On instabilities of deep learning in image reconstruction and the potential costs of AI

Vegard Antun[a], Francesco Renna[b], Clarice Poon[c], Ben Adcock[d], and Anders C. Hansen[a,e,1]

[a]Department of Mathematics, University of Oslo, 0316 Oslo, Norway; [b]Instituto de Telecomunicações, Faculdade de Ciências, Universidade do Porto, Porto 4169-007, Portugal; [c]Department of Mathematical Sciences, University of Bath, Bath BA2 7AY, United Kingdom; [d]Department of Mathematics, Simon Fraser University, Burnaby, BC V5A 1S6, Canada; and [e]Department of Applied Mathematics and Theoretical Physics, University of Cambridge, Cambridge CB3 0WA, United Kingdom

Edited by David L. Donoho, Stanford University, Stanford, CA, and approved March 12, 2020 (received for review June 4, 2019)

Deep learning, due to its unprecedented success in tasks such as image classification, has emerged as a new tool in image reconstruction with potential to change the field. In this paper, we demonstrate a crucial phenomenon: Deep learning typically yields unstable methods for image reconstruction. The instabilities usually occur in several forms: 1) Certain tiny, almost undetectable perturbations, both in the image and sampling domain, may result in severe artefacts in the reconstruction; 2) a small structural change, for example, a tumor, may not be captured in the reconstructed image; and 3) (a counterintuitive type of instability) more samples may yield poorer performance. Our stability test with algorithms and easy-to-use software detects the instability phenomena. The test is aimed at researchers, to test their networks for instabilities, and for government agencies, such as the Food and Drug Administration (FDA), to secure safe use of deep learning methods.

instability | deep learning | AI | image reconstruction | inverse problems

There are two paradigm changes currently happening: 1) Artificial intelligence (AI) is replacing humans in problem solving; however, 2) AI is also replacing the standard algorithms in computational science and engineering. Since reliable numerical calculations are paramount, algorithms for computational science are traditionally based on two pillars: accuracy and stability. This is, in particular, true of image reconstruction, which is a mainstay of computational science, providing fundamental tools in medical, scientific, and industrial imaging. This paper demonstrates that the stability pillar is typically absent in current deep learning and AI-based algorithms for image reconstruction. This raises two fundamental questions: How reliable are such algorithms when applied in the sciences, and do AI-based algorithms have an unavoidable Achilles heel: instability? This paper introduces a comprehensive testing framework designed to demonstrate, investigate, and, ultimately, answer these foundational questions.

The importance of stable and accurate methods for image reconstruction for inverse problems is hard to overestimate. These techniques form the foundation for essential tools across the physical and life sciences such as MRI, computerized tomography (CT), fluorescence microscopy, electron tomography, NMR, radio interferometry, lensless cameras, etc. Moreover, stability is traditionally considered a necessity in order to secure reliable and trustworthy methods used in, for example, cancer diagnosis. Hence, there is an extensive literature on designing stable methods for image reconstruction in inverse problems (1–4).

AI techniques such as deep learning and neural networks (5) have provided a new paradigm with new techniques in inverse problems (6–15) that may change the field. In particular, the reconstruction algorithms learn how to best do the reconstruction based on training from previous data, and, through this training procedure, aim to optimize the quality of the reconstruction. This is a radical change from the current state of the art (SoA) from an engineering, physical, and mathematical point of view.

AI and deep learning have already changed the field of computer vision and image classification (16–19), where the performance is now referred to as super human (20). However, the success comes with a price. Indeed, the methods are highly unstable. It is now well established (21–25) that high-performance deep learning methods for image classification are subject to failure given tiny, almost invisible perturbation of the image. An image of a cat may be classified correctly; however, a tiny change, invisible to the human eye, may cause the algorithm to change its classification label from cat to fire truck, or another label far from the original.

In this paper, we establish the instability phenomenon of deep learning in image reconstruction for inverse problems. A potential surprising conclusion is that the phenomenon may be independent of the underlying mathematical model. For example, MRI is based on sampling the Fourier transform, whereas CT is based on sampling the Radon transform. These are rather different models, yet the instability phenomena happen for both sampling modalities when using deep learning.

There is, however, a big difference between the instabilities of deep learning for image classification and our results on instabilities of deep learning for image reconstruction. Firstly, in the former case, there is only one thing that could go wrong: A small perturbation results in a wrong classification. In image reconstruction, there are several potential forms of instabilities. In particular, we consider three crucial issues: 1) instabilities with respect to certain tiny perturbations, 2) instabilities with respect to small structural changes (for example a brain image with or without a small tumor), and 3) instabilities with respect to changes in the number of samples. Secondly, the two problems are totally unrelated. Indeed, the former problem is, in its simplest form, a decision problem, and hence the decision function ("Is there a cat in the image?") to be approximated is necessarily

This paper results from the Arthur M. Sackler Colloquium of the National Academy of Sciences, "The Science of Deep Learning," held March 13–14, 2019, at the National Academy of Sciences in Washington, DC. NAS colloquia began in 1991 and have been published in PNAS since 1995. From February 2001 through May 2019 colloquia were supported by a generous gift from The Dame Jillian and Dr. Arthur M. Sackler Foundation for the Arts, Sciences, & Humanities, in memory of Dame Sackler's husband, Arthur M. Sackler. The complete program and video recordings of most presentations are available on the NAS website at http://www.nasonline.org/science-of-deep-learning.

Author contributions: B.A. and A.C.H. designed research; V.A., F.R., and C.P. performed research; V.A., F.R., C.P., B.A., and A.C.H. wrote the paper; and V.A., F.R., and C.P. wrote code.

The authors declare no competing interest.

This article is a PNAS Direct Submission.

Published under the PNAS license.

Data deposition: All of the code is available from GitHub at https://github.com/vegarant/Invfool.

[1]To whom correspondence may be addressed. Email: ach70@cam.ac.uk.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.1907377117/-/DCSupplemental.

First published May 11, 2020.

Downloaded at Palestinian Territory, occupied on December 22, 2021

www.manaraa.com

discontinuous. However, the problem of reconstructing an image from Fourier coefficients, as is the problem in MRI, is completely different. In this case, there exist stable and accurate methods that depend continuously on the input. It is therefore paradoxical that deep learning leads to unstable methods for problems that can be solved accurately in a stable way (*SI Appendix*, *Methods*).

The networks we have tested are unstable either in the form of category 1 or 2 or both. Moreover, networks that are highly stable in one of the categories tend to be highly unstable in the other. The instability in form of category 3, however, occurs for some networks but not all. The findings raise two fundamental questions:

1) Does AI, as we know it, come at a cost? Is instability a necessary by-product of our current AI techniques?
2) Can reconstruction methods based on deep learning always be safely used in the physical and life sciences? Or, are there cases for which instabilities may lead to, for example, incorrect medical diagnosis if applied in medical imaging?

The scope of this paper is on the second question, as the first question is on foundations, and our stability test provides the starting point for answering question 2. However, even if instabilities occur, this should not rule out the use of deep learning methods in inverse problems. In fact, one may be able to show, with large empirical statistical tests, that the artifacts caused by instabilities occur infrequently. As our test reveals, there is a myriad of different artifacts that may occur, as a result of the instabilities, suggesting vast efforts needed to answer question 2. A detailed account is provided in *Conclusion* .

The Instability Test

The instability test is based on the three instability issues mentioned above. We consider instabilities with respect to the following.

Tiny Worst-Case Perturbations. The tiny perturbation could be in the image domain or in the sampling domain. When considering medical imaging, a perturbation in the image domain could come from a slight movement of the patient, small anatomic differences between people, etc. The perturbation in the sampling domain may be caused by malfunctioning of the equipment or the inevitable noise dictated by the physical model of the scanning machine. However, a perturbation in the image domain may imply a perturbation in the sampling domain. Also, in many cases, the mathematical model of the sampling reveals that such a sampling process implies an operator that is surjective onto its range, and hence there exists a perturbation in the image domain corresponding to the perturbation in the sampling domain. Thus, a combination of all these factors may yield perturbations that, in a worst-case scenario, may be quite specific, hard to model, and hard to protect against, unless one has a completely stable neural network.

The instability test includes algorithms that do the following. Given an image and a neural network, designed for image reconstruction from samples provided by a specific sampling modality, the algorithm searches for a perturbation of the image that makes the most severe change in the output of the network while still keeping the perturbation small. In a simple mathematical form, this can be described as follows. Given an image $x \in \mathbb{R}^N$ (we interpret an image as a vector for simplicity), a matrix $A \in \mathbb{C}^{m\times N}$ representing the sampling modality (for example, a discrete Fourier transform modeling MRI), and a neural network $f : \mathbb{C}^m \to \mathbb{C}^N$, the neural network reconstructs an approximation $\tilde{x}$ to x defined by $y = Ax$, where $\tilde{x} = f(y)$. The algorithm seeks an $r \in \mathbb{R}^N$ such that

$$\|f(y + Ar) - f(y)\| \text{ is large, while } \|r\| \text{ is small;}$$

see *Methods* for details. However, the perturbation could, of course, be put on the measurement vector y instead.

Small Structural Changes in the Image. By structural change, we mean a change in the image domain that may not be tiny, and typically is significant and clearly visible, but is still small (for example, a small tumor). The purpose is to check whether the network can recover important details that are crucial in, for example, medical assessments. In particular, given the image $x \in \mathbb{R}^N$, we add a perturbation $r \in \mathbb{R}^N$, where r is a detail that is clearly visible in the perturbed image $x + r$, and check whether r is still clearly visible in the reconstructed image,

$$\hat{x} = f(A(x + r)).$$

In this paper, we consider the symbols from cards as well as letters. In particular, we add the symbols ♠, ♡, ♢, ♣ and the letters CAN U SEE IT to the image. The reason for this is that card symbols as well as letters are fine details that are hard to detect, and thus represent a reasonable challenge for the network. If the network is able to recover these small structural changes, it is likely to recover other details of the same size. On the other hand, if the network fails on these basic changes, it is likely to fail on other details as well. The symbols can, of course, be specified depending on the specific application. Our choice is merely for illustration.

An important note is that, when testing stability, both with respect to tiny perturbations and with respect to small structural changes, the test is always done in comparison with an SoA stable method in order to check that any instabilities produced by the neural network are due to the network itself and not because of ill-conditioning of the inverse problem. The SoA methods used are based on compressed sensing and sparse regularization (26–28). These methods often come with mathematical stability guaranties (29), and are hence suitable as benchmarks (see *Methods* for details).

Changing the Number of Samples in the Sampling Device (Such as the MRI or CT Scanner). Typical SoA methods share a common quality: More samples imply better quality of the reconstruction. Given that deep learning neural networks in inverse problems are trained given a specific sampling pattern, the question is, How robust is the trained network with respect to changes in the sampling? The test checks whether the quality of the reconstruction deteriorates with more samples. This is a crucial question in applications. For example, the recent implementation of compressed sensing on Philips MRI machines allows the user to change the undersampling ration for every scan. This means that, if a network is trained on 25% subsampling, say, and, suddenly, the user changes the subsampling ratio to 35%, one would want an improved recovery. If the quality deteriorates or stagnates with more samples, this means that one will have to produce networks trained for each and every combination of subsampling that the machine allows for. Finally, due to the other instability issues, every such network must, individually, be empirically statistically tested to detect whether the occurrence of instabilities is rare or not. It is not enough to test on only one neural network, as their instabilities may be completely different.

Testing the Test

We test six deep learning neural networks selected based on their strong performance, wide range in architectures, and difference in sampling patterns and subsampling ratios, as well as their difference in training data. The specific details about the architecture and the training data of the tested networks can be found in *SI Appendix*.

Downloaded at Palestinian Territory, occupied on December 22, 2021

www.manaraa.com

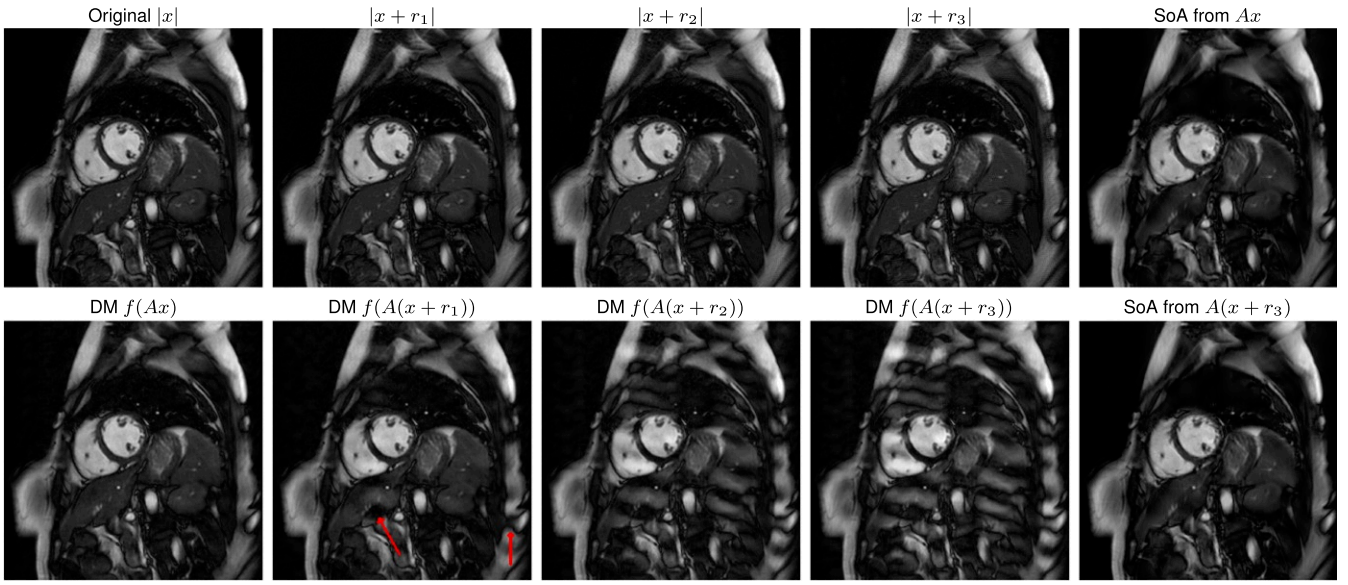

Fig. 1. Perturbations r_j (created to simulate worst-case effect) with $|r_1| < |r_2| < |r_3|$ are added to the image x. (*Top*) Images 1 to 4 are original image x and perturbations $x + r_j$. (Bottom) Images 1 to 4 are reconstructions from $A(x + r_j)$ using the Deep MRI (DM) network f, where A is a subsampled Fourier transform (33% subsampling); see *Methods* and *SI Appendix* for details. (*Top* and *Bottom*) Image 5 is a reconstruction from Ax and $A(x + r_3)$ using an SoA method; see *Methods* for details. Note how the artifacts (red arrows) are hard to dismiss as nonphysical.

An important note is that the tests performed are not designed to test deep learning against SoA in terms of performance on specific images. The test is designed to detect the instability phenomenon. Hence, the comparison with SoA is only to verify that the instabilities are exclusive only to neural networks based on deep learning, and not due to an ill-conditioning of the problem itself. Moreover, as is clear from the images, in the unperturbed cases, the best performance varies between neural networks and SoA. The list of networks is as follows.

AUTOMAP (6) is a neural network for low-resolution single-coil MRI with 60% subsampling. The training set consists of brain images with white noise added to the Fourier samples.

DAGAN (12) is a network for medium-resolution single-coil MRI with 20% subsampling, and is trained with a variety of brain images.

Deep MRI (11) is a neural network for medium-resolution single-coil MRI with 33% subsampling. It is trained with detailed cardiac MR images.

Ell 50 (9) is a network for CT or any Radon transform-based inverse problem. It is trained on images containing solely ellipses (hence the name Ell 50). The number 50 refers to the number of lines used in the sampling in the sinogram.

Med 50 (9) has exactly the same architecture as Ell 50 and is used for CT; however, it is trained with medical images (hence the name Med 50) from the Mayo Clinic database (13). The number of lines used in the sampling from the sinogram is 50.

MRI-VN (14) is a network for medium- to high-resolution parallel MRI with 15 coil elements and 15% subsampling. The training is done with a variety of knee images.

Stability with Respect to Tiny Worst-Case Perturbations

Below follows the description of the test applied to some of the networks where we detect instabilities with respect to tiny perturbations.

For the Deep MRI test, we perturb the image x with a sequence of perturbations $\{r_j\}_{j=1}^{3}$ with $|r_1| < |r_2| < |r_3|$ in order to simulate how the instabilities continuously transform the reconstructed image from a very high-quality reconstruction to an almost unrecognizable distortion. This is illustrated in Fig. 1, *Bottom*. Note that the perturbations are almost invisible to the

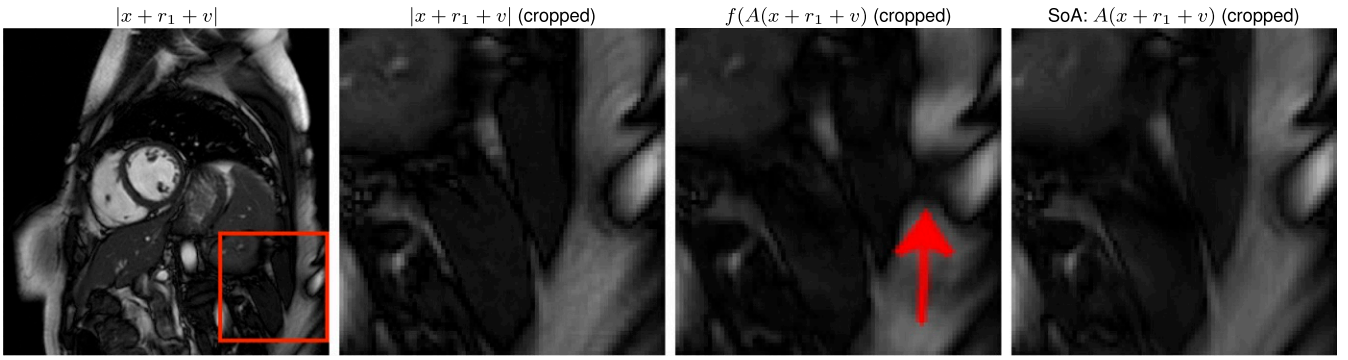

Fig. 2. A random Gaussian vector $e \in \mathbb{C}^m$ is computed by drawing (the real and imaginary part of) each component independently from the normal distribution $\mathcal{N}(0, 10)$. We let $v = A^*e$, and rescale v so that $\|v\|_2 = \frac{1}{4}\|r_1\|_2$, where r_1 is the perturbation from Fig. 1. The Deep MRI network f reconstructs from the measurements $A(x + r_1 + v)$ and shows the same artifact as was seen for r_1 in Fig. 1. Note that, in this experiment, $A \in \mathbb{C}^{m\times N}$ is a subsampled normalized discrete Fourier transform (33% subsampling), so that $AA^* = I$ that is, $e = Av$.

Downloaded at Palestinian Territory, occupied on December 22, 2021
www.manaraa.com

human eye, as demonstrated in Fig. 1., *Top*. The r_j perturbations are created by early stopping of the algorithm iterating to solve for the optimal worst-case perturbation. The purpose of this experiment is to demonstrate how the gradual change in perturbation creates artifacts that may be hard to verify as nonphysical. Indeed, the worst-case perturbation r_3 causes clearly a reconstruction that, in a real-world situation, can be dismissed by a clinician as nonphysical. However, for the smallest r_1, we have a perturbation that is completely invisible to the human eye, yet it results in a reconstruction that is hard to dismiss as nonphysical, and provides an incorrect representation of the actual image. Such examples could, potentially, lead to incorrect medical diagnosis. Note that SoA methods are not affected by the perturbation as demonstrated in the fifth column of Fig. 1. However, although this network is highly unstable with respect to certain tiny perturbations, it is highly stable with respect to small structured changes; see Fig. 5, *Lower Middle*. Note also that the instabilities are actually stable. In particular, in Fig. 2, we demonstrate how a random Gaussian perturbation added to the perturbation r_1 still yields a substantial artifact (see also *SI Appendix*, *Methods*).

The AUTOMAP experiment is similar to the one above; however, in this case, we add $\tilde{r}_1, \ldots, \tilde{r}_4$ to the measurements $y = Ax$, where $|\tilde{r}_1| < |\tilde{r}_2| < |\tilde{r}_3| < |\tilde{r}_4|$, and A is a subsampled discrete Fourier transform. In order to illustrate how small the perturbations are, we have visualized $|x + r_j|$ in Fig. 3, *Top*, where $y + \tilde{r}_j = A(x + r_j)$. To emphasize how the network reconstruction completely deforms the image, we have, inspired by the second test on structural changes, added a small structural change in the form of a heart that gradually disappears completely in the network reconstruction. This is demonstrated in Fig. 3, *Middle*, and Fig. 3, *Bottom* contains the reconstruction done by an SoA method. Note that the worst-case perturbations are completely different than the ones failing the Deep MRI network. Hence, the artifacts are also completely different. These perturbations are white noise-like, and the reconstructions from the network provide a similar impression. As this is a standard artifact in MRI, it is, however, not clear how to protect against the potential bad tiny noise. Indeed, a detail may be washed out, as shown in the experiment (note the heart inserted with slightly different intensities in the brain image), but the similarity to a standard artifact may make it difficult to judge that this is an untrustworthy image.

In the case of MRI-VN, we add one perturbation r_1 to the image, where r_1 is produced by letting the algorithm searching for the worst perturbation run until it has converged. The results are shown in the first two columns of Fig. 4, and the conclusion is the same for the MRI-VN net as for Deep MRI and AUTOPMAP: Perturbations barely visible to the human eye, even when zooming in, yield substantial misleading artifacts. Note also that the perturbation has no effect on the SoA method.

For Med-50, we add a perturbation r_2 that is also produced by running the algorithm until it has converged, and the results are visualized in the last two columns of Fig. 4. The Med-50 network is moderately unstable with respect to tiny perturbations compared to Deep MRI, AUTOMAP, and MRI-VN; however,

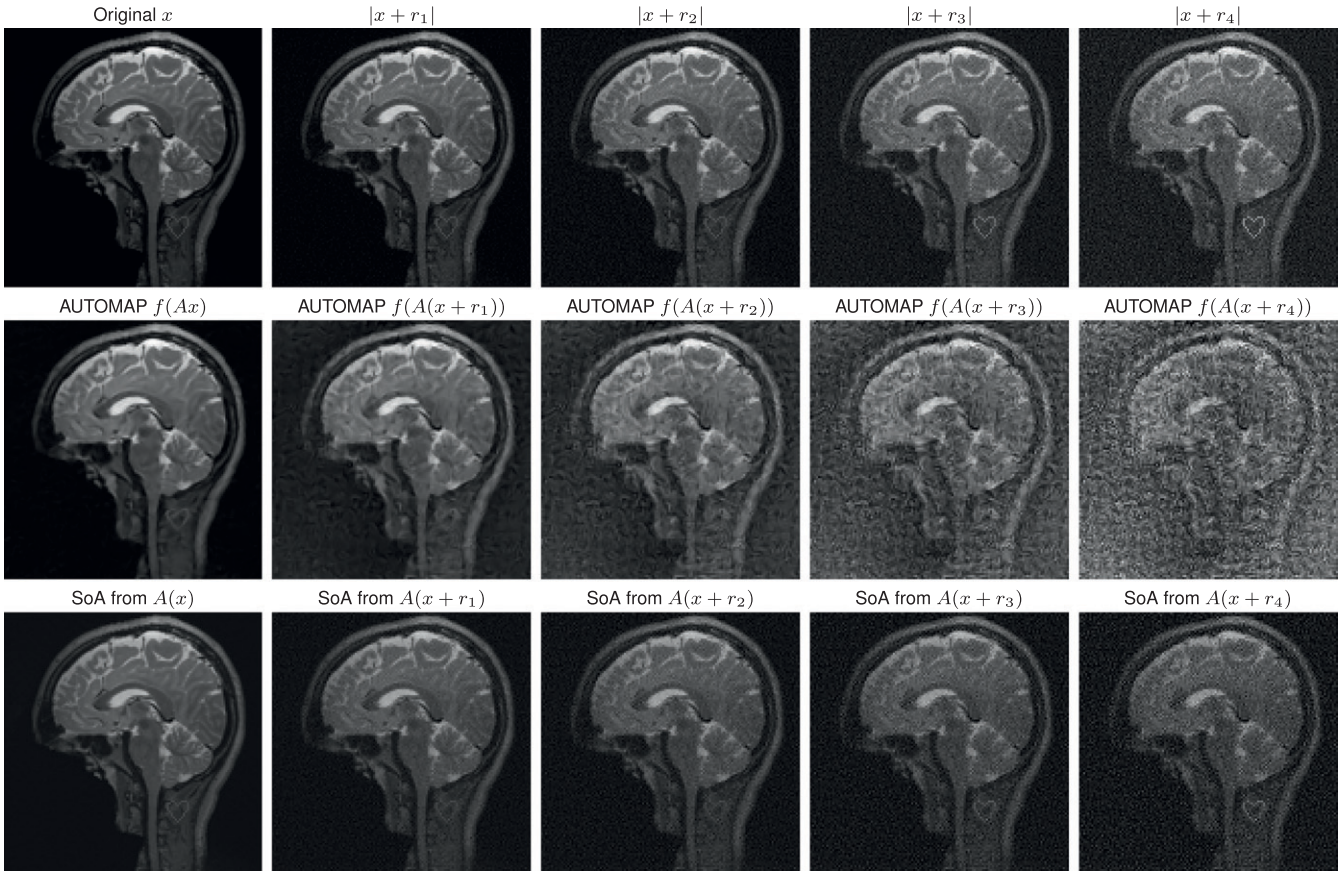

Fig. 3. Perturbations $\tilde{r}_j$ (created to simulate worst-case effect) are added to the measurements $y = Ax$, where $|\tilde{r}_1| < |\tilde{r}_2| < |\tilde{r}_3| < |\tilde{r}_4|$, and A is a subsampled Fourier transform (60% subsampling). To visualize, we show $|x + r_j|$, where $y + \tilde{r}_j = A(x + r_j)$. (*Top*) Original image x with perturbations r_j. (*Middle*) Reconstructions from $A(x + r_j)$ by the AUTOMAP network f. (*Bottom*) Reconstructions from $A(x + r_j)$ by an SoA method (see *Methods* for details). A detail in the form of a heart, with varying intensity, is added to visualize the loss in quality.

Downloaded at Palestinian Territory, occupied on December 22, 2021

www.manaraa.com

severe artifacts are clearly seen. It is worth noting that this network is used for the Radon transform, which is, from a stability point of view, a more unstable operator than the Fourier transform when considering its inverse.

Stability with Respect to Small Structural Changes

Instabilities with respect to small structural changes are documented below.

The Ell-50 network provides a stark example of instability with respect to structural perturbation. Indeed, none of the details are visible in the reconstruction as documented in Fig. 5, *Top*. This may not be entirely surprising, given that the network is trained on ellipses.

The DAGAN network is not as unstable as the Ell-50 network with respect to structural changes. However, as seen in Fig. 5, *Upper Middle*, the blurring of the structural details are substantial, and the instability is still critical.

MRI-VN is an example of a moderately unstable network when considering structural changes. Note, however, how the instability coincides with the lack of ability to reconstruct details in general. This is documented in Fig. 5, *Middle*.

For Deep MRI, to demonstrate how the stability with respect to small structured changes coincides with the ability to reconstruct details, we show how stable the Deep MRI network is. Observe also how well the details in the image are preserved in Fig. 5, *Lower Middle*. Here we have lowered the subsampling ration to 25% even when the network is trained on 33% subsampling ratio. We want to point out that none of the symbols, nor any text, has been used in the training set.

Stability with Respect to More Samples

Certain convolutional neural networks will allow for the flexibility of changing the amount of sampling. In our test cases, all of the networks except AUTOMAP have this feature, and we report on the stability with respect to changes in the amount of samples below and in Fig. 5, *Bottom*.

Ell 50 has the strongest and most fascinating decay in performance as a function of an increasing subsampling ratio. Med 50 is similar, however, with a less steep decline in reconstruction quality.

For DAGAN, the reconstruction quality deteriorates with more samples, similar to the Ell 50/Med 50 networks.

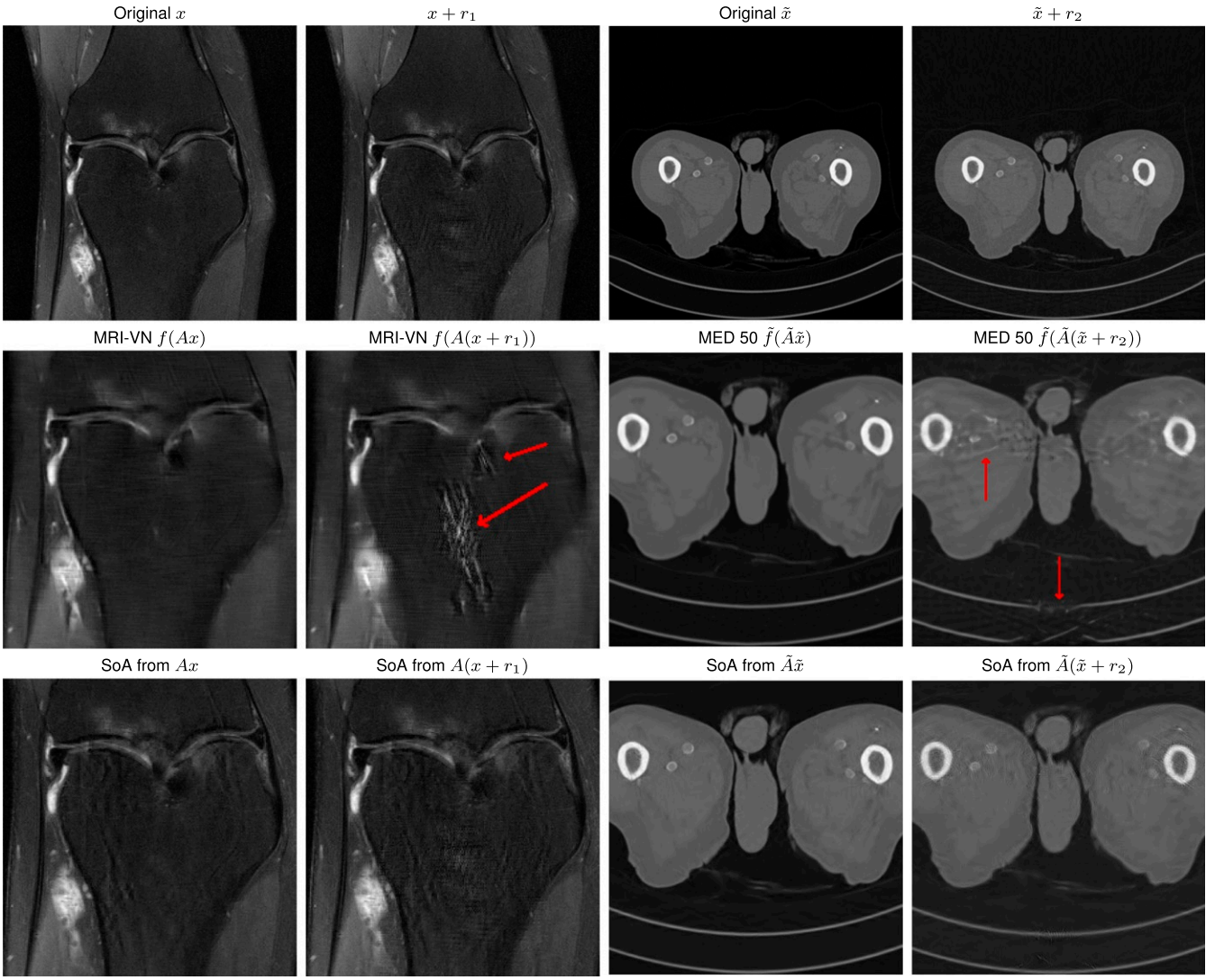

Fig. 4. (*Top*) Perturbations r_1, r_2 (created to simulate worst-case effect) are added to the images x and $\tilde{x}$. (*Middle*) The reconstructions by the network f (MRI-VN), from Ax and $A(x + r_1)$, and the network $\tilde{f}$ (MED 50), from $\tilde{A}\tilde{x}$ and $\tilde{A}(\tilde{x} + r_2)$. A is a subsampled discrete Fourier transform, and $\tilde{A}$ is a subsampled Radon transform. (*Bottom*) SoA comparisons. Red arrows are added to highlight the instabilities.

Downloaded at Palestinian Territory, occupied on December 22, 2021
www.manaraa.com

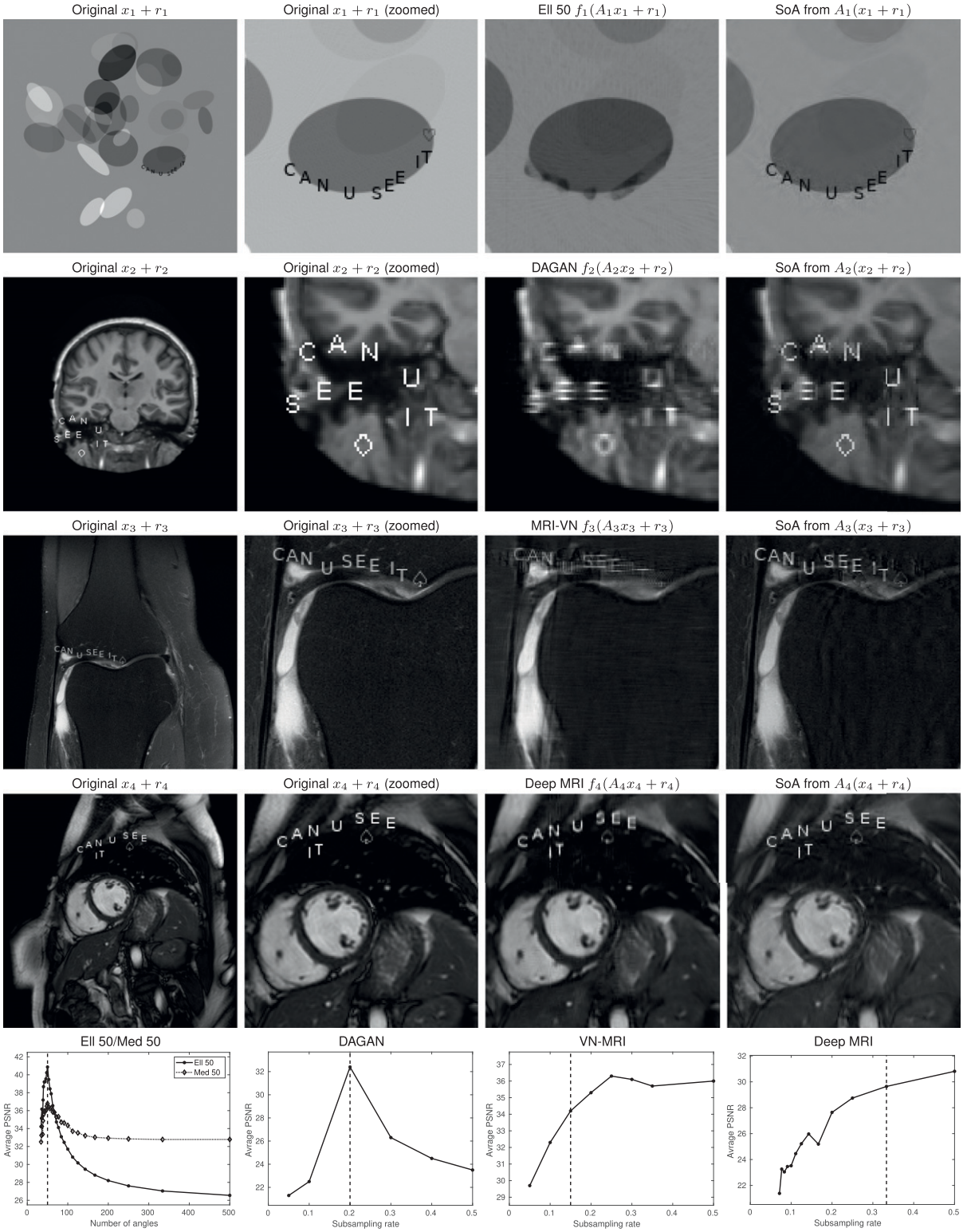

Fig. 5. (*Top*, *Upper Middle*, *Middle*, and *Lower Middle*) Images x_j plus structured perturbations r_j (in the form of text and symbols) are reconstructed from measurements $y_j = A_j(x_j + r_j)$ with neural networks f_j and SoA methods. The networks are $f_1 =$ Ell 50, $f_2 =$ DAGAN, $f_3 =$ MRI-VN, and $f_4 =$ Deep MRI. The sampling modalities A_j are as follows: A_1 is a subsampled discrete Radon transform, A_2 is a subsampled discrete Fourier transform (single coil simulation), A_3 is a superposition of subsampled discrete Fourier transforms (parallel MRI simulation with 15 coils elements), and A_4 is a subsampled discrete Fourier transform (single coil). Note that Deep MRI has not been trained with images containing any of the letters or symbols used in the perturbation, yet it is completely stable with respect to the structural changes. The same is true for the AUTOMAP network (see first column of Fig. 3). (*Bottom*) PSNR as a function of the subsampling rate for different networks. The dashed red line indicates the subsampling ratio that the networks were trained for.

Downloaded at Palestinian Territory, occupied on December 22, 2021

www.manaraa.com

The VN-MRI network provides reconstructions where the quality stagnates with more samples, as opposed to the decay in performance witnessed in the other cases.

The Deep MRI network is the only one that behaves in a way aligned with standard SoA methods and provides better reconstructions when more samples are added.

Conclusion

The new paradigm of learning the reconstruction algorithm for image reconstruction in inverse problem, through deep learning, typically yields unstable methods. Moreover, our test reveals numerous instability phenomena, challenges, and research directions. In particular, we find the following:

1) Certain tiny perturbations lead to a myriad of different artifacts. Different networks yield different artifacts and instabilities, and, as Figs. 1, 3, and 4 reveal, there is no common denominator. Moreover, the artifacts may be difficult to detect as nonphysical. Thus, several key questions emerge: Given a trained neural network, which types of artifacts may the network produce? How is the instability related to the network architecture, training set, and also subsampling patterns?
2) There is variety in the failure of recovering structural changes. There is a great variety in the instabilities with respect to structural changes as demonstrated in Fig. 4, ranging from complete removal of details to more subtle distortions and blurring of the features. How is this related to the network architecture and training set? Moreover, does the subsampling pattern play a role? It is important, however, to observe (as in Fig. 5, *Lower Middle* and the first column of Fig. 3) that there are perfectly stable networks with respect to structural changes, even when the training set does not contain any images with such details.
3) Networks must be retrained on any subsampling pattern. The fact that more samples may cause the quality of reconstruction to either deteriorate or stagnate means that each network has to be retrained on every specific subsampling pattern, subsampling ratio, and dimensions used. Hence, one may, in practice, need hundreds of different network to facilitate the many different combinations of dimensions, subsampling ratios, and sampling patterns.
4) Instabilities are not necessarily rare events. A key question regarding instabilities with respect to tiny perturbations is whether they may occur in practice. The example in Fig. 2 suggests that there is a ball around a worst-case perturbation in which the severe artifacts are always witnessed. This suggests that the set of "bad" perturbations have Lebesgue measure greater than zero, and, thus, there will typically be a nonzero probability of a "bad" perturbation. Estimating this probability may be highly nontrivial, as the perturbation will typically be the sum of two random variables, where one variable comes from generic noise and one highly nongeneric variable is due to patient movements, anatomic differences, apparatus malfunctions, etc. These predictions can also be theoretically verified, as discussed in *SI Appendix*, *Methods*.
5) The instability phenomenon is not easy to remedy. We deliberately choose quite different networks in this paper to highlight the seeming ubiquity of the instability phenomenon. Theoretical insights [see *SI Appendix*, *Methods* on the next generation of methods (30–34)] also support the conclusion that this phenomenon is nontrivial to overcome. Finding effective remedies is an extremely important future challenge.

Code and Data. All of the code is available from https://github.com/vegarant/Invfool.

ACKNOWLEDGMENTS. We thank Kristian Monsen Haug for help with *SI Appendix*, Fig. S3. We thank Dr. Cynthia McCollough, the Mayo Clinic, the American Association of Physicists in Medicine, and the National Institute of Biomedical Imaging and Bioengineering for allowing the use of their data in the experiments. F.R. acknowledges support from the European Union's Horizon 2020 research and innovation program under the Marie Sklodowska-Curie Grant Agreement 655282 and funds through FCT (Fundação para a Ciência e a Tecnologia, I.P.), under the Scientific Employment Stimulus – Individual Call – CEECIND/01970/2017. B.A. acknowledges support from Natural Sciences and Engineering Research Council of Canada (NSERC) Grant 611675. A.C.H. thanks Nvidia for a graphics processing unit (GPU) grant in the form of a Titan X Pascal and acknowledges support from a Royal Society University Research Fellowship, UK Engineering and Physical Sciences Research Council Grant EP/L003457/1, and a Leverhulme Prize 2017.

1. S. F. Gull, G. J. Daniell, Image reconstruction from incomplete and noisy data. *Nature* **272**, 686–690 (1978).
2. V. Studer *et al.*, Compressive fluorescence microscopy for biological and hyperspectral imaging. *Proc. Natl. Acad. Sci. U.S.A.* **109**, 1679–1687 (2011).
3. H. W. Engl, M. Hanke, A. Neubauer, *Regularization of Inverse Problems* (Mathematics and its Applications, Springer Netherlands, 1996).
4. P. C. Hansen, "The l-curve and its use in the numerical treatment of inverse problems" in *Computational Inverse Problems in Electrocardiology*, P. R. Johnston, Ed. (Advances in Computational Bioengineering, WIT Press, 2000), pp. 119–142.
5. Y. LeCun, Y. Bengio, G. Hinton, Deep learning. *Nature* **521**, 436–444 (2015).
6. B. Zhu, J. Z. Liu, S. F. Cauley, B. R. Rosen, M. S. Rosen, Image reconstruction by domain-transform manifold learning. *Nature* **555**, 487–492 (2018).
7. R. Strack, AI transforms image reconstruction. *Nat. Methods* **15**, 309 (2018).
8. M. T. McCann, K. H. Jin, M. Unser, Convolutional neural networks for inverse problems in imaging: A review. *IEEE Signal Process. Mag.* **34**, 85–95 (2017).
9. K. H. Jin, M. T. McCann, E. Froustey, Michael. Unser, Deep convolutional neural network for inverse problems in imaging. *IEEE Trans. Image Process.* **26**, 4509–4522 (2017).
10. M. Mardani *et al.*, "Neural proximal gradient descent for compressive imaging" in *Advances in Neural Information Processing Systems*, S. Bengio *et al.*, Eds. (Curran Associates, Inc., 2018), pp. 9596–9606.
11. J. Schlemper, J. Caballero, J. V. Hajnal, A. Price, D. Rueckert, "A deep cascade of convolutional neural networks for MR image reconstruction" in *International Conference on Information Processing in Medical Imaging*, M. Niethammer *et al.*, Eds. (Springer, 2017), pp. 647–658.
12. G. Yang *et al.*, DAGAN: Deep de-aliasing generative adversarial networks for fast compressed sensing MRI reconstruction. *IEEE Trans. Med. Imag.* **37**, 1310–1321 (2018).
13. C. McCollough, Data from "Low Dose CT Grand Challenge dataset." Mayo Clinic. https://www.aapm.org/GrandChallenge/LowDoseCT/. Accessed 19 April 2018.
14. K. Hammernik *et al.*, Learning a variational network for reconstruction of accelerated MRI data. *Magn. Reson. Med.* **79**, 3055–3071 (2018).
15. A. Lucas, M. Iliadis, R. Molina, A. K. Katsaggelos, Using deep neural networks for inverse problems in imaging: Beyond analytical methods. *IEEE Signal Process. Mag.* **35**, 20–36 (2018).
16. M. Elad, *Deep, Deep Trouble - Deep Learning's Impact on Image Processing, Mathematics, and Humanity* (SIAM News, 2017).
17. R. Girshick, J. Donahue, T. Darrell, J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation" in *IEEE Conference on Computer Vision and Pattern Recognition* (Institute of Electrical and Electronic Engineers, 2014), pp. 580–587.
18. A. Krizhevsky, I. Sutskever, G. E. Hinton, "Imagenet classification with deep convolutional neural networks" in *Advances in Neural Information Processing Systems*, F. Pereira, C. J. C. Burges, L. Bottou, K. Q. Weinberger, Eds. (Curran Associates, Inc., 2012), pp. 1097–1105.
19. B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, A. Oliva, "Learning deep features for scene recognition using places database" in *Advances in Neural Information Processing Systems*, Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, K. Q. Weinberger, Eds. (Curran Associates, Inc., 2014), pp. 487–495.
20. K. He, X. Zhang, S. Ren, J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification" in *IEEE International Conference on Computer Vision* (Institute of Electrical and Electronic Engineers, 2015), pp. 1026–1034.
21. C. Kanbak, S.-M. Moosavi-Dezfooli, P. Frossard, "Geometric robustness of deep networks: Analysis and improvement" in *IEEE Conference on Computer Vision and Pattern Recognition* (Institute of Electrical and Electronic Engineers, 2018), pp. 4441–4449.
22. S.-M. Moosavi-Dezfooli, A. Fawzi, P. Frossard, "Deepfool: A simple and accurate method to fool deep neural networks" in *IEEE Conference on Computer Vision and Pattern Recognition* (Institute of Electrical and Electronic Engineers, 2016).
23. S. Moosavi-Dezfooli, A. Fawzi, O. Fawzi, P. Frossard, "Universal adversarial perturbations" in *IEEE Conference on Computer Vision and Pattern Recognition* (Institute of Electrical and Electronic Engineers, 2017), pp. 86–94.

Downloaded at Palestinian Territory, occupied on December 22, 2021

www.manaraa.com

24. C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. J. Goodfellow, R. Fergus, "Intriguing properties of neural networks" in *International Conference on Learning Representations* (2014). https://openreview.net/forum?id=kklr_MTHMRQjG. Accessed 28 April 2020.
25. A. Fawzi, S.-M. Moosavi-Dezfooli, P. Frossard, The robustness of deep networks–A geometric perspective. *IEEE Signal Process. Mag.* **34**, 11 (2017).
26. E. J. Candès, J. Romberg, T. Tao, Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information. *IEEE Trans. Inf. Theor.* **52**, 489–509 (2006).
27. D. L. Donoho, Compressed sensing. *IEEE Trans. Inf. Theor.* **52**, 1289–1306 (2006).
28. L. I. Rudin, S. Osher, E. Fatemi, Nonlinear total variation based noise removal algorithms. *Phys. Nonlinear Phenom.* **60**, 259–268 (1992).
29. B. Adcock, A. C. Hansen, C. Poon, B. Roman, Breaking the coherence barrier: A new theory for compressed sensing. *Forum Math. Sigma* **5**, 1–84 (2017).
30. H. Gupta, K. H. Jin, H. Q. Nguyen, M. T. McCann, M. Unser, CNN-based projected gradient descent for consistent CT image reconstruction. *IEEE Trans. Med. Imag.* **37**, 1440–1453 (2018).
31. J. Adler, O. Öktem, Solving ill-posed inverse problems using iterative deep neural networks. *Inverse Probl.* **33**, 124007 (2017).
32. K. C. Tezcan, C. F. Baumgartner, R. Luechinger, K. P. Pruessmann, E. Konukoglu, MR image reconstruction using deep density priors. *IEEE Trans. Med. Imag.,* **38**, 1633–1642 (2019).
33. S. A. Bigdeli, M. Zwicker, P. Favaro, M. Jin, "Deep mean-shift priors for image restoration" in *Advances in Neural Information Processing Systems*, I. Guyon *et al.*, Eds. (Curran Associates, Inc., 2017), pp. 763–772.
34. J. H. R. Chang, C.-L. Li, B. Poczos, B. V. K. Vijaya Kumar, A. C. Sankaranarayanan, "One network to solve them all–solving linear inverse problems using deep projection models" in *Proceedings of the IEEE International Conference on Computer Vision* (Institute of Electrical and Electronic Engineers, 2017), pp. 5888–5897.

Downloaded at Palestinian Territory, occupied on December 22, 2021

www.manaraa.com