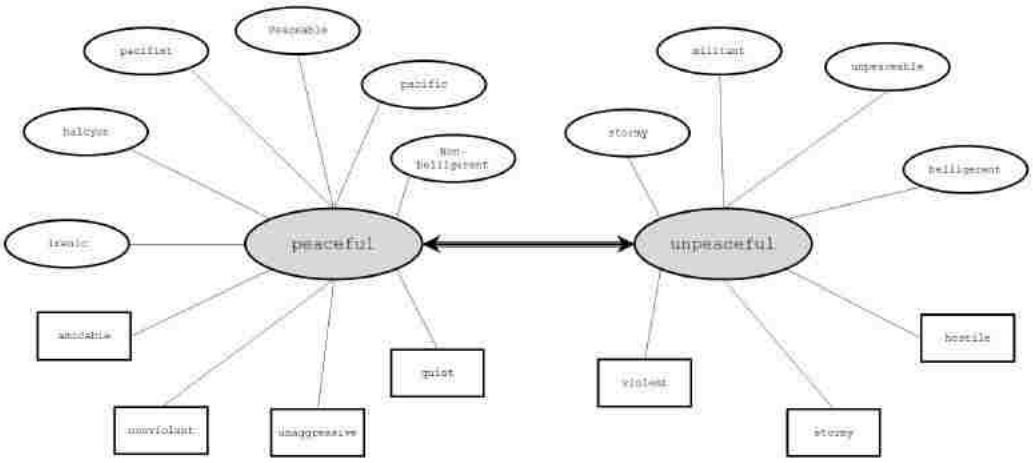


Figure 2.3: The adjective cluster for the antonym pair "peaceful/unpeaceful" as contained in WordNet. "Peaceful" and "unpeaceful" are the head synsets. The synsets in ovals are satellites while the synsets in rectangles are related concepts (technically not part of the adjective cluster).

concepts or concepts that are altogether unrelated, and so on. Unfortunately this ideal scenario is not wholly practical due both to limitations inherent in WordNet, and to the complex nature of meaning in images.

Since WordNet's relationships are determined through a formal analysis of the language by linguistic experts, they don't necessarily agree with colloquial interpretations. For example, in WordNet, "happy" and "sad" are *not* direct antonyms. The antonym of "sad" is "glad", and the antonym of "happy" is "unhappy" (in a distinct adjective cluster from "sad"). "Sad" and "happy" do share related concepts that are antonyms, so there is an indirect connection between the two adjectives. Still, this demonstrates a slight incongruity between commonly assumed meanings (comprising DARCI's hand-labeled training dataset) and meanings defined in WordNet. Furthermore, WordNet's related concepts are arguably not complete as illustrated by the fact that they are not all bidirectional. For example, "glad" is a related concept of "happy", but "happy" is not a related concept of "glad". Finally, certain terms that are technically semantically unrelated, do have semantic connections in people's minds. For example "warm" as in the temperature has a distinctly different meaning than "warm" as in the use of color; however, people relate the two because one suggests the other. WordNet does not usually include these connections.

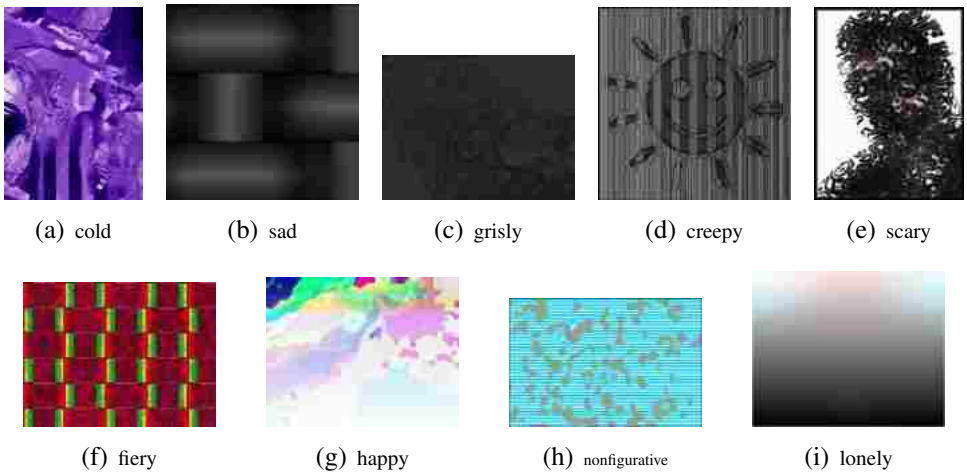
Despite these limitations, there is value in using WordNet to evaluate the way in which DARCI's artifacts are clustered. Having clusters that disagree with WordNet does not necessarily indicate a failure in DARCI since arguably WordNet is not complete; however, agreeing with WordNet *does* signify a degree of success since WordNet's relationships are widely accepted, giving us a somewhat objective measure of a quality that is inherently subjective.

We use the EM (Expectation Maximization) algorithm found in Weka [69] to cluster the images with the same 102 low-level image features used to inform DARCI's neural networks. These features include general measures of color, light, texture, and shape [121]. For these experiments, we specify the number of clusters to be equal to the number of adjectives that are present among the images.

We perform two sets of experiments. In the first set, we analyze how the difference in clustering quality between different groups of artifacts illustrates the expression of meaning in the artifacts. In the second set of experiments, we show how the sequence in which images agglomeratively cluster demonstrates the expression of meaning. In both sets of experiments, DARCI produces images to be clustered using the evolutionary mechanism detailed earlier and in prior work [123].

2.4 Results

We perform two groups of experiments to explore clustering as an evaluation of the expression of semantics. In the first group, we analyze how DARCI's artifacts cluster into specified sets of adjective synsets. We explore different sets of synsets and examine the quality of the resulting clusters. In the second group of experiments, we use agglomerative clustering to build a hierarchy of DARCI's visual metaphors using two different sets of features as data points. Finally, in order to validate the agglomerative clustering results, we survey human volunteers by having them rank image similarity.



(a) cold (b) sad (c) grisly (d) creepy (e) scary

(f) fiery (g) happy (h) nonfigurative (i) lonely

Figure 2.4: Examples of images created by DARCI during practice using various source images.

2.4.1 Cluster Quality

For our first series of experiments we selected two sets of synsets, a *distinct* set and a *similar* set. The distinct set contains five synsets that are either antonymous or conceptually unrelated to each other according to WordNet: "cold", "fiery", "happy", "nonfigurative" (as in abstract or nonfigurative art), and "sad." The similar set contains five synsets that are either conceptually related according to WordNet, or similar in mood to one another: "creepy", "grisly", "lonely", "sad", and "scary." Both sets contain the same synset for "sad." We began by creating a practice gene pool for each synset belonging to these sets. We trained these gene pools for 100 generations and then collected the most fit images created from the last 20 epochs of training. Example images from each of the nine synsets are shown in Figure 2.4. We then performed EM clustering on each set's collection of images. Finally, we analyzed the resulting clusters.

Often clusters favored specific synsets; however, as there are variable source images, and many ways for synsets to be legitimately expressed with DARCI's rendering tools, clusters were usually composites of multiple synsets. It is difficult to draw conclusions from these experiments by simply looking at the breakdown of each cluster. However, we can see how the meaning of synsets is present in DARCI's artifacts by analyzing metrics obtained for each *synset*. We determined the

F1 measure, precision, and recall for each synset in each cluster, and then isolated the best value of each metric for the various synsets. The cluster that the *best value* comes from tells us which cluster best represents each synset. The value itself tells us how well the synset is represented by that cluster. Additionally, we calculated the average cluster entropy and average cluster purity to determine how well the clustering matches the original adjective groupings. We repeated this clustering experiment five times using newly created gene pools each time. Tables 2.2 and 2.3 show the best F1 measure, precision, and recall, as well as average entropy and average purity, for the distinct and similar sets of synsets averaged across the five experiments.

To help understand the results, consider as an example the adjective "cold" in Table 2.2. The recall metric tells us that 57% of the "cold" images clustered together into one cluster (this is the largest grouping of "cold"). The precision metric tells us that in that cluster, 64.5% of the images were "cold." The F1 measure is essentially the weighted average of precision and recall. These metrics give us an overview of how well the images for each individual synset clustered. The metrics entropy and purity tell us how well, as a whole, all the images clustered. For all the metrics (except for entropy), the higher the value the better. It is clear that the distinct synsets have consistently better scores than the similar synsets. This means that more confusion is occurring in the synsets that are semantically more similar—the behavior we would expect in a system that is expressing appropriate meaning in its artifacts.

We performed the same experiment using commissioned images rather than practice images. Using the first of the practice gene pools for each synset just created, we commissioned DARCI with three source images for each synset. Each commission was given 100 epochs to develop. The top 40 images from each commission were collected making a total of 120 images per synset. The empirical results of clustering these commissions are shown in Tables 2.4 and 2.5. Again, the distinct synsets were more effectively clustered than the similar synsets as we would expect. It is interesting to note that "nonfigurative" in Table 2.4 has a value of 1.000 for precision. This tells us that there was a cluster where 100% of images it contained were "nonfigurative." However, the

	F1	Precision	Recall
cold	0.565	0.645	0.570
fiery	0.569	0.614	0.590
happy	0.651	0.578	0.760
nonfigurative	0.520	0.555	0.540
sad	0.660	0.705	0.660
AVERAGES	0.593	0.619	0.624
ENTROPY	0.567	PURITY	0.584

Table 2.2: Best F1 measure, precision, and recall for each synset, as well as average entropy and average purity, after performing EM clustering on practice images for distinct synsets. (Lower is better for entropy.)

	F1	Precision	Recall
creepy	0.398	0.458	0.490
grisly	0.409	0.377	0.480
lonely	0.448	0.708	0.460
sad	0.481	0.463	0.540
scary	0.557	0.538	0.600
AVERAGES	0.459	0.509	0.514
ENTROPY	0.752	PURITY	0.458

Table 2.3: Best F1 measure, precision, and recall for each synset, as well as average entropy and average purity, after performing EM clustering on practice images for similar synsets. (Lower is better for entropy.)

	F1	Precision	Recall
cold	0.784	0.987	0.650
fiery	0.429	0.362	0.525
happy	0.690	0.736	0.650
nonfigurative	0.717	1.000	0.558
sad	0.660	0.557	0.808
AVERAGES	0.656	0.729	0.638
ENTROPY	0.521	PURITY	0.623

Table 2.4: Best F1 measure, precision, and recall for each synset, as well as average entropy and average purity, after performing EM clustering on commissioned images for distinct synsets. (Lower is better for entropy.)

	F1	Precision	Recall
creepy	0.455	0.623	0.450
grisly	0.272	0.500	0.508
lonely	0.469	0.319	0.883
sad	0.429	0.292	0.808
scary	0.711	0.932	0.575
AVERAGES	0.467	0.533	0.645
ENTROPY	0.779	PURITY	0.445

Table 2.5: Best F1 measure, precision, and recall for each synset, as well as average entropy and average purity, after performing EM clustering on commissioned images for similar synsets. (Lower is better for entropy.)

recall metric then tells us that the cluster only contained 55.8% of the "nonfigurative" images. This indicates that just over half of the "nonfigurative" images were particularly distinct.

These results verify those obtained from clustering the practice images. Comparing the results of the commissioned images and the practice images, we see that the commissioned images have better values than practice. This is likely because we started each commission with the practice gene pools for each synset, which essentially gave the commissioned images a head start.

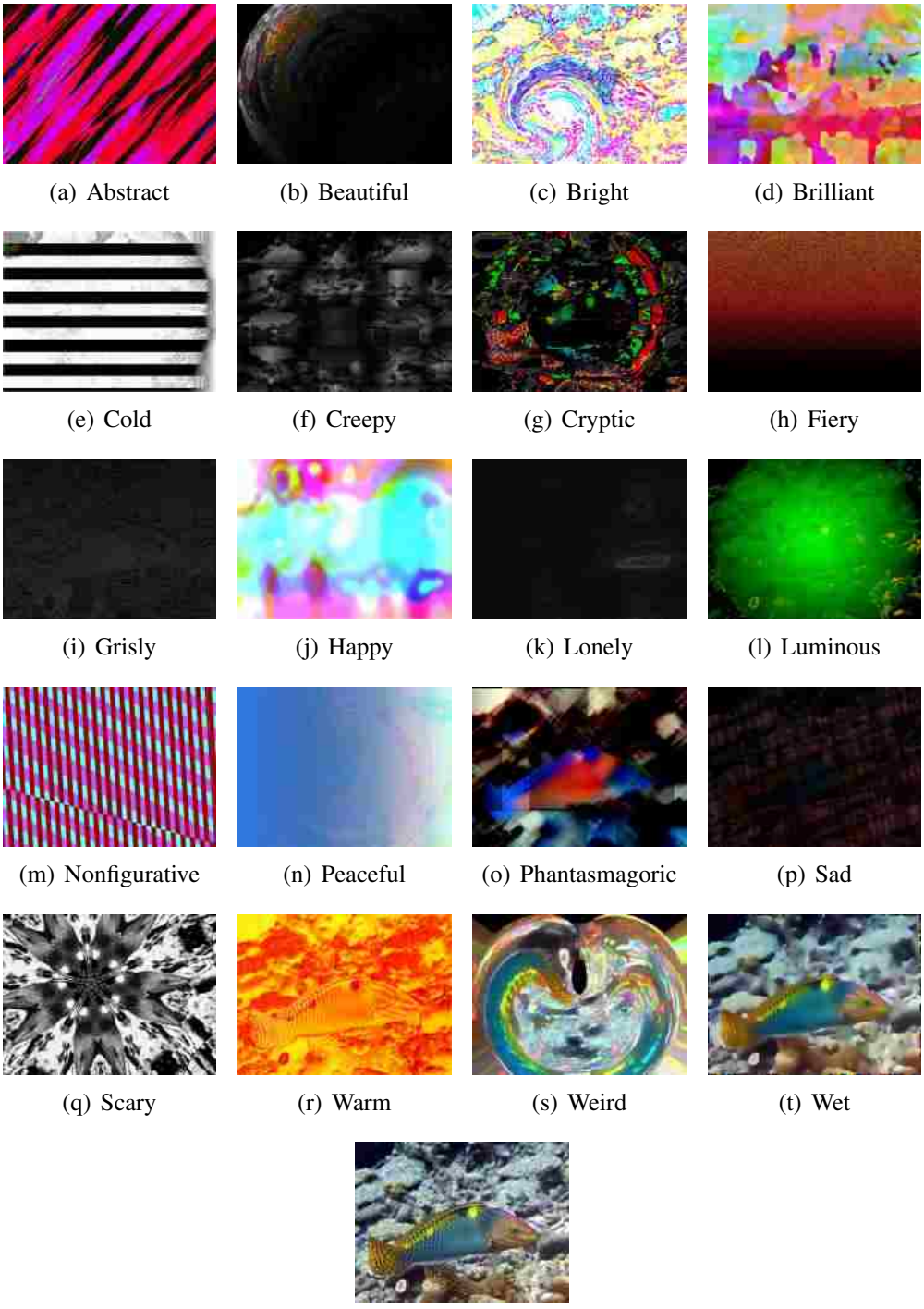
In preparation for the agglomerative clustering experiments, we added eleven synsets to the nine outlined above (see Figure 2.6), and commissioned DARCI with the same three images for

bright	luminous	1	bright	peaceful	7
grisly	scary	1	creepy	phantasmagoric	9
abstract	nonfigurative	2	bright	creepy	-9
abstract	phantasmagoric	5	creepy	luminous	-9
beautiful	bright	6	beautiful	creepy	-5
beautiful	luminous	6	creepy	peaceful	-5
nonfigurative	phantasmagoric	6	happy	sad	-3
peaceful	luminous	7	cold	fiery	-2

Table 2.6: A list of all synset pairs with a semantic distance magnitude less than 10. Synsets are ranked from most similar to most opposite. Note that lower magnitude negative distance indicates a stronger antonymous relationship.

each synset (again 100 epochs). An example commission from each of these 20 synsets for one of the source images is shown in Figure 2.5. It should be noted that the adjective "abstract" in this case means "abstract concept" and is thus in a slightly different synset than "nonfigurative." The synsets were selected to cover a spectrum of meanings with varying degrees of similarity while also being well represented in DARCI's training database. We then determined the number of WordNet related concepts or antonymous relationships between each pair of synsets. We call this value the *semantic distance* and represent positive relationships with a positive value and negative relationships (antonymous relationships) with a negative value. This distance is defined to be a value of 1 for synset pairs that share the same satellite head, and increases in magnitude by 1 for every relationship link (related concept or antonym) crossed. In addition, every time an antonym link is crossed, the distance changes sign. Table 2.6 shows all of the semantic distances between pairs of synsets with a distance magnitude less than 10.

We took each pair of synsets listed in Table 2.6, and performed EM clustering over the two synsets' images. We then calculated the F1 measure, precision, and recall for each synset in each pair, as well as the average entropy and average purity for each pair. Finally, we averaged the metrics across the similar (positive semantic distance) and distinct (negative semantic distance) pairs



(a) Abstract (b) Beautiful (c) Bright (d) Brilliant

(e) Cold (f) Creepy (g) Cryptic (h) Fiery

(i) Grisly (j) Happy (k) Lonely (l) Luminous

(m) Nonfigurative (n) Peaceful (o) Phantasmagoric (p) Sad

(q) Scary (r) Warm (s) Weird (t) Wet

(u) SOURCE IMAGE

Figure 2.5: A sample of commissioned results for the source image used in the clustering experiments. The source image is courtesy of Jan Messersmith (http://www.messersmith.name).

	F1	Precision	Recall	Entropy	Purity
similar	0.726	0.778	0.790	0.763	0.727
distinct	0.803	0.823	0.859	0.596	0.793

Table 2.7: The average F1 measure, precision, recall, entropy, and purity for similar pairs of synsets and distinct pairs of synsets when binary clustering is applied. (Lower is better for entropy.)

of synsets. The results are recorded in Table 2.7. Yet again, the distinct pairs of synsets clustered more effectively than the similar pairs of synsets.

2.4.2 Agglomerative Clustering

In these experiments we use agglomerative clustering based on the EM clustering algorithm and centroids to see if related adjectives will group together before grouping with unrelated adjectives. We performed agglomerative clustering on the 20 adjectives listed in Figure 2.6 in two ways: using all 102 image features, and simplifying the feature space to just our 12 color and lighting features. The original feature set is comprehensive, dealing with various aspects of color, lighting, texture, and shape within an image. For comparison, the smaller set (of 12) uses only color and lighting features, as they are predominately considered the most significant in image comparison research [21]. To perform agglomerative clustering, we first cluster the images into the 20 adjective groups in same way as done in Section 4.1. We then calculate the centroid of each resulting cluster. Note that the centroid of each cluster is the average feature vector of all the images in each cluster. We then decrement the number of clusters by one and recluster the new centroids. We repeat this process, decrementing the number of clusters each time, until there are only two clusters.

The results for each clustering can be seen in Figure 2.6 and Figure 2.7. With few exceptions (like “creepy” and “lonely”), the results of using all 102 features do not agree with the relationships in WordNet, nor with intuition. The results from using only the color and lighting features are considerably better. It is interesting to note that of the five semantically similar adjectives used in the previous experiment, four of them (“creepy”, “lonely”, “grisly”, and “scary”) clustered together before clustering with any other adjective. While the fifth one (“sad”), clustered with them before a

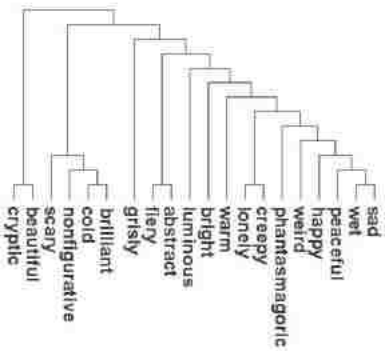


Figure 2.6: Results of agglomerative clustering with 20 adjectives using all 102 images features.

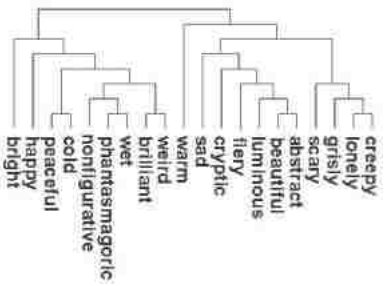


Figure 2.7: Results of agglomerative clustering with 20 adjectives using only 12 color features.

majority of others. In contrast, the adjectives from the previous experiment that were semantically dissimilar ("sad", "happy", "nonfigurative", "fiery", and "cold"), remain distinct far up the cluster hierarchy.

2.4.3 Human Survey

To validate the clustering experiments just described, as well as to further evaluate DARCI's artifacts, we conducted a human survey. In this survey, we used the same images produced by DARCI that were used in the agglomerative clustering experiment for one of the source images. However, we limited the survey to only 10 of the 20 adjectives to avoid overburdening the users. The adjectives in the survey were "happy", "sad", "cold", "bright", "scary", "creepy", "weird", "peaceful", "luminous", and "warm." Recall that there are 40 images produced for each adjective bringing the total to 400 images. In the survey we presented each volunteer with one of DARCI's 400 images randomly chosen. This image acted as a representative for one of the 10 adjectives of interest. Underneath this target image volunteers were shown additional random images (from these 400) for each of the 10 adjectives (in random order). There was no indication of which adjective went with which image. Using a drag and drop interface, the volunteers were required to sort the 10 images below based on how well they thought these images matched the target image above. The

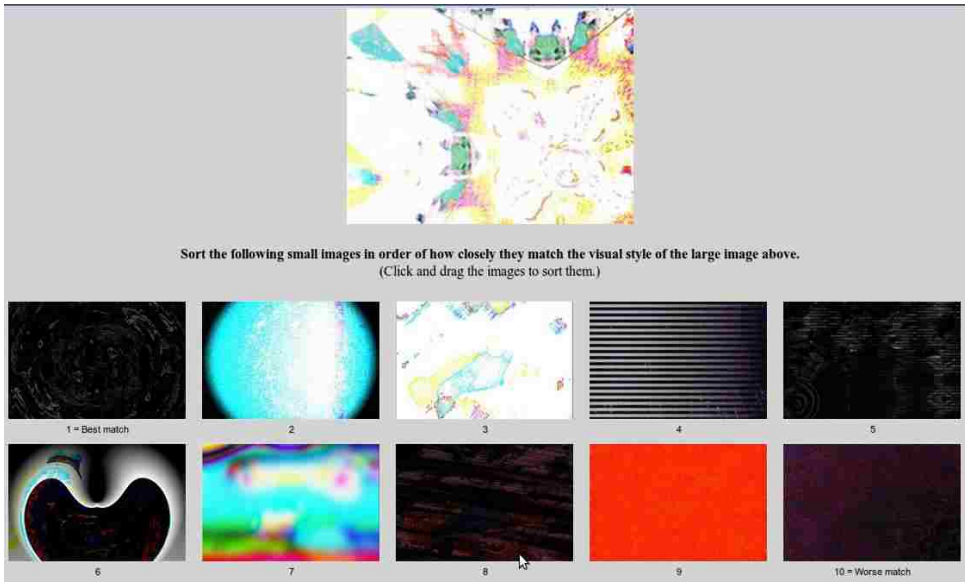


Figure 2.8: An example screenshot of the human survey. The larger image on the top corresponds to a particular adjective. Using a drag and drop interface, the ten smaller images below must be sorted according to how well they match the visual style of the image above.

user repeated this process multiple times, each time with randomly chosen images. An example screen-shot can be seen in Figure 2.8.

There were 70 people that participated in the survey, each completing rankings for, on average, 8.3 target images for a total of 582 entries. Thus, each of the 10 adjectives had an average of 58 entries where each entry consisted of a target image (corresponding with the adjective) and 10 ranked images ranked according to their similarity with the target image. For each adjective we averaged the rankings of the 10 subordinate adjectives (1 being the closet match and 10 being the worst) to determine, overall, how similar each adjective's images were to each other. The results can be seen in Table 2.8.

The first thing to notice is that all 10 adjectives are most closely matched with themselves. This tells us that DARCI is consistent in rendering images that convey each adjective. It also tells us that DARCI's representation of each adjective is distinct from other adjectives. We can also see that adjectives with similar meaning, such as "scary" and "creepy", were consistently ranked closely

Adjective	Ranking (left to right)
bright	bright, happy, peaceful, warm, scary, weird, sad, luminous, cold, creepy
cold	cold, sad, creepy, scary, peaceful, weird, bright, warm, happy, luminous
creepy	creepy, scary, sad, luminous, cold, warm, peaceful, weird, bright, happy
happy	happy, peaceful, weird, bright, scary, luminous, warm, sad, cold, creepy
luminous	luminous, warm, scary, creepy, happy, peaceful, bright, sad, weird, cold
peaceful	peaceful, weird, happy, scary, sad, creepy, luminous, bright, cold, warm
sad	sad, cold, creepy, scary, peaceful, weird, warm, happy, bright, luminous
scary	scary, creepy, luminous, weird, peaceful, sad, bright, happy, cold, warm
warm	warm, luminous, bright, sad, happy, scary, peaceful, cold, weird, creepy
weird	weird, scary, peaceful, sad, happy, cold, creepy, luminous, bright, warm

Table 2.8: Results of the human survey. The adjective lists in the right column indicate the overall ranking of similarity between images rendered with the indicated adjectives and images rendered with the adjective in the left column.

to each other. Conversely, adjectives with dissimilar meaning, such as "happy" and "sad", were consistently ranked farther from each other.

To compare the results of the survey with our clustering results, we performed agglomerative clustering on the survey data by using the average ranking as the distance metric between each adjective. We also reran the agglomerative clustering algorithm on these same 10 adjectives using the 12 color and lighting features. The results for each clustering can be seen in Figure 2.9 and Figure 2.10.

With the exception of "cold", "warm", and "luminous", the two clusterings are very similar in how they cluster. In both cases, we can see that similar adjectives, like "scary", "creepy", and "sad", grouped together first. Likewise, we can also see that dissimilar adjectives, like "happy/sad" and "warm/cold", grouped far apart. This not only validates DARCI's ability to convey distinct

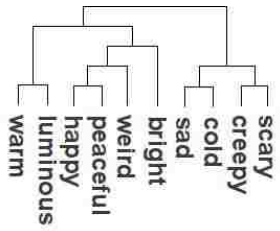


Figure 2.9: Results of agglomerative clustering with 10 adjectives using the human survey data.

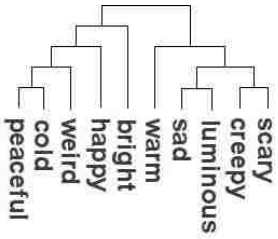


Figure 2.10: Results of agglomerative clustering with 10 adjectives using the 12 color and lighting image features.

meaning in its artwork, but also validates the use of clustering algorithms as an objective metric that can be used to evaluate DARCI's artifacts. Note that in prior work we have done other human surveys that evaluate how well images created by DARCI actually communicate the meaning of a particular adjective compared to images created by human artists [123].

2.5 Conclusions

We have evaluated a system, DARCI, for using visual metaphor to communicate the meaning of specified adjectives in the images it renders. The system learns how to label images with adjectives and, in turn, to render them appropriately by training on eclectic human-labeled data. We have shown that DARCI can produce artifacts that will cluster in ways that reflect the supplied adjectives' relationships according to WordNet. In addition, according to human opinion, DARCI can depict a distinct visual style for each synset the system renders. If we acknowledge that the meaning of words can be at least partially expressed by their relationships to each other, then our results indicate that DARCI is indeed learning how to visually convey the meaning of words. DARCI's ability to generate visual metaphor represents a unique stepping stone in the exploration of metaphor in AI.

We have also demonstrated a unique way to use clustering to evaluate the visual communication of meaning without direct human involvement. To validate this evaluation method, we have shown that it favorably compares with human opinion. Such a metric may be useful in any research interested in the empirical analysis of visual images.

In the future, we will explore the possibility of generating visual metaphor with nouns in addition to adjectives allowing for a more complete use of visual metaphor. This will require the development of higher-level feature extraction methods when building visual-linguistic associations. Further, adding nouns to DARCIs vocabulary suggests a close relationship to content-based image retrieval research and may allow us to leverage techniques from that field.

Chapter 3

Semantic Models as a Combination of Free Association Norms and Corpus-based Correlations[1]

Abstract

We present computational models capable of understanding and conveying concepts based on word associations. We discover word associations automatically using corpus-based semantic models with Wikipedia as the corpus. The best model effectively combines corpus-based models with preexisting databases of free association norms gathered from human volunteers. We use this model to play human-directed and computer-directed word guessing games (games with a purpose similar to *Catch Phrase* or *Taboo*) and show that this model can measurably convey and understand some aspect of word meaning. The results highlight the fact that human-derived word associations and corpus-derived word associations can play complementary roles in semantic models.

[1]Derrall Heath, David Norton, Eric Ringger, and Dan Ventura, Semantic Models as a Combination of Free Association Norms and Corpus-based Correlations, *Proceedings of the 7th IEEE International Conference on Semantic Computing*, pp. 48-55, 2013

3.1 Introduction

Language is a critical component of human intelligence, and the development of computer systems that can understand and communicate through language is an important problem in the field of artificial intelligence. Building computational models that provide meaning to words is a step in that direction. Most words are a representation of a concept, and it is the concept itself in which we are interested and thus, the terms 'concept' and 'word' will be used interchangeably throughout the paper.

The study of word meaning and conceptual knowledge is called *lexical semantics* (in linguistics), *semantic memory* (in cognitive psychology), or *cognitive semantics* (in cognitive linguistics). This question of what gives words meaning has been debated for years; however, it is commonly agreed that a word, at least in part, is given meaning by how the word is used in conjunction with other words (i.e., its context) [51, 96]. Many computational semantic models consist of building associations between words [40, 152]. These word associations essentially form a large graph that is typically referred to as a *semantic network*.

Word associations are commonly acquired in one of two ways: explicitly from people and automatically by inferring them from a corpus. ConceptNet [101] and WordNet [55] are examples of semantic networks that have been created explicitly by hand (or through crowd sourcing). In these networks, words are linked by specific types of relationships that are often intended for specific purposes. Although these networks have been applied to problems such as common sense reasoning [102], they are often either limited in their vocabulary, limited in their variety of word associations, or do not provide any notion of relationship strength.

Corpus-based semantic models (CSMs) are a class of computational models that attempt to learn semantic information from patterns of word co-occurrences in a corpus. These models are based on the idea that similar words will occur in similar contexts and words that are often associated together will often co-occur close together. CSMs have been successfully used on a variety of tasks such as information retrieval [143], multiple choice vocabulary tests [43], multiple

choice synonym questions from the TOEFL test [137], and multiple choice analogy questions from the SAT test [158].

Free Association Norms (FANs) are a common means of gathering word associations from people and are considered to be one of the best methods for understanding how people, in general, associate words in their own minds [118]. Thus, the corpus-based models are often compared directly with FANs as a way to evaluate the quality of word associations discovered from corpora [131, 168]. FANs are rarely used themselves to help solve word similarity tasks and exist only as a baseline metric or to be analyzed directly by cognitive scientists.

Most CSMs have been applied to word similarity tasks such as the previously mentioned multiple choice synonymy test and clustering words into predefined groups. However, it would be beneficial for a semantic model to be able to determine a concept given a description of that concept without multiple choice options. For example, if the model were given a description of a word, it should be able to determine what that word is. Conversely, given a word, the model should be able to provide a description of that word that makes sense to humans.

We introduce a new task that involves using CSMs to play word guessing games (similar to *Catch Phrase* or *Taboo*) with people online. The idea is to evaluate how well a computer system that uses word associations can understand and convey concepts to humans. These word guessing games are designed for two purposes: for evaluation, and as a novel way of collecting new viable word associations. Thus, these guessing games are also contributions to the growing field of *Games with a Purpose* [166]. We play these word guessing games using Free Association Norms, using three common CSMs, and using a hybrid approach combining FANs with CSMs. We show that using a hybrid approach improves the ability of the system to play these guessing games and therefore can, at least partially, convey meaning to humans.

We will first describe our methodology in using FANs and our method for building the three CSMs used in this paper. We will then outline the initial experiments and results used to evaluate the models. Finally, we will detail the online word guessing game and discuss the results and applications.

3.2 Methodology

We want to create a system that can communicate and understand concepts. Given a concept, we want the system to provide a description that will enable a human to know what the given concept is. Conversely, given a description, we want the system to know to which concept the description is referring. For example, suppose the given concept is 'space', the system could provide a description in the form of other words associated with 'space' such as 'planet', 'astronaut', 'star', 'rocket', 'dark', and 'mysterious'. Conversely, if the system is given a collection of words as a description such as 'soldier', 'guns', 'bomb', 'death', 'fight', 'sorrow', and 'courage', then the system should infer that the collection of words most likely represents the concept 'war'.

We use the term 'description' here to mean 'a collection of other words'. In this paper we are not concerned with how words are structured together. We acknowledge that structuring of sentences, relationship types, and word order contribute to the meaning of concepts. Indeed much of the recent work on CSMs deal with trying to automatically infer additional semantic information such as word order, sentence structure, and word relationship types [5, 6, 84]. However, we are interested in the degree to which simple word associations can accomplish the task at hand and leave these more advanced CSMs to future work.

We present a computational semantic model that combines human free association norms with common corpus-based approaches. The idea is to use the FANs to capture general knowledge and then fill in the gaps using a CSM. Here we describe each individual model, initial testing results, and how the individual models will be combined.

Lemmatization and Stop Words

We use the standard practice of removing stop words (words like 'the' and 'of') and lemmatizing (representing different forms of the same word with the word's morphological lemma) as we build word associations. WordNet maintains a database of word forms and hence, we use WordNet to perform the lemmatization [55]. It should be noted, however, that lemmatization with WordNet has its limits. For one, we cannot lemmatize a word across different parts of speech (noun,

verb, adjective, etc). For example, 'redeem' and 'redeeming' will remain separate words because 'redeeming' could be the gerund form of the verb 'redeem' or it could be an adjective (i.e., 'a redeeming quality'). Since the part of speech is not provided for individual words encountered, we must account for all parts of speech, hence words like 'relax', 'relaxing' and 'relaxation' remain separate words.

3.2.1 Free Association Norms

We use two preexisting databases of human word associations: The Edinburgh Associative Thesaurus [91] and University of Florida's Word Association Norms [118]. These databases were built by asking hundreds of human volunteers to provide the first word that comes to mind when given a cue word. This technique, called *free association*, is able to capture many different types of word associations including word co-ordination (pepper, salt), collocation (trash, can), super-ordination (insect, butterfly), synonymy (starving, hungry), and antonymy (good, bad).

We build a semantic model from this data as follows: The association strength between two words is simply a count of the number of volunteers that said the second word given the first word. We also consider word associations to be undirected. In other words, if word A is associated with word B, then word B is associated with word A. Hence, when we encounter data in which word A is a cue for word B and word B is also a cue for word A, we combine them into a single association pair by adding their respective association strengths. Between these two databases, there are a total of 19,310 (lemmatized) unique words and 288,069 unique associations. From now on, we will refer to this model as *FANs*.

3.2.2 Corpus-based Semantic Models

One of the most popular CSMs is *Latent Semantic Analysis* (LSA) [42, 96]. LSA is based on the idea that similar words will appear in similar documents (or contexts). LSA builds either a term $\times$ document or a term $\times$ term matrix from a corpus and then performs Singular Value Decomposition (SVD), which reduces the given large sparse matrix to a low-rank approximation of that matrix

along with a set of vectors, each representing a word (as well as a set of vectors for each document). These vectors also represent points in semantic space, and the closer a word's vector is to another in this space, the closer they are in meaning (and the stronger the association between words). Starting with a term $\times$ term matrix is considered advantageous because the size of the matrix is invariant to the size of the corpus. It is also argued by some that it is more congruent to human cognition than the term $\times$ document matrix used in some implementations of LSA [18, 168]. We implement the same version of LSA that is used in [137], which uses a term $\times$ term matrix and a co-occurrence window of ± 2.

Another popular method is the *Hyperspace Analog to Language* (HAL) model [104]. This model is based on the same idea as LSA, except the notion of word order is partially captured in the co-occurrence matrix (with a co-occurrence window of ± 10), and HAL then uses the co-occurrence counts directly as vectors representing each word in semantic space. We use the same implementation as specified in the original paper [104].

The third CSM we use is constructed from the direct co-occurrence counts (DCC) obtained from the corpus. We build a term $\times$ term co-occurrence matrix M using a co-occurrence window of ± 30. To account for the fact that common words will have generally higher co-occurrence counts, we scale these counts by weighting each element of the matrix by the inverse of the total frequency of both words at each element. This is done by considering each element $M_{i,j}$, then adding the total number of occurrences of each word (i and j), subtracting out the value at $M_{i,j}$ (to avoid counting it twice), then dividing $M_{i,j}$ by this computed number, as follows:

$$M_{i,j} \leftarrow \frac{M_{i,j}}{(\sum_i M_{i,j} + \sum_j M_{i,j} - M_{i,j})} \tag{3.1}$$

The result could be a very small number and hence, we then also normalize the values between 0 and 1. Once the co-occurrence matrix is built from the corpus, we use the weighted/normalized co-occurrence values themselves as association strengths between words.

	AllQ	VocabQ	AssocQ
FANs	0.300(24/80)	0.511(24/47)	**0.923(24/26)**
HAL	0.388(31/80)	0.660(31/47)	0.660(31/47)
DCC	0.438(35/80)	0.745(35/47)	0.745(35/47)
LSA	0.513(41/80)	0.872(41/47)	0.872(41/47)
DCC2	0.738(59/80)	0.747(59/79)	0.747(59/79)
LSA2	0.888(71/80)	0.899(71/79)	0.899(71/79)
FAN-LSA2	**0.900(72/80)**	**0.911(72/79)**	0.911(72/79)

Table 3.1: The TOEFL synonym test scores for the different models. AllQ is the raw score for all 80 questions. VocabQ is the score based on the limited vocabulary and AssocQ is the score based on existing associations in each model. LSA2 uses an extended vocabulary and performs the best when combined with FANs (FAN-LSA2). FANs performs the best when there exists an association.

Note that the DCC model only captures first order relationships between words, while HAL and LSA capture higher order relationships. That is to say, DCC will only associate words that co-occur often (e.g., 'dog', 'kennel'), while HAL and LSA can associate words that co-occur with similar words even if those words never co-occur directly (e.g., 'dog', 'puppy'). For each of the models (LSA, HAL, and DCC), we use the Wikipedia corpus as it is large, easily accessible, and covers a wide range of human knowledge [44]. Initially, for comparison we limit the vocabulary to the same 19,310 (lemmatized) words that exist in the FANs database.

3.2.3 TOEFL Synonymy Test

We take a detour to conduct an initial test to compare the performance of each model on the standard TOEFL multiple choice synonym test [96]. We consider three versions of the results. The first is simply the number of questions correctly answered out of the 80 total questions (AllQ). The second is limited to only the questions in which the word being considered and the correct answer exist in the vocabulary (VocabQ). The third is limited to only the questions in which the model has an association between the word being considered and the correct answer (AssocQ). Note that if a model cannot answer the question, it is counted as wrong; we do not adjust the score for random guessing. The results for FANs, HAL, DCC, and LSA can be seen in the first four rows of Table 3.1.

The first thing to note is that for all questions (AllQ), the four models perform poorly (the human standard for actual TOEFL test takers is 0.645). The obvious reason for the poor results is

that the vocabulary is limited to the 19,310 lemmatized words. When throwing out questions that are not in the vocabulary (VocabQ), the scores improve considerably. The vocabulary for the CSMs can be easily expanded and since LSA and DCC perform the best, we expand their vocabulary to 43,578 lemmatized words, called LSA2 and DCC2. When we do this, the TOEFL scores for AllQ are increased to 0.888 and 0.738 respectively. LSA2's score is comparable to previous TOEFL results for this implementation of LSA, which achieved a score of 0.925 (the differences likely due to different corpus, and not accounting for random guessing) [137].

FANs perform the worst except when considering only questions in which an association score between the words exists (AssocQ). The AssocQ scores for the three CSM models are the same as the VocabQ scores because a similarity score can be computed between any pair of words. FANs are limited in their number of associations because obtaining them is a tedious process of receiving input from people. However, when an association does exist, FANs achieve the best score of 0.923.

We can use the FANs to augment the LSA2 model and build a hybrid model (FAN-LSA2) that improves the results (see last row in Table 3.1). FAN-LSA2 simply defers to the FANs model first and if no association exists between the words in question, then the LSA2 model is used.

3.2.4 Combining Models

From the TOEFL test we can see that FANs and CSMs have different strengths and weaknesses. FANs are limited in vocabulary, are limited in the number of associations between words, and are difficult to acquire. However, the associations that do exist are meaningful and they capture the most relevant associations. CSMs, on the other hand, can automatically discover associations with a large vocabulary, but it is difficult to tell how meaningful the associations are. Other studies have shown that FANs and CSMs each provide different types of word associations [168]. A combination of these methods into a single model has the potential to take advantage of the strengths of each method, as indicated by the improved performance of FAN-LSA2 in the TOEFL test. The hypothesis is that

the combined model will better communicate meaning to a person than either model individually because it presents a wider range of associations.

Combining Method

This method merges two separate databases of word associations into a single database before querying it for associations. This method assumes that FANs contain more valuable word associations than the CSMs because FANs are typically used as the gold standard in the literature. However, CSMs do contain some valuable associations not present in the FANs. The idea is to add the top n associations for each word from one of the CSMs to the FANs but to weight the association strength low. This is beneficial for two reasons. First, if there are any associations that overlap, adding them again will strengthen the association in the combined database. Second, new associations not present in the FANs will be added to the combined database and provide a greater variety of word associations. We keep the association strength low because we want the CSM data to reinforce, but not dominate, the FANs.

We first copy all word associations from the FANs to the combined database. Next, let W be the set of all unique words in the vocabulary, let $A_{i,n} \subseteq W$ be the set of the top n words associated with word $i \in W$ from the CSM, let $score_{i,j}$ be the association strength between words i and j from the CSM, let max_i be the maximum association score present in the FANs for word i, and let θ be a scaling factor. Now for each $i \in W$ and for each $j \in A_{i,n}$, the new association score between words i and j is computed as follows:

$$new\_score_{i,j} \leftarrow (max_i \cdot \theta) \cdot score_{i,j} \tag{3.2}$$

This equation scales $score_{i,j}$ (which is already normalized) to lie between 0 and a certain percentage (θ) of max_i. The n associated words from the CSM are then added to the combined database with the updated scores ($new\_score_{i,j}$). If the word pair is already in the database, then the updated score is added to the score already present. For the results presented in this paper we use $n = 20$ and $\theta = 0.2$, which were determined based on preliminary experiments.

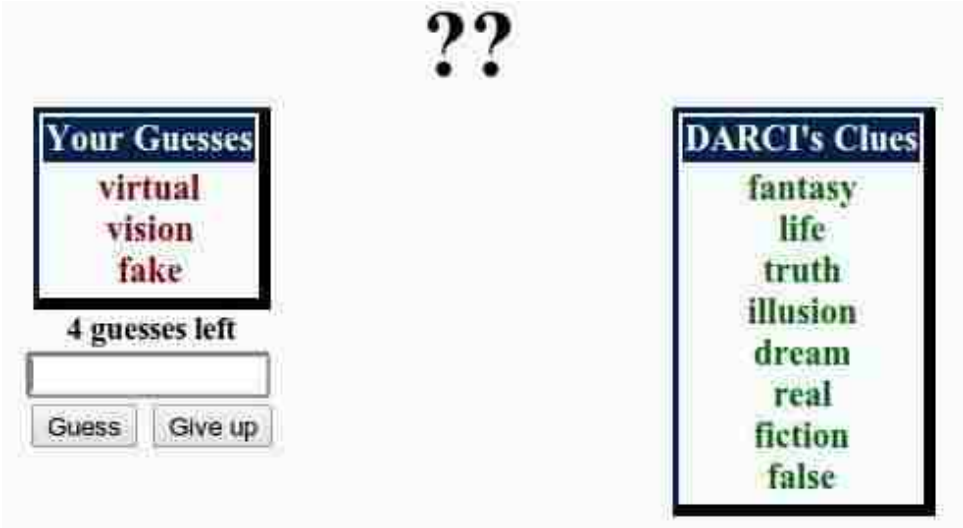


Figure 3.1: User interface for the user_guess mode. On the left the user attempts to guess the concept the system is trying to communicate through the word associations on the right.

3.3 Word Guessing Game

The models are evaluated by playing a word guessing game called *Wordlery* (similar to *Catch Phrase* or *Taboo*), which is accessed through an online interface (http://darci.cs.byu.edu/Wordlery/). There are two modes: one in which the user must guess the word (user_guess) and one in which the system (or model) must guess the word (system_guess). In the user_guess mode, the system presents the user with a set of eight words. The user then has seven chances to guess the concept that the words represent. We record whether or not the user is able to guess the word and how many guesses it took. Figure 3.1 shows the user interface for user_guess mode. The eight words presented by the system (on the right in the figure) are the top n word associations for the hidden word, where $n = 8$. This mode is similar to one of the evaluation metrics used in another study, in which human volunteers had a single chance to guess the word that generated a list of associated words [104].

In the system_guess mode, the user is presented with a word/concept and can then provide up to seven other words one at a time as clues to the system. Figure 3.2 shows the user interface for this mode. After each word provided by the user (on the left), the system gives its current guess (on the right along with its previous guesses) until either it guesses the correct word or the

Figure 3.2: User interface for the system_guess mode. The system attempts to guess the concept ('food') from user provided word association clues on the left.

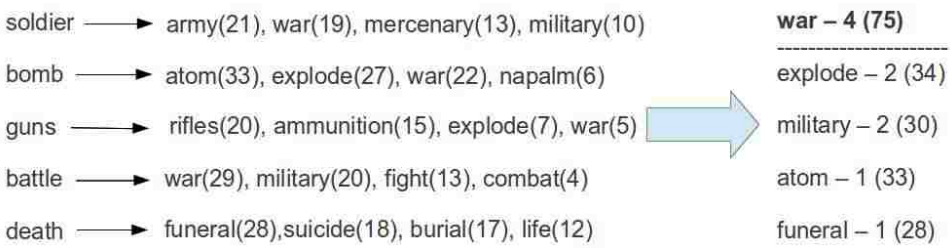


Figure 3.3: A simplified example of how the model guesses a concept given a set of words. The words on the left are the user-provided clues, while the words to the right of the arrows are the lists of associated words for each clue, with their association strength. The system then sorts the associated words by their frequency (the number of different "clue-relation" lists in which a word appears), then by their total association strength. The top word is returned as the guess.

number of allotted clues has been reached. System guesses are generated in a few simple steps. First, the system retrieves the words associated with each user-provided clue. Second, for each associated word retrieved, the system counts the number of user-provided words with which it is associated. The system also sums the association strengths between each associated word and each of the user-provided clues for a total association strength. In the third step, the system ranks the associated words first by their frequency, then by their total association strength. Finally, the top word (that hasn't already been guessed) is returned. See Figure 3.3 for an example of this process.

1 Guess/Clue	**HAL**	**DCC**	**DCC2**	**LSA**	**LSA2**	**FAN**	**FAN-DCC**	**FAN-DCC2**	**FAN-LSA**	**FAN-LSA2**
Overall	0.063	0.186	0.188	0.178	0.171	0.329	0.347	**0.353**	0.344	0.344
User_guess	0.081	0.146	0.145	0.173	0.165	0.295	**0.314**	0.313	0.311	0.311
System_guess	0.045	0.227	0.231	0.182	0.178	0.364	0.380	**0.392**	0.377	0.378
7 Guesses/Clues	**HAL**	**DCC**	**DCC2**	**LSA**	**LSA2**	**FAN**	**FAN-DCC**	**FAN-DCC2**	**FAN-LSA**	**FAN-LSA2**
Overall	0.153	0.374	0.365	0.381	0.368	0.563	0.578	0.577	0.582	**0.589**
User_guess	0.196	0.315	0.309	0.349	0.346	0.499	0.508	0.509	0.515	**0.527**
System_guess	0.111	0.433	0.421	0.413	0.391	0.627	0.649	0.645	0.649	**0.650**

Table 3.2: The win/loss record for several semantic models for each mode of the Wordlery game (higher is better). The combined models perform the best for each mode of the game. This table corresponds to the results in Figure 3.4.

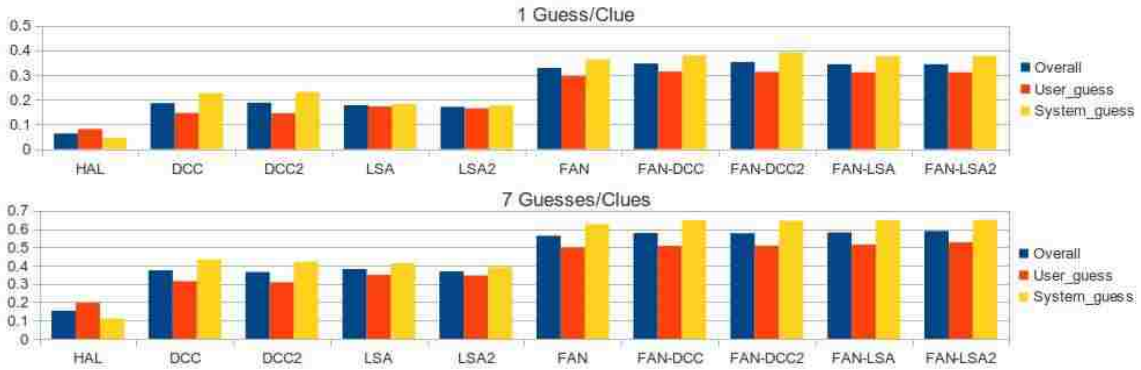


Figure 3.4: The win/loss record for several semantic models for each mode of the Wordlery game (higher is better). The combined models perform the best for each mode of the game. This chart corresponds to the data in Table 3.2.

At each round of the Wordlery game, the system randomly selects a word from the vocabulary of available words. The system then randomly selects a mode, either user_guess or system_guess. Finally, the system randomly selects one of the semantic models being evaluated, and the word guessing game is played by the user. To evaluate the effectiveness of each method we use the *win/loss record*, or the proportion of games in which the correct word was guessed to the total number of games played. We consider the win/loss record allowing all seven guesses/clues as well as the win/loss record allowing for only the first guess/clue.

In addition to evaluation, another purpose for playing these word guessing games is to provide a new method for gathering word associations from people. In the system_guess mode, the user provides clue words that are associated with the given word and each word entered is saved in a separate database as being associated with the given word. The idea is to gather word associations

that more accurately reflect how a person thinks about a concept as opposed to simply the first other word that comes to mind as is done with FANs. We have opted to not save any word associations for the user_guess mode as user guesses tend to be more inconsistent, especially when the user has no definite idea of what to guess.

There exist other online games for the purpose of collecting semantic information from people. For example, a game called *Wikispeedia* is an online game for inferring semantic distances between concepts [174]. Wikispeedia is played by randomly selecting two unrelated Wikipedia articles and having the user reach one of the articles from the other by clicking through hyperlinks in the articles encountered. The path the user takes is analyzed to derive a semantic distance between the concepts represented by the starting and ending articles.

3.4 Results

Initially over 2500 games were played by a variety of anonymous individuals through the online interface. This resulted in 7868 word associations gathered from the game. This data provides a snapshot of how people play the online game, and we can use this data to objectively evaluate the models by having each model play against the collected data (reenacting the human input). We first evaluate several variations of the models using this collected data. We then select the best models from this initial phase and have them play against humans.

3.4.1 Wordlery with Collected Data

We randomly selected 1500 unique words from the collected database and had each model play both modes of the word guessing game using the collected data to simulate the human input. Table 3.2 and Figure 3.4 shows the results for FANs, DCC, DCC2 (extended vocabulary), HAL, LSA, and LSA2 (extended vocabulary) as well as the results for the combination of FANs with DCC, DCC2, LSA, and LSA2 (named FAN-DCC, FAN-DCC2, FAN-LSA, and FAN-LSA2).

When considering only the individual models, FANs perform the best by a considerable margin, as expected, since this version of the game consists of simulating real human responses.

However, all the combined models perform better than the FANs. This shows that combining FANs and CSMs successfully takes advantage of their respective strengths. Surprisingly, increasing the vocabulary size (for LSA and DCC) has very little influence on the performance of the CSMs. This result is likely due to the fact that people tend to provide more common words as guesses/clues. These common words are likely to be included in the 19,310 words from the smaller vocabulary, and extending the vocabulary makes little difference.

Another surprising result to note is that DCC performs slightly better than LSA for the system_guess mode and the FAN-DCC combination performs slightly better than the FAN-LSA combination for both modes of the 1st guess version of the game. This suggests the possibility that the 1st order word correlations that DCC captures (e.g., 'dog','kennel') are better than (or at least comparable to) the higher order word correlations that LSA captures (e.g., 'dog','puppy') for this type of semantic task. The differences are not significant, and LSA has the edge for the 7 guesses version of the game, but the suggestion is there. Most semantic tasks in the literature deal with semantic similarity (like the TOEFL test), in which models like LSA usually perform well because they explicitly try to capture word synonymy/similarity. The word guessing game requires the ability to come up with an answer (free response) as opposed to multiple choice, which additional types of word associations (beyond just synonymy) can help facilitate.

3.4.2 Wordlery with People

We selected FANs, DCC, LSA, FAN-DCC2, and FAN-LSA2 to play Wordlery with people online since they represented the top models from the collected data experiments. Approximately 900 games were played by a variety of anonymous individuals through the online interface. On average, each of the five semantic models participated in 180 of the total rounds played. Table 3.3 and Figure 3.5 shows the results for the overall win/loss record, the user_guess win/loss record and the system_guess win/loss record.

1 Guess/Clue	DCC	LSA	FANs	FAN-DCC2	FAN-LSA2
Overall	0.232	0.325	0.354	**0.440**	0.418
User_guess	0.421	0.472	**0.577**	0.576	0.511
System_guess	0.089	0.113	0.161	0.296	**0.308**
7 Guesses/Clues	**DCC**	**LSA**	**FANs**	**FAN-DCC2**	**FAN-LSA2**
Overall	0.678	0.656	0.756	**0.837**	0.811
User_guess	0.684	0.607	0.722	**0.812**	0.728
System_guess	0.673	0.726	0.786	0.864	**0.910**

Table 3.3: The win/loss record for each of the five semantic models for each mode of the Wordlery game (higher is better). The FAN-DCC2 model performs the best overall, while FAN-LSA2 is a close second. However, FANs performs better on the single guess User_guess mode. Underlined scores denote statistical significance compared to the FANs model using the z proportionality test. This table corresponds to the results in Figure 3.5.

The first thing to note is that, overall, the FAN-DCC2 model achieves the best win/loss score (for both 1 guess and 7 guesses). The FAN-LSA2 model performs second best, with FANs not far behind. The CSMs (LSA and DCC) by themselves are significantly inferior to the FANs and combined models. The FANs do well presumably because those associations come directly from humans and hence can convey concepts back to human players. The combined methods take advantage of those human associations and then supplement them with the corpus inferred associations, which results in better performance. However, when allowing for only one guess on the the User_guess mode, FANs perform the best. We note again that the DCC-based models slightly outperform (or are close to) the LSA-based models on the user_guess mode, which promotes the usefulness of 1st-order correlations for certain semantic tasks such as this game.

When allowing for all 7 guesses/clues, the User_guess scores are generally lower than System_guess which suggests that, of the two modes, user_guess is harder. This makes sense since this mode only provides a static set of words as clues from which the user has to make a finite number of guesses. The interactive nature of the System_guess mode is likely one of the reasons for the better performance. However, the reason could also be that humans are good at providing relevant words as clues. Hence, we are collecting these word/clue pairs provided by the users for future studies. Perhaps the user_guess mode could be enhanced to allow the system to adapt its clues based on the user's guesses as one might do in a human-human game. Note that when only 1

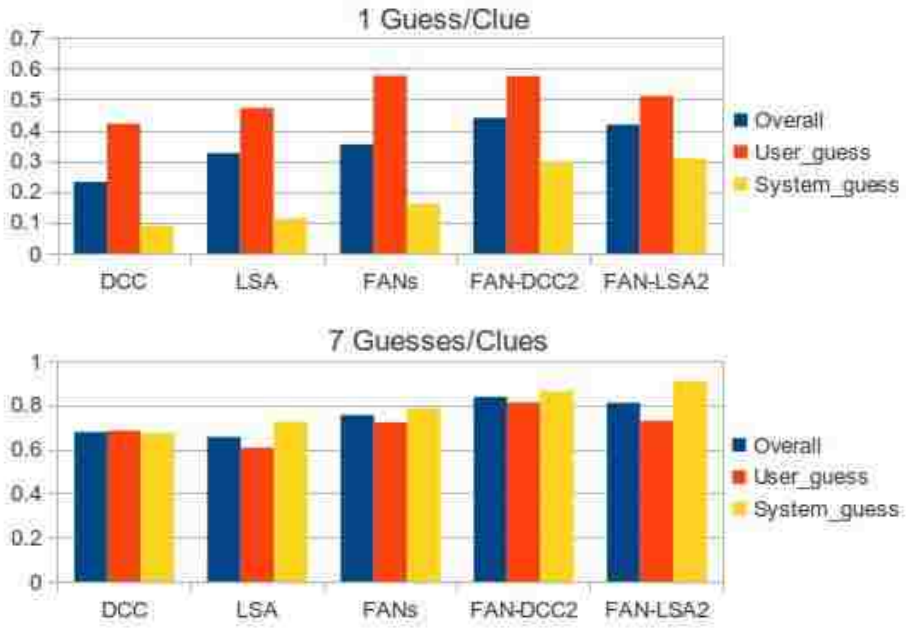


Figure 3.5: The win/loss record for each of the five semantic models for each mode of the Wordlery game (higher is better). The FAN-DCC2 model performs the best overall, while FAN-LSA2 is a close second. However, FANs performs better on the single guess User_guess mode. This chart corresponds to the data in Table 3.3.

guess/clue is allowed, the scores for the System_guess mode are very low. This is expected since the model is allowed only a single guess based on one clue, for which the FAN-LSA2 and FAN-DCC2 models performs the best by a significant margin.

3.5 Applications

The results of these games show that word associations can convey some aspect of the meaning of words. Such associations allow the system to communicate concepts to humans and allow humans to communicate concepts to the system, which is an important step in building a computational model that is capable of understanding language, processing input, making decisions, and interacting with people.

For example, in information retrieval, a user may not know what word to use in a query. The user could formulate a query using other words that describe (i.e., are associated with) the concept.

	DCC	LSA	FANs	FAN-DCC2	FAN-LSA2	Human
Score	0.09	0.15	0.18	0.18	0.18	**0.41**

Table 3.4: The results for determining the correct word given only its definition (higher is better). The low scores confirm the difficulty of the task.

In the query expansion process, the system could then automatically infer the concept and provide search results accordingly with possible improvements to both recall and precision.

To provide another example, and as an additional experiment, we randomly selected 100 word definitions from a dictionary. The system tokenized each definition into individual words (ignoring stop words), which essentially became "clue" words for the definition's corresponding word. Using the same method as in the System_guess mode of the Worldery game, each semantic model had to guess the word (or a synonym) that corresponded to each definition (allowing only one guess). For a baseline comparison, a couple of human volunteers performed the same task for all 100 definitions. Table 3.4 shows the results for the same models used in the Wordlery game and the average score for the human volunteers.

Keep in mind that the purpose of this experiment is not for rigorous testing but is a proof of concept to demonstrate a difficult task that requires semantic modeling. Even the human volunteers were not able to correctly determine the word for nearly 60% of the definitions. Although the semantic models performed relatively poorly, they do show potential. The ability of a computer system to recall a concept given a description shows some level of language understanding. On a larger scale this is analogous to topic modeling. For example, instead of the system answering the question, "to what is this definition referring?", it could answer the question, "what is this document about?".

Future applications could go beyond word-to-word associations and build associations between words and other objects (such as images), which can potentially expand the ability of the system to communicate and understand meaning in a variety of ways. For example, we plan to build a system that is capable of communicating ideas through visual art. For a concept such as

'freedom' the system will use the word and image associations to automatically compose an image that conveys the meaning of 'freedom' to the viewer.

3.6 Conclusions

We have experimented with two methods for obtaining word associations: through human free association norms (FANs) and by inferring them from a corpus (CSMs). We have also introduced a new semantic task to evaluate word associations by playing word guessing games. We compare the word associations from these methods and conclude that the FANs generally provide better quality associations than CSMs alone. Obviously, corpus-based approaches are heavily dependent on the corpus used. In future work, this influence could be assessed by evaluating word associations using the word guessing game from a variety of corpora. Our findings seem to be consistent with other studies that also show the superiority of FANs [131, 168]. It would seem that a universal corpus would be needed to discover word associations that are of the same quality as free association norms. But does a universal corpus exist? Or is it possible to create a model that can extract quality word associations from a standard corpus such as Wikipedia?

We have outlined a way to combine FANs with corpus-based semantic models. We use the word guessing game to show that combining the two methods of forming word associations is superior to each of the methods in isolation. This tells us that the CSMs have value and can complement the FANs. Perhaps for domain-specific tasks, preexisting databases of free association norms could provide a core of common human knowledge, while a domain-specific corpus and a CSM could be used to enhance the associations.

The word guessing games also provide a new way of gathering word associations from people. Once enough data has been collected, we will reevaluate the associations generated from the game comparing them with free association norms to see if they provide better quality associations for communicating meaning. In future work, we intend to incorporate more advanced corpus-based semantic models that take into account additional semantic information such as word order, sentence

structure, and relationship types. We believe this will improve the results on certain semantic tasks such as the definition-based word recall experiment.

Chapter 4

Autonomously Communicating Conceptual Knowledge Through Visual Art[1]

Abstract

In visual art, the communication of meaning or intent is an important part of eliciting an aesthetic experience in the viewer. Building on previous work, we present three additions to DARCI that enhances its ability to communicate concepts through the images it creates. The first addition is a model of semantic memory based on word associations for providing meaning to concepts. The second addition composes universal icons into a single image and renders the image to match an associated adjective. The third addition is a similarity metric that maintains recognizability while allowing for the introduction of artistic elements. We use an online survey to show that the system is successful at creating images that communicate concepts to human viewers.

[1]Derrall Heath, David Norton, and Dan Ventura, Autonomously Communicating Conceptual Knowledge Through Visual Art, *Proceedings of the 4th International Conference on Computational Creativity*, pp. 97-104, 2013

4.1 Introduction

DARCI (Digital ARtist Communicating Intention) is a system for generating original images that convey meaning. The system is part of ongoing research in the subfield of computational creativity, and is inspired by other artistic image generating systems such as AARON [109] and The Painting Fool [25]. Central to the design philosophy of DARCI is the notion that the communication of meaning in art is a necessary part of eliciting an aesthetic experience in the viewer [36]. DARCI is unique from other computationally creative systems in that DARCI creates images that explicitly express a given concept.

DARCI is composed of two major subsystems, an *image analysis* component, and an *image generation* component. The image analysis component learns how to annotate images with adjectives by training a series of neural networks with labeled images. The specific inputs to these neural networks, called *appreciation networks*, are global features extracted from each image, including information about the general occurrence of color, lighting, and texture in the images [121]. The image generation component uses a genetic algorithm, governed partly by the analysis component, to render a *source image* to visually convey an adjective [123]. While often effective, excessive filtering and extreme parameters can leave the source image unrecognizable.

In this paper we introduce new capabilities to DARCI—primarily, the ability to produce original source images rather than relying upon pre-existing, human-provided images. DARCI composes these original source images as a collage of iconic concepts in order to express a range of concepts beyond adjectives, similar to a recently introduced system for The Painting Fool that creates collages from the text of web documents [93]. However, in contrast to that system, ours creates collages from conceptual icons discovered with a semantic memory model. The resulting source images are then rendered according to an adjective discovered with this same semantic memory model. In order to preserve the content of the collages after rendering them, we introduce a variation on DARCI's traditional image rendering technique. Figure 4.1 outlines the two major components and their interaction, including the new elements presented in this paper. By polling

online volunteers, we show that with these additions, DARCI is capable of creating images that convey selected concepts while maintaining the aesthetics achieved with filters.

4.2 Methodology

Here we introduce the improvements to DARCI that enhance the system's capability to communicate intended meaning in an aesthetic fashion: a semantic memory model for broadening the range of concepts the system can communicate, an image composer for composing concrete representations of concepts into source images to be rendered, and a new metric for governing the evolution of the rendering process. We also describe an online survey that we use to evaluate the success of these additions.

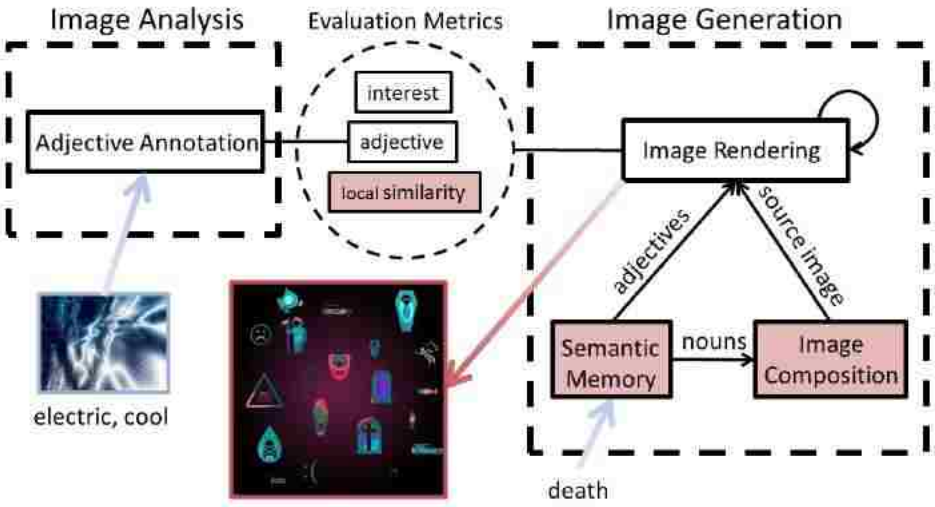


Figure 4.1: A diagram outlining the two major components of DARCI. *Image analysis* learns how to annotate new images with adjectives using a series of *appreciation networks* trained with labeled images. *Image generation* uses a *semantic memory* model to identify nouns and adjectives associated with a given concept. The nouns are composed into a source image that is rendered to reflect the adjectives, using a genetic algorithm that is governed by a set of evaluation metrics. The final product is an image that reflects the given concept. Additions from this paper are highlighted.

4.2.1 Semantic Memory Model

In cognitive psychology, the term *semantic memory* refers to the memory of meaning and other concept-based knowledge that allows people to consciously recall general information about the world. It is often argued that creativity requires intention (and we are certainly in this camp). In this context, we mean creativity in communicating a concept, and at least one part of this can be accommodated by an internal knowledge of the concept (i.e, a semantic memory).

The question of what gives words (or concepts) meaning has been debated for years; however, it is commonly agreed that a word, at least in part, is given meaning by how the word is used in conjunction with other words (i.e., its context) [51]. Many computational models of semantic memory consist of building associations between words [40, 152], and these word associations essentially form a large graph that is typically referred to as a*semantic network.* Associated words provide a level of meaning to a concept (word) and can be used to help convey its meaning.

Word associations are commonly acquired in one of two ways: from people and automatically by inferring them from a corpus. Here we describe a computational model of semantic memory that combines human free association norms with a simple corpus-based approach. The idea is to use the human word associations to capture general knowledge and then to fill in the gaps using the corpus method.

Lemmatization and Stop Words

In gathering word associations, we use the standard practice of removing stop words and lemmatizing. The latter process is accomplished using WordNet's [55] database of word forms; it should be noted, however, that lemmatization with WordNet has its limits. For example, we cannot lemmatize a word across different parts of speech. As a result, words like 'redeem' and 'redeeming' will remain separate concepts because 'redeeming' could be the gerund form of the verb 'redeem' or it could be an adjective (i.e., the act of 'a redeeming quality').

Free Association Norms

One of the most common means of gathering word associations from people is through *Free Association Norms* (FANs), which is done by asking hundreds of human volunteers to provide the first word that comes to mind when given a cue word. This technique is able to capture many different types of word associations including word co-ordination (pepper, salt), collocation (trash, can), super-ordination (insect, butterfly), synonymy (starving, hungry), and antonymy (good, bad). The association strength between two words is simply a count of the number of volunteers that said the second word given the first word. FANs are considered to be one of the best methods for understanding how people, in general, associate words in their own minds [118]. In our model we use two preexisting databases of FANs: The Edinburgh Associative Thesaurus [91] and the University of Florida's Word Association Norms [118].

Note that in this model we consider word associations to be undirected. In other words, if word A is associated with word B, then word B is associated with word A. Hence, when we encounter data in which word A is a cue for word B and word B is also a cue for word A, we combine them into a single association pair by adding their respective association strengths. Between these two databases, there are a total of 19,327 unique words and 288,069 unique associations. We refer to these associations as *human data*.

Corpus Inferred Associations

Discovering word associations from a corpus is typically accomplished using a family of techniques called *Vector Space Models* [159], which uses a matrix for keeping track of word counts either co-occurring with other words (term $\times$ term matrix) or within each document (term $\times$ document matrix).

One of the most popular vector space models is *Latent Semantic Analysis* (LSA) [42], based on the idea that similar words will appear in similar documents (or contexts). LSA builds a term$\times$ document matrix from a corpus and then performs Singular Value Decomposition (SVD), which essentially reduces the large sparse matrix to a low-rank approximation of that matrix along with

a set of vectors, each representing a word (as well as a set of vectors for each document). These vectors also represent points in semantic space, and the closer words are to each other in this space, the closer they are in meaning (and the stronger the association between words).

Another popular method is the *Hyperspace Analog to Language* (HAL) model [104]. This model is based on the same idea as LSA, except the notion of context is reduced more locally to a word co-occurrence window of ± 10 words instead of an entire document. Thus, the HAL model builds a term $\times$ term matrix of word co-occurrence counts from a corpus. HAL then uses the co-occurrence counts directly as vectors representing each word in semantic space. The size of the term $\times$ term matrix is invariant to the size of the corpus and has been argued to be more congruent to human cognition than the term $\times$ document matrix used in LSA [18, 168].

The corpus component of our model is constructed similarly to HAL but with some important differences. We restrict the model to the same number of unique words as the human-generated free associations, building a 19,327 $\times$ 19,327 (term $\times$ term) co-occurrence matrix M using a co-occurrence window of ± 50 words. To account for the fact that common words will have generally higher co-occurrence counts, we scale these counts by weighting each element of the matrix by the inverse of the total frequency of both words at each element. This is done by considering each element $M_{i,j}$, then adding the total number of occurrences of each word (i and j), subtracting out the value at $M_{i,j}$ (to avoid counting it twice), then dividing $M_{i,j}$ by this computed number, as follows:

$$M_{i,j} \leftarrow \frac{M_{i,j}}{(\sum_i M_{i,j} + \sum_j M_{i,j} - M_{i,j})} \tag{4.1}$$

The result could be a very small number, and therefore we then also normalize the values between 0 and 1.

For our corpus we use Wikipedia, as it is large, easily accessible, and covers a wide range of human knowledge [44]. Once the co-occurrence matrix is built from the entire text of Wikipedia, we use the weighted/normalized co-occurrence values themselves as association strengths between words. This approach works, since we only care about the strongest associations between words,

and it allows us to reduce the number of irrelevant associations by ignoring any word pairs with a co-occurrence count less than some threshold. We chose a threshold of 100 (before weighting), which provides a good balance of producing a sufficient number of associations, while reducing the number of irrelevant associations. When looking up a particular word, we return the top n other words with the highest weighted/normalized co-occurrence values. This method, which we will call *corpus data* from now on, gives a total of 4,908,352 unique associations.

Combining Word Associations

Since each source (human and corpus) provide different types of word associations, a combination of these methods into a single model has the potential to take advantage of the strengths of each method. The hypothesis is that the combined model will better communicate meaning to a person than either model individually because it presents a wider range of associations.

Our method merges the two separate databases into a single database before querying it for associations. This method assumes that the human data contains more valuable word associations than the corpus data because the human data is typically used as the gold standard in the literature. However, the corpus data does contain some valuable associations not present in the human data. The idea is to add the top n associations for each word from the corpus data to the human data but to weight the association strength low. This is beneficial for two reasons. First, if there are any associations that overlap, adding them again will strengthen the association in the combined database. Second, new associations not present in the human data will be added to the combined database and provide a greater variety of word associations. We keep the association strength low because we want the corpus data to reinforce, but not dominate, the human data.

To do this, we first copy all word associations from the human data to the combined database. Next, let W be the set of all 19,327 unique words, let $A_{i,n} \subseteq W$ be the set of the top n words associated with word $i \in W$ from the corpus data, let $score_{i,j}$ be the association strength between words i and j from the corpus data, let max_i be the maximum association score present in the

human data for word i, and let θ be a weight parameter. Now for each $i \in W$ and for each $j \in A_{i,n}$, the new association score between words i and j is computed as follows:

$$score_{i,j} \leftarrow (max_i \cdot \theta) \cdot score_{i,j} \tag{4.2}$$

This equation scales $score_{i,j}$ (which is already normalized) to lie between 0 and a certain percentage (θ) of max_i. The n associated words from the corpus are then added to the combined database with the updated scores. If the word pair is already in the database, then the updated score is added to the score already present. For the results presented in this paper we use $n = 20$ and $\theta = 0.2$, which were determined based on preliminary experiments. After the merge, the combined database contains 443,609 associations.

4.2.2 Image Composer

The semantic memory model can be considered to represent the meaning of a word as a (weighted) collection of other words. DARCI effectively makes use of this collection as a decomposition of a (high-level) concept into simpler concepts that together represent the whole, the idea being that in many cases, if a (sub)concept is simple enough, it can be represented visually with a single icon (e.g., the concept 'rock' can be visually represented with a picture of a 'rock'). Given such collection of iconic concepts, DARCI composes their visual representations (icons) into a single image. The image is then rendered to match some adjective associated with the original (collective) concept.

To represent these "simple enough" concepts, DARCI makes use of a collection of icons provided by *The Noun Project*, whose goal is to build a repository of symbols/icons that can be used as a visual language [154]. The icons are intended to be simple visual representations of any noun and are published by various artists under the Creative Commons license. Currently, The Noun Project provides 6,334 icons (each 420 $\times$ 420 pixels) representing 2,535 unique nouns and is constantly growing.

When given a concept, DARCI first uses the semantic memory model to retrieve all words associated with the given concept, including itself. These word associations are filtered by returning only nouns for which DARCI has icons and adjectives for which DARCI has appreciation networks. The nouns are sorted by association strength and the top 15 are kept. For each noun, multiple icons are usually available and one or two of these icons are are chosen at random to create a set of icons for use in composing the image.

The icons in the set are scaled to between 25% and 100% of their original size according to their association strength rank. Let I be the set of icons, and let $r : I \rightarrow [0, |I| - 1]$ be the rank of icon $i \in I$, where the icon with rank 0 corresponds to the noun with the highest association strength. Finally, let ϕ_i be the scaling factor for icon i, which is computed as follows:

$$\phi_i \leftarrow 1 - \frac{0.75}{|I| - 1} r(i) \tag{4.3}$$

An initial blank white image of size 2000 $\times$ 2000 pixels is created and the set of scaled icons are drawn onto the blank image at random locations, the only constraints being that no icons are allowed to overlap and no icons are allowed to extend beyond the border of the image. The result is a collage of icons that represents the original concept. DARCI then randomly selects an adjective from the set returned by the semantic memory model weighted by each adjective's association strength. DARCI uses its adjective rendering component, described in prior work, to render the collage image according to the selected adjective [76, 123, 124]. The final image will both be artistic and in some way communicate the concept to the viewer. Figure 4.1 shows how this process is incorporated into the full system.

4.2.3 Similarity Metric

To render an image, DARCI uses a genetic algorithm to discover a combination of filters that will render a source image (in this case, the collage) to match a specified adjective. The fitness function for this process combines an *adjective metric* and an *interest metric*. The former measures how effectively a potential rendering, or *phenotype*, communicates the adjective, and the latter measures

the "difference" between the phenotype and the source image. Both metrics use only global image features and so fail to capture important local image properties correlated with image content.

In this paper we introduce a third metric, *similarity*, that borrows from the growing research on *bag-of-visual-word* models [37, 149] to analyze local features, rather than global ones. Typically, these interest points are those points in an image that are the most surprising, or said another way, the least predictable. After an interest point is identified, it is described with a vector of features obtained by analyzing the region surrounding the point. *Visual words* are quantized local image features. A dictionary of visual words is defined for a domain by extracting local interest points from a large number of representative images and then clustering them (typically with k-means) by their features into n clusters, where n is the desired dictionary size. With this dictionary, visual words can be extracted from any image by determining which clusters the image's local interest points belong. A bag-of-visual-words for the image can then be created by organizing the visual word counts for the image into a fixed vector. This model is analogous to the bag-of-words construct for text documents in natural language processing.

For the new *similarity metric*, we first create a bag-of-visual-words for the source image and each phenotype, and then calculate the Euclidean distance between these two vectors. This metric has the effect of measuring the number of interest points that coincide between the two images.

We use the standard SURF (Speeded-Up Robust Features) detector and descriptor to extract interest points and their features from images [8]. SURF quickly identifies interest points using an approximation of the difference of Gaussians function, which will often identify corners and distinct edges within images. To describe each interest point, SURF first assigns an orientation to the interest point based on surrounding gradients. Then, relative to this orientation, SURF creates a 64 element feature vector by summing both the values and magnitudes of Haar wavelet responses in the horizontal and vertical directions for each square of a four by four grid centered on the point.

We build our visual word dictionary by extracting these SURF features from the database of universal icons mentioned previously. The 6334 icons result in more than two hundred thousand interest points which are then clustered into a dictionary of 1000 visual words using Elkan k-

means [49]. Once the Euclidean distance, d, between the source image's and the phenotype's bags-of-visual-words is calculated, the metric, S, is calculated to provide a value between 0 and 1 as follows:

$$S = MAX(\frac{d}{100}, 1)$$

where the constant 100 was chosen empirically.

4.2.4 Online Survey

Since our ultimate goal is a system that can create images that both communicate intention and are aesthetically interesting, we have developed a survey to test our most recent attempts at conveying concepts while rendering images that are perceived as creative.

The survey asks users to evaluate images generated for ten concepts across three rendering techniques. The ten concepts were chosen to cover a variety of abstract and concrete topics. The abstract concepts are 'adventure', 'love', 'music', 'religion', and 'war'. The concrete concepts are 'bear', 'cheese', 'computer', 'fire', and 'garden'.

We refer to the three rendering techniques as *unrendered*, *traditional*, and *advanced*. For *unrendered*, no rendering is applied—these are the plain collages. For the other two techniques, the images are rendered using one of two fitness functions to govern the genetic algorithm. For *traditional*, the fitness function is the average of the adjective and interest metrics. For *advanced* rendering, the new similarity metric is added. Here the adjective metric is weighted by 0.5, while the interest and similarity metrics are each weighted by 0.25. For each rendering technique and image, DARCI returned the 40 highest ranking images discovered over a period of 90 generations. We then selected from the pools of 40 for each concept and technique, the image that we felt best conveyed the intended concept while appearing aesthetically interesting. An example image that we selected from each rendering technique can be seen in Figure 4.2.

To query the users about each image, we followed the survey template that we developed previously to study the perceived creativity of images rendered with different adjectives [124]. In this study, we presented users with six five-point Likert items [100] per image; volunteers were

(a) unrendered (b) traditional (c) advanced

Figure 4.2: Example images for the three rendering techniques representing the concept ‘garden’.

(a) unrendered (b) traditional (c) advanced

Figure 4.3: Example dummy images for the concept ‘water’ that appeared in the survey for the indicated rendering techniques.

asked how strongly they agreed or disagreed (on a five point scale) with each statement as it pertained to one of DARCI’s images. The six statements we used were (abbreviation of item in parentheses):

I like the image. (*like*)

I think the image is novel. (*novel*)

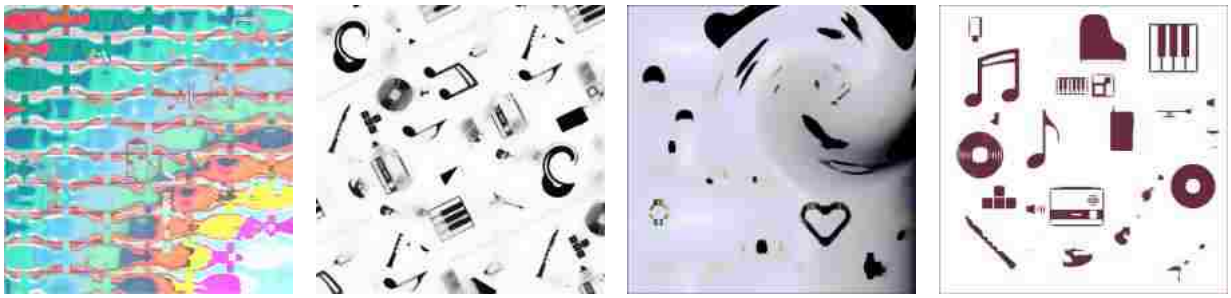
I would use the image as a desktop wallpaper. (*wallpaper*)

Prior to this survey, I have never seen an image like this one. (*never seen*)

I think the image would be difficult to create. (*difficult*)

I think the image is creative. (*creative*)

In previous work, we showed that the first five statements correlated strongly with the sixth, “I think the image is creative” [124], justifying this test as an accurate evaluation of an image’s



(a) (b) (c) (d)

Figure 4.4: The images that were rated the highest on average for each statement. Image (a) is the advanced rendering of 'adventure' and was rated highest for *like*, *novel*, *difficult*, and *creative*. Image (b) is the traditional rendering of 'music' and was rated highest for *wallpaper*. Image (c) is the advanced rendering of 'love' and was rated highest for *never seen*. Image (d) is the advanced rendering of 'music' and was rated highest for *concept*.

subjective creativity. In this paper, we use the same six Likert items and add a seventh to determine how effective the images are at conveying their intended concept:

I think the image represents the concept of "____." (*concept*)

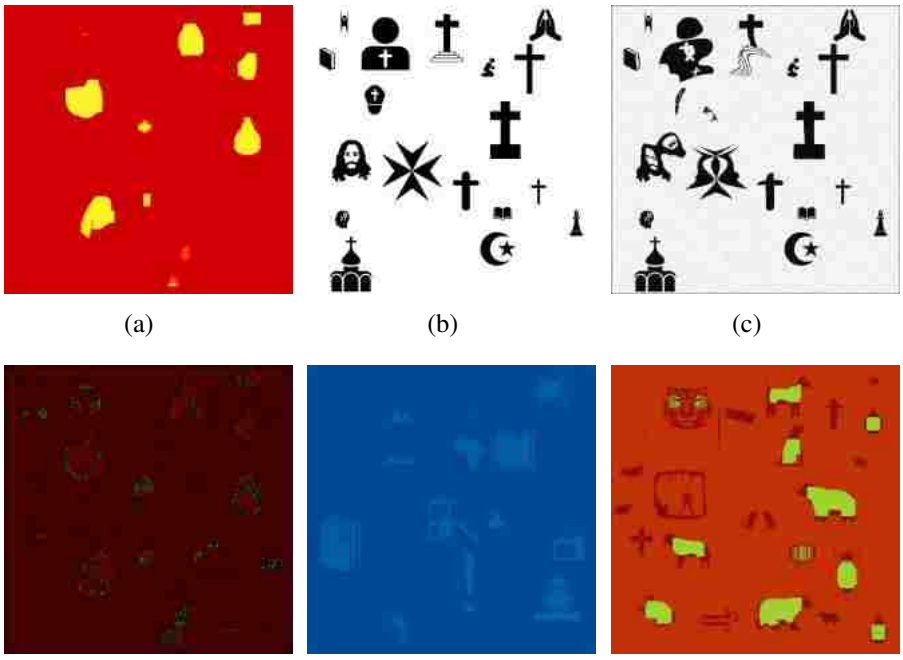
To avoid fatigue, volunteers were only presented with images from one of the three rendering techniques mentioned previously. The technique was chosen randomly and then the images were presented to the user in a random order. To help gauge the results, three dummy images were introduced into the survey for each technique. These dummy images were created for arbitrary concepts and then assigned different arbitrary concepts for the survey so that the image contents would not match their label. Unfiltered dummy collages were added to the unrendered set of images, while traditionally rendered versions were added to the traditional and advanced sets of images. The three concepts used to generate the dummy images were: 'alien', 'fruit', and 'ice'. The three concepts that were used to describe these images in the survey were respectively: 'restaurant', 'water', and 'freedom'. To avoid confusion, from here on we will always refer to these dummy images by their description word. The dummy images for the concept of 'water' are shown in Figure 4.3. In total, each volunteer was presented with 13 images.

4.3 Results

A total of 119 anonymous individuals participated in the online survey. Volunteers could quit the survey at anytime, thus not evaluating all 13 images. Each person evaluated an average of 9 images and each image was evaluated by an average of 27 people. The highest and lowest rated images for each question can be seen in Figures 4.4 and 4.5 respectively.

The three dummy images for each rendering technique are used as a baseline for the concept statement. The results of the dummy images versus the valid images are show in Figure 4.6. The average concept rating for the valid images is significantly better than the dummy images, which shows that the intended meaning is successfully conveyed to human viewers more reliably than an arbitrary image. These results confirm that the intelligent use of iconic concepts is beneficial for the visual communication of meaning. Further, it is suggestive that the ratings for the other statements are generally lower for the dummy images than for the valid images. Since the the dummy images were created for a different concept than the one which they purport to convey in the survey, this may be taken as evidence that successful conceptual or intentional communication is an important factor for the attribution of creativity.

The results of the three rendering techniques (unrendered, traditional, and advanced) for all seven statements are shown in Figure 4.7. The unrendered images are generally the most successful at communicating the intended concepts. This is likely because the objects/icons in the unrendered images are left undisturbed and are therefore more clear and discernible, requiring the least perceptual effort by the viewer. The rendered images (traditional and advanced) often distort the icons in ways that make them less cohesive and less discernible and can thus obfuscate the intended meaning. The trade-off, of course, is that the unrendered images are generally considered less likable, less novel, and less creative than the rendered images. The advanced images are generally considered more novel and creative than the traditional images, but the traditional images are liked slightly more. The advanced images also convey the intended meaning more reliably than



(a) (b) (c)

(d) (e) (f)

Figure 4.5: The images that were rated the lowest on average for each statement. Image (a) is the advanced rendering of 'fire' and was rated lowest for *difficult* and *creative*. Images (b) and (c) are the unrendered and advanced version of 'religion' and were rated lowest for *neverseen* and *wallpaper* respectively. Images (d), (e), and (f) are the traditional renderings of 'fire', 'adventure', and 'bear', respectively, and were rated lowest for *like*, *novel*, and *concept* respectively.

the traditional images, which indicates that the similarity metric is finding a better balance between adding artistic elements and maintaining icon recognizability.

The difference between the traditional and advanced rendering was minimized by the fact that we selected the image (out of DARCI's top 40) from each group that best conveyed the concept while also being aesthetically interesting. Out of all the traditional images, 39% had at least one recognizable icon, while 74% of the advanced images had at least one recognizable icon. This difference demonstrates that the new similarity metric helps to preserve the icons and provides a greater selection of good images from which to choose, which is consistent with the results of the survey. For comparison, Figure 4.8 shows some example images (both traditional and advanced) that were not chosen for the survey.

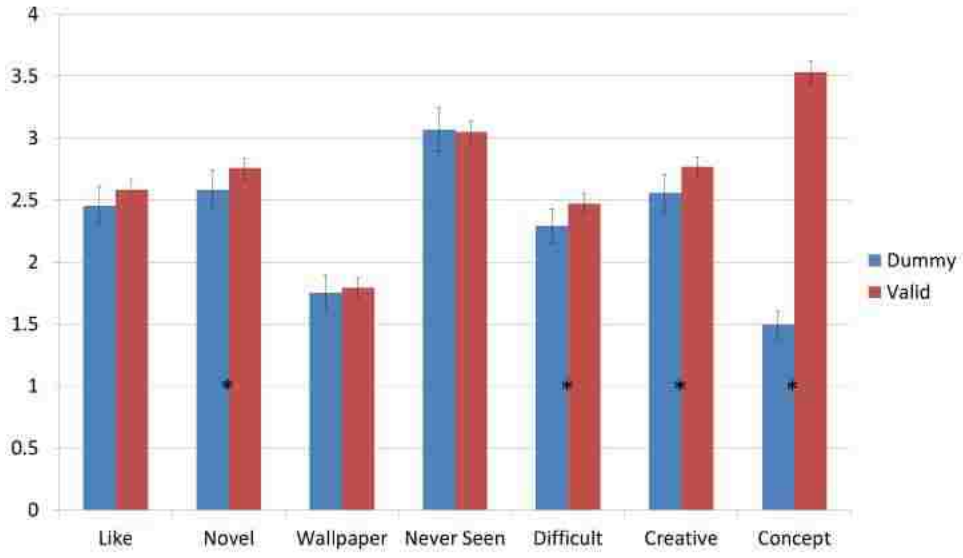


Figure 4.6: The average rating from the online survey for all seven statements comparing the dummy images with the valid images. The valid images were more successful at conveying the intended concept than the dummy images by a significant margin. Results marked with an asterix (*) indicate statistical significance using the two tailed independent t-test. The lines at the top of each bar show the 95% confidence interval for each value. The sample sizes for dummy and valid images are 251 and 818 respectively.

The results comparing the abstract concepts with the concrete concepts are shown in Figure 4.9. For all seven statements, the abstract concepts are, on average, rated higher than the concrete concepts. One possible reason for this is that concrete concepts are not easily decomposed into a collection of iconic concepts because, being concrete, they are more likely to be iconic themselves. For concrete concepts, the nouns returned by the semantic memory model are usually other related concrete concepts, and it becomes difficult to tell which object is the concept in question. For example, the concept 'bear' returns nouns like 'cave', 'tiger', 'forest', and 'wolf', which are all related, but don't provide much indication that the intended concept is 'bear'. A person might be inclined to generalize to a concept such as 'wildlife'. Another possible reason why abstract concepts result in better survey results than do concrete concepts is because abstract concepts allow a wider range of interpretation and are generally more interesting. For example, the concept 'cheese' would generally be considered straightforward to most people, while the concept 'love' could have variable meanings to different people in different circumstances. Hence, the images generated for

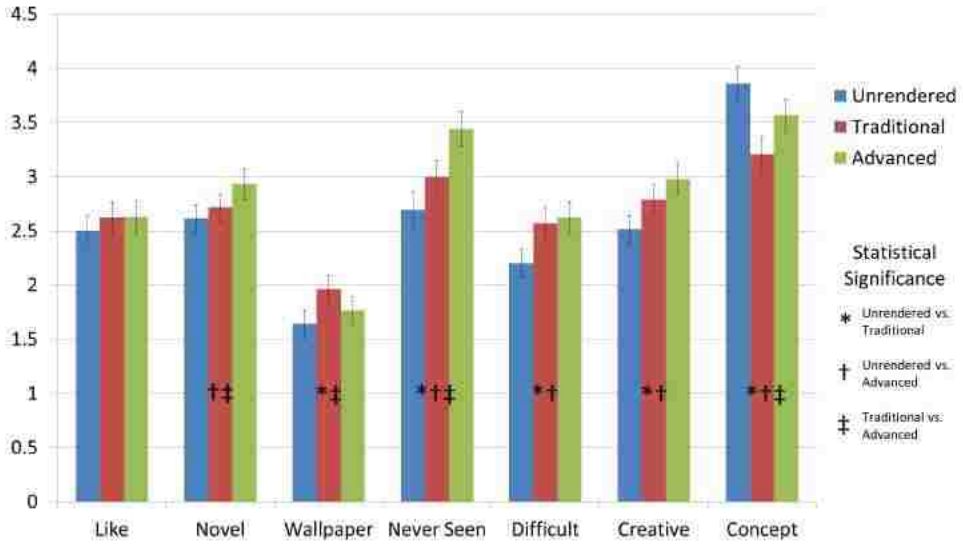


Figure 4.7: The average rating from the online survey for all seven statements comparing the three rendering techniques. The unrendered technique is most successful at representing the concept, while the advanced technique is generally considered more novel and creative. Statistical significance was calculated using the two tailed independent t-test. The lines at the top of each bar show the 95% confidence interval for each value. The sample sizes for the unrendered, traditional, and advanced techniques are 256, 285, and 277 respectively.

abstract concepts are generally considered more likable, more novel, and more creative than the concrete images.

4.4 Conclusions and Future Work

We have presented three additions to the computer system, DARCI, that enhance the system's ability to communicate specified concepts through the images it creates. The first addition is a model of semantic memory that provides conceptual knowledge necessary for determining how to compose and render an image by allowing the system to make decisions and reason (in a limited manner) about common world knowledge. The second addition uses the word associations from a semantic memory model to retrieve conceptual icons and composes them into a single image, which is then rendered in the manner of an associated adjective. The third addition is a new similarity metric used

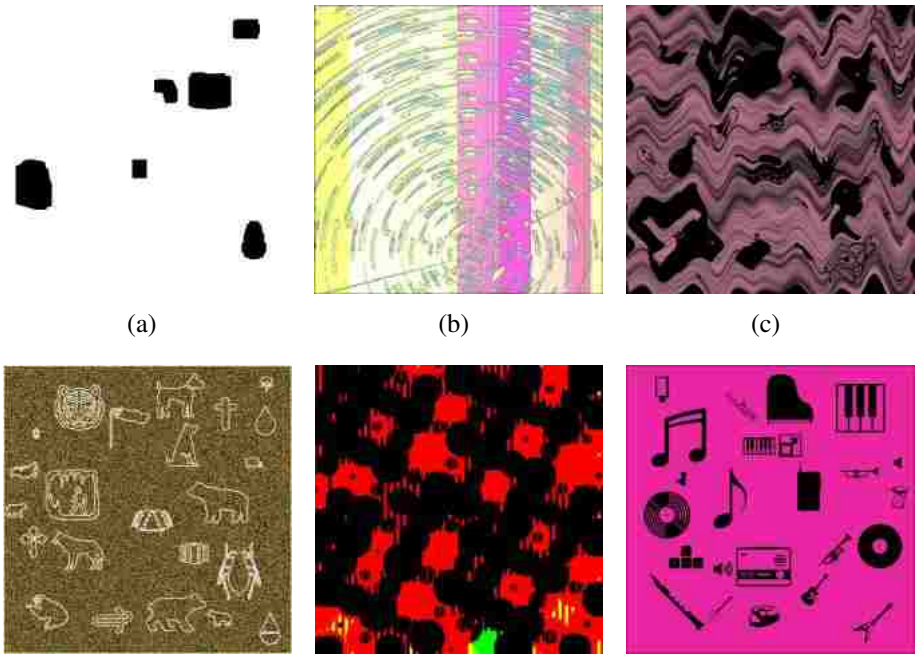


Figure 4.8: Sample images that were not chosen for the online survey. Images (a), (b), and (c) are traditional renderings of ‘adventure’, ‘love’, and ‘war’ respectively. Images (d), (e), and (f) are advanced renderings of ‘bear’, ‘fire’, and ‘music’ respectively.

during the adjective rendering phase that preserves the discernibility of the icons while allowing for the introduction of artistic elements.

We used an online survey to evaluate the system and show that DARCI is significantly better at expressing the meaning of concepts through the images it creates than an arbitrary image. We show that the new similarity metric allows DARCI to find a better balance between adding interesting artistic qualities and keeping the icons/objects recognizable. We show that using word associations and universal icons in an intelligent way is beneficial for conveying meaning to human viewers. Finally, we show that there is some degree of correlation between how well an image communicates the intended concept and how well liked, how novel, and how creative the image is considered to be. To further illustrate DARCI’s potential, Figure 4.10 shows additional images encountered during various experiments with DARCI that we thought were particularly interesting.

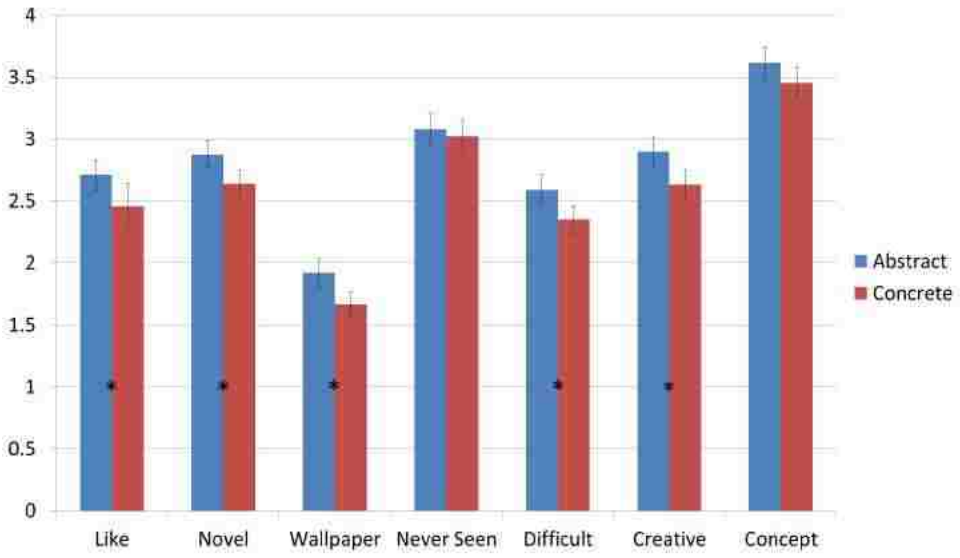


Figure 4.9: The average rating from the online survey for all seven statements comparing the abstract concepts with the concrete concepts. The abstract concepts generally received higher ratings for all seven statements. Results marked with an asterix (*) indicate statistical significance using the two tailed independent t-test. The lines at the top of each bar show the 95% confidence interval for each value. The sample sizes for abstract and concrete concepts are 410 and 408 respectively.

In future research we plan to do a direct comparison of the images created by DARCI with images created by human artists and to further investigate how semantic memory contributes to the creative process. We plan to improve the semantic memory model by going beyond word-to-word associations and building associations between words and other objects (such as images). This will require expanding DARCI's image analysis capability to include some level of image noun annotation. The similarity metric presented in this paper is a step in that direction. An improved semantic memory model could also help enable DARCI to discover its own topics (i.e., find its own inspiration) and to compose icons together in more meaningful ways, by intentional choice of absolute and relative icon placement, for example.

(a) bear (b) murder (c) war

Figure 4.10: Notable images rendered by DARCI during various experiments and trials.

Chapter 5

Imagining Imagination: A Computational Framework Using Associative Memory Models and Vector Space Models[1]

Abstract

Imagination is considered an important component of the creative process, and many psychologists agree that imagination is based on our perceptions, experiences, and conceptual knowledge, recombining them into novel ideas and impressions never before experienced. As an attempt to model this account of imagination, we introduce the Associative Conceptual Imagination (ACI) framework that uses associative memory models in conjunction with vector space models. ACI is a framework for learning conceptual knowledge and then learning associations between those concepts and artifacts, which facilitates imagining and then creating new and interesting artifacts. We discuss the implications of this framework, its creative potential, and possible ways to implement it in practice. We then demonstrate an initial prototype that can imagine and then generate simple images.

[1]Derrall Heath, Aaron Dennis, and Dan Ventura, Imagining Imagination: A Computational Framework Using Associative Memory Models and Vector Space Models,*Proceedings of the 6th International Conference on Computational Creativity*, pp. 244–251, 2015

5.1 Introduction

The concept of imagination is not often talked about in cognitive psychology without reference to creativity [59, 167]. In fact, the term 'imaginative' is many times used as a synonym for 'creative'. Defining imagination, like creativity, is difficult because the word is used broadly and depends on the audience, the level of granularity, and the context [151]. In cognitive psychology, imagination is commonly generalized as thinking of something (real or not) that is not present to the senses [10]. In terms of creativity, it is being able to conceive of and conceptualize novel ideas. Imagination, thus it seems, should be an important consideration when developing creative systems.

In the field of computational creativity, imagination is discussed explicitly only on rare occasions, such as Colton's creative tripod [24]. Most creative systems incorporate imagination implicitly and do not model it directly. In this paper, we propose a computational framework that attempts to explicitly model imagination in order to perform creative tasks. Our framework, called the Associative Conceptual Imagination (ACI) framework, uses associative memory models (AMMs) combined with vector space models (VSMs) to enable the system to imagine and then create novel and interesting artifacts.

We begin by looking more closely at the psychology literature in order to establish a cognitive basis for imagination, which will motivate the design of our framework. We then consider how current computational models of creativity both succeed and fail at addressing imagination. We then outline in detail the ACI framework for imagination and demonstrate an initial implementation (proof-of-concept) in the domain of visual art. Finally, we discuss the possibilities this framework can afford us in building creative systems and talk about questions regarding its application.

5.1.1 Psychology of Imagination

Imagination is ubiquitous in everyday life. We can visually imagine a world described through narrative, or imagine how to get to the grocery store, or imagine what it would be like to be a celebrity. We can imagine what a lion crossed with an eagle could look like, or imagine new ways to express meaning through art. Although most often thought of as visualizing in the mind, we can

imagine in conjunction with any of our senses. Indeed, we can talk about imagination across the whole range of human experience. Imagination is a broad term with many different taxonomies and ways to interpret it. We restrict our view to two major types of imagination that are commonly used by psychologists [38].

The first type of imagination is *sensory* (or reproductive) imagination. This is mentally recalling past experience, which is directly related to our memories. For example, one can imagine what their favorite food tastes like without actually tasting the food, or imagine their mother's face when she is not present, or imagine an annoying song that is stuck in one's head. This type of imagination can be thought of as creative in the sense of *recreating* in one's mind a previous experience.

The second type of imagination is *creative* (or productive) imagination. It is the ability to combine ideas in different ways never before observed, or the ability to think about the world from a different perspective than previously experienced. For example, one can imagine what a hairy banana monster could look like, or what life would be like if born in another country, or imagine how to compose music that is happy and uplifting. This type of imagination is more clearly tied to creativity and some have argued that it forms a necessary basis for creativity [167], while others have argued that imagination is merely a tool used in the creative process [59].

Most psychologists agree that our senses, our conceptual knowledge, and our memories form the bases of imagination [7, 10]. As we perceive the world and have experiences, we create memories by establishing and strengthening connections in our mind. These connections form concepts, which are in turn interconnected. Memories are often argued to be distributed and content addressable across groups of neurons [58]. This means that multiple neurons respond in varying strengths to certain experiences, different experiences may activate overlapping neurons, and similar experiences will have more overlapping neurons than dissimilar experiences. This distributed memory allows the brain to implicitly associate concepts and experiences together.

Thus we have associations between concepts (e.g., rain is related to water) and between what we perceive and these concepts (e.g., apples look round and are typically reddish in color). Creative

imagination cannot make something out of nothing, nor is it random; everything we imagine is anchored to things we have actually experienced in the past and on their connections [167]. The novelty is in combining these experiences in different ways. When a chef imagines new recipes, she uses her knowledge of existing recipes, ingredients, methods, and kitchen tools. The new recipe is essentially a recombination of this previous information in a novel and (hopefully) delicious way.

A computational model of imagination should address the abilities to perceive, to create memories, and to learn associations between concepts. Such a model should then be able to reconstruct this information (sensory imagination), as well as recombine this information in novel ways to create new and interesting things never before experienced (creative imagination).

5.1.2 Related Work

In accounting for creativity in computational systems, Colton was one of the first to explicitly mention imagination as part of the creative process [24]. In order for a system to have imagination, it should be able to produce artifacts that are novel. Others have mentioned imagination in relation to a creative system that produces narratives [178].

A computational system that explicitly tries to model imagination is SOILIE (Science Of Imagination Laboratory Imagination Engine) [15]. SOILIE maintains a large database of labeled images, and words are associated together when they appear as co-occurring labels. For example, a picture of a face could be labeled with 'face', 'ear', 'mouth', etc. and the system learns to associate those labels together. A word is given to the system which then finds 5-10 associated words and creates a collage out of images that have been labeled with those associated words. This system demonstrates a rudimentary form of sensory imagination in which it tries to recreate an image of the inputed word. SOILIE is similar to one of the abilities of the Painting Fool, which can extract key words from a text document and create a collage by finding images of those key words in a database [93].

Creative imagination was partially demonstrated in a system that used recurrent neural networks to produce melodies according to a set of other melodies arranged on a 2D plane [155].

Each of the melodies in the training set were tied to a specific 2D location, and the model was trained to reproduce each melody at their respective locations. After training, the system would be given a new location on the 2D plane and could essentially interpolate a new melody according to its proximity to the original set of melodies. This is the beginnings of creative imagination in that the system is blending melodies together according to spacial proximity.

Imagination has been mentioned in conjunction with systems that perform conceptual blending to produce metaphors and narratives [41, 162, 178]. Conceptual blending is the process of taking two input mental spaces (representing concepts) and mixing them together to make a blended mental space that is novel, meaningful, and has emergent structure (e.g., lightsaber is a blend of sword and laser) [54]. Computational models of conceptual blending have been used to produce narrative [132], poetry [71], and even mathematical axioms [108].

Conceptual blending certainly has potential for imagination as it explicitly attempts to blend conceptual knowledge into novel ideas. Although there are still many technical challenges in autonomously blending input spaces, conceptual blending does seem to address creative imagination. Unfortunately, most implementations do not consider sensory information and the input spaces are typically hand engineered, so the system does not learn from experience and cannot imagine sensory type artifacts. However, one computational system does try to implement conceptual blending with images [150]. The system takes two pictures that each represent a concept and blends them by extracting commonly shaped objects in one image and pasting them over similarly shaped objects in the other image (e.g., a globe in one image is pasted over a bicycle tire in another image).

Evolutionary computation is a common method incorporated into creative systems because of its innate ability to yield unpredictable yet acceptable results [61]. Indeed, evolutionary computation seems to at least partially model creative imagination in that it recombines and modifies existing artifacts through crossover/mutation and can, thus, diverge and discover novel artifacts. The fitness function also guides the evolutionary process to converge on quality results. Many systems incorporate the use of evolutionary techniques to produce artifacts in domains such as visual art [46, 106, 124], music [113], and semantic networks [9].

Evolutionary computation appears to have potential in addressing both sensory and creative imagination. However, the creative intent seems to reside solely in the fitness function, which is separated from the actual generation of artifacts. The creation of artifacts is an independently random event that is not connected to any associations learned through experience (except for maybe the population of artifacts themselves). The act of imagination in this case is mostly a selection/filtering process, which, although viable, doesn't seem to capture the complete picture. In its basic form at least, there is no notion of associations between concepts and artifacts.

5.2 Associative Conceptual Imagination

We attempt to explicitly model imagination through a computational framework called the Associative Conceptual Imagination (ACI) framework. ACI uses ideas from other domains in a novel way that is capable of both sensory and creative imagination. ACI is composed of two major types of components, a vector space model and associative memory models as shown in Figure 5.1. We will discuss the major components of the ACI framework, how they interact to perform various imaginative tasks, and the creative potential of systems built using this framework.

5.2.1 Vector Space Model

Creativity is valued not just because of the novelty of things created, but also because of their utility. For example, in domains such as visual art, the value is in how the art conveys meaning to the viewers [36]. There is an element of intentionality as an artist purposefully expresses meaning through art. How can an artist intentionally express meaning without having knowledge of the world and of what things mean? Conceptual knowledge helps to provide a foundation for the ability to imagine and create. Incorporating conceptual knowledge into a creative system can potentially be achieved through Vector Space Models (VSMs) [159].

It is commonly agreed that a word (or concept), at least in part, is given meaning by how the concept is used in conjunction with other words (i.e., its context) [96]. Vector space models take

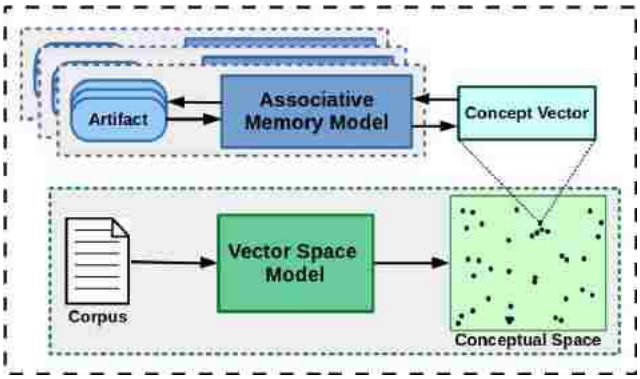


Figure 5.1: An overview of the Associative Conceptual Imagination framework. The vector space model learns, from a large corpus, how to encode semantic information into concept vectors that populate conceptual space. Multiple associative memory models can then learn associations between these concept vectors and example artifacts from various domains, such as art, music, or recipes. These associative memory models are bi-directional and can not only discriminate, but also generate artifacts according to a given concept vector. The semantic structure encoded in the concept vectors allows the framework to facilitate the imagining of artifacts according to concepts for which it has never seen examples.

advantage of this by analyzing large corpora and learning multi-dimensional vector representations for each concept that encode such semantic information. These models are based on the idea that similar words will occur in similar contexts and words that are often associated together will often co-occur close together. These models reduce words to a vector representation that can be compared to other word vectors. VSMs have been successfully used on a variety of tasks such as information retrieval [143], multiple choice vocabulary tests [43], TOEFL multiple choice synonym questions [137], and multiple choice analogy questions from the SAT test [158].

Concepts similar in meaning will have vectors that are close to each other in "vector space", which we will refer to as *conceptual space*. Associations between concepts are implicitly encoded by their proximity in conceptual space. Figure 5.2 shows relationships between example word vectors that correspond to various topics projected onto a 2D plane. These concept vectors capture other interesting semantic relationships that are consistent with arithmetic operations. For example, $vector(\text{“}king\text{”}) - vector(\text{“}man\text{”}) + vector(\text{“}woman\text{”})$ results in a vector that is closest to $vector(\text{“}queen\text{”})$.

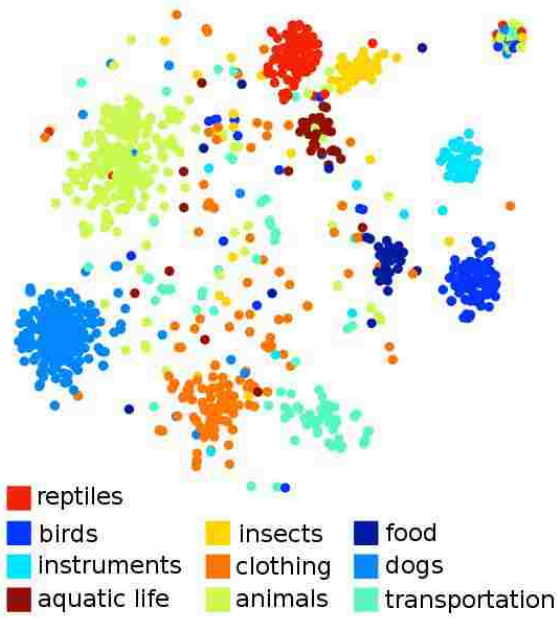


Figure 5.2: A 2D visualization (projected from high dimensional space) of several word vectors color coded by topics. These concept vectors were learned using the skip-gram VSM, which was incorporated into the DeViSE model (visualization courtesy of Frome et al. 2013). Note that concepts from similar topics generally cluster together because the concept vectors encode semantic relationships.

The potential of VSMs in creative systems has been discussed before, and we aim to make use of them in this framework [110]. The semantic information encoded in the vectors provides a form of conceptual knowledge to the ACI framework, which will help provide a basis for imagination.

5.2.2 Associative Memory Models

In addition to knowing how concepts relate to each other, the ACI framework needs to allow understanding of how concepts relate to actual artifacts. In other words, ACI systems should be able to perceive and observe the world (i.e., to be grounded in sensory information). ACI incorporates Associative Memory Models (AMMs) to learn how to associate artifacts with concept vectors. For example, models built using ACI can learn what a 'cat' looks like by observing pictures of 'cats', or learn what a 'car' sounds like by listening to sound files of 'cars'.

Here we use "associative memory model" as a generic term that refers to any computational model or algorithm that is capable of learning bi-directional relationships between artifacts and concept vectors. Not only should the AMM be capable of predicting the appropriate concept vector given an artifact, but it should also be capable of going the other direction and producing an artifact given a concept vector. Of course, the quality of learning will be dependent on the quality and quantity of labeled training data, as well as on the characteristics of the particular associative memory model that is chosen.

Bidirectional associative memory models (BAMs) seem like an obvious possible choice to implement the AMM [92]. A BAM is a type of recurrent neural network that learns to bidirectionally map one set of patterns to another set of patterns. Given an artifact (encoded into a pattern), a BAM could return the appropriate concept vector. Conversely, given a concept vector, a BAM could return an appropriate artifact, which can essentially be thought of as performing sensory imagination. Variations of BAMs have been used in computational creativity to associate input patterns to features in order to model the phenomenon of surprise [12].

Another family of algorithms that have potential use in the ACI framework are probabilistic generative models. These models learn a joint distribution for observed data and their respective labels/classes. Once trained, not only can these models classify new data, but they can also be used generatively to create new instances of data that correspond to a particular label. For example, a Deep Belief Network (DBN) is a generative model that can also be thought of as a deep neural network in which several layers of nodes (or latent variables) are connected by weights from neighboring layers, while nodes of the same layer are not connected [78]. Hinton *et al.* used DBNs to classify images of handwritten digits (0-9) by training on several examples and then used them generatively to "imagine" what a 2 looks like by creating several images that each uniquely looked like a handwritten two, thus demonstrating a form of sensory imagination.

Another generative model uses a hierarchical approach to recognize and then generate unique images of handwritten symbols, again demonstrating sensory imagination [95]. Sum Product Networks (SPNs) have also been used to learn bidirectional associations between patterns [135].

Given a picture of half a face, SPNs were able to infer (or imagine) the other half. These generative models can often be applied directly to the raw inputs (i.e., directly to pixels in an image) and thus seem to exhibit advanced perceptual abilities and in turn can generate artifacts directly.

The associative memory model implementation is not limited to a single model, but could be split into separate discriminative and generative parts. A machine learning algorithm could be the discriminative part and be trained to predict a given artifact's concept vector (e.g., given a 'sad' melody, the learning algorithm predicts the 'sad' vector). The generative part could be implemented by a genetic algorithm that uses the discriminative model as the fitness function. For example, a genetic algorithm could be given the 'sad' vector to imagine a 'sad' melody, and the discriminative model knows what characteristics a 'sad' melody should have and could then guide the evolutionary process.

Other specific associative memory models could be incorporated depending on the domain, its representation, and available training data. Additionally, multiple AMMs for different domains could be incorporated into the framework simultaneously (i.e., one model learns images while another learns sounds for each concept), with the AMMs then indirectly related through conceptual space.

5.2.3 Performing Imagination

Once an implementation of the ACI framework has its components in place and properly trained, it is ready to imagine, and even create, artifacts. To perform sensory imagination, an ACI model can generate artifacts for a particular concept that it has previously learned. For instance, after having seen example images of 'cats', the system has learned an internal representation for what a 'cat' looks like. The associative memory model can then start with the 'cat' concept vector and generate a unique image that would likely be associated with the 'cat' vector, presumably an image of a 'cat' (see Figure 5.3(a)). In the case of using probabilistic generative models, the probabilistic nature of the model and the distribution of various poses, angles, and colors learned from the many example 'cat' images allow the system to generate a unique 'cat' image each time.

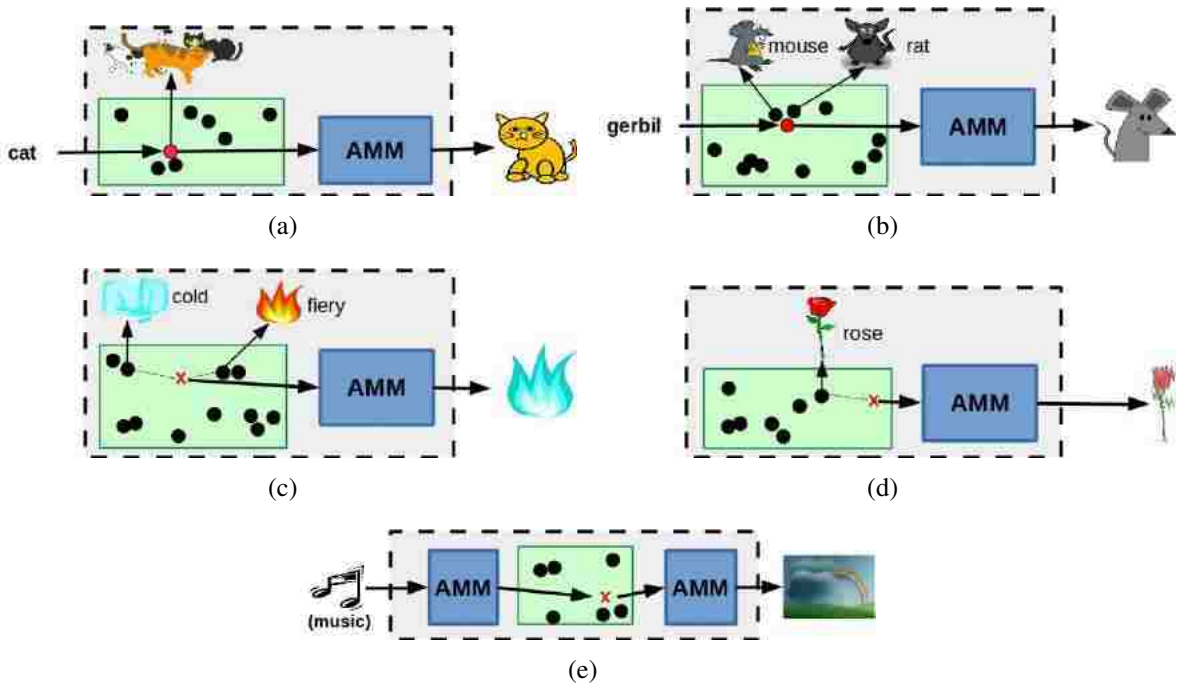


Figure 5.3: Different ways the Associative Conceptual Imagination framework can be used to imagine artifacts. The green rectangle with black dots represents concept vectors in conceptual space, which are learned from a vector space model. The Associative Memory Model (AMM) associates concept vectors to artifacts. The framework allows the imagining of artifacts for concepts it has previously observed (a). It can facilitate the imagining of artifacts for concepts it has not previously observed but that are similar to other concepts that is has observed (b). The framework allows the imagining of artifacts that are combinations of two (or more) previously observed concepts (c). Models based on ACI can imagine changes to a previously observed concept (d). Finally, the framework can facilitate imagination across different domains by observing an artifact in one domain and then imagining a related artifact in another domain (e).

To perform creative imagination, the framework takes inspiration from the DeViSE model, which uses VSMs to aid in correctly recognizing images of objects [57]. The DeViSE model first learns word vectors from a large corpus using a VSM. The model is then trained with raw image pixels using a deep convolutional neural network that learns to predict the correct labels' vector (instead of the label directly). Cosine similarity is performed between the predicted vector and the other word vectors to determine what the correct label should be. Since the vectors encode semantic relationships between concepts, the model can successfully label an image with a word for which it has never seen example images (called zero-shot prediction). For example, the system may have

been trained on images of 'rats' and 'mice' but not on images labeled 'gerbil'. Given a picture of a 'gerbil' the model can still successfully label it as such because a 'gerbil' is similar (according to the VSM) to a 'rat' and a 'mouse'.

Replacing the convolutional neural network with, say, a probabilistic generative model could allow the system to act in reverse. We could input the vector for 'gerbil' and the system could imagine what a 'gerbil' looks like without having ever seen a picture of a 'gerbil', because of the semantic knowledge encoded in the vectors (see Figure 5.3(b). Similarly, the system could take advantage of the semantic structure of the VSM and imagine what a concept sounds like without having heard any example sounds for that concept. For example, the system could have been trained on sounds for 'horses', 'tractors', 'dogs', and 'trumpets', but not have been exposed to any sounds for 'donkeys'. Yet, the system could still generate a unique sound for a 'donkey'. The result may not sound exactly like a 'donkey', but it will sound closer to a 'horse' than to the other concepts because the system knows that 'donkeys' are more similar to 'horses' than to the other concepts. An ACI model can imagine its own 'donkey' sound in a way that is novel, yet still reasonable by leveraging semantic information gained through the VSM and transferring it to the task of generating sound.

In another situation, a system based on ACI can imagine what a combination of concepts could look like by starting with a vector that is in between concepts in conceptual space. As shown in Figure 5.3(c), the system could imagine what a 'cold' and 'fiery' image looks like by starting with a vector part-way between the 'cold' and 'fiery' vectors. The system should generate a novel image that is some blending of the two concepts (and perhaps other surrounding concepts). The system is essentially imagining what new combinations of concepts look like, while being anchored in past experience.

ACI could facilitate the imagining of distortions to existing concepts by gradually venturing away from a concept's vector along different dimensions (see Figure 5.3(d)). The system could generate images of 'roses' starting with the 'rose' vector, but then gradually move away from the 'rose' vector. The resulting images should become distorted depending on the direction and distance from the original vector.

Finally, an ACI model could generate artifacts across different domains. The system could learn, using separate associative memory models, what concepts look *and* sound like. Given a picture of a 'dog', the system could then imagine what the 'dog' sounds like. The ACI model simply uses the AMM for images to predict the vector associated with the 'dog' picture and then feeds that predicted vector into the AMM for audio and has it generate a unique sound. The system could also be given a melody and then imagine an image to go with it, the two domains being tied together through the conceptual space as shown in Figure 5.3(e).

The ACI framework provides potential for these types of imaginative (and creative) abilities. It has been designed to model imagination by learning conceptual knowledge, perceiving concepts (artifacts), and generating novel artifacts never before experienced in several ways. Of course, this is only a framework, and the actual power of it depends on the abilities of the specific VSM and AMM implementations chosen for each domain (and their training data). Current state-of-the-art models are probably not yet capable of generating (or even classifying) large, detailed images of arbitrary concepts at the pixel level. Nor are they likely yet able to perceive sophisticated music in the general case. However, these capabilities do seem to be on the horizon with the advent of generative deep learning systems (such as DBNs).

5.3 Imagining Images

In order to show how the ACI framework could work in practice, we created a simple toy implementation that can imagine basic binary images. Instead of using a vector space model, we manually specified the conceptual space as a 2D plane in order to more easily visualize how images at various vector locations relate to one another. We then chose four vectors in the 2D conceptual space that are spatially located at four corners. The four vectors are $\vec{tl} = (0.0, 0.1)$, $\vec{tr} = (1.0, 1.0)$, $\vec{bl} = (0.0, 0.0)$, and $\vec{br} = (1.0, 0.0)$, to which we will refer as the *known* vectors.

We then generated four sets of training images for each of the four known vectors that are 32×32 pixels in dimension and are binary (i.e., black and white). The training images were pictures of actual corners, and example images for each of the four known vectors can be seen in

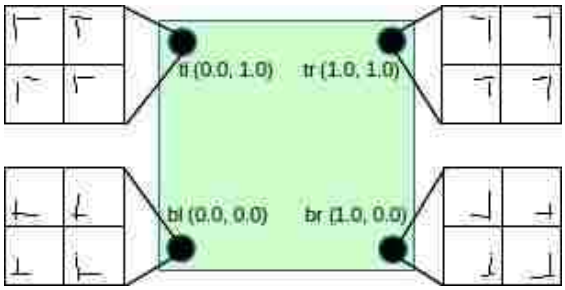


Figure 5.4: Example training images for each of the four known 2D vectors shown in conceptual space.

Figure 5.4. We implemented the associative memory model using a sum product network (SPN) and trained the SPN using corner images paired with their associated known-vectors (perturbed slightly using Gaussian noise). To learn the structure and parameters of the SPN, we used a modified version of the LEARNSPN algorithm that is able to accommodate both categorical and continuous random variables [60]. The result was a model that represents a joint probability distribution over image-vector pairs. We used the efficient, exact-inference capabilities of the SPN to generate novel images by sampling from the conditional probability distribution of images, conditioned on the concept vector. This was done by clamping the concept vector to a specific value and sampling the image pixel variables.

The model can perform sensory imagination by generating images for each of the four known vectors that it has learned. The bottom set of images in Figure 5.5 are example images imagined for the $\vec{br} = (1.0, 0.0)$ vector. Notice how each imagined image is unique yet still looks like the training images in Figure 5.4.

The system can also perform creative imagination by generating images for vectors for which it has never seen example images. These imagined images should look more similar to nearby known vectors than to known vectors farther away. The top set of images in Figure 5.5 were produced for the vector $(0.8, 0.2)$. These images are indeed similar to the images at vector $\vec{br} = (1.0, 0.0)$ (bottom set), which is the closest known vector. Although the system was never

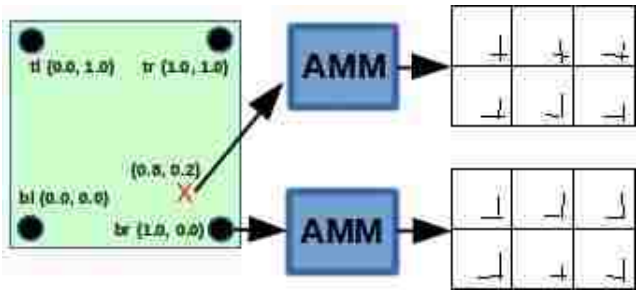


Figure 5.5: The bottom set of images were imagined for the vector $\vec{br} = (1.0, 0.0)$, which is one of the four vectors on which the system had been trained. The top set of images were imagined for the vector $(0.8, 0.2)$, which is a vector on which the system was not trained. The top images are similar to the bottom images because the vector $(0.8, 0.2)$ is close, in conceptual space, to the known vector $\vec{br} = (1.0, 0.0)$.

shown images for vector $(0.8, 0.2)$, it could still imagine what the images could look like by leveraging the information represented by the vectors in conceptual space (in this simple case just spatial information).

To further illustrate the imagining capabilities in this simple example, we had the system generate images at vector locations all over the 2D plane in 0.1 increments. In order to help visualize how the various generated images transition along conceptual space, we generated 100 images at each vector location and averaged them into a single image. We then arranged each averaged image on the plane according to their respective 2D vector (see Figure 5.6).

Moving from corner to corner on the 2D plane essentially shows the known images morphing into each other. The center image becomes a blend of all four corner shapes, while the images in the middle of the edges are a blend of the two corners on that edge. The model has only seen images for the corner vectors, which provide a basis for the other vectors in the 2D plane. The model cannot imagine images that do not relate to the four known corner images, which the results seem to confirm.

Admittedly, this toy example with a small 2D conceptual space and simplistic binary images is not visually impressive. It may be hard to ascribe imagination to a model that just seems to be doing a form of interpolation. Keep in mind that this example is only intended to be a proof-of-concept that demonstrates how the framework could work to generate actual artifacts. This

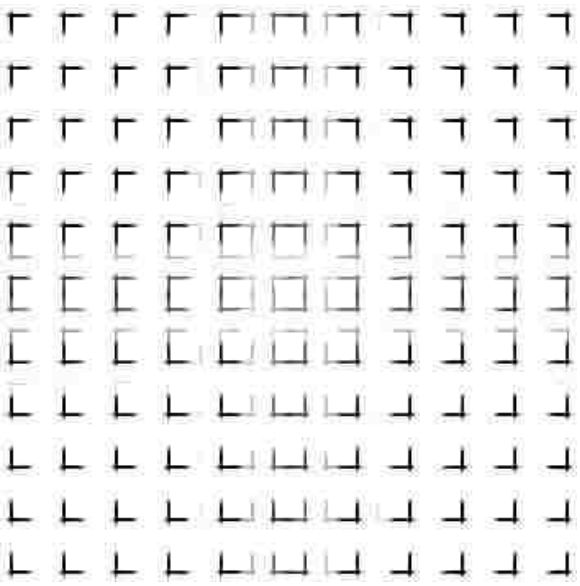


Figure 5.6: The average of 100 rendered images for each 2D vector in conceptual space at 0.1 increments. The system was trained on example images only for the vectors located at the four corners and then the system had to imagine what images at vectors in the middle would look like based on the images observed for each of the four corner vectors. Note how the images start to blend together as their corresponding vector approaches the middle of the space.

example also allows us to understand why the model is generating the images that it does—because of the training images (perceived artifacts) and the spacial arrangements of the vectors (conceptual relationships). A full implementation of this framework would be dealing with thousands of concepts in a conceptual space hundreds of dimensions in size, which is a much richer representation of conceptual knowledge. Also working with real artifacts, such as actual visual art or music, has the potential to yield much more impressive results.

5.4 Conclusions and Future Work

We have outlined the Associative Conceptual Imagination framework, which models how imagination could occur in a computational system that generates novel artifacts. The ACI framework accounts for the cognitive processes of learning conceptual knowledge and concept perception (via artifacts). The framework proposes using vector space models to learn associations between

different concepts, and using associative memory models to learn associations between concepts and artifacts. This network of associations can be leveraged by the system to produce novel artifacts.

We have demonstrated a basic implementation of ACI and applied it to simple binary images. We showed that the system could perform both sensory and creative imagination through the images it was able to produce.

The ACI framework poses some interesting questions. How will this framework perform when applied to real artifacts? What implementation and corpus should be used for the VSM? What models are appropriate to use for the AMMs? Does the choice of the model depend on the domain? Does the choice of the model depend on the artifact's representation (e.g., an image could be represented by raw pixels, extracted image features, or parameters to a procedural algorithm)? Research needs to be done to implement and refine this framework for various domains in order to explore these questions, and we are confident that the ACI framework will be useful for computationally creative systems.

In future work, we plan to apply the ACI framework to DARCI, a system designed to generate original images that convey meaning [76]. We plan to use the *skip-gram* VSM [112] trained on Wikipedia, which will learn vectors for 40,000 concepts in 300 dimensional space. Initially, we intend to implement the AMM using a discriminative model and a genetic algorithm. We will use 145 descriptive concepts (e.g., 'violent', 'strange', 'colorful', etc) to train the discriminative model to recognize those concepts in images. For example, the model will learn to predict the 'scary' vector when given a 'scary' image.

Once trained, the discriminative model will act as the fitness function to the genetic algorithm, which can then render images in ways that convey descriptive concepts (i.e., it can render a 'sad' image). The system will also be able to render images that convey concepts on which it has not been trained (beyond the 145) because of the semantic relationships encoded in the vectors. In other words, it will be able to imagine what other concepts would look like based on past experience and conceptual knowledge.

This framework could also be extended to include ideas involving conceptual blending. As it stands, the conceptual space does not change once the VSM learns the concept vectors and blending occurs through the associations between concepts and artifacts. It could be interesting to find ways to blend the concepts themselves together to produce new concepts that can then be expressed through artifacts.

Chapter 6

Creating Images by Learning Image Semantics Using Vector Space Models[1]

Abstract

When dealing with images and semantics, most computational systems attempt to automatically extract meaning from images. Here we attempt to go the other direction and autonomously create images that communicate concepts. We present an enhanced semantic model that is used to generate novel images that convey meaning. We employ a vector space model and a large corpus to learn vector representations of words and then train the semantic model to predict word vectors that could describe a given image. Once trained, the model autonomously guides the process of rendering images that convey particular concepts. A significant contribution is that, because of the semantic associations encoded in these word vectors, we can also render images that convey concepts on which the model was not explicitly trained. We evaluate the semantic model with an image clustering technique and demonstrate that the model is successful in creating images that communicate semantic relationships.

[1]Derrall Heath and Dan Ventura, Creating Images by Learning Image Semantics Using Vector Space Models, *Proceedings of The Thirtieth AAAI Conference on Artificial Intelligence*, 2016

6.1 Introduction

When considering the relationship between images and meaning (or semantics), most computational systems focus on extracting meaning from images. For example, image annotation [169] and content-based image retrieval [103] are two major topics in computer vision whose goal is to automatically understand the semantics within images. Here we focus on going the other direction, that is to *generate* images based on semantics.

There are few systems we know of that attempt to autonomously generate images that communicate meaning. The WordsEye system tries to generate 3D scenes based on written descriptions [34]. The Story Picture Engine [85] and the Text-to-Picture Synthesis System [179] are both systems built to do automatic text illustration (i.e., to visually tell a story or to graphically communicate the gist of text). AARON [109] and The Painting Fool [25] are both systems designed to autonomously create visual art in ways meaningful to human viewers.

Our own system, DARCI, is designed to create novel, artistic images that explicitly express a given concept [126]. Central to the design philosophy of DARCI is the notion that the communication of meaning in visual art is a necessary part of eliciting an aesthetic experience in the viewer [36]. In this paper we present a sophisticated semantic model that allows DARCI to internally represent the meaning of concepts and to express these concepts through images in novel ways.

It is commonly agreed that a word (or concept), at least in part, is given meaning by how the word is used in conjunction with other words (i.e., its context) [51, 96]. Vector Space Models (VSMs) are common methods for automatically learning vector representations of word meaning from a large corpus [159]. These models are based on the idea that similar words will occur in similar contexts and words that are often associated together will often co-occur close together. These models reduce words to a vector representation that can be compared to other word vectors.

VSMs have been successfully used on a variety of tasks such as information retrieval [143], multiple choice vocabulary tests [43], multiple choice synonym questions from the TOEFL test [137], multiple choice analogy questions from the SAT test [158], and object recognition systems [57].

Additionally, an approach called the ACI (Associative Conceptual Imagination) framework has recently been proposed as a way to use VSMs for imaginative and generative tasks [77].

We apply the ACI framework to DARCI by incorporating a VSM and building a visual semantic model that uses a large neural network to learn associations between low-level image features and adjective vectors from the VSM. This visual semantic model allows DARCI to create images that convey the meaning of adjectives to the viewer. It also allows DARCI to take advantage of the semantic structure between words and render images according to adjectives on which it was never explicitly trained. For example, DARCI could be trained on 'scary' and 'dark' images, but not 'creepy' images. DARCI could then "imagine" what a 'creepy' image would look like because 'creepy' is similar in meaning to 'scary' and 'dark'. Even higher level concepts (e.g., 'love', 'freedom') can be partially expressed through the images DARCI renders.

Clustering techniques have been previously developed to measure how well rendered images convey descriptive concepts [76]. We apply these clustering methods here and show that the new semantic model successfully enables DARCI to render images that convey a larger variety of concepts in ways that accurately reflect their semantic relationships.

6.2 Methodology

DARCI is composed of two major subsystems, a *semantic model* and an *image generator* as shown in Figure 6.1. We outline in detail the semantic model, which includes a state-of-the-art VSM that learns semantic relationships between words, and an artificial neural network that does multi-target regression to associate image features with the word vectors inferred from the VSM. We then describe the image generator and how it interacts with the semantic model to create meaningful images. Note that the image generator is not the focus of this paper, and further details, including extensive evaluation, can be found in prior work [126].

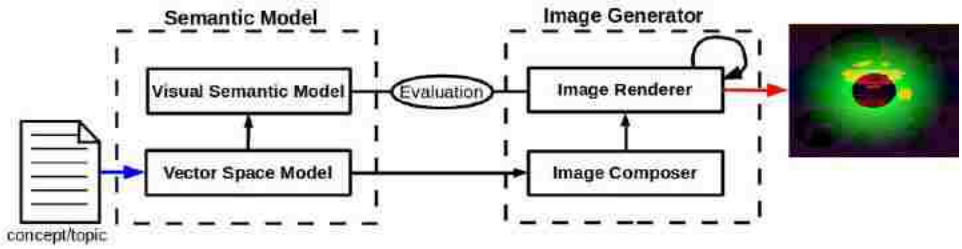


Figure 6.1: The two major components of DARCI. The *semantic model* first learns vector representations of words by analyzing a corpus (vector space model). The visual semantic model then learns to predict these word vectors using a neural network trained with labeled images. The *image generator* uses the vector space model to identify other words associated with a given concept. The nouns are composed into a source image (image composer) that is rendered to convey the original concept using a genetic algorithm (image renderer) that is governed in part by the visual semantic model. The final product is an image that reflects the given concept.

6.2.1 Semantic Model

In order for DARCI to express semantic information through pictures, it must first have its own semantic knowledge that can influence the images it creates. Our goal is to leverage semantic information gained through written text and transfer it to the task of meaningful image generation.

Vector Space Model

We use a state-of-the-art VSM, called the *skip-gram model* [112]. The skip-gram model is a neural architecture that analyses a large corpus and learns to predict the surrounding words given a current word. During training, the skip-gram model consequently learns vector representations for each word, which encode semantic information. Words similar in meaning will have vectors that are close to each other in "vector space". These word vectors capture other interesting semantic relationships that are consistent with arithmetic operations. For example, $vector(\text{“}king\text{”}) - vector(\text{“}man\text{”}) + vector(\text{“}woman\text{”})$ results in a vector that is closest to $vector(\text{“}queen\text{”})$.

These semantic vectors allow DARCI to find concepts related to a given word and to assess the similarity in meaning between words, which will aid DARCI in creating meaningful images. We use a publicly available implementation of the skip-gram model[2] and a lemmatized Wikipedia corpus to learn the word vectors [44]. The skip-gram implementation is used with out-of-the-box

parameters except for the vector size, which is set to 300. The choice of 300 provides a balance between encoding enough semantic information to be useful and ease of prediction when associating the vectors with image features.

Visual Semantic Model

In order for DARCI to leverage the word vectors for image creation, it must learn to associate image qualities with the semantic vectors. Currently, we limit the associated words to vectors representing adjectives and use a neural network model to predict the adjective vector for a given image.

We maintain a dataset of approximately 15,000 images that have either been explicitly hand labeled or automatically retrieved through Google image search. Once an adjective has enough labeled images (20 positive and 20 negative), we begin learning that adjective. As of this paper, there are 145 adjectives that meet this threshold. We extract from each image 51 global and local features representing attributes like color, lighting, texture, and local interest points, and have been shown to work well for emotional and descriptive labels [127].

We train two separate neural networks, one with the positively labeled images, and one with the negatively labeled images. The positive network tries to predict what adjective an image IS, while the negative network tries to predict what adjective an image IS NOT. These networks learn to predict the appropriate adjective *vector* given an image. We treat this as a multi-target regression problem and initialize each neural network with 300 output nodes (one for each vector element). The inputs are the 51 image features, the hidden layer is non-linear (sigmoid), and the output layer is linear. The parameters for the neural networks were determined through experimentation (see the Evaluation Section for the metrics used) and include a learning rate of 0.01, a momentum of 0.1, and 100 hidden nodes. We use standard backpropagation with drop-out regularization to initially train the weights. Since the output layer is linear, we improved each model by solving for the least-squares solution as the final training step.

[2]https://code.google.com/p/word2vec/

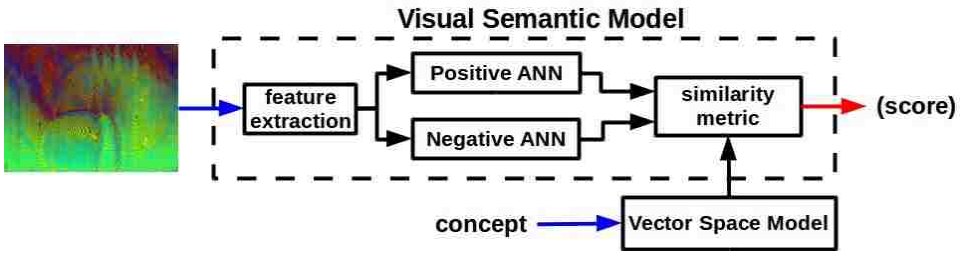


Figure 6.2: This diagram illustrates how the visual semantic model determines to what degree an image matches a given concept. It first extracts features from the image which are passed to both the positive neural network and the negative neural network. The word vector for the given concept is retrieved from the vector space model and compared via cosine similarity to the predicted vectors from the two neural networks. The similarity scores are combined and normalized for an overall score.

Figure 6.2 shows how the networks are used to determine how well an image matches an adjective. Let $\vec{v}_p$ and $\vec{v}_n$ be the vectors predicted by the positive and negative networks, respectively. Let $\vec{v}_a$ be the vector for adjective a from the VSM and let $sim(\vec{v}_1, \vec{v}_2)$ compute the cosine similarity between two vectors. Given an image, we can compute its score for a particular adjective using the following formula:

$$score = \frac{(sim(\vec{v}_p, \vec{v}_a) - sim(\vec{v}_n, \vec{v}_a)) + 1.0}{2.0} \tag{6.1}$$

Learning to predict an adjective's vector is a harder task than learning to predict the adjective directly and, thus, introduces a few trade-offs. First, labeling images with adjectives is a multi-label classification problem (i.e., an image can be described by more than one valid adjective) and our new model can only predict one vector at a time, while normally each adjective could be predicted independently. The second trade-off is that our visual semantic model is predicting a 300 dimensional vector and has to account for every adjective. This means that it may not predict the 145 adjectives as accurately as using separate models for each adjective. The main advantage of learning the vectors, however, is that we can do zero-shot prediction. In other words, it is not limited to the 145 adjectives for which it was explicitly trained and can predict vectors for any adjective. The model can assess, how 'glad' an image is even if it has never seen a 'glad' picture because the semantic relationships of many adjectives are encoded in the vectors.

6.2.2 Image Generator

With the vector space and visual semantic models in place, DARCI can now produce images. Figure 6.1 shows how this process works. First, a concept/word/topic is given to the system and the VSM finds semantically related concepts. DARCI effectively makes use of these word associations as a decomposition of a (high-level) concept into simpler concepts that together represent the whole. The idea being that in many cases, if a (sub)concept is simple enough, it can be represented visually with a single icon (e.g., the concept 'rock' can be visually represented with a picture of a 'rock'). Given such a collection of iconic concepts, DARCI composes their visual representations (icons) into a single image. This *source* image is then passed to the image renderer, which uses a genetic algorithm to render the image in an artistic way that conveys the meaning of the original concept. During rendering, the visual semantic model acts as the fitness function to guide the rendering process.

For example, suppose the original concept given to DARCI was 'war'. The vector space model would send related words like 'soldier', 'army', 'conflict', and 'battle' to the image composer. The resulting source image would be some composition of simple iconic images of the related words. The image renderer would then render this source image according to the visual semantic model. In this case the visual semantic model is telling the image renderer to create images that are close to the semantic vector for 'war'. However, since the visual semantic model was only trained on the 145 adjectives, this results in a rendering based on the adjectives that are semantically related to 'war' (in this case, 'bloody', 'violent', 'lonely', etc).

DARCI can also forgo the image composer and go straight to the image renderer, in which case the image produced will be an *abstract* rendering of the given concept. The user could also provide DARCI a source image of their own, like a photograph, and DARCI will re-render the photograph in an artistic way that expresses the given concept. Since the new semantic model is the focus of this paper, the image generator is simplified to create only abstract images (by skipping the image composer) for all experiments.

Rendering images based on predicting word vectors instead of predicting the adjectives directly makes it more difficult for the image renderer to match a given adjective. However, the power comes in taking advantage of the learned semantic structure encoded in the vectors. DARCI can render images to convey any adjective that has at least some semantic relationship with any of the 145 explicitly trained adjectives. Even non-adjectives, such as 'war', can be rendered this way and essentially get interpreted as an adjective (i.e., 'war-like').

6.3 Evaluation and Results

We start with evaluating how well the semantic modeling component learns to predict word vectors from images. We then use clustering techniques to determine how well the images that DARCI produces actually reflect their intended adjective. Finally, we evaluate how clusters of images relate to each other and to the word vectors on which they are based.

6.3.1 Semantic Model Evaluation

We consider two metrics: *coverage* and *ranking loss*. For each adjective, the model ranks each test image by the similarity score obtained from the visual semantic model (Eq. 6.1). Images labeled with the adjective (positive images) should be ranked higher than images that are negatively examples of the adjective. Coverage represents how far to go down the list of ranked images in order to cover all positive images (normalized between 0 and 1). Ranking loss represents the percentage of negative images that are ranked higher than positive images. These metrics are averaged across all 145 adjectives. We compare our visual semantic model (Vector) with a binary relevance model (Binary) using 10-fold cross validation. The results can be seen in Table 6.1.

As expected, our visual semantic model performs worse than binary relevance on the 145 adjectives. However, the benefit is that our new model can rank images based on adjectives on which it was never trained. We chose 10 additional adjectives for which DARCI had not been trained and created a hold-out set of test images for them. We evaluated how well the model ranked

	Cross Validation				Zero-shot	
	Random	*Binary*	*Vector*		*Random*	*Vector*
Coverage	0.768	**0.533**	0.628		0.709	**0.444**
Ranking Loss	0.502	**0.297**	0.357		0.502	**0.199**

Table 6.1: The 10-fold cross validation image ranking results of learning the 145 adjectives (lower scores are better). We compare our visual semantic model (Vector) with a binary relevance model (Binary) that learns the adjectives directly. The binary method performs better on the 145 adjectives. However, the vector method allows the system to rank images based on adjectives it has never been trained on (Zero-shot), which we test using a hold-out set for 10 adjectives the model has never seen.

images based on these new adjectives (Zero-shot in Table 6.1). The results show that the visual semantic model is successful (i.e., better than random) at ranking the test images for the 10 new adjectives.

6.3.2 Image Evaluation

Evaluating how well an image conveys an adjective is a subjective task, especially for a system that is also trying to generate novel images. Usually, a human survey is necessary to arrive at a general consensus in measuring the semantic quality of images, but even then such a consensus is not always possible (or desirable).

Clustering techniques have been developed for evaluating how well images convey semantic relationships [76]. The idea is that images should cluster in ways that reflect the semantic similarity of the adjectives on which they were based. For example, ‘scary’ and ‘creepy’ images should cluster together more closely (i.e., be harder to tell apart) than ‘cold’ and ‘happy’ images because ‘scary’ and ‘creepy’ are more similar in meaning than ‘cold’ and ‘happy’. By using clustering, we may not be able to objectively tell if a *specific* image conveys a *particular* adjective, but we can objectively see how well the system in general is creating images that reflect the semantic relationships learned by the vector space model. Heath et al. showed that their clustering methods were consistent with human evaluators.

Let *SEEN* refer to the set of 145 adjectives that DARCI was trained on and let *UNSEEN* refer to adjectives not of those 145. We selected two sets of 5 adjectives from SEEN. The first set consisted of semantically *similar* adjectives, while the second set consisted of semantically *distinct*

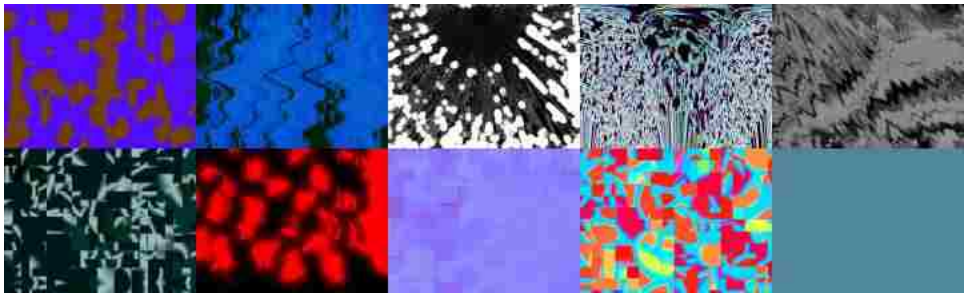


Figure 6.3: Example abstract images created for the adjectives referenced in Table 6.2. The top row (from left to right) corresponds to the semantically similar adjectives ‘creepy’, ‘ghastly’, ‘scary’, ‘strange’, and ‘weird’. The bottom row corresponds to the distinct adjectives ‘cold’, ‘fiery’, ‘peaceful’, ‘vibrant’, and ‘wet’.

	Similar	Distinct
Entropy	0.857	**0.714**
Purity	0.360	**0.440**

Table 6.2: The cluster entropy and purity results from clustering images of semantically similar adjectives compared to clustering images of semantically distinct adjectives (the adjectives are listed in Figure 6.3). Lower entropy is better, while higher purity is better. These results confirm that it is harder to cluster the images of similar adjectives than it is to cluster the images of distinct adjectives.

adjectives. We had DARCI render 10 separate images for each adjective using the abstract rendering method (i.e., no source image). The two sets of 5 adjectives are listed and example images for each can be viewed in Figure 6.3.

The 51 global and local features from the visual semantic model were extracted from each rendered image. We used the EM (Expectation Maximization) algorithm found in WEKA [69] to cluster each set’s collection of images (using the extracted features). We then applied two metrics, average *entropy* and average *purity*, to evaluate the quality of the clusters. The results can be seen in Table 6.2.

The results verify that the images for the similar set of adjectives are harder to correctly cluster than are images for the distinct set of adjectives. This is evidence that DARCI is successful at rendering images that convey the meaning of adjectives relative to each other. In this paper, we especially want to focus on how well DARCI can render images based on UNSEEN adjectives, or any word for which it has never seen images.

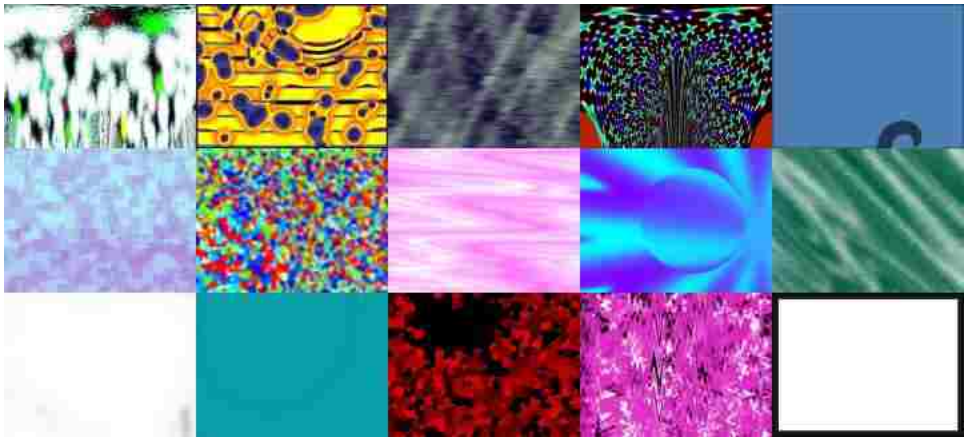


Figure 6.4: Example abstract images created for adjectives DARCI was never trained on and that correspond to the results in Table 6.3. The images of the first two rows from left to right convey the adjectives ‘bizarre’, ‘brilliant’, ‘freezing’, ‘frightening’, ‘frigid’, ‘hazy’, ‘lively’, ‘lovely’, ‘luminous’, and ‘somber’. The images of the third row convey the non-adjectives ‘Alaska’, ‘crying’, ‘fear’, ‘love’, and ‘winter’.

	Adjectives			**Non-adjectives**	
	Similar	*Dissimilar*		*Similar*	*Dissimilar*
Entropy	0.691	**0.480**		0.828	**0.479**
Purity	0.775	**0.850**		0.680	**0.840**

Table 6.3: The average cluster entropy and purity results from clustering images conveying adjectives (and non-adjectives) on which the system was never trained. The adjectives and non-adjectives used are listed in Figure 6.4. Lower entropy is better, while higher purity is better. The results show that it is harder to cluster images from semantically similar words than images from dissimilar words. This is evidence that DARCI is successfully rendering images that convey the intended word, even when it has never seen an example image of that word before.

We chose 10 UNSEEN adjectives and had DARCI render 10 separate abstract images for each. We also chose five non-adjectives and again had DARCI render 10 abstract images for each. The words are listed and example images for each are shown in Figure 6.4. We again used clustering to evaluate the the semantic quality of the rendered images. For each of the 10 UNSEEN adjectives, we took a SEEN adjective that was semantically similar and one that was dissimilar and had DARCI generate 10 images for each of them. We then clustered the images for the UNSEEN adjective and the images for the SEEN *similar* adjective, while separately clustering the UNSEEN images and the SEEN *dissimilar* images. Finally, we averaged the metrics of the 10 UNSEEN adjectives. We repeated this process for the five non-adjectives and the results are shown in Table 6.3.

The similar images are harder to cluster than the dissimilar ones for both UNSEEN adjectives and non-adjectives. This indicates that DARCI is successfully rendering the images to convey the intended words relative to each other, even though DARCI has never seen any example images of the words. DARCI is able to use the semantic structure learned from the vector space model to interpolate, or more colloquially "imagine", what images of these unseen adjectives could look like. The clustering results give us a measurable indication of DARCI's ability to render images consistent with the semantic structure of the words for which they were rendered.

6.3.3 Image Cluster Visualization

We can also visualize how the clusters of images relate to one another and to the semantic vectors on which they are based. We created a 2D visualization of how the words cluster in *vector space* and compared it to a 2D visualization of how their respective images cluster in *image feature space*. We created the 2D visualizations by using agglomerative clustering combined with multi-dimensional scaling [134]. We took the 10 UNSEEN adjectives and 5 non-adjectives from the previous experiment and chose an additional 15 SEEN adjectives that had variable semantic similarity to the 15 UNSEEN words.

For the visualization in vector space, we used multi-dimensional scaling to find an approximate 2D plot (from 300 dimensions) of the distances between each word's vector. We then did agglomerative clustering (using EM) with the 30 word vectors and drew the resulting clusters on the 2D plot. For the visualization in image feature space, we calculated the average feature vector of the 10 separately rendered images for each of the 30 words. We then performed multi-dimensional scaling (from 51 dimensions) and agglomerative clustering in the same way we did with the word vectors. Both visualizations can be seen in Figure 6.5.

In vector space, note the distinct clusters of similar words. Also note that the non-adjectives are generally more distant from the larger groups of adjectives, likely due to their having closer similarities to some other non-adjectives. Overall, the image clusters roughly correspond to the word clusters. In both visualizations there exist relative groupings for scary type words/images,

and groupings for temperature type words/images. Even in the clusters that don't match exactly, the relative positions of most words are similar. For example, 'bright', 'luminous' and 'glowing' are still generally near each other, even though they were absorbed into different neighboring clusters. Differences between the word clusters and the image clusters are to be expected as the visual semantic model learns from noisy data and multi-dimensional scaling has to approximate 2D positions from a high dimensional space. Also keep in mind that DARCI, while trying to convey the adjective in the image, is also trying to innovate and create novel images. For example, Figure 6.6 shows variations across different renderings of the same adjective ('fiery').

It should be noted that visual differences between words don't always correspond to their semantic differences. For example, we would expect the adjectives 'warm' and 'cold' to have distinct visual qualities. However, in Figure 6.5(a) we see that 'warm' and 'cold' are semantically similar and so DARCI's renderings of these adjectives can look similar. Figure 6.7 shows five of the 10 rendered images for 'cold'. Notice that a few of them could easily be confused with a 'warm' image. This seems unfortunate, but it is actually another indication that DARCI is accurately generating images according to the semantic relationships learned by the VSM.

6.4 Conclusions and Future Work

We have introduced a sophisticated semantic model into DARCI that enables it to create images that convey a wide variety of concepts. We have shown that the similarity of the resulting images correspond to the semantic similarity of the concepts on which they were based, which is evidence that the images do reflect their intended adjective. We have also shown that DARCI can render adjectives (and even non-adjectives) that it has never seen example images of. This ability is a rudimentary form of imagination and is analogous to a person being able to imagine, say, what a 'majestic' image might look like when told that 'majestic' is similar to 'powerful' and 'beautiful', even though the person may have never experienced the word 'majestic' before.

This simple form of imagination is not limited to images and could be applied to practically any domain. For example, a system could generate music based on the same word vectors (e.g., compose a 'happy' song), and could then produce new music to match previously unheard concepts. The VSM could even act as a bridge between different domains. A system could listen to a 'sad' song, which would be mapped near the 'sad' vector, and the system could then "imagine" a sad-like image inspired by the song.

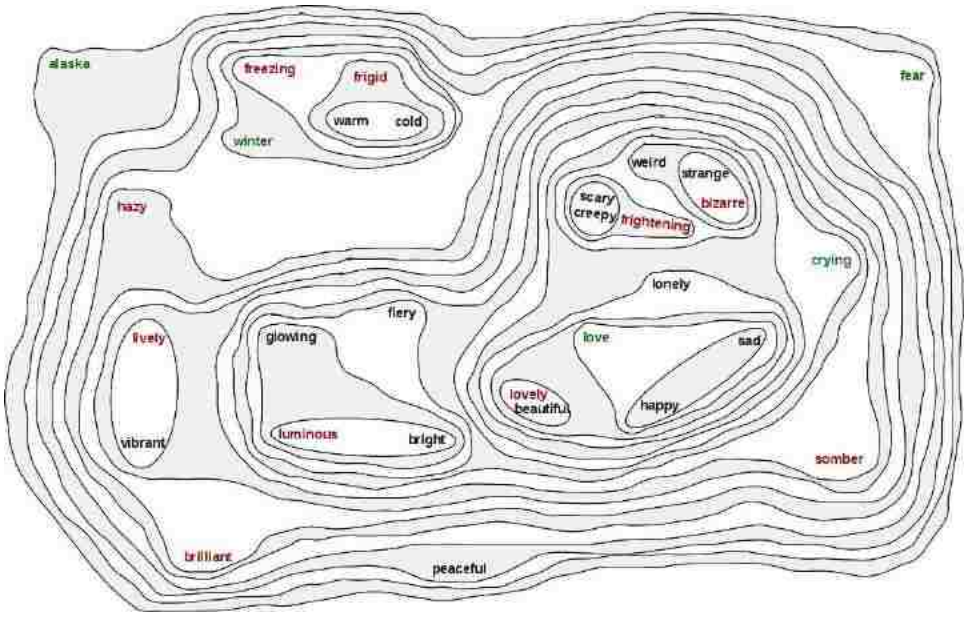
Using a VSM for these types of creative learning problems allows for more freedom, more autonomy, and demonstrates a more robust form of intelligence. In classical machine learning, models are typically rigidly confined to the concept(s) explained by available training data, performing poorly outside this scope. In contrast, the semantic model presented here attempts a form of transfer learning from written text to image understanding/generation, which gives our system a chance to perform reasonably even for concepts it has not explicitly learned. This flexibility is especially useful for problems in the field of computational creativity, where there may not be a "best" or "right" answer.

With the success of the semantic model, we can consider the system as a whole and move beyond abstract images by having DARCI create more sophisticated art that conveys meaning for more advanced concepts. For example, a user could provide a source image and DARCI could then re-render the source image to convey any given concept. Figure 6.8 shows several examples of this method of rendering.

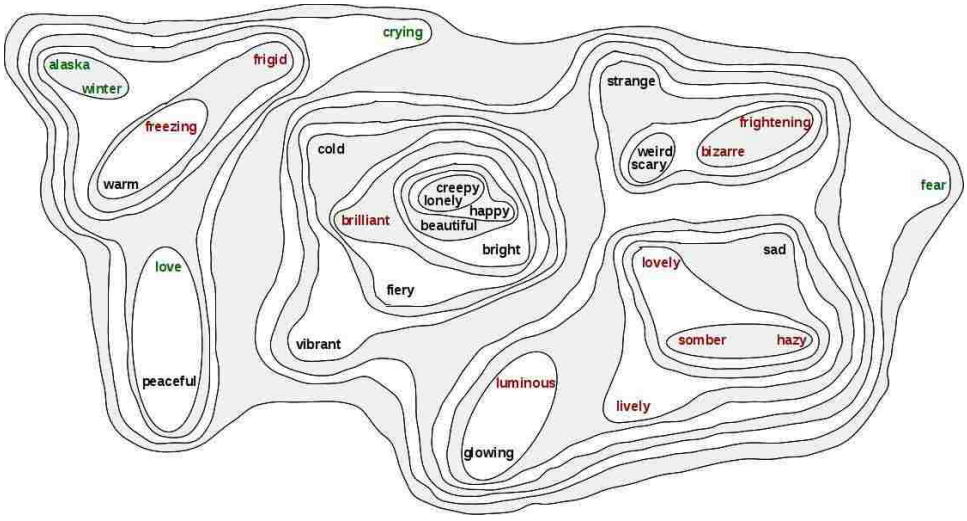
As outlined in the Image Generator Section, DARCI can also create a source image of simple icons by finding nouns semantically related to the provided concept. This collage of icons can then be artistically rendered to communicate the original concept, as shown in Figure 6.9. DARCI chooses to include icons based on what it has learned through the vector space model, and the result is an original image that conveys the given concept. We intend to evaluate the DARCI system as a whole to determine its creative ability to communicate meaning through visual art.

We noted that semantic differences between words don't always correspond to visual differences. One idea to overcome this is to use a hierarchical approach that locates different densities or clusters within the word vector space: a top-level visual semantic model that learns to identify different clusters, and separate visual semantic models for each cluster that focus on distinguishing among the individual words in a given cluster.

We would eventually like to extend the ideas in this paper beyond adjectives to include nouns. We want to enable DARCI to create actual (non-abstract) pictures of nouns without relying on a provided source image or a database of icons. This will most likely require a deep learning system that leverages semantic information to discriminate between pictures of nouns, as done in other studies [57]. A deep generative model could potentially generate images by visualizing how the model has learned features at various levels. Recently, deep neural systems have already had success in automatically generating images [45, 68, 97].



(a) Vector Space (300 dimensions)



(b) Image Feature Space (51 dimensions)

Figure 6.5: A 2D visualization of the spacial relationships between the word vectors (a), compared to the spatial relationships of their respective images (b). Red words are adjectives on which DARCI was never trained, while green words are non-adjectives. The image clusters/positions roughly correspond to the word clusters/positions. This demonstrates that DARCI was able to render images that at least partially convey the meaning of adjectives, and even of words on which DARCI was never trained, including non-adjectives.

Figure 6.6: Five of the 10 abstract images rendered for the adjective ‘fiery’. Notice the variation between different renderings as DARCI is trying to innovate, in addition to conveying the adjective.

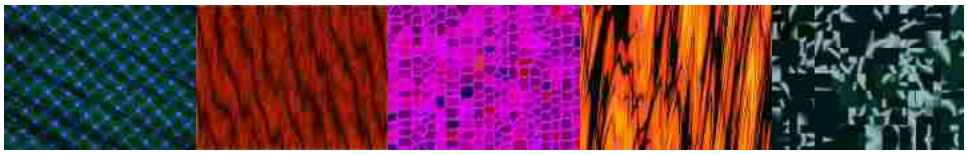


Figure 6.7: Five of the 10 abstract images rendered for the adjective ‘cold’. Notice that some of the images could easily be confused with ‘warm’ due to ‘cold’ being semantically related to ‘warm’.

Figure 6.8: Images DARCI rendered (bottom row) after being provided a source image (top row) and a concept. From left to right, the concepts are ‘fiery’, ‘Alaska’, and ‘hunchback’. Although the source image was given, DARCI discovered its own way to render the image to convey the given concept.

Figure 6.9: Images that DARCI has rendered after being given only a concept. From left to right, the concepts are ‘bizarre’, ‘war’, ‘art’, ‘murder’ and ‘hunger’.

Chapter 7

Before A Computer Can Draw, It Must First Learn To See[1]

Abstract

Most computationally creative systems lack adequate means of perceptually evaluating the artifacts they produce and are therefore not fully grounded in real world understanding. We argue that perceptually grounding such systems will increase their creative potential. Having adequate perceptual abilities can enable computational systems to be more autonomous, learn better internal models, evaluate their own artifacts, and create artifacts with intention. We draw from the fields of cognitive psychology, neuroscience, and art history to gain insights into the role that perception plays in the creative process. We use examples and methods from deep learning on the task of image generation and pareidolia to show the creative potential of systems with advanced perceptual abilities. We also discuss several issues and philosophical questions related to perception and creativity.

[1]Derrall Heath and Dan Ventura, Before A Computer Can Draw, It Must First Learn To See, *Proceedings of The 7th International Conference on Computational Creativity*, to appear, 2016

7.1 Introduction

Some people seem to have a natural talent for drawing, while others only wish they could draw well. Many of these people have turned to books and teachers to help them develop their drawing skills. One of the most widely used and consistently successful books for teaching people how to draw is titled *Drawing on the Right Side of the Brain* [48]. This book uses insights from neuroscience to help potential artists improve their drawing skills. One of the main premises in the book is that drawing is not a skill of hand, paper or pencil, but a skill of perception. To quote from the book:

> "The magic mystery of drawing ability seems to be, in part at least, an ability to make a shift in brain state to a different mode of seeing/perceiving. When you see in the special way in which experienced artists see, then you can draw... Drawing is not very difficult. Seeing is the problem, or, to be more specific, shifting to a particular way of seeing."

This idea can extend to any kind of creative ability. Before one can create visual art, compose music, or invent recipes, one must first learn to see, hear/listen, or taste, respectively. Even creative tasks like writing poetry must ultimately be grounded to what has been experienced through perception. Our ability to perceive influences how and what we create. Just as drawing is really about perceptual skills, our ability to think creatively and do creative things heavily depends on how we perceive and understand the world.

In his book, *The Anthropologist On Mars*, Oliver Sacks recounts the story of Shirley Jennings, who had been blind since early childhood and had surgically regained his sight at the age of 50 [142]. After the operation, he could not immediately see, could not recognize his family, could not pick out separate objects, and struggled with depth perception. Over several months, his brain had to learn how to see and make sense of an incredible amount of new information. It was a slow and difficult process reconciling his non-visual mental model of the world with this new form of perception. As he learned to make use of his new sense, things that were aesthetically beautiful to him differed

from those others found pleasing. He eventually took painting lessons and created paintings that demonstrate his unique taste in visual art[2].

Shirley's case, along with several other vision disorders and anomalies like blindsight [173], Capgras syndrome [50], and agnosia [53], has helped to uncover the work and learning that our brain undertakes in order for us to perceive and understand the world. In this paper we argue that the ability to perceive is a necessary and influential piece of the creative process. It enables us to learn a mental model of the world, to understand and continually evaluate our own creations, and to infuse what we produce with meaning. Indeed, even perception itself is a creative act that our brains regularly perform, although often subconsciously. The necessity of perception applies to the field of computational creativity, in which one of the goals is to build computational systems that can autonomously create art. Before a system can learn to create art, we argue that it must first learn to perceive.

We proceed by exploring the relationship between perception and creativity and then discuss the role of perception in computational systems. We then consider how state-of-the-art computer vision methods can enhance the creative potential of systems designed for visual art. We demonstrate, using deep neural networks, how perceptual skills facilitate imagination and can lead directly to generating novel images. We then elaborate further on why perception itself is a creative process and demonstrate a form of creative perception, called pareidolia, using deep neural networks. Finally, we discuss philosophical issues and the implications of our ideas and elaborate on what more advanced perceptual abilities could mean for the future of computational creativity.

7.2 Perception and Creativity

When talking about visual art, Csíkzentmihályi says, "...the aesthetic experience occurs when information coming from the artwork interacts with information already stored in the viewer's mind..." [36]. In other words, the viewer's appreciation (perception) of art is determined by his current mental model of the world. Likewise, the artist has her own mental model of the world and

[2]http://www.atfirstsightthebook.com/shirls-paintings.html

created the artwork to convey meaning according to that mental model. How was that mental model established? It's reasonable to say that it was learned through a lifetime of experiences, and people experience the world through perception. Everything we know and understand about the world has come through our senses. Every memory and every thought we have is in terms of what we have experienced in the past [7].

It is difficult to comprehend what life would be like without perception because it is so fundamental to how we think. Would it even be possible to think, imagine, or create anything at all without some kind of input? There is no definite answer to that question; however, studies of long term sensory deprivation and solitary confinement suggest significant mental deterioration [1, 65]. Perception directly influences our ability to think and understand, and the better and more varied our perceptual abilities are, the more we are able to think about, imagine, and ultimately, create. We can take this idea further and say that, with our current senses, there are thoughts we cannot think simply because we lack additional (or adequate) senses to know how to think them. To quote Richard Hamming [70]:

> "Just as there are odors that dogs can smell and we cannot, as well as sounds that dogs can hear and we cannot, so too there are wavelengths of light we cannot see and flavors we cannot taste. Why then, given our brains wired the way they are, does the remark 'Perhaps there are thoughts we cannot think,' surprise you? Evolution, so far, may possibly have blocked us from being able to think in some directions; there could be unthinkable thoughts."

We need perception (i.e., input) in order to build a mental model that can facilitate thinking, which can then facilitate creativity. Indeed, as noted earlier in the case of Shirley Jennings [142], the mental model itself is what does the perceiving. Our eyes merely translate light into brain signals, but it is our brain that must learn to make sense of that information, which then allows us to think in those terms.

Imagination is clearly tied to this idea and is closely linked with creativity in cognitive psychology literature [59]. Imagination is typically generalized as thinking of something (real

or not) that is not present to the senses. Most psychologists agree that our perceptions (senses), our conceptual knowledge, and our memories make up our mental model and form the bases of imagination [10]. As we perceive the world and have experiences, our mental model is formed by establishing and strengthening connections in our mind. These connections form concepts, which are in turn interconnected. Creative imagination is achieved by combining these connections and experiences in different ways that produce novel results.

7.2.1 Thinking Beyond Natural Perception

It is possible for us to indirectly experience things outside of our perceptual abilities by translating other modalities into our range of senses. For example, we visualize infrared light by shifting it into the visible spectrum. We create charts and graphs that represent data we cannot observe directly, like barometric pressure, or electromagnetic fields. In this way we can vicariously think in terms of other modalities and perhaps even be creative in those modalities.

This idea is applied explicitly in the case of *sensory substitution* [3], where one sense can take the place of another that has been lost. For example, devices have been made that can allow blind people to literally "see" with their tongue. They work by mounting a video camera on the blind person's forehead, which sends video data to a plate that sits on the person's tongue. This plate contains a grid of "pixels" consisting of pressure points. These pressure pixels correspond to grayscale video by pressing harder where the image is brighter and pressing lighter where the image is darker. The tongue can then "feel" the video information and, after several months of training, a blind person's brain starts to see images in their mind. It's certainly not high resolution, but it's enough to allow a blind person to read large print text and navigate new terrain.

Another way that we humans can communicate and understand things that we ourselves have not perceived directly is through language. In other words, through verbal/written communication we can experience by proxy what others have directly experienced [181]. In this case, language acts as an analogy between two people's experiences. Our interpretation of a described experience must still be grounded by our personal perceptions and experiences [7]. For example, it is very difficult

to describe colors to a congenitally blind person because colors are inherently visual and the blind person has no visual grounding at all. This is why even creative literature and poetry also require perception—the writer must have experiences to write about and the reader must have experiences with which to interpret the writing.

Art, whether it be visual, written, musical, etc, acts as a metaphor between the experiences of the artist and the experiences of the receiver. Successful artists are creative because they have a unique perspective on the world that they are trying to communicate through their art, and people appreciate art that helps them gain new perspectives. In other words, having unique experiences and perceiving the world differently plays a role in the creative process. It has been postulated in cognitive psychology that creative people literally see the world differently [11], which is in turn why they tend to think differently and can produce novel things and ideas.

7.2.2 Quality of Perception Affecting Visual Art

There have been studies analyzing several famous artists with documented visual impairments [107]. For example, Claude Monet developed cataracts, while Edgar Degas began to suffer from retinal disease. These studies point out that the earlier works of these painters (when they had good eyesight) were better formed and detailed, while later works (made with poorer eyesight) became more and more abstract. These studies generally conclude that the failing eyesight of the artists *did* have a large impact on the quality and style of their work. Although some researchers say that this was not necessarily a bad thing, and some artists would use their visual impairments to their advantage by removing their corrective lenses for certain paintings.

These artists had issues with just their eyes, but what about *cognitive* impairments involving vision? How do different cognitive disorders of the brain affect artists' work? This question was explored by Anjan Chatterjee, where he reported on multiple studies analyzing the drawing ability of several artists with various cognitive disorders, including spacial neglect, visual agnosia, epilepsy, TBI, etc [20]. The results for many of the disorders, like epilepsy were mixed, but artists with

disorders more specific to vision, like agnosia, had some notable peculiarities to their drawing ability.

For example, one artist with a type of visual agnosia could create beautiful drawings as long as the item he was drawing was continuously right in front of him. However, if he was asked to draw from memory (e.g., draw a 'bus'), then his drawings were simplistic and often unrecognizable. Another artist, with a traumatic brain injury, produced drawings that were more abstract and "expressive" than drawings produced before the accident. Although these studies appear anecdotal due to the rarity of many of these disorders, it is apparent that how the brain sees and understands the world does affect the ability to draw.

There are several cases of successful artists who are blind, and one artist in particular has received a lot of attention because he is *congenitally* blind [2]. This blind man has a remarkable ability to paint and draw pictures that are consistently meaningful to sighted people. He uses special paper and pencils that form ridges that he can feel as he draws. He first explores an object with his hands and then, remarkably, can draw it from different perspectives. He's never been able to see, yet can understand perspective. His case provides insight into how the brain perceives and builds invariant mental representations of the world.

The blind artist's case is related to sensory substitution, where a blind person can "see" through touch, and further supports the idea that the brain is what processes and makes sense of perceptual input. Researchers who study blindness and visual art have indicated that vision and touch are linked and make use of similar processes and similar features in the brain [86, 89]. The brain can do remarkable things even when the quality of the input signal is disrupted or re-routed. Perception is really about being able to build these mental models and using them to interact with the world. It's not that *visual* art requires *vision*, but that creating visual art requires *some* form of perception that establishes and continually informs the artist's mental model.

7.3 Perception and Computational Creativity

We've discussed the role of perception in *human* creativity, but what about computational creativity? Certainly, there's no requirement that computers can only be creative in the same way as humans. However, we are positing that perception is *fundamentally* a necessary component of the creative process. So, just as perception is important for human creativity, perceptual ability is also important for computational creativity. The exact methods of perceiving and creating may be different than those of humans, but some form of perceptual grounding is requisite for a truly creative system.

Colton proposed the creative tripod as necessary criteria for a creative system [24]. A creative system must have imagination, which is analogous to producing novel artifacts; it must have skill, which corresponds to generating quality artifacts; and it must have appreciation, which is the ability to recognize the novelty and quality of its own artifacts (i.e., self-evaluation). In other words, there must be a perceptual component that directs the creative process by helping the system explore new ideas (imagination), and understanding which ideas are worth pursuing (appreciation).

Many creative systems exist across several domains that can generate novel artifacts. Most of these systems, however, are merely mimicking example human-created artifacts without understanding or appreciating what they are producing, like a parrot mimicking human speech. For example, the PIERRE system generates new crockpot recipes according to a model trained on user ratings of existing recipes, but it has no sense of what the recipes actually taste like, only that humans have liked similar recipes [116]. In music, there are several systems that analyze patterns and n-grams from existing melodies, then probabilistically draw from those distributions or construct grammars when producing music [33, 129]. Likewise, poetry systems are also often based on corpora and n-gram distributions, without much understanding of what the words actually represent [30, 119].

Other existing creative systems produce artifacts according to hand engineered metrics and databases, where the ability to appreciate and perceive what is produced is limited to those explicit metrics. For example, some musical systems rely on rules and metrics based on musical theory in order to produce and evaluate melodies [47, 111]. Visual art systems often use some form of evolutionary algorithm for producing art, which involves a fitness function by which the art is

evaluated at each iteration. The fitness functions in these systems are usually based on models trained using extracted image features in order to evaluate aesthetic quality or novelty [46, 106]. In these cases the perceptual ability is ultimately limited to those specific features.

There are some creative systems that do attempt to incorporate a sophisticated model of perceptual ability. For example, there is a system that invents recipes based on actual chemical properties of the individual ingredients [160]. It at least has some understanding of what would actually taste good in a recipe and isn't limited to just producing something that is mimicking human examples. The DARCI system extracts various image features and trains neural networks to evaluate how well the images convey the meaning of particular adjectives [73]. Although DARCI still relies on extracting specific low level features, it at least attempts to learn the semantic qualities related to those features (in the form of adjectives). In this way DARCI, more than other visual art systems, is able to at least partially perceive *meaning* in the art that is produced.

The example systems just described can produce interesting and novel artifacts. However, without advanced perceptual abilities, the systems lack any notion of understanding and intentionality. The systems can produce something, but can't necessarily tell us why, or what it means. They are instances of Searle's Chinese room [144], that simply follow rules and algorithms, with no comprehension of what is taking place. Just as humans cannot think beyond our perceptions, computational systems cannot think beyond theirs. Some have argued that even human thought and creativity is subject to the Chinese room analogy at the biological (cellular) level. This may be true, but if we aim to build systems that can be creative at a human level, then they must at least have human-level perception.

Somewhat surprisingly, in the case of visual art, current creative systems rarely use state-of-the-art computer vision techniques, like deep neural networks. Certainly having more advanced perceptual abilities would improve the quality of their art by enabling these systems to understand more concrete things. For example, a system could conceivably create an original image of a dog, if it knew how to see and recognize dogs. It seems, then, that incorporating advanced computer vision

techniques, especially ones tied to semantic understanding, should be a high priority in the field of computational creativity.

7.3.1 Visual Art and Deep Learning

The last few years have shown a resurgence of *deep neural networks* (DNNs), especially for computer vision tasks, where they hold current records for several vision benchmarks [52, 153]. Deep learning has the potential to significantly improve visually creative systems as well. A key advantage of DNNs is that they are capable of learning their own image features, while the visual art systems described above all rely on manually engineered features. Thus, deep learning models can provide more advanced perceptual abilities by building better "mental" models of the world.

Some of these deep learning models can already be used directly to improve current artistic systems. In recent work on DARCI, we built a sophisticated semantic model that uses a shallow neural network to associate image features with a vector space model [73]. Here we can show significant improvement by replacing the shallow neural network and extracted features, with a DNN (and the raw image pixels as input). Specifically, we used a deep learning framework, called Caffe [82], and started with the CaffeNet model, which was first trained to recognize 1000 different items using the ImageNet 2012 competition data [141]. We then further trained and fine-tuned the model on DARCI's image-adjective dataset (with a vector space model).

The DARCI system is capable of zero-shot prediction (using the vector space model), meaning it can successfully evaluate images for adjectives that it was not explicitly trained on. We compare DARCI's original results [73] with our deep neural network version in Table 7.1. The results show significant improvement using the DNN to evaluate images, and fully incorporating a deep model into the DARCI system will likely help it to produce more semantically relevant images.

In fact, DNNs have already been used to generate images directly [45, 68, 97]. One particular method, called *gradient ascent* [147], works by essentially using the DNN in reverse. The trained network starts with a random noise image and tries to maximize the activation of the output node corresponding to the desired class to generate. The network then backpropagates the error into the

	Random	*DARCI*	*Deep Network*
Coverage	0.709	0.444	**0.202**
Ranking Loss	0.502	0.199	**0.102**

Table 7.1: Zero-shot image ranking results comparing the DARCI system with our modification of DARCI that uses a deep neural network (lower scores are better). We used the same test set from the original DARCI paper [73]. The use of a DNN improves the system's ability to perceive and understand adjectives in images.

Figure 7.1: Four images generated using gradient ascent from the deep neural network trained on the DARCI dataset. From left to right the images were generated for the adjectives 'vibrant', 'cold', 'fiery', and 'peaceful'. These images are essentially visualizations of the features that the model has learned and demonstrate a form of imagination.

image itself (keeping the network weights unchanged) and the image is slightly modified at each iteration to look more and more like the desired class.

We demonstrate gradient ascent using the same deep model that we trained with the DARCI image-adjective data set, and the resulting images can be seen in Figure 7.1. These images can be thought of as visualizations of the features learned by the model for each adjective. Each adjective's features seem fairly general, except in the case of 'peaceful', where the visualized features are consistent with the fact that most of the training images depict calm beaches. It is theorized that imagination in humans can be partially thought of as running our vision processing systems in reverse [7], in which case our deep neural network is analogously demonstrating its own kind of imagination.

The generated images seem fairly abstract, which is expected for adjectives, especially since the DARCI data set contains a wide variety of scenes, objects, genres, and styles for each adjective label. Deep neural networks are becoming powerful enough to render actual recognizable objects using the gradient ascent method. The ImageNet 2012 competition consists of classifying 1000 different categories of objects ranging from various animals, to clothing, to household items. We

Figure 7.2: Images generated using gradient ascent from the CaffeNet model and the GoogleNet model, both trained on the 2012 ImageNet challenge data. The first two rows of images are from CaffeNet and, from left to right, were generated for 'pool table', 'broccoli', 'flamingo', 'goldfish', 'bald eagle', 'lampshade', 'starfish', and 'volcano'. The last row of images are from GoogleNet and were generated for 'bald eagle', 'tarantula', 'starfish', and 'ski mask'. These original images are certainly not photo realistic, but it is still fairly easy to identify each image's subject. Notice that the two models have different styles because they have learned different features.

took the CaffeNet model (used as the base for the DARCI model), as well as another successful model called GoogleNet [153], and generated several images depicting objects from the 1000 possible categories. The resulting images for several objects can be seen in Figure 7.2.

While the images are not photo-realistic, they are original and do resemble the intended item. Notice how the two models generated images with different styles as each model learned different features. The generation of images using DNNs is currently an active area of computer vision and machine learning research, and several researchers have produced impressive results [45, 97]. The field of computational creativity has yet to significantly leverage the potential of deep learning, although some have alluded to it [77]. However, some researchers have already begun incorporating

deep learning into evolutionary art systems that are capable of rendering images that resemble concrete objects, with interesting results [120].

7.4 To See Is To Create

We have argued that perception is an important aspect of creativity and that more advanced perceptual abilities can lead to more sophisticated creative systems. We also argue that perception is a creative act in its own right. When light hits a person's eyes, it is converted into signals, which travel to the visual cortex via the optic nerve. The brain itself does not receive any light, only information about the light. The brain must then learn to make sense of that information, and an image in the mind is fabricated, and that is what a person "sees". Our brain over our lifetime has built a mental model of the world through the various signals it has received from our senses. This mental model is what determines our personal reality, and it is an impressively creative act [79, 133].

We don't think of perception itself as a creative act because it happens instantly, constantly, and seemingly without effort. We take for granted how difficult perception is because it is an ordinary part of life, and we've become desensitized to it. However, even the most advanced state-of-the-art computer intelligence cannot process visual information as well as a child can almost instantaneously. The case of Shirley Jennings [142], in which he spent months learning how to see for the first time at age 50, and other cognitive visual disorders, shed light on the tremendous amount of work that goes into vision.

Optical illusions also provide insights on how the brain understands visual input and constructs images in the mind [79]. Different people given the same input, experience it differently. A person's subjective experience is unique to them, an act of novelty by their creative brain. This idea became even more evident when a particular image of a dress sparked huge debate on social media over the color of the dress [94]. Some saw white/gold, others saw blue/black, because our brains construct differing realities based on our mental models.

If we accept the idea that our brains are doing the actual creating of the images we see, then what is the artist doing when she paints a picture? The artist is providing a set of constraints, in

the form of a painting, that viewers use to create an image in their minds. The more realistic a painting is, the more it constrains the viewer to mentally create it a certain way. The more abstract or ambiguous the painting is, the less it constrains the viewer, and the more variety and novelty in the individual aesthetic experiences.

7.4.1 Pareidolia

Attributing creativity to a system just because it has some perceptual abilities doesn't appear very compelling. However, there are some perceptual tasks that seem more creative than others. Pareidolia is the phenomenon of perceiving a familiar pattern where none actually exists. For example, seeing constellations in the stars, faces in ordinary things, objects in blotches of ink, or shapes in the clouds. Sometimes these are considered mistakes or optical illusions, but they can actually be a deliberate act of creativity. When a child says a cloud looks like a particular animal, we admire her imagination, especially when we can then see the shape too. We obviously know it's a cloud, but we have chosen to see it as something else.

Pareidolia is a creative act because it is is not about seeing things for what they are but seeing things for what they could be. Creative systems capable of pareidolia may have applications in visual communication, advertising, story telling, illustration, and non-photo-realistic art. As it stands, there are few computational systems developed for automatically performing pareidolia. One group of researchers developed a system for recognizing "faces" in ordinary pictures and then automatically determining the emotion expressed by the "face" [80]. Here we demonstrate how deep learning can be used for pareidolia and argue that it is a form of creativity because the model is interpreting images in novel ways.

Finding Faces

Seeing faces in objects is by far the most common type of pareidolia and provides a simplified version of the problem to begin with. The initial task is to use a deep neural network (DNN) to identify what aspects of an image could make up a face. We then have the DNN iteratively

emphasize those features, using gradient ascent, until the "face" that the network sees emerges in the image. We use two different DNN models trained on faces. The first is called VGG-Face and was trained to recognize the faces of over 2500 different celebrities [130]. The second model, which we'll call AGE-Face, was trained to determine the age of a person (one of eight age ranges) in a provided image [98].

We perform pareidolia by having each network, when given an image, determine the output node (corresponding to a class) with the highest activation. The model then performs the gradient ascent algorithm in an attempt to increase that node's activation further, thus emphasizing the strongest features it found initially. Figure 7.3 shows example pareidolia images generated with both the VGG-Face and AGE-Face networks. The networks generally do a decent job of drawing (cartoony) faces on the source images in ways that make sense, although some are harder to appreciate. The VGG-Face model tends to draw more realistic facial features (i.e., eyes, nose, etc) than the AGE-Face model. However, the VGG-Face model will often highlight isolated facial features (especially when a face in the source image is not apparent to humans), while the AGE-Face model tends to keep the facial features together for a full face.

Finding Objects

We now move on to a harder version of pareidolia in which we ask the model to find and highlight any kind of object in an image. We again use the CaffeNet model that was trained on the 1000 category 2012 ImageNet data; thus the model could potentially see any of those 1000 items in an image. We use the same method as just described in the faces version. The model is given a source image, then performs gradient ascent on the source image in order to further maximize the highest activated output node. We applied this method to several source images, and the results can be seen in Figure 7.4.

For some of the examples it is easy to see why the model did what it did. For instance, it is understandable how the [Figure 7.4, 1st] source image looks like a mask, and we can see how

Figure 7.3: Images created for face pareidolia using deep neural networks. The top row are the source images, the second row are faces highlighted by the VGG-Face model, and the third row are faces highlighted by the AGE-Face model.

the modified image came from it. However, it is more difficult to appreciate how the model saw an 'arctic fox' in the [Figure 7.4, 2nd] source image. Other examples are hard to relate to initially, but on inspection, we can start to see the connection. For example, the [Figure 7.4, 4th] source image looks, to most humans, like a spider, but the model saw it as a ringworm. After considering the resulting image, we can at least appreciate why the model thought ringworm.

This leads to an interesting discussion about perception and creativity. If a person says that a particular cloud looks like a horse, then *if we can also see it*, we think the person has imagination. However, if we can't see it ourselves, then we don't necessarily praise the person's imagination. Conversely, if a person says that a photo of a horse looks like a horse, we also don't admire the imagination, and we end up wondering why they bothered to say something so obvious. We appreciate creativity when it is different from the norm, but not so different that we can't connect.

When it comes to visual art, how a person sees will influence their art; thus, people that see things differently (but not too differently) can potentially be more creative with their art. Using

Figure 7.4: Images for object pareidolia using CaffeNet, trained on the 2012 ImageNet data for 1000 object categories. From left to right, the items highlighted in the images (bottom row) from each source image (top row) are ‘mask’, ‘arctic fox’, ‘scorpion’, and ‘ringworm’.

deep learning models for pareidolia helps us to understand how these models are actually seeing, and it helps us to visualize what features are being learned. The features learned by each model are likely different than the features that human brains use when processing visual input. This is why the CaffeNet model sees an arctic fox in the [Figure 7.4, 2nd] source image, but most humans would say it looks like an elephant.

If a computational system perceives things differently than a human, and accordingly produces different kinds of art, then is the art only viable if we humans can relate to it? It has been suggested that the most creative and influential people are ones that see (and therefore think) differently [11], and Colton argues that computational systems that see differently than humans have enhanced creative potential [31], but is that true only to an extent? Could a computational system that perceives differently (even radically differently) than humans actually help us to extend our notions of what constitutes good art?

To go even further, could we build a system capable of understanding and creating art beyond the capabilities of current human perception? For example, could we build a system that creates infrared art? Or electromagnetic field art? Or gravity art? Or some other kind of art? Would there be any purpose in doing so? Or perhaps augmenting computational systems with other forms

of perception could help them gain a richer, deeper understanding of the world, and allow them to create visual art that can be even more meaningful to humans.

7.5 Conclusion

We have argued that perceptual abilities are fundamental to the creative process. We have discussed the relationship between perception and creativity from a cognitive psychology perspective and also in terms of computational systems. We have even asserted that perception itself is a creative act and that perceiving things differently can facilitate creative thinking. We've demonstrated how state-of-the-art deep neural networks can be used to create images and perform certain types of imagination, and we've also demonstrated how they can see creatively through pareidolia.

As with humans, advanced perceptual abilities can provide a foundation on which computational systems can think, imagine, and create. In the field of artificial general intelligence, current trends and ideas are also advocating the need for perception, and recent general AI systems are learning to perform intelligent tasks exclusively from raw inputs [72, 114]. They argue that having a system learn from the ground up, with raw inputs (e.g., raw pixel values), is essential for general/adaptable intelligence. Perceiving and understanding various raw inputs can act as a basis for a large variety of intelligence tasks, and learning how to perceive and perform for one task should transfer to additional tasks. Furthermore, advanced cognitive ability, such as language and reasoning, could emerge naturally from these perceptual primitives as they form connections and hierarchies of understanding.

The idea of perceptual primitives can also be applied to a general notion of computational creativity. Ideally, we would like to develop a universal creative process, which allows for connections to form across multiple domains, experiences, and knowledge. Perceptual abilities for multiple modalities establish an internal mental model of the world, which can provide a system with freedom and adaptability to be creative in any of its modalities or combination of modalities. For example, a system trying to invent recipes could benefit from visually recognizing ingredients (in addition to understanding how they taste) and could invent new recipes by substituting similar

looking ingredients. It is possible that developing and incorporating advanced perceptual abilities in computational systems will not only increase the creative potential of those systems but may also facilitate the abstraction of a domain-independent, general creativity "algorithm".

Chapter 8

Autonomously Conveying Inspiration In Visual Art Using Deep Neural Networks[1]

Abstract

In visual art, the communication of meaning or intent is an important part of eliciting an aesthetic experience in the viewer. Building on previous work, we present several enhancements to the DARCI system that extend its ability to express meaning and intent through the images it creates. We first improve DARCI's semantic model by incorporating deep neural networks that have been trained to assess the aesthetic quality, the artistic style, and the semantic content of images. This improved semantic model also allows DARCI to find inspiration by looking at pre-existing images. Additionally, we have built an aesthetic rendering component that enables DARCI to more consistently produce aesthetically pleasing images. Finally, we added a title/description generator that allows DARCI to provide additional insight into how and why it creates art the way it does. We use an online survey to show that the system is successful at creating visual art that expresses meaning and is generally appreciated by human viewers.

[1]Derrall Heath, and Dan Ventura, Autonomously Conveying Inspiration In Visual Art Using Deep Neural Networks, *International Journal of Semantic Computing* , to submit, 2016

8.1 Introduction

DARCI (Digital ARtist Communicating Intention) is a system for generating original images that convey meaning and is part of ongoing research in the subfield of computational creativity. Central to the design philosophy of DARCI is the notion that the communication of meaning in art is a necessary part of eliciting an aesthetic experience in the viewer [36]. There are few systems we know of that attempt to autonomously generate images that communicate meaning. The WordsEye system tries to generate 3D scenes based on written descriptions [34]. The Story Picture Engine [85] and the Text-to-Picture Synthesis System [179] are both systems built to do automatic text illustration (i.e., to visually tell a story or to graphically communicate the gist of text). AARON [109] and The Painting Fool [25] are both systems designed to autonomously create visual art in ways meaningful to human viewers. DARCI is unique from other computationally creative systems in that DARCI creates novel, artistic images that explicitly convey concepts. By incorporating a sophisticated semantic model grounded in perception, DARCI is able to internally represent the meaning of concepts, which facilitates the expression of these concepts through images in innovative ways.

It has recently been argued that in order for computational systems to be considered autonomously creative, they must have adequate perceptual capabilities in their respective domain [74]. Creative systems must be able to "understand" to some degree what it is they are producing in order for the system to truly be considered intentional. In this paper, we improve DARCI's semantic model by incorporating deep neural networks that have been trained to assess the aesthetic quality, the artistic style, and the semantic content of images. This improved semantic model allows DARCI to better evaluate and understand its own artifacts in order to produce semantically relevant and aesthetically pleasing images. It also allows DARCI to evaluate other existing images in order to find inspiration and generate its own semantically related images in response.

Although breakthroughs in deep learning have demonstrated state-of-the-art performance on many computer vision tasks [52, 153], they are still generally inferior to human perception. Even if computational systems could visually perceive as well, and even surpass, humans, *how* they perceive would be different than humans because they use different methods and features.

In systems producing visual art, these differences in vision could lead to differences in aesthetic and semantic preferences. However, Colton argues that computational systems that see differently than humans have enhanced creative potential [31]. In cognitive psychology, it has been suggested that creative people literally see the world differently [11], which is in turn why they tend to think differently and can produce novel things and ideas.

Nevertheless, if a computational system perceives things differently than humans, and accordingly produces different kinds of art, then is the art only viable if we humans can relate to it? When attributing creativity, there is generally a balance between novelty and typicality [138]. A creative artifact should be novel and different than what has been produced before, but it should still be a typical instance in the domain. If it is too different from the norm, then people may not be able to connect and relate to it, and thus they may not appreciate its value.

DARCI's perceptual abilities are different than humans, but we want human viewers to be able to appreciate DARCI's art. With this in mind, we built a new component into DARCI that allows it to create a title for its artwork, which will help to communicate DARCI's intent in creating the art. Additionally, DARCI can provide a detailed explanation of each step of the image generation process, which will allow human viewers to understand how DARCI is perceiving and why it did what it did in producing a particular image. Our hypothesis is that the extent in which people understand what/how DARCI is actually perceiving and why DARCI makes certain artistic choices, will influence the extent in which people appreciate and admire the art DARCI produces.

We begin by providing an overview of the DARCI system and then go into detail for individual new components. We then demonstrate the full DARCI system and show several artifacts that DARCI has produced. We then describe our evaluation methods and report on and analyze the results. Finally, we provide a summary, discuss our conclusions, and elaborate on future work.

8.2 The DARCI System

DARCI operates by first being given as input a preexisting source image, which can be a painting, a photograph, or any kind of image. DARCI then analyzes the source image and, in response, creates

its own artistic, novel image that is somehow related or inspired by the original source image. The goal is for DARCI to evaluate and extract semantic content from the source image, in order to acquire some kind of inspiration, and then generate an image that conveys whatever inspiration it found. Ideally, a human viewer should be able to appreciate the resulting image and see how it was influenced by the source image.

At a high level, DARCI is composed of two major subsystems, a *semantic model* and an *image generator* as shown in Figure 8.1. The semantic model includes a vector space model (VSM) that learns semantic relationships between words, and several deep neural networks trained to assess the artistic style, the aesthetic quality, and the semantic content of images. The image generator uses the image composer to create an initial sketch, then uses the semantic renderer to render that sketch to reflect certain semantic qualities. Finally the aesthetic renderer is used to make the final image more aesthetically appealing. Additionally, a title/description generator creates a title (and optionally a description) for the the final image. Here we outline each component individually and focus primarily on the components and algorithms that have been added and modified since the last iteration of DARCI. We will then explain how each piece works together to produce meaningful art.

8.2.1 Vector Space Model

To enable DARCI to express semantic information through pictures, it must first have its own semantic knowledge that can influence the images it creates. Using a vector space model (VSM), we can leverage semantic information gained through written text and transfer it to the task of meaningful image generation. In this iteration of DARCI we use the same VSM as was used in the previous iteration. We provide a simple overview, but more details can be found in prior work [73].

We use a VSM, called the *skip-gram model* [112]. The skip-gram model is a neural architecture that analyses a large corpus and learns to predict the surrounding words given a current word. During training, the skip-gram model consequently learns vector representations for each word, which encode semantic information. Words similar in meaning will have vectors that are close to each other in "vector space". These word vectors capture other interesting semantic relationships

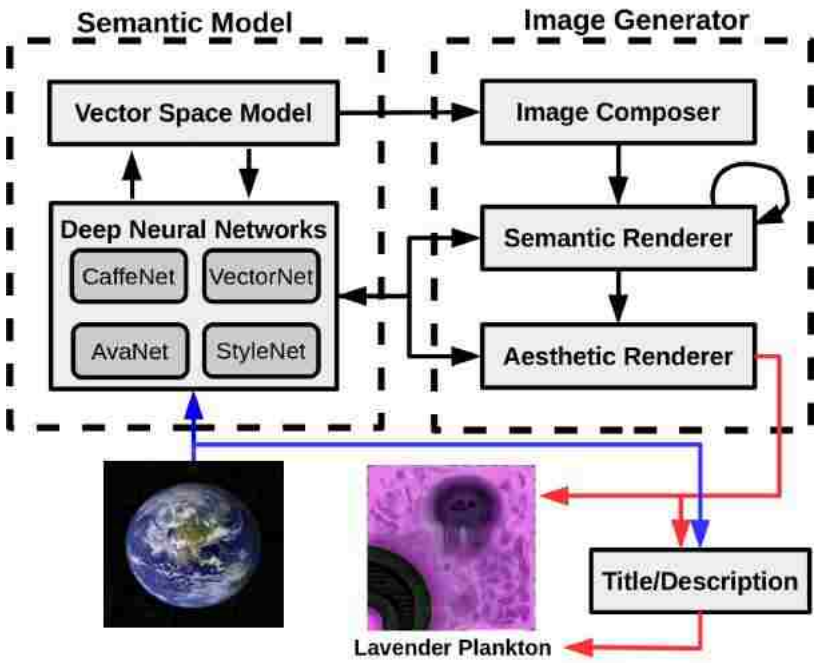


Figure 8.1: A high level overview of the DARCI system. The *semantic model* takes a preexisting source image and the vector space model is used in conjunction with several deep neural networks to evaluate the semantic content and style of the image. The *image generator* then uses the vector space model to identify concepts associated with the semantic content of the source image. The concepts are composed into a source image (image composer) that is rendered to convey the discovered semantic content using a genetic algorithm (semantic renderer) that is governed in part by the deep neural networks. The resulting image is then further rendered to be more aesthetically pleasing and to resemble the source image's style (aesthetic renderer). The final product is a novel image that is inspired by the original source image. A title (and optionally a description) is also generated based on both the original source image and the final image.

that are consistent with arithmetic operations. For example, $vector(\text{“}king\text{”}) - vector(\text{“}man\text{”}) + vector(\text{“}woman\text{”})$ results in a vector that is closest to $vector(\text{“}queen\text{”})$.

These semantic vectors allow DARCI to find concepts related to a given word and to assess the similarity in meaning between words, which will aid DARCI in creating meaningful images. We use a publicly available implementation of the skip-gram model[2] and a lemmatized Wikipedia corpus to learn the word vectors [44]. The skip-gram implementation is used with out-of-the-box parameters except for the vector size, which is set to 300. The choice of 300 provides a balance

between encoding enough semantic information to be useful and ease of prediction when associating the vectors with images.

8.2.2 Deep Neural Networks

In order for DARCI to leverage the word vectors for image creation, it must learn to associate image qualities with the semantic vectors. In prior work, DARCI used a single shallow neural network that did multi-target regression to associate various low-level image features with the word vectors inferred from the VSM [73]. Training was limited to vectors representing adjectives and the neural network model could be used to predict the adjective vector for a given image. In this iteration of DARCI, we replace the shallow neural network and low-level features with four deep neural networks, which we will call *CaffeNet*, *VectorNet*, *AvaNet*, and *StyleNet*. These models were all implemented and trained using a deep learning framework called Caffe [82].

CaffeNet

The CaffeNet model is a pre-trained network that was trained on the ImageNet 2012 competition data [141], which consists of classifying 1000 different categories of objects ranging from various animals, to clothing, to household items. This model gives DARCI the capability to perceive and recognize any of those 1000 items in any image as it finds inspiration for its own artwork. It should be noted that the labels for some of the 1000 categories were modified to be consistent with our vector space model. For example, several of the ImageNet categories are various types of dogs that are not present in DARCI's VSM vocabulary (e.g., 'Maltese dog'), so they were simplified to share the label 'dog'.

VectorNet

The VectorNet is trained to evaluate how well a given image conveys particular adjectives. For example, VectorNet can give a score between 0 and 1 for how 'happy' or how 'scary' an image is.

[2]https://code.google.com/p/word2vec/

It is different from the other deep networks in that it does not do classification. Instead of trying to predict an image's adjective label, it does multi-target regression to predict an image's adjective *vector* (from the VSM), which is essentially mapping the image into vector space. To determine how much the image conveys a particular adjective, cosine similarity is used between the image's predicted vector and the adjective's vector, which yields a similarity score between 0 and 1.

We maintain a dataset of approximately 15,000 images that have either been explicitly hand labeled or automatically retrieved through Google image search. Once an adjective has enough labeled images (20 positive and 20 negative), we begin learning that adjective. As of this paper, there are 145 adjectives that meet this threshold. Our dataset is still very small for training deep neural networks, and so instead of training from scratch, we start with the pre-trained CaffeNet model. CaffeNet has been trained with millions of images and has thus already learned various useful and general image features.

We take this pre-trained model and fine-tune it for adjective vector prediction by replacing the 1000 dimensional softmax output layer of CaffeNet with a 300 dimensional linear output layer, where each output node corresponds to each of the 300 dimensions in the vector space model. Instead of training with the typical softmax loss used for classification, the VectorNet uses a Euclidean loss for regression (or multi-target regression in this case). The VectorNet model was trained for 10,000 iterations, with a base learning rate of 0.001 (divided by two every 2000 iterations), a momentum of 0.9, and a weight decay of 0.0005.

Learning to predict an adjective's vector is a harder task than learning to predict the adjective directly and, thus, introduces a few trade-offs. First, labeling images with adjectives is a multi-label classification problem (i.e., an image can be described by more than one valid adjective) and our new model can only predict one vector at a time, while normally each adjective could be predicted independently. The second trade-off is that VectorNet is predicting a 300 dimensional vector and has to account for every adjective. This means that it may not predict the 145 adjectives as accurately as using separate models for each adjective.

	Random	*Former Model*	*VectorNet*
Coverage	0.709	0.444	**0.202**
Ranking Loss	0.502	0.199	**0.102**

Table 8.1: Zero-shot image ranking results comparing our new VectorNet model that uses a deep neural network compared to the model used in the previous iteration of DARCI [73] (lower scores are better). The use of a deep neural network improves the system's ability to perceive and understand adjectives in images.

The main advantage of learning the vectors, however, is that we can do zero-shot prediction. In other words, VectorNet is not limited to the 145 adjectives for which it was explicitly trained. It allows DARCI to take advantage of the semantic structure between words and to evaluate images according to adjectives it never explicitly learned. For example, DARCI could be trained on 'scary' and 'dark' images, but not 'creepy' images. DARCI could then "understand" what a 'creepy' image looks like because 'creepy' is similar in meaning to 'scary' and 'dark'. Even higher level concepts (e.g., 'love', 'freedom') can be partially understood by VectorNet through the word vectors.

In prior work, DARCI successfully used a shallow neural network and extracted image features to associate the adjective vectors with images [73]. The original model was evaluated using image ranking metrics, and we compare DARCI's original results with our deep neural network version in Table 8.1. The results show significant improvement using the deep VectorNet for zero-shot image ranking. VectorNet will be used by DARCI to analyze the source image for inspiration, and it will also be used to guide the semantic rendering process to generate images with particular semantic qualities.

AvaNet

The AvaNet model is a binary classification model trained to determine the aesthetic value of an image. AvaNet is trained using the AVA dataset, which is a large-scale database for aesthetic visual analysis [117]. This dataset consists of 255,000 images from the online photography site `www.dpchallenge.com`, and each image has aesthetic ratings (from 1 to 10) from hundreds of professional and amateur photographers.

We prepared the dataset by first averaging the aesthetic ratings for each image. If the average rating was less than 5.0, it was considered not aesthetically pleasing, and if the average rating was greater than 6.0, then it was considered aesthetically pleasing. The rest of the images were considered ambiguous and were discarded from the training set, but were still *included* in the test set (with 5.5 as a threshold). Using this method we obtained a training set of 112,033 images and a test set of 25,585 images.

We again started with the pre-trained CaffeNet model and fine-tuned it for aesthetic classification by replacing the 1000 dimensional softmax layer of CaffeNet with a 2 dimensional softmax layer. The AvaNet model was trained for 100,000 iterations, with a base learning rate of 0.001 (divided by ten every 15,000 iterations), a momentum of 0.9, and a weight decay of 0.0005. The AvaNet model was able to achieve a classification accuracy of 69.6% on the test set, compared to the original AVA paper, which reached a maximum accuracy of 67.0% using SVMs and SIFT/color features [117].

DARCI will use AvaNet to evaluate final artistic modifications to its images in the aesthetic rendering component. One of the weaknesses of AvaNet is that it was trained with photographs and not with paintings (or other non-photorealistic art), which can make it focus and converge on a smaller variety of artistic modifications. The fourth deep neural network, StyleNet, is used in conjunction with AvaNet to help overcome this problem.

StyleNet

The StyleNet model is a pre-trained model from the Caffe deep learning library that was trained on the Flickr Style dataset [82, 88]. The model classifies images into one of twenty different image styles ranging from painting styles (e.g., impressionism, cubism) to photography styles (e.g., macro, long exposure). DARCI uses this model in the aesthetic rendering component in conjunction with AvaNet to evaluate and choose final artistic rendering modifications to its images. This model is actually used as a style similarity metric between the original source image (the inspiring image) and the images DARCI creates. We take the 20 output activations of the top layer of StyleNet

for each image as a vector and perform cosine similarity between them to determine how similar in style they are. DARCI favors artistic modifications that make its images closer in style to the original source image and is aesthetically pleasing according to AvaNet.

8.2.3 Finding Inspiration

In order for DARCI to create its own artwork, it needs somewhere to start, it needs some concept or idea that it can express through its art. The deep neural networks just described give DARCI some ability to perceive the world through images. With this perception it can look at preexisting images and partially understand them and use that understanding as inspiration for its own art. A preexisting source image is inputed into the system, and DARCI uses the top two object labels from CaffeNet and the top two adjective labels from VectorNet for inspiration. CaffeNet provides a concrete idea, while VectorNet provides a descriptive/emotional connotation. We opt to use the top two labels from each model (as opposed to just the top) in order to facilitate variation from the original source image. We want DARCI's image to be related to the source image, but also allow for some leeway.

Figure 8.2 shows three example source images with the top two labels from the CaffeNet and VectorNet models, as well as a training image from ImageNet that, according to CaffeNet, closely resembles each source image. In the first source image, CaffeNet has no trouble recognizing it as an 'eagle' because 'eagle' is one of the 1000 categories it was trained on (as demonstrated by its respective training image). For the second source image, however, 'earth' or 'planet' is not one of the 1000 categories and CaffeNet misclassifies the image as a 'jellyfish'. At first this seems completely wrong, until the example training image of 'jellyfish' is considered and it can be understood why CaffeNet would think 'jellyfish'. Likewise, the third image is non-photorealistic and CaffeNet sees it as a 'petri dish', which is justifiable according to the example training image.

DARCI's perception of concrete objects is limited to 1000 items. Furthermore, the CaffeNet model was trained on photographs, not paintings or other non-photorealistic images, which makes accurately recognizing objects in such images difficult. This seems like a severe limitation for

(a) eagle (kite), menacing (enigmatic)

(b) jellyfish (slug), dark (bright)

(c) petri_dish (fan), simple (cute)

(d)

(e)

(f)

Figure 8.2: The top row shows three example source images inputed into the DARCI system. The second row shows example training images from the 2012 ImageNet dataset that most closely resembles its respective source image (according to CaffeNet). Each source image is labeled with the top category/adjective determined by both CaffeNet and VectorNet respectively (words in parenthesis are the second place labels). CaffeNet labels the eagle picture correctly because eagle is one of the 1000 object categories that it knows. The other two source images depict an object not of those 1000 categories, or is non-photorealistic. In this case, CaffeNet determines the closest matching category of the 1000 it knows and displays an example training image for justification.

DARCI, however, it can make things interesting from a creativity standpoint. Just as we humans tend to look for shapes in the clouds or constellations in the stars, DARCI will look for those 1000 items in the various source images it encounters. DARCI will find inspiration according to its unique perceptual abilities and experiences, which has the potential to enhance its creative ability.

8.2.4 Image Composer

With the vector space model and deep visual models in place, DARCI can begin to produce images. The image generation process starts with the image composer, which makes heavy use of the vector space model to create an initial sketch. The image composer works similar to how it was done in prior work [75], with a few modifications that we will outline here. DARCI finds inspiration in the

form of two noun concepts and two adjective concepts, and the image composer uses the vector space model to retrieve their respective vectors and average them together into a single vector.

The image composer gets the concept nearest to the averaged vector (using cosine similarity) and finds an image on the internet, based on that nearest concept, to be used as a background. Additionally, DARCI maintains a collection of iconic images for approximately 5,000 concepts, any of which can be arranged on the background image [154]. At this point, the image composer uses the VSM to retrieve three to five of these 5,000 iconic concepts that are semantically associated to the averaged vector. The idea is that DARCI can effectively make use of these word associations as a decomposition of a (high-level) concept into simpler concepts that together represent the whole. In many cases, if a (sub)concept is simple enough, it can be represented visually with a single icon (e.g., the concept 'rock' can be visually represented with a picture of a 'rock').

Each of these icons is scaled linearly according to how semantically similar its respective concept is to the averaged vector. The most semantically related icon is placed on top of the background image near the center and subsequent icons are placed randomly spiraling out from the center, avoiding overlap if possible. To prevent the background image from dominating and to help the icons blend in smoothly, the background transparency is set to 75%, while each icon transparency is set to 25%. The result is a collage of icons on a faded background image that is conceptually related to the original inspiration (the four words) found from the source image. For example, suppose the closest concept to the averaged vector was 'snow'. The vector space model would send related words like 'snowflake', 'ice', 'snowman', and 'ski' to the image composer. The resulting image would be some composition of simple iconic images of the associated words overlayed on a background related to 'snow'.

8.2.5 Semantic Renderer

The semantic renderer takes the initial sketch from the image composer and uses a genetic algorithm to discover a sequence of image filters for rendering the sketch so that it will artistically reflect the meaning of the averaged word vector. In this iteration of DARCI, the genetic algorithm now

incorporates VectorNet in its fitness function as opposed to a simpler shallow network and extracted images features used in past iterations. At each generation, each candidate image is evaluated by VectorNet to determine how close (using cosine similarity) its predicted vector is to the averaged word vector. Other aspects of the fitness function include measures of similarity to the original sketch (the sketch image should remain somewhat recognizable) and measures of similarity to what it has produced before (it should be trying different things). The semantic renderer is the oldest part of DARCI and has been evaluated and described extensively in prior work. In this paper, the semantic renderer is implemented and used exactly as done in previous iterations (with the exception of incorporating VectorNet), and further details can be found in prior work [126].

8.2.6 Aesthetic Renderer

The output of the semantic renderer can often be raw and not very consistent with typical artistic styles. The aesthetic renderer takes the semantic renderer output and further modifies and refines the image with special artistic filters. DARCI has a collection of 10 such artistic filters that simulate various mediums and styles such as colored pencil, chalk, finger paint, etc. These filters are computationally expensive and, thus, it was not practical to incorporate them into the semantic renderer's genetic algorithm.

The aesthetic renderer applies every filter with all their possible parameter settings and picks the best one, which results in a total of 76 filter/setting combinations. The best filter/setting combination is determined by both AvaNet and StyleNet. AvaNet gives each combination a score between 0 and 1 for how aesthetically pleasing the resulting image is, while StyleNet gives a score between 0 and 1 for how closely the resulting image's style matches the original source image's style. These two scores are averaged together and the filter/parameter combination with the highest average score is chosen for the last rendering step. The output of the aesthetic renderer is the final image generated by the DARCI system.

8.2.7 Title/Description Generator

The title/description generator produces a title for the resulting image that helps to communicate its connection with the original source image, as well as provide insight as to its intent in creating the image. The title/description generator can also produce a more detailed explanation of DARCI's "though process" and walks the viewer through each main step of creating the image. If a human viewer can understand why the system created art the way it did, then that person can better appreciate and enjoy the art.

We use a simple template strategy to generate a title, where the system randomly chooses from a list of, one to two word, title templates. These templates specify what part-of-speech (e.g., noun, verb, adjective) should be inserted where. Example templates include "adjective noun", "noun verb", and "adverb". Once a template is chosen, the system must insert appropriate words in the template. These words should relate to both the inspiration DARCI found in the source image, as well as to how the final image actually turned out.

We start by gathering potential words to fill in the template, which initially include the top noun and the top adjective from the original inspiration. We then use CaffeNet and VectorNet to evaluate the final image and get an additional noun and adjective. At this point, our initial set of words include two nouns and two adjectives. To handle verbs and adverbs, to add variety, and to generate more interesting titles, we use the vector space model to find alternative semantically relevant words as we fill in the template.

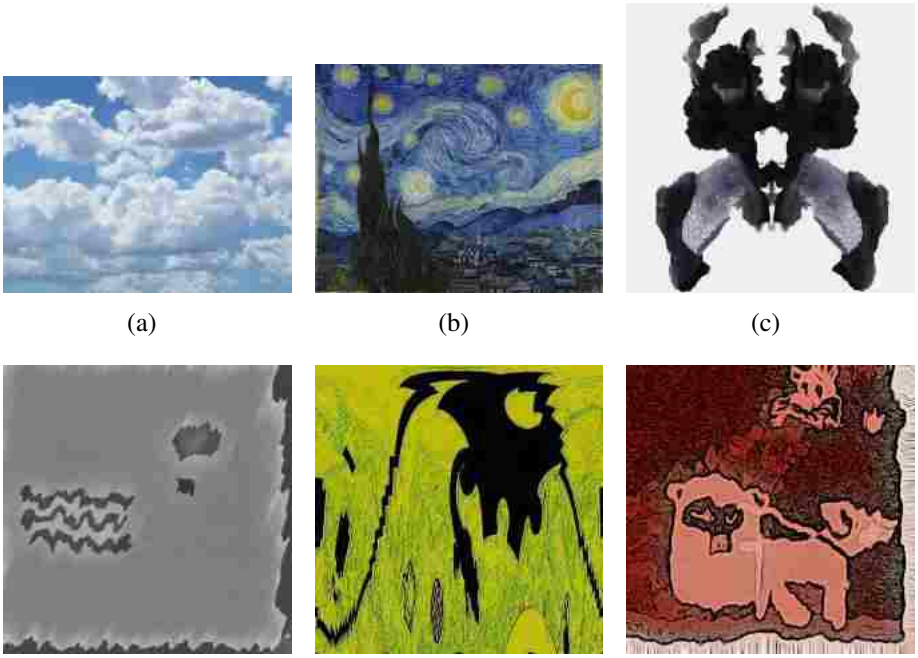
For example, when we encounter an "adjective", the system randomly picks one of the two initial adjectives, randomly picks one of the top 5 semantically similar adjectives from the VSM, and inserts it in place of the "adjective" placeholder. For "noun" we randomly pick one of the two initial nouns and then randomly pick from the top 5 semantically similar nouns. For "verb", we randomly pick from the top 5 semantically similar verbs (based on one of the 2 nouns). For "adverb", we randomly pick from the top 5 semantically similar adverbs (based on one of the 2 adjectives). The result is a title for the final image that is relevant to the initial inspiration, and/or the final image.

Optionally, DARCI can generate a more detailed description of its "thought process" as it generated the art. This description shows the original source image and explains what inspiration it found. It then justifies what it saw by showing a similar image it has seen before from the CaffeNet training data (see Figure 8.2 for examples). The system then explains how it created an initial sketch with a background image and icons. It then shows the modified sketch after the semantic rendering. Finally, the description shows the final image after the aesthetic renderer and displays the title.

8.2.8 End-to-End Image Generation

DARCI's end-to-end image generation process has several steps and follows the diagram presented in Figure 8.1. Here we summarize the entire process at a high level and show some example images.

- A preexisting source image is inputed into the system.
- The source image is analyzed by both CaffeNet and VectorNet to find inspiration.
- CaffeNet returns the top 2 nouns it sees in the source image and VectorNet returns the top 2 adjectives.
- The word vectors (from the Vector Space Model) for each of the these words are averaged together into a single vector.
- The Image Composer creates an initial sketch by placing semantically associated icons on top of a related background image.
- The Semantic Renderer uses a genetic algorithm and VectorNet to render the initial sketch in a style consistent with the averaged word vector.
- The Aesthetic Renderer further modifies the image by applying a special artistic filter that is chosen based on AvaNet and StyleNet.
- Finally, a title (and optionally a detailed description) is generated based on the initial inspiration and the final image.



(a) (b) (c)

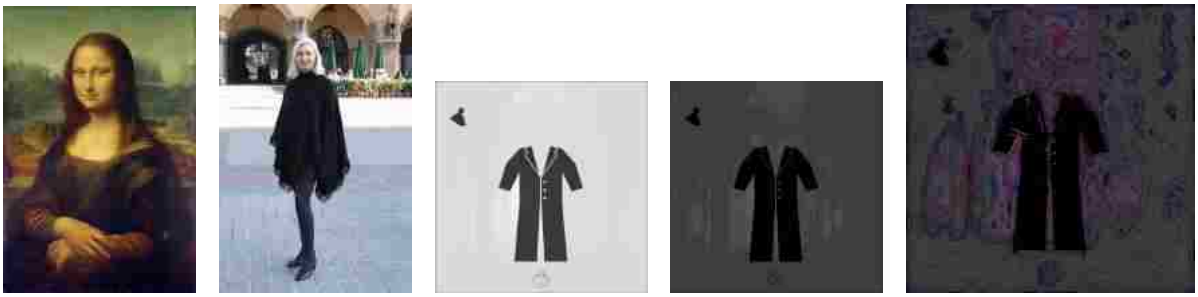
(d) Overcast Slop (e) Whimsical Puzzle (f) Edgy Pushover

Figure 8.3: The first row shows three example source images inputed into the DARCI system. The second row shows the final images DARCI produced for each source image, including their respective titles.

Figure 8.3 shows example images generated by DARCI with just the source image, the final image, and the title. Figure 8.4 shows example intermediate images from the entire image creation process, including a full description generated by DARCI.

8.3 Evaluation

DARCI is a large system with many components that work together to autonomously produce visual art, and each has been evaluated extensively in prior work [73, 75, 126]. Here we focus on the DARCI system as a whole and evaluate how well the system can create visual art that people can appreciate. We want to gain insights into the correlation between perceived meaning expressed through DARCI's art and its perceived creativity. Additionally, DARCI's perceptual and artistic abilities are different than humans, which can potentially facilitate the creation of more interesting



(a) source (b) example (c) sketch (d) rendered (e) **Overdress**

Figure 8.4: An example showing the intermediate images of each step of DARCI's image creation process. A generated description is as follows: "I was looking for inspiration from this image (a), And it made me feel **gloomy** and **dreamy**. It also made me think of this image that I've previously seen (b), which is a picture of a **poncho**. So I started an initial image of my own by searching for a background image on the Internet based on **poncho**, **gloomy**, and **dreamy**. Then I took basic iconic images associated with those concepts and resized/placed them on the background according to how relevant they were. This was the result (c). I then modified it in a style related to **poncho**, **gloomy**, and **dreamy**, which resulted in this image (d). I did a final modification based on aesthetic quality and how closely the style related to the original image (e). The end result perhaps looks more like a **cloak** or a **vestment**, and it feels particularly **gloomy**. It is called **Overdress**."

and novel images. However, if the art is too different, then human viewers may not be able to enjoy the art. Thus we want to assess how much people enjoy and appreciate DARCI's art across different degrees of understanding DARCI's "thought process".

Will knowing more about DARCI's process for creating art help people to appreciate its art more or less? If DARCI has a chance to "explain itself", then will people relate to its artwork better? Will they be more willing to say that DARCI's images are creative and that the system has intention? In order to evaluate this, we use use an online survey and ask human viewers to assess DARCI's art.

8.3.1 Online Survey

We had DARCI generate 30 images inspired by a large variety of source images, ranging from photographs, to famous artwork, to other computer generated art. Each survey participant evaluated four of DARCI's images, which were randomly selected from the pool of 30 for each user. To start the survey, each user was given instructions and a brief overview of DARCI. This overview

explained that DARCI is a computer system that creates visual art that conveys meaning, and that it works by looking at a source image for inspiration.

The participants of the survey were randomly assigned to one of three groups, which we will call BASIC, DESC, and BOTH. The users in the BASIC group were only shown the source image, the final image, and the title (see Figure 8.3), and then asked to evaluate the final image. The users of the DESC group were shown the full description generated by DARCI (see Figure 8.4), and then asked to evaluate the final image. In the BOTH group, the first two images the user saw were the full description, while the last two images were just the source image, the final image, and the title. After evaluating the four images, users were asked a few questions about the DARCI system in general.

When querying users about each image, we followed a survey template similar to one we developed in a previous study to measure the perceived creativity of images and to assess how well they communicated an intended concept [75]. Our here survey consisted of seven Likert items (7 point scale) [100], where volunteers were asked how strongly they agreed or disagreed with each statement as it pertained to one of DARCI's images. The seven statements we used were (abbreviation of item in parentheses):

I like the image. (*like*)

Prior to this survey, I have never seen an image like this one. (*never seen*)

I think the image would be difficult to create. (*difficult*)

I can see how the image relates to the source image. (*relate*)

I think DARCI created the image with intent. (*intent*)

I think the image conveys meaning. (*meaning*)

I think the image is creative. (*creative*)

For comparison, *like*, *never seen*, *difficult*, and *creative* are the same questions used in our previous study, while the other three questions were added for this survey. After the users evaluated four of DARCI's images, they were asked the following (Likert) questions about the DARCI system in general:

I can appreciate DARCI's art. (*appreciate*)

I can understand why DARCI creates art the way it does. (*understand*)

DARCI Images			**DARCI System**	
Omitted Item	*Alpha Value*		*Omitted Item*	*Alpha Value*
None	0.830		None	0.744
Like	0.806		Appreciate	0.626
Never seen	0.843		Understand	0.795
Difficult	0.799		Relate To	0.613
Relate	0.825		Being Creative	0.683
Intent	0.802			
Meaning	0.788			
Creative	0.779			

Table 8.2: Alpha values measuring consistency of survey questions for both the image questions and the DARCI system questions. The lower the alpha value, the more consistent the omitted item is with the rest of the items. For the image questions, *creative* is the most important question, while *relate to* is most important for the DARCI system questions.

I can relate to how DARCI creates art. (*relate to*)

DARCI is being creative. (*being creative*)

8.3.2 Results

We received a total of 307 participants who took the survey, with 103, 101, and 103 participants in groups BASIC, DESC, and BOTH, respectively. Some survey participants never fully completed the survey, but we still used their partial responses. This led to each image being evaluated by an average of 27 people in total across all three groups.

Question Consistency

In a Likert scale survey, it is important that the items correlate with each other to some degree. Since we are interested in understanding how conveying meaning and intent influences the perceived creativity of DARCI and its images, the Likert questions should be consistent with each other. We calculated the Cronbach alpha of the survey with respect to these items [35]. Our survey received Cronbach alphas of 0.830 and 0.744 for the image questions and DARCI system questions, respectively, indicating a high degree of consistency. In order to determine which questions were the most pertinent to the consistency of the survey, we calculated the Cronbach alpha with each question omitted. The results for the image questions (across all images and groups) and the DARCI system questions are shown in Table 8.2.

From this we see that the two most important items for the image questions, in terms of consistency with the other items, are the statements: "I think the image is creative", and "I think the image conveys meaning". Removing these statements resulted in the greatest drop in alpha values. The least important item was "Prior to this survey, I have never seen an image like this one." Removing this question actually resulted in a higher alpha than had the statement remained. In any case, every question omission still resulted in a satisfactory alpha value. From this we conclude that all of the items are valuable to our survey. In addition, since the most consistent items were one directly asking about creativity and one asking about conveying meaning, we conclude that the average survey results do indeed offer a valid measure of creativity and that images that better convey meaning are thought to be more creative.

For the DARCI system questions, the most important item is the statement: "I can relate to how DARCI creates art". The least important is the statement: "I can understand why DARCI creates art the way it does". This is interesting because it indicates that being able to relate to DARCI's art is important, but understanding the details of DARCI's creative process does not matter that much. Indeed, it may even be detrimental to the perceived creativity of the system as we'll see when we consider the results for the different groups of participants.

Group Comparison

We compare the average ratings of each group of survey participants for the image questions in Figure 8.5 and the DARCI system questions in Figure 8.6. For all image questions, there were no statistically significant differences between the groups. Although the *intent* question was slightly higher for the DESC group, and the *like* question was slightly higher for the BASIC group. For the DARCI system questions, the *understand* question had the only statistically significant difference, with the DESC group rating it higher than the BASIC group. The *being creative* and *appreciate* questions, however, were rated slightly higher by the BASIC group participants.

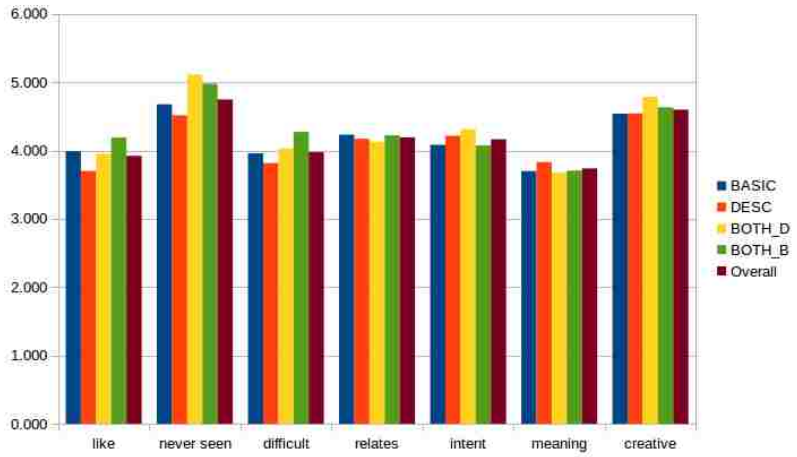


Figure 8.5: Average ratings for all seven image questions comparing each group of survey participants. The BOTH group was broken down into the first 2 images with the full description (BOTH_D) and the last 2 images with the basic description (BOTH_B). There is no statistically significant difference between the groups. However, the *intent* question was slightly higher for the DESC group, and the *like* question was slightly higher for the BASIC group.

The DESC group participants were given a non-technical explanation of DARCI's "thought process" for each image it created, while the BASIC participants were only given the source image, final image, and title. Thus, it makes sense that DESC users would rate the *understand* question higher than BASIC users. However, having the full description made the DARCI system, in general, seem less creative and its art less liked, which seems counterintuitive. A reason for this could be that providing people with an explanation removes the "mystery" behind the artwork. Without the full description, people had to come up with their own explanations as to why DARCI did what it did. The people in the BASIC group had to work harder at discovering meaning in the art, while the DESC users had it partially given to them.

This seems to be consistent with Csíkzentmihályi's notion that "...the aesthetic experience occurs when information coming from the artwork interacts with information already stored in the viewer's mind..." [36]. The absence of the full description allowed people in the BASIC group to more freely interact with the artwork, and were thus more likely to have a positive aesthetic experience. This in turn is why the BASIC group liked the images more and rated the DARCI system as being more creative than the DESC group.

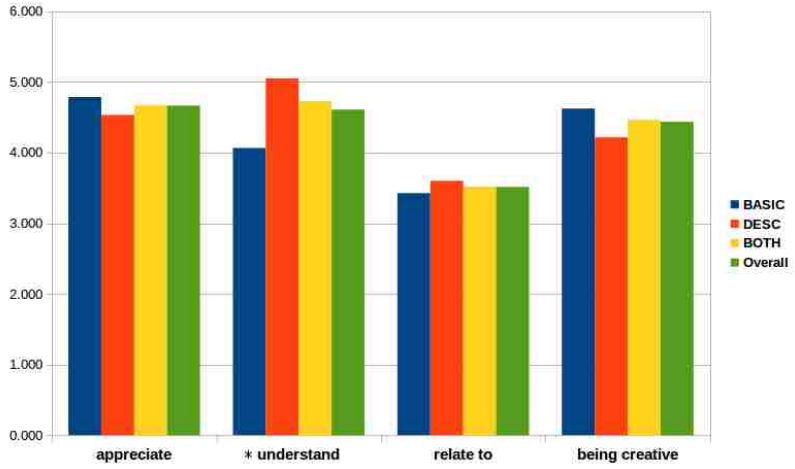


Figure 8.6: Average ratings for all four DARCI system questions comparing each group of survey participants. The only question with statistical significance between the groups was the *understand* question. The DESC group understands DARCI's art creation process better than the BASIC group. However, the BASIC group thinks DARCI is slightly more creative than the DESC group.

Comparison With Prior Results

In Figure 8.7, we compare the overall average ratings of the *like*, *never seen*, *difficult*, and *creative* image questions with the ratings for the same questions in a prior study evaluating a previous iteration of DARCI [75]. The previous study used a 5 point scale, while we currently use a 7 point scale, so the results show the previous study adjusted to a 7 point scale for comparison. We see that the current version of DARCI is generally rated higher for all questions than the older version, which demonstrates that current enhancements to the DARCI system are improving its ability to create visual art that is generally more liked and thought to be more creative by human viewers.

Best and Worst Images

Here we show the best and worst of the 30 images in the survey according to each question across all groups. Figure 8.8 shows the highest rated images, while Figure 8.9 shows the lowest rated images for each question. Because the image questions are generally consistent with each other, often the same image would be rated the highest or lowest for multiple questions.

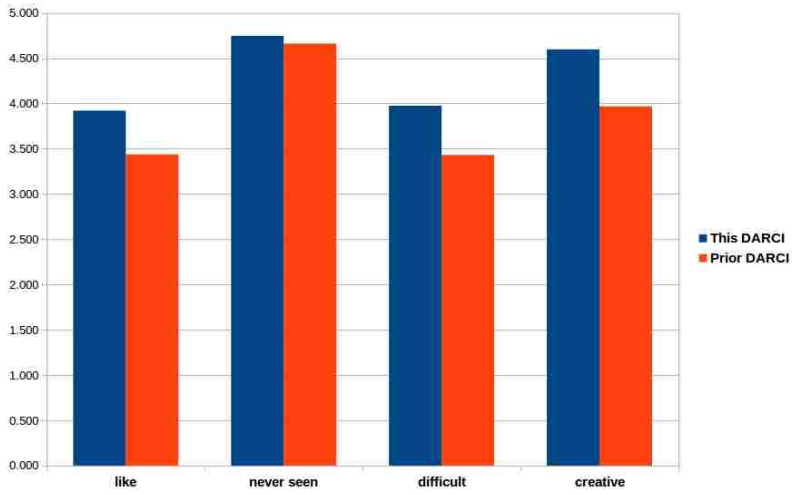


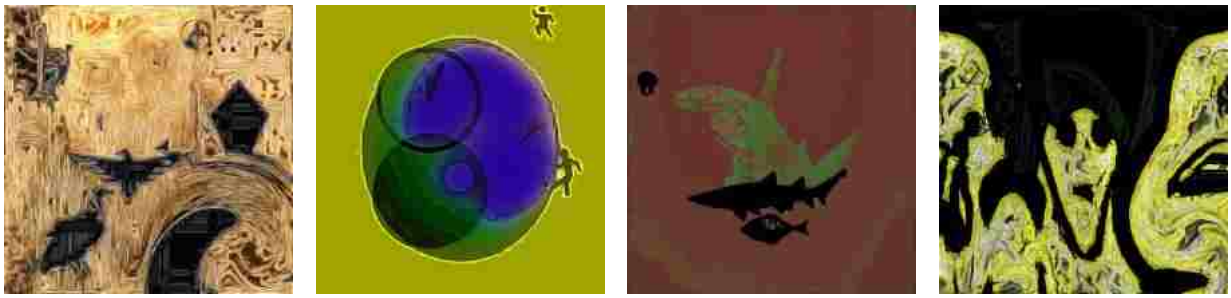
Figure 8.7: Average ratings for four of the image questions compared to the ratings of the same questions on a previous iteration of DARCI. These results indicate improvement in that people generally prefer the artwork created by our most recent version of DARCI.

8.4 Conclusion

We have outlined the DARCI system and elaborated on new components of the system in detail. Specifically, we improved DARCI's semantic model by incorporating deep neural networks that have been trained to assess the aesthetic quality, the artistic style, and the semantic content of images. This improved semantic model allows DARCI to better evaluate and understand its own artifacts in order to produce semantically relevant and aesthetically pleasing images. It also allows DARCI to evaluate other existing images in order to find inspiration and generate its own semantically related images in response. Additionally, we included an aesthetic rendering component that enabled DARCI to more consistently produce aesthetically pleasing images, and a title/description generator that allows DARCI to provide insight into how and why it creates art the way it does.

We evaluated DARCI through an online survey, where participants rated DARCI's artwork on various criteria related to creativity, expression of meaning, and likability. We concluded that images that are perceived to convey meaning are also perceived as more creative. The results also indicate that people who had a better understanding of DARCI's creative process tended to appreciate the artwork less. This is likely due to removing the "mystery" behind DARCI's art and restricting the viewers ability to find their own meaning in the artwork. Additionally, we compared

(a) like, difficult, creative (b) never seen (c) relate (d) intent, meaning



(e) Melancholy Bungalow (f) Chime Oversleep (g) Trout (h) Disturbingly

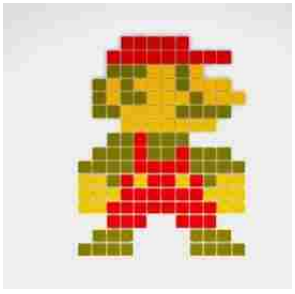
Figure 8.8: The highest rated of DARCI's artwork corresponding to each image question. The top row shows the inputed source image along with the question(s) DARCI's image was rated highest for. The bottom row shows the corresponding image DARCI created along with its title.

the results of the survey with results from a previous iteration of DARCI. We found that DARCI's newer work is generally more liked and thought to be more creative than the past iteration, which demonstrates improvement.

In future work, we would like to extend DARCI's abilities in generating and modifying images. Currently, it is limited to arranging collages of icons and Photoshop-like image filters. We want to provide DARCI with a more varied set of artistic skills, which could include the ability to make individual brush strokes, to do procedural/algorithmic art, or to incorporate 3D models and visual effects. We would also like to more fully incorporate the deep neural networks into the image rendering process. These models are capable of not only understanding images, but also directly generating novel and semantically relevant images through various state-of-the-art techniques [45, 97]. Such methods would allow DARCI to directly make use of its perceptual abilities to facilitate imagination and to potentially generate actual concrete objects and scenes instead of relying on collections of icons.

(a) relate

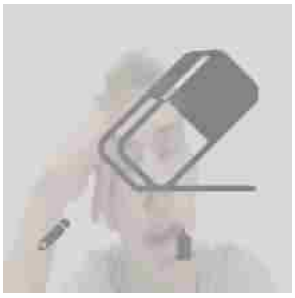
(b) intent, meaning



(c) like, never seen, difficult, creative

(d) Bluish Schnapps

(e) Mouse



(f) Textured

Figure 8.9: The lowest rated of DARCI's artwork corresponding to each image question. The top row shows the inputed source image along with the question(s) DARCI's image was rated lowest for. The bottom row shows the corresponding image DARCI created along with its title.

Chapter 9

Exhibitions, Galleries, and Art Community Collaborations

Over the years we have participated in several exhibitions, galleries and collaborations with other artists/systems. These events have been in collaboration with David Norton and his early work on DARCI. DARCI has had the opportunity to act as an art juror, an artist, and a collaborator, and we summarize each of these events here.

9.1 Fitness Function

A semester long collaboration with Brigham Young University's Visual Arts program culminated in an interactive art exhibit held in the BYU Harris Fine Arts Center from March 19th to the 30th, 2010 [122]. The exhibit was called *Fitness Function*[1], and DARCI acted as the curator. Initially the exhibit consisted of a room with only DARCI (installed on a computer), a printer, some instructions, and a blank wall. Visitors could submit their art to DARCI, and DARCI evaluated it in a way similar to how it evaluates its own images. The submissions that DARCI scored high enough were printed, and the visitor could hang their artwork on the wall. Figure 9.1 shows photographs of the gallery after it received several qualifying submissions. Reactions to the exhibit were varied, ranging from outrage, to intrigue, to excitement, and it challenged many notions about how art should be judged and about what makes art good. In that regard, DARCI's involvement in *Fitness Function* was a success as it caused a re-evaluation of what art really means.

[1] A website documenting Fitness Function can be found at: `http://darci.cs.byu.edu/fitnessfunction/`

(a) (b)

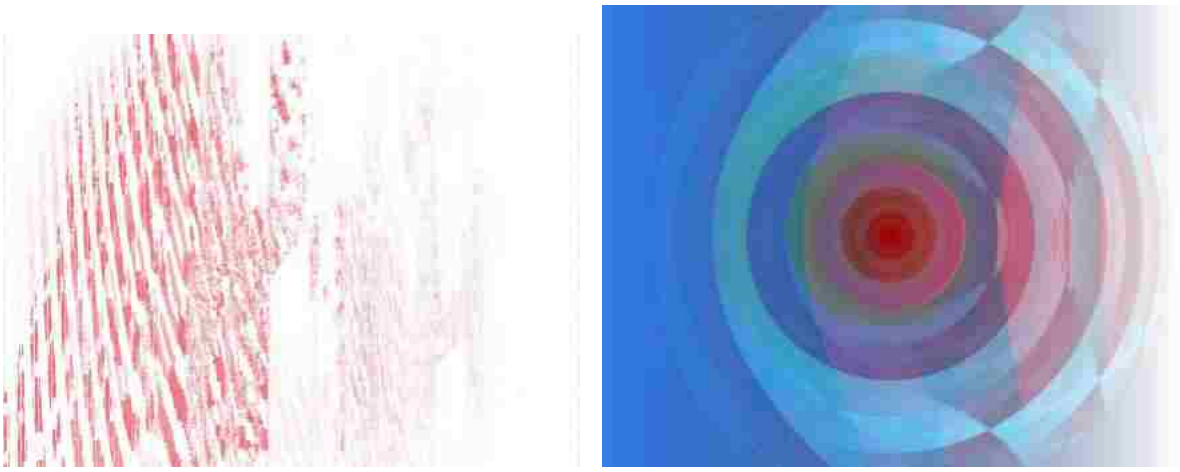
Figure 9.1: Photographs of the *Fitness Function* art exhibit at the BYU Harris Fine Arts Center.

The success of *Fitness Function* inspired us to repeat the exhibit for an evening at the*ACM Conference on Creativity and Cognition* in Atlanta, GA. The second *Fitness Function* exhibit was held in the High Museum of Art and covered by NPR's Studio 360[2]. The exhibit again caused visitors to question their notions about art curation and about what makes art meaningful.

9.2 Utah County Art Gallery

In 2011, we participated in the Utah County Art Gallery: Fall Photography and Digital Art Show by submitting two digital images created by DARCI. These images were selected from a pool of images DARCI rendered using a blank source image for the adjectives 'happy', 'peaceful' and 'scary'. Because we chose the final two images from several created by DARCI, these submissions were more of a collaboration between us and DARCI. The first image we submitted we created using the adjective 'peaceful' and a blank black source image (Figure 9.2(a)). The second image was also rendered using the adjective 'peaceful', but with a blank white source image (Figure 9.2(b)). The judges of the art show had no knowledge that the art was created by a computer program, and the image in Figure 9.2(a) won second place in the Digital Art category.

[2] http://www.wnyc.org/story/designing-computer-great-taste/



(a) "Peaceful on Black 4-3" (b) "Cold Avian"

Figure 9.2: Two images created by DARCI that were submitted to the Utah County Art Gallery Fall Photography and Digital Art Show in 2011 (with provided titles). "Peaceful on Black 4-3" won second place in the Digital Art category.

9.3 Evolutionary Art, Design, and Creativity Competition

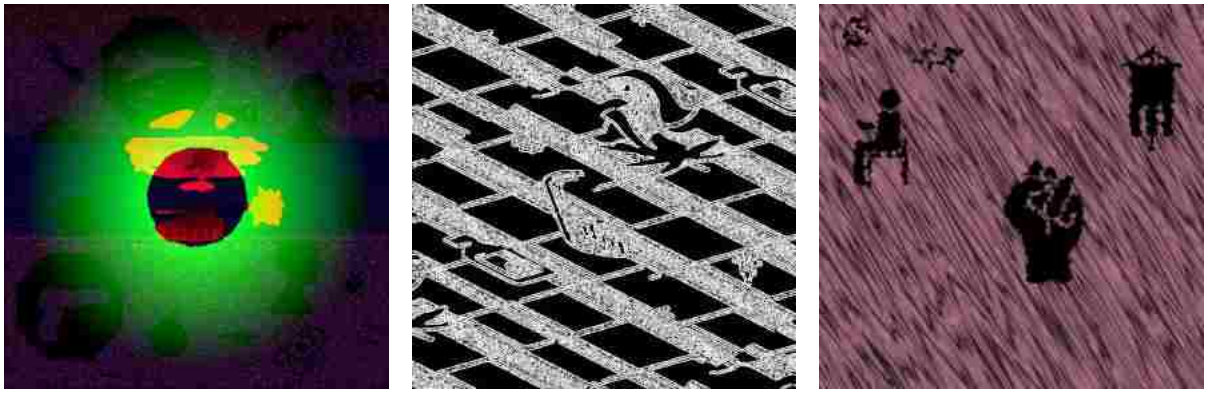
In 2013, we submitted several images to the *GECCO Evolutionary Art, Design, and Creativity Competition* held in Amsterdam. The version of DARCI described in Chapter 4 is capable of rendering original images for any arbitrary concept by composing a collage of iconic images and then rendering the collage according to semantically related adjectives. We had DARCI render several different images for several different high level concepts, and we then selected the best images to include in the competition. The submitted images, along with their source concept and a short description we provided for the competition to explain icon and adjective choices, are shown in Figure 9.3. These images were exhibited as finalists for the competition.

9.4 You Can't Know My Mind

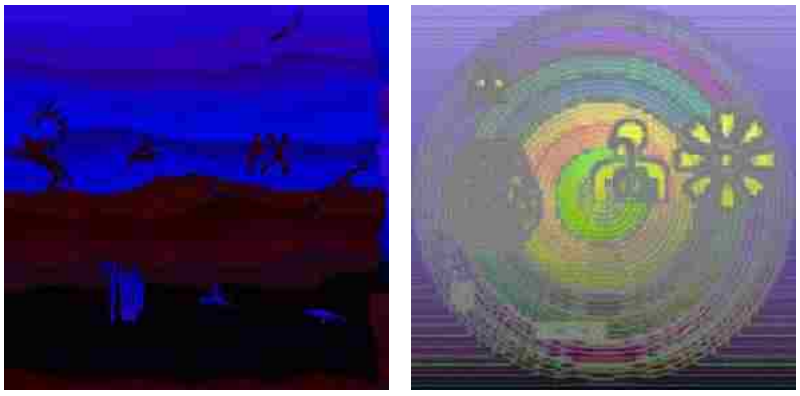
Also in 2013, we joined with Simon Colton at the Galerie Oberkampf in Paris, France for a festival celebrating computational creativity [27]. The festival was called *You Can't Know My Mind*, and

demonstrated several computationally creative systems. Colton's The Painting Fool [25] used DARCI's adjective neural networks, which enabled it to justify its own artwork and reflect upon it.

Participants in the festival could have their photographs taken for reference. At which point, The Painting Fool/DARCI hybrid system would figure out its "mood" using sentiment analysis of current news stories from *The Guardian*. The "mood" was then used to select an adjective that the system would reference in producing an artistic portrait of the participant. DARCI would select a non-photorealistic image filter and abstract background that best conveyed the chosen adjective, and the filter would then be applied to the background. Next, The Painting Fool would artistically paint the photograph on the filtered background in a style dictated by its "mood". DARCI would then analyze the resulting image to see how well the final image conveyed the selected adjective. Finally, using basic language templates, the system would return a description of the portrait including what it meant to create, why it painted what it did, and how it felt about the final product.



(a) War (b) Epic Drug Scandal (c) Guilty Protest



(d) Murder (e) Artificial Intelligence

Figure 9.3: Images submitted to the Evolutionary Art, Design, and Creativity Competition held in Amsterdam as part of GECCO (Genetic and Evolutionary Computation Conference). The concepts used to create the pieces are listed under each image. (a) *War* incorporates concepts such as tank, gun, bomber and atom and renders them with a style that suggests explosive and bloody. (b) *Epic Drug Scandal* weaves a dizzy conception of icons such as pill, marijuana, medicine and syringe. (c) *Guilty Protest* combines concepts such as student, banner, crime and jail in a rendering designed to evoke a feeling of sadness. (d) *Murder* offers a dark, oppressive evocation of the grim reaper, electrocution and weapons. (e) *Artificial Intelligence* offers a quirky mix of conceptual proxies for intelligence, such as brain and school, with elements associated with artificial, like flower and lung, rendered to evoke the idea of light.

Chapter 10

Conclusion

Throughout this dissertation we have presented enhancements to DARCI, which is a system designed to autonomously create visual art that conveys meaning to human viewers. Through designing, implementing, and evaluating DARCI, we have been able to explore computational creativity and make several contributions to the field. Broadly, these contributions consist of the development of a one-of-a-kind autonomously creative visual art system (DARCI); general models, frameworks, theories, and philosophical ideologies related to computational creativity and other domains; and influence on the art community through participating in several art galleries, art exhibits, and collaborations. Specifically, the contributions are listed as follows:

- We have continued the development of DARCI, which is currently the only system that can, in full autonomy, create visual art that explicitly conveys meaning. It is the only system of its kind with a sophisticated cognitive model that initiates and drives its creative process by giving it advanced perceptual abilities and broad semantic understanding.
- We have developed novel methods for automatically evaluating the semantic qualities of images using clustering techniques (Chapters 2 and 6). We also have developed human surveys for evaluating DARCI's ability to generate meaningful images that can be considered creative (Chapters 4 and 8).
- Through these surveys we have been able to find evidence suggesting that images that successfully communicate an intended meaning are perceived as more likable and creative than images that don't seem to convey any particular meaning (Chapters 4 and 8).

- We have developed a novel semantic model based on a combination of human free association norms and automatically derived corpus-based word associations. We have shown that not only can this model be used by DARCI to create visual art that conveys high-level concepts, but also it can be used to automatically solve several standardized word analogy/similarity tasks and play word guessing games online (Chapter 3).
- We designed a domain independent computational framework for facilitating imagination, called the Associative Conceptual Imagination (ACI) framework, which is based on cognitive theories of human imagination (Chapter 5).
- The ACI framework was successfully applied to DARCI, which allowed the system to "imagine" how to render images for concepts on which it was never explicitly trained (Chapter 6). This was also a contribution to the field of computer vision, as it enabled DARCI to do zero-shot shot image ranking by taking advantage of semantic structure learned through written text.
- We developed a general theory regarding the role of perception as a necessary component of the creative process (Chapter 7). We explored the fields of cognitive psychology, neuroscience, art history, and artificial general intelligence to argue the need for advanced perceptual abilities in creative systems.
- We then incorporated state-of-the-art deep learning models into the DARCI system to provide it with more advanced perceptual abilities (Chapter 8). This allowed DARCI to better evaluate and understand its own artifacts in order to produce semantically relevant and aesthetically pleasing images and to find inspiration in preexisting images.
- DARCI has impacted the art community by being involved in several art galleries, art exhibits, and collaborations (Chapter 9). This resulted is meaningful discussions about what makes good art, how art is judged, how the creative process works, and the role of computers in the visual arts.

Although DARCI has seen much success, it is not without its criticisms. Many people who have evaluated DARCI's art have sometimes called it simple, childish, not great, or just plain ugly. Many people also point out other computational systems (such as The Painting Fool) that can consistently produce artwork that is usually more aesthetically pleasing than DARCI's art. We certainly acknowledge these criticisms, and having DARCI produce better looking images is one of our research goals. However, it is important to keep in mind that our goal is for DARCI to be a fully autonomous system, and it has been built from the ground up to be so. It is making its own creative decisions based on what it has learned about art and semantics through various machine learning techniques.

Other image producing systems are not at the same level of autonomy as DARCI. They either require human judgment/intervention (i.e., used as a tool for human artists), or they have been hand engineered for very specific types of art (e.g., fractal art). Even other visual art systems in the field of computational creativity, such as The Painting Fool, rarely incorporate any self-evaluation and have not reached the same level of autonomy as DARCI. Some systems do have self-evaluation, but it is typically based on some static hand engineered metric and not based on semantic real world understanding.

While other systems may focus on the quality of the art itself, we focus on the quality of the *process* in creating the art. In doing so, we are subject to what is called the *latent heat effect*—as the creative responsibility given to systems increases, the value of its output does not initially increase (and may even decrease) [28]. DARCI is taking on more creative responsibility than other generative art systems, and so we should not be surprised that the initial quality of the art is negatively affected. As we have considered ways to truly model and understand the creative process, it has always come down to the system's ability to perceive and understand its own artifacts in the context of real-world, human-relatable concepts. Even with the perceptually grounded semantic models we have built into DARCI, it is still limited in scope and ability compared to humans. Thus, we cannot expect DARCI to consistently and autonomously produce human-level artwork until its perceptual abilities and semantic understanding have also reached human levels.

10.1 Future Work

The research goals for DARCI are ambitious and go beyond the scope of one dissertation, thus there is much work yet to be done on DARCI. Future work certainly includes improving the quality of DARCI's artwork, which can be achieved in a number of ways. One way is by improving and expanding the semantic and perceptual models incorporated into DARCI. Many of the current limitations of these models are limitations in other fields such as computer vision and computational semantics, and advancements in these other fields can be included in DARCI. Current state-of-the-art computer vision models, such as deep neural networks, have yet to be fully utilized by DARCI (or by any computationally creative system). These models are capable of not only understanding images, but also directly generating novel and semantically relevant images.

Another way to improve DARCI's artwork would be to expand DARCI's capabilities in the actual creation and modification of images. Currently, it is limited to arranging collages of icons and Photoshop-like image filters. We want to provide DARCI with a more varied set of artistic skills, which could include the ability to make individual brush strokes, to do procedural/algorithmic art, or to incorporate 3D models and visual effects.

We would like DARCI to continue collaborating with other creative systems and to participate in art galleries and exhibits. For example, DARCI could work with a system that generates stories by creating illustrations that help tell the story. DARCI could work with systems that create music to produce relevant visualizations or use music for inspiration. There are creative systems designed to automatically create video games, and DARCI could be used to create backgrounds or textures. We also want to hold another art exhibit that showcases the work of human artists alongside DARCI's work, where they each create artwork in response to the other's art.

Ultimately, we are interested in understanding and formalizing the creative process, independent of a specific domain. The knowledge we have gained through working with DARCI can potentially be applied to other domains, such as music, recipes, and video games. Building a sophisticated cognitive model with semantic understanding grounded in perception is a general design that can be applied to multiple domains. In fact, a system that is capable of working in

multiple domains has greater flexibility to innovate and find inspiration across different modes of experience, which would increase its creative potential.

References

[1] Cara B Allen, Tansu Celikel, and Daniel E Feldman. Long-term depression induced by sensory deprivation during cortical map plasticity in vivo. *Nature Neuroscience*, 6(3):291–299, 2003.

[2] Amir Amedi, Lotfi B Merabet, Joan Camprodon, Felix Bermpohl, Sharon Fox, Itamar Ronen, Dae-Shik Kim, and Alvaro Pascual-Leone. Neural and behavioral correlates of drawing in an early blind painter: a case study. *Brain Research*, 1242:252–262, 2008.

[3] Paul Bach-y Rita and Stephen W Kercel. Sensory substitution and the human–machine interface. *Trends in Cognitive Sciences*, 7(12):541–546, 2003.

[4] Shumeet Baluja, Dean Pomerleau, and Todd Jochem. Towards automated artificial evolution for computer-generated images. *Connection Science*, 6:325–354, 1994.

[5] Marco Baroni and Alessandro Lenci. Distributional memory: A general framework for corpus-based semantics. *Computational Linguistics*, 36(4):673–721, December 2010.

[6] Marco Baroni, Brian Murphy, Eduard Barbu, and Massimo Poesio. Strudel: A corpus-based semantic model based on properties and types. *Cognitive Science*, 34:222–254, 2010.

[7] Lawrence W Barsalou. Perceptual symbol systems. *Behavioral and Brain Sciences*, 22(04): 637–660, 1999.

[8] Herbert Bay, Andreas Ess, Tinne Tuytelaars, and Luc Van Gool. Speeded-up robust features (SURF). *Computer Vision and Image Understanding*, 110:346–359, 2008.

[9] Atılım Güneş Baydin, Ramon López de Mántaras, and Santiago Ontañón. A semantic network-based evolutionary algorithm for computational creativity. *Evolutionary Intelligence*, 8(1):3–21, 2014.

[10] Michael Beaney. *Imagination and Creativity*. Open University Milton Keynes, UK, 2005.

[11] Gregory Berns. *Iconoclast: A neuroscientist reveals how to think differently*. Harvard Business Press, 2008.

[12] Shashank Bhatia and Stephan K Chalup. A model of heteroassociative memory: Deciphering surprising features and locations. In *Proceedings of the 4th International Conference on Computational Creativity*, pages 139–146, 2013.

[13] Kim Binsted, Helen Pain, and Graeme Ritchie. Children's evaluation of computer-generated punning riddles. *Pragmatics & Cognition*, 5(2):305–354, 1997.

[14] Margaret A. Boden. *Handbook of Creativity*, chapter 18. Press Syndicate of the University of Cambridge, 1999.

[15] Vincent Breault, Sébastien Ouellet, Sterling Somers, and Jim Davies. SOILIE: A computational model of 2d visual imagination. In *Proceedings of the 12th International Conference on Cognitive Modeling*, pages 95–100, 2013.

[16] Blain Brown. *Cinematography: Theory and Practice*, chapter 5, pages 67–75. Focal Press, 2011.

[17] Elia Bruni, Jasper Uijlings, Marco Baroni, and Nicu Sebe. Distributional semantics with eyes: Using image analysis to improve computational representations of word meaning. In *Proceedings of the 20th ACM International Conference on Multimedia*, pages 1219–1228, New York, NY, USA, 2012. ACM.

[18] Curt Burgess. From simple associations to the building blocks of language: Modeling meaning in memory with the HAL model. *Behavior Research Methods, Instruments, & Computers*, 30:188–198, 1998. ISSN 0743-3808.

[19] Eugene Charniak and Yorick Wilks. *Computational semantics: an introduction to artificial intelligence and natural language comprehension*. Fundamental studies in computer science. North-Holland, 1976.

[20] Anjan Chatterjee. The neuropsychology of visual artistic production. *Neuropsychologia*, 42 (11):1568–1583, 2004.

[21] Ryszard S. Choras. Image feature extraction techniques and their applications for cbir and biometrics systems. *International Journal of Biology and Biomedical Engineering*, 1:6–16, 2007.

[22] Allan Collins and M. Quillian. Retrieval time from semantic memory. *Journal of Verbal Learning and Verbal Behavior*, 8(2):240–247, April 1969.

[23] Allan M. Collins and Elizabeth F. Loftus. A spreading-activation theory of semantic processing. *Psychological Review*, 82(6):407–428, 1975.

[24] Simon Colton. Creativity versus the perception of creativity in computational systems. *Creative Intelligent Systems: Papers from the AAAI Spring Symposium*, pages 14–20, 2008.

[25] Simon Colton. The painting fool: Stories from building an automated painter. In J. McCormack and M. d'Inverno, editors, *Computers and Creativity*. Springer-Verlag, 2011.

[26] Simon Colton. *Automated theory formation in pure mathematics*. Springer Science & Business Media, 2012.

[27] Simon Colton and Dan Ventura. You can't know my mind: A festival of computational creativity. In *Proceedings of the Fifth International Conference on Computational Creativity*, pages 351–354, 2014.

[28] Simon Colton and Geraint Wiggins. Computational creativity: The final frontier? In *European Conference on Artificial Intelligence*, volume 12, pages 21–26, 2012.

[29] Simon Colton, Jeremy Gow, Pedro Torres, and Paul Cairns. Experiments in objet trouvé browsing. *Proceedings of the 1st International Conference on Computational Creativity*, pages 238–247, 2010.

[30] Simon Colton, Jacob Goodwin, and Tony Veale. Full face poetry generation. In *Proceedings of the Third International Conference on Computational Creativity*, pages 95–102, 2012.

[31] Simon Colton, Jakob Halskov, Dan Ventura, Ian Gouldstone, Michael Cook, P Blanca, et al. The Painting Fool sees! new projects with the automated painter. In *Proceedings of the 6th International Conference on Computational Creativity*, pages 189–196, 2015.

[32] Michael Cook and Simon Colton. Ludus ex machina: Building a 3d game designer that competes alongside humans. In *Proceedings of the 5th International Conference on Computational Creativity*, volume 380, 2014.

[33] David Cope. *Experiments in musical intelligence*, volume 12. AR editions Madison, WI, 1996.

[34] Bob Coyne and Richard Sproat. WordsEye: An automatic text-to-scene conversion system. In *Proceedings of the 28th Annual Conference on Computer Graphics and Interactive Techniques*, pages 487–496, New York, NY, USA, 2001. ACM. ISBN 1-58113-374-X.

[35] Lee Cronbach. Coefficient alpha and the internal structure of tests. *Psychometrika*, 16(3): 297–334, September 1951.

[36] Mihaly Csíkzentmihályi and Rick E. Robinson. *The Art of Seeing*. The J. Paul Getty Trust Office of Publications, 1990.

[37] Gabriella Csurka, Christopher R. Dance, Lixin Fan, Jutta Willamowski, and Cdric Bray. Visual categorization with bags of keypoints. In *Proceedings of the Workshop on Statistical Learning in Computer Vision*, pages 1–22, 2004.

[38] Gregory Currie and Ian Ravenscroft. *Recreative Minds: Imagination in Philosophy and Psychology*. Oxford University Press, 2002.

[39] Ritendra Datta, Dhiraj Joshi, Jia Li, and James Z. Wang. Studying aesthetics in photographic images using a computational approach. *Lecture Notes in Computer Science*, 3953:288–301, 2006.

[40] Simon De Deyne and Gert Storms. Word associations: Norms for 1,424 Dutch words in a continuous task. *Behavior Research Methods*, 40(1):198–205, feb 2008.

[41] Tom De Smedt. *Modeling Creativity: Case Studies in Python*. University Press Antwerp, 2013.

[42] Scott Deerwester, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, and Richard Harshman. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6):391–407, 1990.

[43] Guy Denhière and Benoît Lemaire. A computational model of children's semantic memory. In *Proceedings of the 26th Conference of the Cognitive Science Society*, pages 297–302, Mahwah, NJ, 2004. Lawrence Erlbaum Associates.

[44] Ludovic Denoyer and Patrick Gallinari. The Wikipedia XML corpus. In *INEX Workshop Pre-Proceedings*, pages 367–372, 2006.

[45] Emily L Denton, Soumith Chintala, Rob Fergus, et al. Deep generative image models using a Laplacian pyramid of adversarial networks. In *Advances in Neural Information Processing Systems*, pages 1486–1494, 2015.

[46] Steve DiPaola and Liane Gabora. Incorporating characteristics of human creativity into an evolutionary art algorithm. *Genetic Programming and Evolvable Machines*, 10(2):97–110, 2009.

[47] Kemal Ebcioğlu. An expert system for harmonizing four-part chorales. *Computer Music Journal*, 12(3):43–51, 1988.

[48] Betty Edwards. *Drawing on the Right Side of the Brain*. New York: Tarcher, 1989.

[49] Charles Elkan. Using the triangle inequality to accelerate *k*-means. In *Proceedings of the Twentieth International Conference on Machine Learning*, pages 147–153, 2003.

[50] Hadyn D Ellis and Michael B Lewis. Capgras delusion: a window on face recognition. *Trends in Cognitive Sciences*, 5(4):149–156, 2001.

[51] Katrin Erk. What is word meaning, really? (and how can distributional models help us describe it?). In *Proceedings of the 2010 Workshop on Geometrical Models of Natural Language Semantics*, pages 17–26, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics.

[52] Clement Farabet, Camille Couprie, Laurent Najman, and Yann LeCun. Learning hierarchical features for scene labeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8):1915–1929, 2013.

[53] Martha J Farah. *Visual Agnosia*. MIT press, 2004.

[54] Gilles Fauconnier and Mark Turner. Conceptual integration networks. *Cognitive Science*, 22 (2):133–187, 1998.

[55] Christiane Fellbaum, editor. *WordNet: An Electronic Lexical Database*. The MIT Press, 1998.

[56] Charles Forceville. *Pictorial Metaphor in Advertising*. New York: Routledge, 1996.

[57] Andrea Frome, Greg Corrado, Jon Shlens, Samy Bengio, Jeffrey Dean, Marc'Aurelio Ranzato, and Tomas Mikolov. DeViSE: A deep visual-semantic embedding model. In *Advances In Neural Information Processing Systems*, pages 2121–2129, 2013.

[58] Liane Gabora and Apara Ranjan. *How Insight Emerges in Distributed, Content-addressable Memory*. Oxford University Press, 2013.

[59] Berys Gaut. Creativity and imagination. *The Creation of Art*, pages 148–173, 2003.

[60] Robert Gens and Pedro Domingos. Learning the structure of sum-product networks. In *Proceedings of the 30th International Conference on Machine Learning* , volume 28, pages 873–880, 2013.

[61] John S. Gero. Creativity, emergence, and evolution in design. *Knowledge-Based Systems*, 9: 435–448, 1996.

[62] Pablo Gervás. Computational approaches to storytelling and creativity. *AI Magazine*, 30: 49–63, 2009.

[63] Theo Gevers and Arnold Smeulders. Combining color and shape invariant features for image retrieval. *IEEE Transactions on Image Processing*, 9:102–119, 2000.

[64] Andrew B. Goldberg, Xiaojin Zhu, Charles R. Dyer, Mohamed Eldawy, and Lijie Heng. Easy as ABC?: Facilitating pictorial communication via semantically enhanced layout. In *Proceedings of the Twelfth Conference on Computational Natural Language Learning*, pages 119–126, Stroudsburg, PA, USA, 2008. Association for Computational Linguistics.

[65] Stuart Grassian and Nancy Friedman. Effects of sensory deprivation in psychiatric seclusion and solitary confinement. *International Journal of Law and Psychiatry*, 8(1):49–65, 1986.

[66] Gary Greenfield. Color dependent computational aesthetics for evolving expressions. In Reza Sarhangi, editor, *Bridges: Mathematical Connections in Art, Music, and Science; Conference Proceedings*, pages 9–16. Winfield, KS: Central Plains Book Manufacturing, 2002.

[67] Gary Greenfield. Co-evolutionary methods in evolutionary art. In Juan Romero and Penousal Machado, editors, *The Art of Artificial Evolution: A Handbook on Evolutionary Art and Music*, pages 357–380. Berlin: Springer, 2007.

[68] Karol Gregor, Ivo Danihelka, Alex Graves, and Daan Wierstra. DRAW: A recurrent neural network for image generation. In *Proceedings of the 32nd International Conference on Machine Learning*, pages 1462–1471, 2015.

[69] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. The WEKA data mining software: an update. *ACM SIGKDD Explorations Newsletter*, 11:10–18, November 2009.

[70] Richard Wesley Hamming. The unreasonable effectiveness of mathematics. *American Mathematical Monthly*, 87:81–90, 1980.

[71] D Fox Harrell. Shades of computational evocation and meaning: The GRIOT system and improvisational poetry generation. In *Proceedings of the Sixth Digital Arts and Culture Conference*, pages 133–143, 2005.

[72] Jeff Hawkins and Sandra Blakeslee. *On intelligence*. Macmillan, 2007.

[73] Derrall Heath and Dan Ventura. Creating images by learning image semantics using vector space models. In *Proceedings of The Thirtieth AAAI Conference on Artificial Intelligence*, 2016.

[74] Derrall Heath and Dan Ventura. Before a computer can draw, it must first learn to see. In *Proceedings of the 7th International Conference on Computational Creativity*, page to appear, 2016.

[75] Derrall Heath, David Norton, and Dan Ventura. Autonomously communicating conceptual knowledge through visual art. In *Proceedings of the 4th International Conference on Computational Creativity*, pages 91–104, 2013.

[76] Derrall Heath, David Norton, and Dan Ventura. Conveying semantics through visual metaphor. *ACM Transactions on Intelligent Systems and Technology*, 5(2):31:1–31:17, April 2014.

[77] Derrall Heath, Aaron Dennis, and Dan Ventura. Imagining imagination: A computational framework using associative memory models and vector space models. In *Proceedings of the 6th International Conference on Computational Creativity*, pages 244–251, 2015.

[78] Geoffrey Hinton, Simon Osindero, and Yee-Whye Teh. A fast learning algorithm for deep belief nets. *Neural Computation*, 18(7):1527–1554, 2006.

[79] Donald D. Hoffman. *Visual Intelligence: How We Create What We See*. W.W. Norton, 2000.

[80] Kenny Hong, Stephan K Chalup, Robert King, Michael J Ostwald, et al. Scene perception using pareidolia of faces and expressions of emotion. In *IEEE Symposium on Computational Intelligence for Creativity and Affective Computing*, pages 79–86, 2013.

[81] Eric H. Huang, Richard Socher, Christopher D. Manning, and Andrew Y. Ng. Improving word representations via global context and multiple word prototypes. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2012.

[82] Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell. Caffe: Convolutional architecture for fast feature embedding. In *Proceedings of the ACM International Conference on Multimedia*, pages 675–678, 2014.

[83] Daniel S Johnson and Dan Ventura. Musical motif discovery in non-musical media. In *Proceedings of the 5th International Conference on Computational Creativity*, pages 91–99, 2014.

[84] Michael N. Jones and Douglas J. K. Mewhort. Representing word meaning and order information in a composite holographic lexicon. *Psychological Review*, 114:1–37, 2007.

[85] Dhiraj Joshi, James Z. Wang, and Jia Li. The story picturing engine—a system for automatic text illustration. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 2(1):68–89, February 2006. ISSN 1551-6857.

[86] Hesham M Kamel and James A Landay. A study of blind drawing practice: creating graphical information without the visual channel. In *Proceedings of the Fourth International ACM Conference on Assistive Technologies*, pages 34–41, 2000.

[87] Stuart Kaplan. Visual metaphors in print advertising for fashion products. In Ken Smith, Sandra Moriarty, Gretchen Barbatsis, and Keith Kenney, editors, *Handbook of Visual Communication: Theory, Methods, and Media*, chapter 11, pages 167–177. New Jersey: Lawrence Erlbaum Associates, Publishers, 2005.

[88] Sergey Karayev, Matthew Trentacoste, Helen Han, Aseem Agarwala, Trevor Darrell, Aaron Hertzmann, and Holger Winnemoeller. Recognizing image style. *arXiv preprint arXiv:1311.3715*, 2013.

[89] John Miller Kennedy. *Drawing & the Blind: Pictures to Touch*. Yale University Press, 1993.

[90] Irwin King. Distributed content-based visual information retrieval system on peer-to-pear(p2p) network. http://appsrv.cse.cuhk.edu.hk/~miplab/discovir/.

[91] G. R. Kiss, C. Armstrong, R. Milroy, and J. Piper. An associative thesaurus of English and its computer analysis. In A. J. Aitkin, R. W. Bailey, and N. Hamilton-Smith, editors, *The Computer and Literary Studies*. University Press, Edinburgh, UK, 1973.

[92] Bart Kosko. Bidirectional associative memories. *IEEE Transactions on Systems, Man and Cybernetics*, 18(1):49–60, 1988.

[93] Anna Krzeczkowska, Jad El-Hage, Simon Colton, and Stephen Clark. Automated collage generation — with intent. In *Proceedings of the 1st International Conference on Computational Creativity*, pages 36–40, 2010.

[94] Rosa Lafer-Sousa, Katherine L Hermann, and Bevil R Conway. Striking individual differences in color perception uncovered by 'the dress' photograph. *Current Biology*, 25(13): R545–R546, 2015.

[95] Brenden M Lake, Ruslan R Salakhutdinov, and Josh Tenenbaum. One-shot learning by inverting a compositional causal process. In *Advances in Neural Information Processing Systems*, pages 2526–2534, 2013.

[96] T.K. Landauer and S.T. Dumais. A solution to Plato's problem: The latent semantic analysis theory of acquisition induction and representation of knowledge. *Psychological Review*, 104 (2):211–240, 1997.

[97] Alexander S. Ecker Leon A. Gatys and Matthias Bethge. A neural algorithm of artistic style. *Computing Research Repository*, 2015.

[98] Gil Levi and Tal Hassner. Age and gender classification using convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 34–42, 2015.

[99] Congcong Li and Tsuhan Chen. Aesthetic visual quality assessment of paintings. *IEEE Journal of Selected Topics in Signal Processing*, 3:236–252, 2009.

[100] R. Likert. A technique for the measurement of attitudes. *Archives of Psychology*, 22(140): 1–55, 1932.

[101] H Liu and P Singh. ConceptNet—a practical commonsense reasoning tool-kit. *BT Technology Journal*, 22:211–226, 2004.

[102] Hugo Liu and Push Singh. Commonsense reasoning in and over natural language. In Mircea Negoita, Robert Howlett, and Lakhmi Jain, editors, *Knowledge-Based Intelligent Information and Engineering Systems*, volume 3215 of *Lecture Notes in Computer Science*, pages 293–306. Springer Berlin / Heidelberg, 2004.

[103] Ying Liu, Dengsheng Zhang, Guojun Lu, and Wei-Ying Ma. A survey of content-based image retrieval with high-level semantics. *Pattern Recognition*, 40(1):262–282, January 2007. ISSN 0031-3203.

[104] Kevin Lund and Curt Burgess. Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods, Instruments, & Computers*, 28:203–208, 1996. ISSN 0743-3808.

[105] Penousal Machado and Amilcar Cardoso. All the truth about NEvAr. *Applied Intelligence, Special Issue on Creative Systems*, 16:101–119, 2002.

[106] Penousal Machado, Juan Romero, and Bill Manaris. Experiments in computational aesthetics: An iterative approach to stylistic change in evolutionary art. In Juan Romero and Penousal Machado, editors, *The Art of Artificial Evolution: A Handbook on Evolutionary Art and Music*, pages 381–415. Berlin: Springer, 2007.

[107] Michael F Marmor and James Ravin. *The Artist's Eyes*. Harry N Abrams Incorporated, 2009.

[108] Maricarmen Martinez, Tarek Besold, Ahmed Abdel-Fattah, Kai-Uwe Kuehnberger, Helmar Gust, Martin Schmidt, and Ulf Krumnack. Towards a domain-independent computational framework for theory blending. In *2011 AAAI Fall Symposium Series*, 2011.

[109] Pamela McCorduck. *AARON's Code: Meta-Art, Artificial Intelligence, and the Work of Harold Cohen*. W. H. Freeman & Co., 1991.

[110] Stephen McGregor, Geraint Wiggins, and Matthew Purver. Computational creativity: A philosophical approach, and an approach to philosophy. In *Proceedings of the 5th International Conference on Computational Creativity*, pages 254–262, 2014.

[111] Andrés F Melo and Geraint Wiggins. A connectionist approach to driving chord progressions using tension. In *Proceedings of the AISB Symposium on Creativity in Arts and Science*, 2003.

[112] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. In *Proceedings of the International Conference on Learning Representations*, 2013.

[113] Eduardo Reck Miranda and Al Biles. *Evolutionary Computer Music*. Springer, 2007.

[114] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–533, 2015.

[115] Kristine Monteith, Tony Martinez, and Dan Ventura. Automatic generation of melodic accompaniments for lyrics. In *Proceedings of the 3rd International Conference on Computational Creativity*, pages 87–94, 2012.

[116] Richard G Morris, Scott H Burton, Paul M Bodily, and Dan Ventura. Soup over bean of pure joy: Culinary ruminations of an artificial chef. In *Proceedings of the 3rd International Conference on Computational Creativity*, pages 119–125, 2012.

[117] Naila Murray, Luca Marchesotti, and Florent Perronnin. AVA: A large-scale database for aesthetic visual analysis. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2408–2415. IEEE, 2012.

[118] Douglas L. Nelson, Cathy L. McEvoy, and T. A. Schreiber. The University of South Florida word association, rhyme, and word fragment norms. http://www.usf.edu/FreeAssociation/, 1998.

[119] Yael Netzer, David Gabay, Yoav Goldberg, and Michael Elhadad. Gaiku: Generating haiku with word associations norms. In *Proceedings of the Workshop on Computational Approaches to Linguistic Creativity*, pages 32–39, 2009.

[120] Anh Mai Nguyen, Jason Yosinski, and Jeff Clune. Innovation engines: Automated creativity and improved stochastic optimization via deep learning. In *Proceedings of the Genetic and Evolutionary Computation Conference*, pages 959–966, 2015.

[121] David Norton, Derrall Heath, and Dan Ventura. Establishing appreciation in a creative system. In *Proceedings of the 1st International Conference on Computational Creativity*, pages 26–35, 2010.

[122] David Norton, Derrall Heath, and Dan Ventura. An artistic dialogue with the artificial. In *Proceedings of the 8th ACM Conference on Creativity and Cognition*, pages 31–40, New York, NY, USA, 2011. ACM.

[123] David Norton, Derrall Heath, and Dan Ventura. Autonomously creating quality images. In *Proceedings of the 2nd International Conference on Computational Creativity*, pages 10–15, 2011.

[124] David Norton, Derrall Heath, and Dan Ventura. Finding creativity in an artificial artist. *Journal of Creative Behavior*, 47(2):106–124, 2013.

[125] David Norton, Derrall Heath, and Dan Ventura. Autonomously managing competing objectives to improve the creation and curation of artifacts. In *Proceedings of the 5th International Conference on Computational Creativity*, 2014.

[126] David Norton, Derrall Heath, and Dan Ventura. Accounting for bias in the evaluation of creative computational systems: An assessment of DARCI. In *Proceedings of the 6th International Conference on Computational Creativity*, pages 31–38, 2015.

[127] David Norton, Derrall Heath, and Dan Ventura. Annotating images with emotional adjectives using features that summarize local interest points. *IEEE Transactions on Affective Computing, Under Review*, 2016.

[128] David Oranchak. Evolutionary synthesis of photographic artwork using human fitness function derived from web-based social networks. In *Proceedings of the 9th Annual Conference on Genetic and Evolutionary Computation*, page 2264, 2007.

[129] François Pachet and Pierre Roy. Markov constraints: Steerable generation of Markov sequences. *Constraints*, 16(2):148–172, 2011.

[130] Omkar M Parkhi, Andrea Vedaldi, and Andrew Zisserman. Deep face recognition. *British Machine Vision*, 1(3):6, 2015.

[131] Yves Peirsman and Dirk Geeraerts. Predicting strong associations on the basis of corpus data. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, pages 648–656, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics.

[132] Justin Permar and Brian Magerko. A conceptual blending approach to the generation of cognitive scripts for interactive narrative. In *Proceedings of the Ninth Artificial Intelligence and Interactive Digital Entertainment Conference*, pages 44–50, 2013.

[133] Eila M Peterson. Creativity in music listening. *Arts Education Policy Review*, 107(3):15–21, 2006.

[134] Christian Pich. *Applications of Multidimensional Scaling to Graph Drawing*. PhD thesis, University of Konstanz, 2009.

[135] Hoifung Poon and Pedro Domingos. Sum-product networks: A new deep architecture. In *Proceedings of the Twenty-Seventh Annual Conference on Uncertainty in Artificial Intelligence*, pages 337–346. AUAI Press, 2011.

[136] James Pustejovsky. *The Generative Lexicon*. Bradford Books. MIT Press, 1998. ISBN 9780262661409.

[137] Reinhard Rapp. Word sense discovery based on sense descriptor dissimilarity. In *Proceedings of the Ninth Machine Translation Summit*, pages 315–322, 2003.

[138] Graeme Ritchie. Some empirical criteria for attributing creativity to a computer program. *Minds and Machines*, 17:67–99, 2007.

[139] Steven Rooke. Eons of genetically evolved algorithmic images. In Peter J. Bentley and David W. Corne, editors, *Creative Evolutionary Systems*, chapter 13, pages 339–365. Morgan Kaufmann Publishers, 2002.

[140] Eleanor Rosch. Principles of categorization. In E. Rosch and B. Lloyd, editors, *Cognition and Categorization*, pages 27–48. Erlbaum, Hillsdale, New Jersey, 1978.

[141] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015.

[142] Oliver Sacks. *An Anthropologist on Mars*. New York: Knopf, 1995.

[143] G. Salton. *The SMART Retrieval System—Experiments in Automatic Document Processing*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 1971.

[144] John R Searle. Minds, brains, and programs. *Behavioral and Brain Sciences*, 3(03):417–424, 1980.

[145] Jimmy Secretan, Nicholas Beato, David B. D'Ambrosio, Adelein Rodriguez, Adam Campbell, Jeremiah T. Folsom-Kovarik, and Kenneth O. Stanley. Picbreeder: A case study in collaborative evolutionary exploration of design space. *Evolulationary Computation*, 19(3): 373–403, September 2011. ISSN 1063-6560.

[146] Phillip C.-Y. Sheu. *Semantic Computing*, pages 1–9. John Wiley & Sons, Inc., 2010. ISBN 9780470588222.

[147] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013.

[148] Karl Sims. Artificial evolution for computer graphics. *Computer Graphics*, 25(4):325–327, 1991.

[149] Josef Sivic, Bryan C. Russell, Alexei A. Efros, Andrew Zisserman, and William T. Freeman. Discovering objects and their location in images. *International Journal of Computer Vision*, 1:370–377, 2005.

[150] Alexa Steinbrück. Conceptual blending for the visual domain. Master's thesis, University of Amsterdam, 2013.

[151] Leslie F. Stevenson. Twelve conceptions of imagination. *British Journal of Aesthetics*, 43(3): 238–59, 2003.

[152] Ron Sun. *The Cambridge Handbook of Computational Psychology*. Cambridge University Press, New York, NY, USA, 1st edition, 2008. ISBN 0521674107, 9780521674102.

[153] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–9, 2015.

[154] Scott Thomas, Edward Boatman, Sofya Polyakov, Jeremy Mumenthaler, and Chris Wolff. The noun project. `http://thenounproject.com`, 2013.

[155] Peter M Todd. A connectionist system for exploring melody space. In *Proceedings of the International Computer Music Conference*, pages 65–68. International Computer Music Association, 1992.

[156] Steven J. P. Todd and William Latham. *Evolutionary art and computers*. Academic Press, 1992.

[157] Grigorios Tsoumakas and Ioannis Katakis. Multi-label classification: An overview.*International Journal of Data Warehousing and Mining*, 3(3):1–13, 2007.

[158] Peter D. Turney. Similarity of semantic relations. *Computational Linguistics*, 32(3):379–416, September 2006. ISSN 0891-2017.

[159] Peter D. Turney and Patrick Pantel. From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research*, 37:141–188, 2010.

[160] Lav R Varshney, Florian Pinel, Kush R Varshney, Angela Schörgendorfer, and Yi-Min Chee. Cognition as a part of computational creativity. In *Procceedings of the 12th IEEE International Conference on Cognitive Informatics and Cognitive Computing*, pages 36–43, 2013.

[161] Tony Veale. Creativity as pastiche: A computational treatment of metaphoric blends, with special reference to cinematic "borrowing". *Proceedings of Mind II: Computational Models of Creative Cognition*, 1997.

[162] Tony Veale. From conceptual mash-ups to bad-ass blends: A robust computational model of conceptual blending. In *Proceedings of the 3rd International Conference on Computational Creativity*, pages 1–8, 2012.

[163] Tony Veale. Less rhyme, more reason: Knowledge-based poetry generation with feeling, insight and wit. In *Proceedings of the 4th International Conference on Computational Creativity*, pages 152–159, 2013.

[164] Tony Veale and Yanfen Hao. Comprehending and generating apt metaphors: A web-driven, case-based approach to figurative language. *AAAI Proceedings of the 22nd national conference on Artificial Intelligence*, 2, 2007.

[165] Tony Veale and Guofu Li. Creative introspection and knowledge acquisition: Learning about the world through introspective questions and exploratory metaphors. *Proceedings of the 25th AAAI Conference on Artificial Intelligence*, 1991.

[166] L. von Ahn. Games with a purpose. *Computer*, 39(6):92–94, June 2006.

[167] Lev Vygotsky. Imagination and Creativity in Childhood. *Journal of Russian and East European Psychology*, 42(1):7–97, 2004.

[168] Tonio Wandmacher, Ekaterina Ovchinnikova, and Theodore Alexandrov. Does latent semantic analysis reflect human associations? In *Proceedings of the ESSLLI Workshop on Distributional Lexical Semantics*, pages 63–70, 2008.

[169] Feichao Wang. A survey on automatic image annotation and trends of the new age. *Procedia Engineering*, 23(0):434 – 438, 2011.

[170] Wei-Ning Wang and Qianhua He. A survey on emotional semantic image retrieval. *Proceedings of the International Conference on Image Processing*, pages 117–120, 2008.

[171] Wei-Ning Wang, Ying-Lin Yu, and Sheng-Ming Jiang. Image retrieval by emotional semantics: A study of emotional space and feature extraction. *IEEE International Conference on Systems, Man, and Cybernetics*, 4:3534–3539, 2006.

[172] Wei-Ning Wang, Ying-Lin Yu, and Sheng-Ming Jiang. Image retrieval by emotional semantics: A study of emotional space and feature extraction. In *Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics*, pages 3534–3539. IEEE, 2006.

[173] Lawrence Weiskrantz. Blindsight revisited. *Current Opinion In Neurobiology*, 6(2):215–220, 1996.

[174] Robert West, Joelle Pineau, and Doina Precup. Wikispeedia: an online game for inferring semantic distances between concepts. In *Proceedings of the 21st International Joint Conference on Artificial Intelligence*, pages 1598–1603, San Francisco, CA, USA, 2009.

[175] Geraint A Wiggins, George Papadopoulos, Somnuk Phon-Amnuaisuk, and Andrew Tuson. Evolutionary methods for musical composition. *International Journal of Computing Anticipatory Systems*, 1998.

[176] Ken Xu, James Stewart, and Eugene Fiume. Constraint-based automatic placement for scene composition. In *Procceedings of Graphics Interface*, pages 25–34, May 2002.

[177] Rafael Pérez y Pérez. Employing emotions to drive plot generation in a computer-based storyteller. *Cognitive Systems Research*, 8(2):89–109, 2007.

[178] Jichen Zhu and D Fox Harrell. Daydreaming with intention: Scalable blending-based imagining and agency in generative interactive narrative. In *AAAI Spring Symposium: Creative Intelligent Systems*, volume 156, 2008.

[179] Xiaojin Zhu, Andrew B. Goldberg, Mohamed Eldawy, Charles R. Dyer, and Bradley Strock. A text-to-picture synthesis system for augmenting communication. In *Proceedings of the 22nd National Conference on Artificial Intelligence - Volume 2*, pages 1590–1595. AAAI Press, 2007.

[180] J. Zujovic, L. Gandy, S. Friedman, B. Pardo, and T.N. Pappas. Classifying paintings by artistic genre: An analysis of features & classifiers. In *Proceedings of the IEEE International Workshop on Multimedia Signal Processing*, pages 1–5, Rio de Janeiro, Brazil, October 2009.

[181] Rolf A Zwaan and Michael P Kaschak. Language in the brain, body, and world. *The Cambridge Handbook of Situated Cognition*, pages 368–381, 2008.