Lehigh University
Lehigh Preserve

Theses and Dissertations

2017

Caption Aided Action Recognition Using Single Images

Adam Kafka
Lehigh University

Recommended Citation
Kafka, Adam, "Caption Aided Action Recognition Using Single Images" (2017). *Theses and Dissertations*. 2653.
http://preserve.lehigh.edu/etd/2653

Caption Aided Action Recognition Using Single Images

by

Adam Kafka

A Thesis

Presented to the Graduate and Research Committee

of Lehigh University

in Candidacy for the Degree of

Master of Science

In

Computer Science

Lehigh University

May 2017

This thesis is accepted and approved in partial fulfillment of the requirements for the Master of Science.

Date

Thesis Advisor: Mooi Choo Chuah

Chairperson of Department: Daniel Lopresti

Acknowledgements:

I would like to thank Professor Mooi Choo Chuah and PhD candidate Xin Li for working with me throughout this project. Your advice was indispensable.

Table of Contents:

List of Tables:

List of Figures:

Abstract:

In this paper, we attack the problem of classifying human actions from a single, static image. We propose that leveraging an automatic caption generator for this task will provide extra information when compared to a traditional convolutional neural network based classifier. The architecture consists of two stages, caption generation and caption classification, used sequentially to a proposed human action class label from a single image. Evaluation is performed of our system and it is evident that caption generation is the limiting factor in accuracy. We propose fixes to both the dataset and the caption generator, in order to improve the model. Finally, it is discovered that caption classification is significantly improved by concatenating all captions from a single image together, to produce one input vector.

1. Introduction:

Classification of objects within images is a well explored and understood topic [6]. State-of-the-art object classifiers are Convolutional Neural Networks (CNN), that learn high-level location independent features, which are used for classification in Fully Connected (FC) layers. However, classification of human actions within an image still remains a challenge [2]. The image features from CNNs must be processed further before being classified, in order to achieve strong action classification in an image.

In addition to object classification, automatic caption generation is an active area of academic research. Leading architectures combine a CNN with Long-Short Term Memory (LSTM) layers, a type of recurrent neural network layer, to generate descriptive sentences from an image. Accurate captions are quite information dense: they use natural language to relate objects to one another, giving context to what is happening in the image. For this reason, the architecture used to generate captions, may be applicable towards classifying human actions in an image. A caption generator could be used to build an action classifier, simply by appending a model that classifies captions into actions. We believe that this architecture would work better than generating an action classification directly from an image, using only a CNN like AlexNet.

Classifying actions from a single image will involve less computational power and information than classifying actions from a video. This is particularly useful on mobile devices, where computational power is limited and energy consumption is of concern. An example use case for this would be a baby monitor device, that checks what the baby is doing every couple of seconds, to make sure it is safe. Alternatively, it could be used with a personal assistant, that changes its behaviour based on what it detect people are doing within its field of view.

There are two data sets we identified that are commonly being used to evaluate action classification performance. Human3.6M consists of 3.6 million 3D frames, spanning 17 human actions [4]. The Volleyball Activity Dataset 2014, is comprised of 6 videos of volleyball games, with 7 classes of annotated activities for the players [10]. Both datasets supply videos annotated with action labels, however, they provide no ground truth captions that we could train a caption generator on. Instead of using these, we decide to leverage a dataset with ground truth captions, and extract human action labels from the captions.

A common dataset used to evaluate caption generation is the Microsoft COCO (MSCOCO) dataset [7]. MSCOCO provides over 100,000 images, each with at least five ground truth captions. We will use this dataset in two ways. First, we create a caption generator, trained on the entire MSCOCO training set. We will then use simple natural language processing (NLP) techniques on the ground truth captions to identify which images can be used for human action classification. We will then fine tune the caption generating model on this subset, and finally train a classifier using these computed action classes. This human action subset of MSCOCO will be referred to as HA.

The remainder of this paper is organized as follows. Section 2 will give an overview of previous work in this area. We present our architecture in Section 3 and describe the methodology. Section 4 discusses our evaluation of the action classifier. In Section 5, we discuss these results. And finally, Section 6 will discuss future work and conclude.

2. Related Work:

Automatic generation of captions from an image has been considered a very difficult task, until recently. Modern developments in both computer vision and natural language formation have created the scaffolding for impressively accurate caption generators. In their paper, Vinyals et al. [9] propose a new architecture for automatically describing images using natural language. The authors combine recent advances in computer vision and machine translation to produce a generative model, based on a deep recurrent architecture. Training is performed with the objective of maximizing the likelihood of the ground truth description, given the training image. The architecture is an end-to-end system that is fully trainable using stochastic gradient descent. Their model, named Show and Tell, is shown both quantitatively and qualitatively, to produce state-of-the-art descriptions of images. To attest to this, their model won first in the 2015 Microsoft COCO competition, both in automatic metrics, and human evaluation. The captions generated are both grammatically correct and accurate descriptions of the image. We choose to use this model to generate captions in our architecture.

In 2013, Gupta et al. [3] set about to investigate human action classification in depth images. Depth images, sometimes referred to as RGBD images, were collected from a Microsoft Kinect. They presented a human action classifier that leveraged cues from only the depth channel of images. In their model, they first segment out the human silhouettes from the depth images, then, compute pose descriptors, invariant to both scale and depth. They then cluster these descriptors together, into what are referred to as distinct ‘codewords.’ Finally, the codewords are used to classify the action present in the image. This work serves as an example of current approaches towards classifying human action from an image. In this case, they are using depth images, however.

3. Methodology:

The architecture for single image action classification is depicted in Figure 1, and can be split into two main subtasks: caption generation and caption classification. Additionally, design decisions including encoding the caption and creating image-level classifications must be made. The following subsections will describe the components outlined in the figure and architecture.

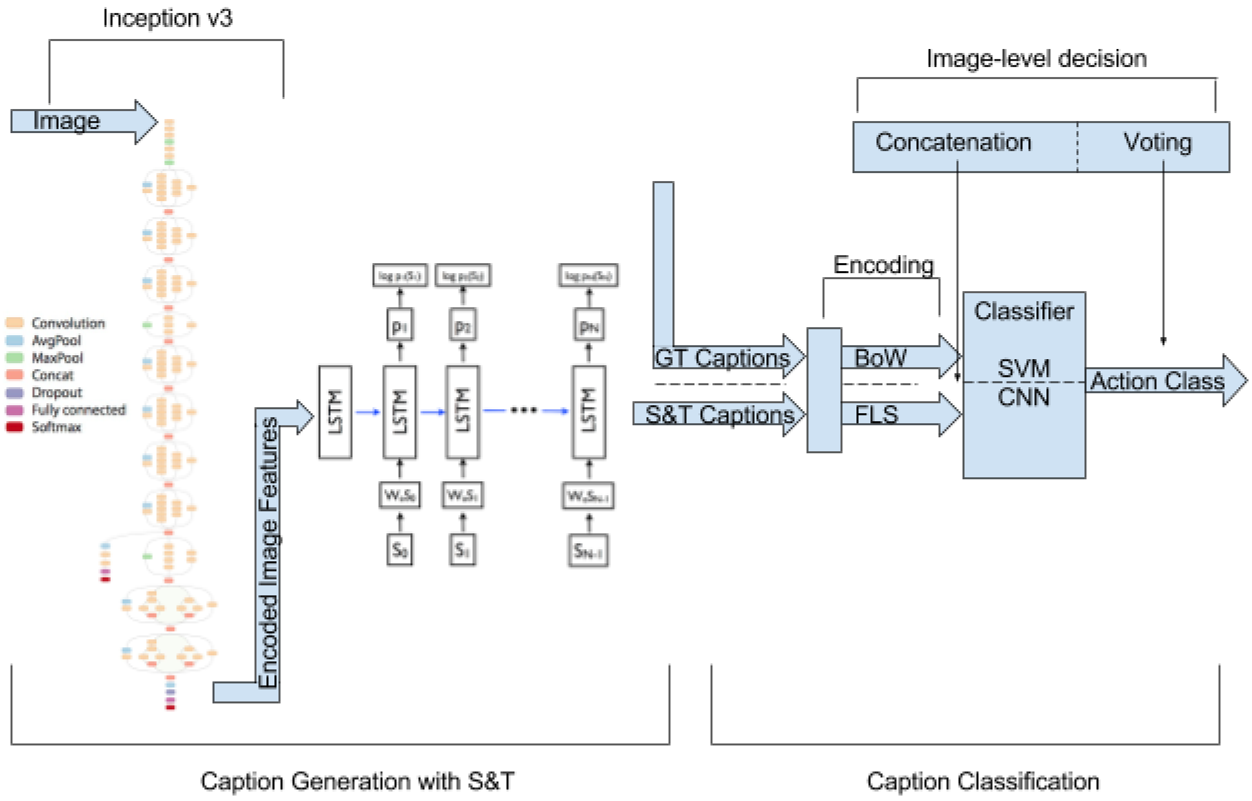


Figure 1: Architecture of the Action Classifier. Dashed lines denote different options.

3.1 Caption Source

Captions are provided to the classifier from two sources: ground truth captions and automatically generated captions. Ground truth captions (GT) will be utilized to isolate the performance of the caption classifier from that of the caption generator.

Automatic caption generation will be achieved using the *Show and Tell* model (S&T) from Vinyals et al.'s [9] work. S&T is an encoder-decoder network; the image is encoded into a feature vector, then decoded into natural language descriptions. The image encoding subsystem used is *Inception v3*, a deep convolutional neural network, pre-trained on the ILSVRC-2012-CLS image classification dataset [9]. Decoding is achieved through the use of LSTM layers. The LSTM network is trained as a language model, conditioned on the encoded feature vectors.

3.2 Caption Encoding

We explore two mechanisms to input the captions into the classifier. A bag of words (BoW) encoding passes the caption as a vector with length equal to the size of the vocabulary. The value of the at position *i* in the vector is the number of occurrences of the *i*th vocabulary word in the sentence. A downside to using BoW representations, is that we lose the order of the words in the sentence. An alternative method to pass the sentence to the classifier is to use a fixed-length sentence (FLS). In this case, all captions are a vector, whose size is the maximum caption length, and thus, maintain the order of words in a sentence. Fixed-length sentences are a mapping from words to the model's vocabulary list. An example of these two methods is demonstrated in Figure 2.

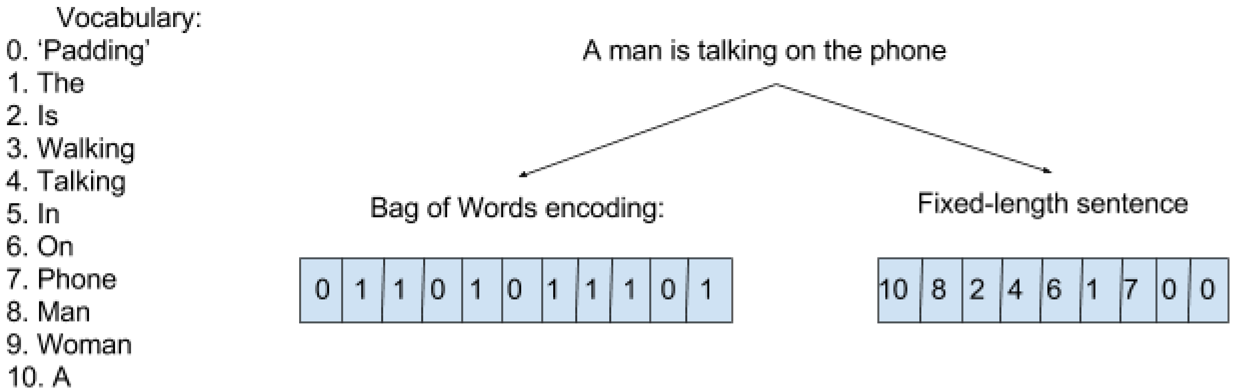


Figure 2: Two different methods of passing as input the sentence 'A man is talking on the phone'.

3.3 Caption Classification

Caption classification will be tested using two classifiers, a Support Vector Machine (SVM) and a Convolutional Neural Network (CNN). The particular SVM chosen was NuSVC, part of the python scikit-learn package [8]. This SVM was chosen because of its simple interface and ability to control the '*nu*' parameter. This parameter specifies the upper bound on the fraction of training errors and a lower bound of the fraction of support vectors. A value of *nu=0.35* was chosen experimentally.

On the other hand, the CNN classifier is based off the implementation by Kim et al. [5]. First, it transforms each word in the supplied fixed-length sentence into a vector that represents its meaning, using learned word embeddings. Then, convolution is performed over the matrix of word embeddings, generating high-level features from the caption. Finally, classification is realized through fully connected layers. This model is much more powerful than a SVM, partly because it incorporates the order of words in a caption.

3.4 Image-level Decision

Both GT and S&T provide multiple captions per image, with GT giving five captions per image, and S&T producing three captions per image. As a result, we must decide how to report the results of classification, especially at the image-level. We explore three methods for evaluation: per caption, voting, and concatenation. In per caption classification, we evaluate the correctness of every caption individually. This configuration is used solely for comparison with the other two methods. With voting classification, the classifier casts a 'vote' after each caption, and at the end, chooses the class with the most votes. Finally, we also explore concatenating all of the captions into one paragraph, then classifying the paragraph as a whole.

4. Evaluation

In this section, we describe our steps taken to evaluate the performance of our architecture.

4.1 Dataset

Since MSCOCO dataset contains images that do not involve human actions, we must first identify a subset of MSCOCO images which involve human actions and determine their action class labels.. To decide this, we will use NLP to parse out the action verbs in all of the captions. We start by compiling a list of the most used verbs in the captions, then, we hand choose a list of human actions to comprise the classes of HA, such that there are no synonyms.

Once this list is settled upon, ground truth labels are generated by checking every caption of an image for the presence of a verb that matches one of the chosen classes in HA. Only images whose captions have a single class label are used, so there are no images with two class labels. The last step is to normalize the distribution of images. The number of images (*N*) of the least populous class is used to determine the number of images in all classes. The dataset is comprised of *N* images per class, chosen randomly from the set of total images identified to be in the class. This ensures that no class occurs more often than another, allowing the classifier to make fair decisions. The test/val split is maintained by using the same split as MSCOCO.

Two class sets were used during evaluation, a 14-class set (HA-14), and a 20-class set (HA-20). HA-14 has 364 training images and 169 test images per class, while HA-20 has 252 training images and 114 test images per class. The class labels chosen for both sets are shown below in Table 1.

Play	Walk	Lay	Eat	Drive
Smile	Ski	Jump	Talk	Throw
Catch	Paint	Cook	Drink	Fall
Tie	Point	Give	Write	Stack

Table 1: Human action classes used. Underlined classes were those added to make the 20-class set (HA-20).

4.2 Metrics

We will be using two metrics to measure performance. Firstly, accuracy will measure what percent of the predicted action classes agree with the ground truth action classes. Additionally, confusion matrices will be used to gain deeper insight into the performance. Confusion matrices report an NxN (N classes in dataset) grid of predictions vs. ground truth. They are especially useful for understanding class-based trends.

4.3 Setup and Baseline

Our S&T model was pretrained on the MSCOCO dataset for 1 million iterations. Then, we fine-tuned the model on our human action subset for 400,000 iterations. The final, fine-tuned model will be referred to explicitly as S&T+FT. The model is implemented in Python, using Tensor Flow [1].

To evaluate our architecture, we performed a series of tests, varying parameters to gain insight into each one's influence on the system. These parameters are[1]: dataset (**HA-14** or HA-20), encoding (**BoW** or FLS), classifier (**SVM** or CNN), and evaluation scheme (per caption, **voting**, concatenation).

Configuration	**Accuracy**
S&T	57.3%
S&T with per caption	56.2%
S&T+FT	58.5%
S&T+FT with per caption	57.9%
S&T+FT with captions concatenated	**59.1%**
GT	77.3%
GT with FLS	48.2%
GT with CNN	66.0%
GT with per caption	71.6%
GT with captions concatenated, HA-20	93.9%
GT with captions concatenated	**94.5%**

Table 2: Accuracy results using different parameters. Bold indicates the highest accuracy for that caption source.

4.4 Caption Classification

Caption classification will be evaluated independently from caption generation, by using GT captions. The two classifiers, SVM and CNN will be trained on HA-14 or HA-20 training captions, then evaluated with the respective test data.

[1] Bold indicates the de facto parameter, unless otherwise specified during testing.

4.5 Results

Figure 3 shows the confusion matrices for S&T and S&T+FT, and Figure 4 shows confusion matrices for GT captions, using per image and concatenation image-level classification. All relevant accuracy results are reported in Table 2.

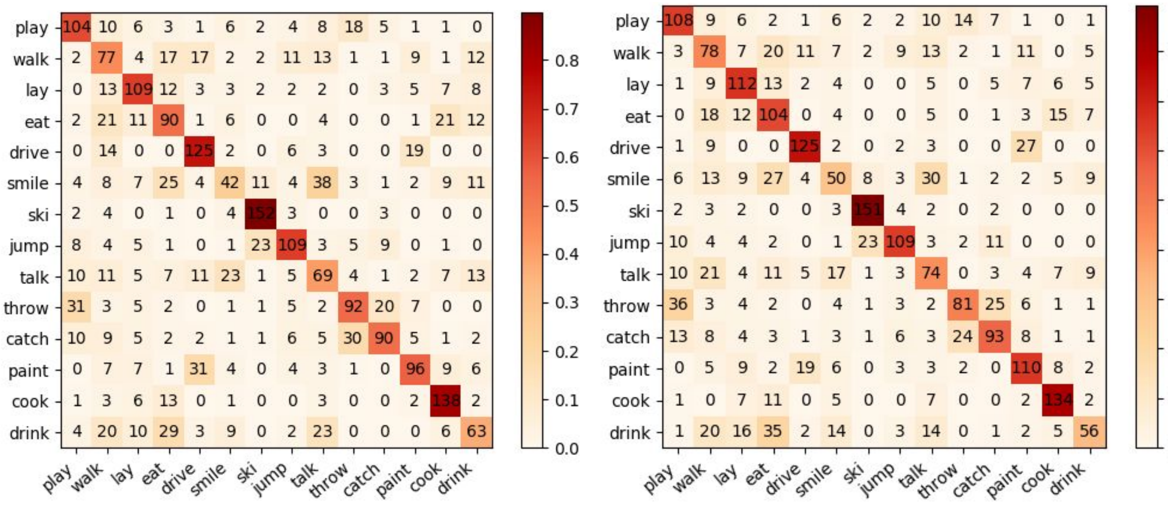


Figure 3: Left: S&T, accuracy: 57.3%. Right: S&T+FT, accuracy: 58.5%

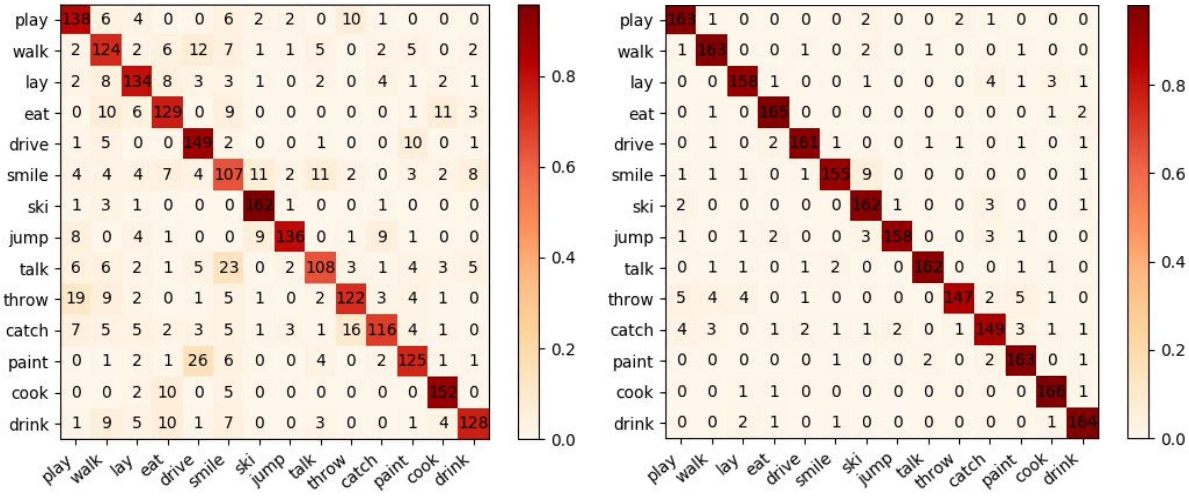


Figure 4: Left: GT, accuracy: 77.3%. Right: GT, captions concatenated, accuracy: 94.5%

For caption classification, we find that the ideal set of parameters is the HA-14 dataset, using BoW encoding with the SVM classifier, and concatenating all captions together to produce one input caption. This configuration was able to achieve an accuracy of almost 95% on GT captions. For captions generated with S&T and S&T+FT, the trends are the same, but less significant. For example, changing from per-image classifications to concatenating them all together gave an accuracy boost of 32% using GT captions, but only 1% using S&T+FT.

5. Discussion

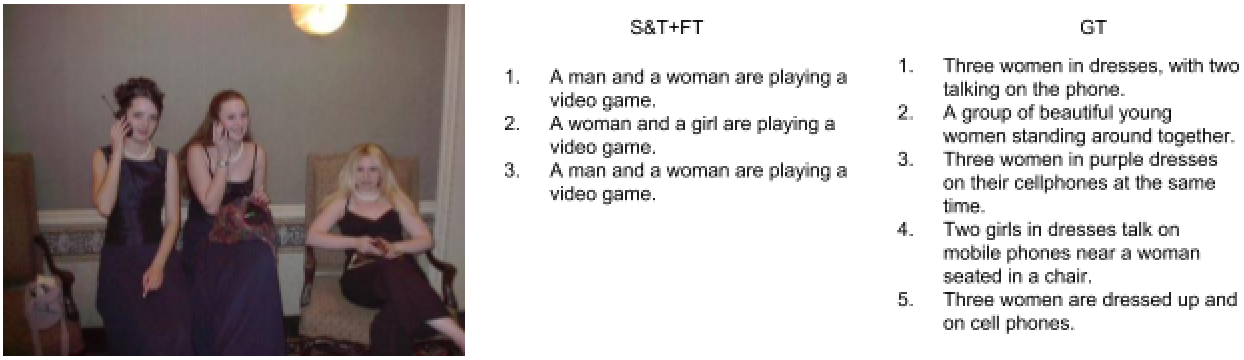
Here, we discuss what was learned from the evaluation of both major parts of the architecture: caption generation and caption classification. As well, we discuss how parameters affected the accuracy of the model. We also propose some improvements to the architecture to address its shortcomings.

5.1 Quality of Automatically Generated Captions

The results from the previous section show that classification on captions from S&T performed significantly worse than those from GT, which indicates inaccurate captions from S&T. Figure 5 shows an example image from the test set, along with the captions generated, supporting this hypothesis. Not only are the captions generated not relevant, they are also quite similar. This is in contrast to GT, where the five captions are very different descriptions of the same scene. Similarity within captions makes the caption classifier's job harder, when it comes to making an image-level decision. Similar captions will encode redundant information, whereas dissimilar captions will encode new, additional information.

Another takeaway from S&T analysis, is that although the 400,000 iterations of fine tuning improved performance, its impact was minimal, only raising accuracy be 1.2%. The best performance with automatically generated captions achieved just under 60% accuracy, which is certainly suboptimal.

The inaccuracy of S&T can be primarily attributed to the dataset. With only 5,096 images in the training set, S&T has very little data to train on. For reference, the full MSCOCO training set has 82,783 images, meaning our dataset is a measly 6.2% of the full dataset. Evidently, this action dataset is simply too small to get adequate performance out of. In addition to the dataset's size, it was found that some of the classifications were of questionable quality. Upon investigation, it was discovered that the analysis performed did not take into consideration what the subject of the sentence is (the noun performing the action verb). So, although we tried to limit the dataset to *human* actions, many irrelevant images and captions leaked into the dataset.



S&T+FT

1. A man and a woman are playing a video game.
2. A woman and a girl are playing a video game.
3. A man and a woman are playing a video game.

GT

1. Three women in dresses, with two talking on the phone.
2. A group of beautiful young women standing around together.
3. Three women in purple dresses on their cellphones at the same time.
4. Two girls in dresses talk on mobile phones near a woman seated in a chair.
5. Three women are dressed up and on cell phones.

Figure 5: Left: An image (id=536) with label 'Talk. Center: captions from S&T+FT, Right: captions from GT.

An example of this is with an image from the test set that depicts a table full of food. One GT caption for this image is: "A table laid out with food…", which we incorrectly flag as being in the class 'Lay'. Additionally, there are many images used where an animal is the subject of the captions. Removing these incongruous images would improve performance, but would further decrease the size of the dataset. This hints at the tradeoff of dataset size vs quality.

5.2 Improvements to Caption Generation

With S&T identified as the bottleneck in the architecture, we introduce three proposals to enhance the performance of our automatic caption generator.

5.2.1 Dataset Improvements

A dataset with less noise and more images, would allow the caption generator to produce higher quality captions. Quality of the dataset can be enhanced by using more meticulous NLP techniques. Specifically, the noun performing the verb must be identified. Only if that noun is a human, should the image be logged as belonging to the action class. Alternatively, we could eliminate all captions that don't specifically refer to a human. Though these changes would help clean the data set, we will still need additional labeled data to achieve higher performance.

One way to solve the issue of data scarcity, would be to use less classes in HA. Decreasing from 14 classes to 6 classes would increase the training set size from 5096 to 6528, a 28% increase in size. Counterintuitively, decreasing the number of classes increases the size of the whole dataset, due to the normalization step during dataset processing. However, this increase would likely be insufficient. At this point, we seem to have exhausted the capabilities of MSCOCO, so we must turn to another source to continue to grow the data.

We could further increase data size by using a photo-sharing site (such as flickr) to download images given the class keyword. We would then need to generate ground truth captions for these images, perhaps using Amazon Mechanical Turk. Additionally, data sets such as Human3.6M can be used in a similar way, by augmenting the ground truth actions with new ground truth captions. The downside to this is that creating ground truth captions will be expensive and will involve many human hours.

5.2.2 Caption Diversity

We previously identified that the captions created by S&T are very similar, and that adversely affected the classification accuracy. This issue could be addressed by redefining the loss function computed during training as a linear combination of the current loss function and the mutual difference between the sentences generated. This new loss function will encourage the caption generator to learn to generate multiple captions that all encode different information, while still converging to captions close to the ground truth. As we saw in GT, diverse captions will greatly improve accuracy with the caption classifier, especially when captions are concatenated with one another.

5.2.3 Deeper Model

Furthermore, with a larger dataset, a deeper model could be used. We could merge the two stages, caption generation and caption classification, into a single model that could perform end-to-end training. The architecture would start with a caption generator, such as S&T. After fine tuning using the loss function above, convolutional layers would be added to the end of the model to perform the classification (similar to the CNN text classifier that suffered from overfitting). The entire model could then be trained on the ground truth action labels. This differs from our model, because the gradient only propagates through the classifier at this stage, not both the caption generator and classifier. End-to-end learning is likely to give higher accuracy here, because merging the two components results in a single, trainable neural network. A downside to this approach is that it will be impossible to produce accurate captions after end-to-end learning. The caption generator is no longer judged on the accuracy of its captions, only on the accuracy of the final classification.

5.3 Overfitting with CNN

In Table 2, we notice a surprisingly low score for CNN caption classification. It seems that because CNN is a more powerful model than SVM, it suffered from overfitting, a phenomenon that occurs when training accuracy is much higher than test accuracy. See Figure 6 for an example of the train/test accuracy during learning. To try to mitigate the overfitting, we used pre-trained word2vec embeddings. We noticed that the accuracy rose much quicker, but still suffered from overfitting. Additionally, dropout and L2 regularization were introduced, but as well, did not make enough of an improvement. We speculate that with a larger dataset, overfitting would dissolve and CNN text classification would outperform SVM classification.

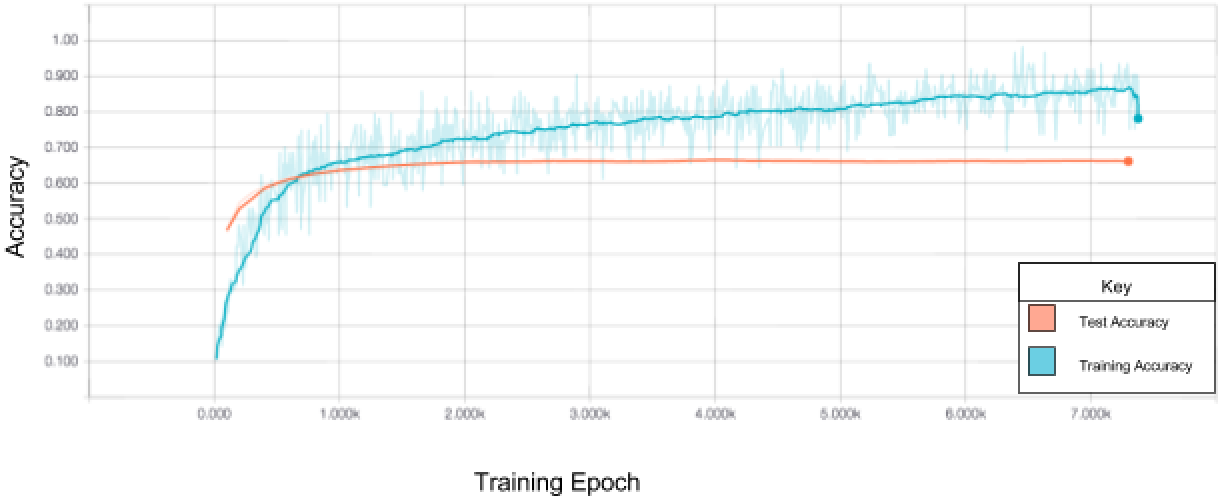


Figure 6: Overfitting of CNN text classification model.

5.4 Comparison of Voting and Concatenation

We found one parameter especially impactful for classification: evaluation scheme. As predicted, per-caption classification performed the worst because it generates a classification with the least information, one caption. However, only a relatively small increase was noticed when image-level caption classifications were generated with voting. On the other hand, a significant performance boost was measured when captions were concatenated with one another.

These observations suggests that some captions have more significance than others, of the same image. With voting, each caption is given the same influence on the final classification, a single vote. Evidently, this is suboptimal. For example, one of the five captions may contain the class label itself as a word, which is a strong indicator of the class of the image. Another caption may describe a different part of the scene that is irrelevant to the action class. However, in the case of an irrelevant caption, the classifier must still make a vote for one of the classes, and may end up swaying the decision because of this. In contrast, by concatenating the captions together, the classifier casts a single vote with all information at hand.

Voting could be improved by introducing one more class during testing, that indicates ‘unknown’ class. If the caption is classified as unknown, the vote is thrown away and not considered in the image-level classification. We can get an idea of how confident the SVM classifier is with a particular prediction by calculating the distance of the sample to the separating hyperplane, computed in training. A threshold could then be applied to this distance to decide if the classification is noisy, or confident enough to be cast as a vote. We postulate that with a proper threshold set, voting could achieve similar performance as concatenation.

The spike in performance gained by concatenating captions for GT, was interestingly much smaller when used on S&T captions. This is likely due to the lack of diversity in captions produced from the S&T model mentioned before. A caption generator that produces more diverse captions achieve closer accuracy to what we observed with GT captions.

5.5 Datasets

Comparing datasets, we find that HA-14 achieves higher accuracy than HA-20. Despite having almost the same number of images in the training set (5096 and 5040 respectively), HA-14 has more data points *per class*. However, the difference in performance is quite small, under a 1% improvement in accuracy. This minor drop in accuracy hints that the caption classifier could scale to more classes, given sufficient data.

5.6 Class by Class Analysis

Investigating performance on a per-class basis can be done by analyzing the confusion matrices in Figures 3 and 4. There are three pairs of classes that are most commonly confused: Drink & Eat, Smile & Talk, and Throw & Play. It is no coincidence that these pairs involve actions that visually appear similar. Another trend is that captions classified from S&T are often correct with images from the Ski class. Comparing Ski to the rest of the classes, it is the only action likely to take place in a white, winter scene, which is very visually distinguishable for the other scenes.

6. Conclusion

Due to time constraints, some insightful tests were unable to be performed. One of these would be to compare our results to a typical CNN classifier, like that of AlexNet or Inception v3. This would help us understand if our approach of using captions to gain context in an image is useful, or unnecessary. Additionally, more testing on the S&T learning parameters may have improved performance. Specifically, pre-training on the entirety of MSCOCO training for an additional 2-3 million iterations, then fine-tuning for 1 million iterations may have helped produce more accurate captions. However, this would have taken weeks to run on the local machine.

Future effort should be focused on addressing the improvements set forth in the previous section. Firstly, expanding the dataset and raising its ground truth accuracy will improve performance of S&T. Furthermore, using a deeper model with a better loss function that can be trained end-to-end, will go a step further.

In this paper, we explore a subset of MSCOCO with the goal of achieving human action classification. HS-14 and HS-20 are our attempts at creating datasets that provide both ground truth captions, and ground truth human action labels. Our architecture to generate action classification consists of a caption generator and a caption classifier. During evaluation, we discovered that the caption generator (S&T) cannot provide sufficient information to the caption classifier. We show that this is primarily due to the similarity of captions generated for one image and the small, noisy datasets. A modified loss function is outlined to address the similarity issue, and ideas are provided to grow and clean the datasets.

However, the caption classifier achieved an accuracy of almost 95% when receiving the ground truth captions as input, which indicates the validity of the architecture when accurate captions are provided. Additionally, modification of parameters provided insight into how to optimally classify multiple captions. Most interestingly, we found that concatenating captions together provides a significant boost in classification performance. Finally, we introduce a more powerful architecture that could perform end-to-end learning, given an improved dataset.

List of References (MLA):

[1] Abadi, Martín, et al. "Tensorflow: Large-scale machine learning on heterogeneous distributed systems." arXiv preprint arXiv:1603.04467 (2016).

[2] Guo, Guodong, and Alice Lai. "A survey on still image based human action recognition." Pattern Recognition 47.10 (2014): 3343-3361.

[3] Gupta, Raj, Alex Yong-Sang Chia, and Deepu Rajan. "Human activities recognition using depth images." Proceedings of the 21st ACM international conference on Multimedia. ACM, 2013.

[4] Ionescu, Catalin, et al. "Human3. 6m: Large scale datasets and predictive methods for 3d human sensing in natural environments." IEEE transactions on pattern analysis and machine intelligence 36.7 (2014): 1325-1339.

[5] Kim, Yoon. "Convolutional neural networks for sentence classification." arXiv preprint arXiv:1408.5882 (2014).

[6] Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton. "Imagenet classification with deep convolutional neural networks." Advances in neural information processing systems. 2012.

[7] Lin, Tsung-Yi, et al. "Microsoft coco: Common objects in context." European Conference on Computer Vision. Springer International Publishing, 2014.

[8] Pedregosa, Fabian, et al. "Scikit-learn: Machine learning in Python." Journal of Machine Learning Research 12.Oct (2011): 2825-2830.Vinyals, Oriol, et al. "Show and tell: A neural image caption generator." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015.

[9] Vinyals, Oriol, et al. "Show and tell: A neural image caption generator." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015.

[10] Waltner, Georg, Thomas Mauthner, and Horst Bischof. "Indoor Activity Detection and Recognition for Sport Games Analysis." arXiv preprint arXiv:1404.6413 (2014).

Vita:

Born in 1993 to the family of Martin and Karen Kafka in Newton, MA, Adam has been pursuing his passion for computer science from an early age. Adam graduated from Lehigh University with highest honors as an Undergrad in May 2016, majoring in Computer Science. During this time, he and earned the Presidential Scholarship, an opportunity to stay an extra year in pursuit of a Master’s degree.