Adversarial Examples are a Manifestation of the Fitting-Generalization Trade-off

O. Deniz, N. Vallez, G. Bueno

VISILAB, Universidad de Castilla-La Mancha. E.T.S.Ingenieros Industriales. Avda. Camilo Jose Cela s/n. 13071 Ciudad Real. Spain
Oscar.Deniz@uclm.es

Abstract. In recent scientific literature, some studies have been published where recognition rates obtained with Deep Learning (DL) surpass those obtained by humans on the same task. In contrast to this, other studies have shown that DL networks have a somewhat strange behavior which is very different from human responses when confronted with the same task. The case of the so-called "adversarial examples" is perhaps the best example in this regard. Despite the biological plausibility of neural networks, the fact that they can produce such implausible misclassifications still points to a fundamental difference between human and machine learning. This paper delves into the possible causes of this intriguing phenomenon. We first contend that, if adversarial examples are pointing to an implausibility it is because our perception of them relies on our capability to recognise the classes of the images. For this reason we focus on what we call *cognitively adversarial examples*, which are those obtained from samples that the classifier can in fact recognise correctly. Additionally, in this paper we argue that the phenomenon of adversarial examples is rooted in the inescapable trade-off that exists in machine learning (including DL) between fitting and generalization. This hypothesis is supported by experiments carried out in which the robustness to adversarial examples is measured with respect to the degree of fitting to the training samples.

Keywords: adversarial examples, deep learning, bioinspired learning

1 Introduction

While advances in Deep Learning (DL) have been significant and have propelled AI into the spotlight, many researchers know that the abilities of this methodology are being at times overestimated [1]. In recent scientific literature, some studies have been published where recognition rates obtained with DL surpass those obtained by humans on the same task. In contrast to this, some studies have shown that DL networks have a somewhat strange behavior which is very different from human responses when confronted with the same task. Perhaps the best example to describe it is the case of the so-called "adversarial examples" [2], see the following figure. Adversarial examples are apparently identical to the original example versions except for a very small change in pixels of the

image. Despite being perceived by humans as completely equal to the originals, DL techniques fail miserably at classifying them.

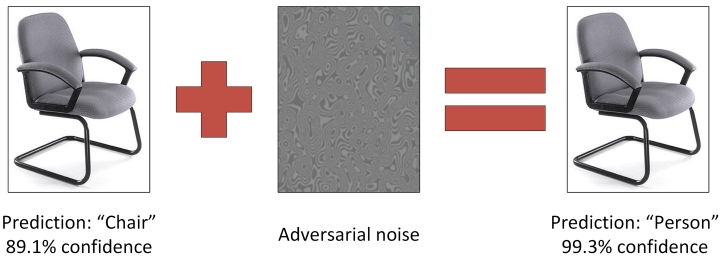

Fig. 1. Adversarial example. "Person" is the so-called *target* class (AE generation or "'attack"' methods aim at making the original class be confused with a target class).

Thus, while apparently having *superhuman* capabilities, DL also seems to have weaknesses that are not coherent with human performance. Not only that, from the structure of DL (essentially an interconnected network of neurons with numerical weights), it is unclear what gives rise to that behavior. The problem is not also in maliciously-selected noise, since some transformations involving rescaling, translation, and rotation produce the same results [3].

The contributions of this paper are as follows. We first contend that, if adversarial examples are pointing to an implausibility it is because our perception of them relies on our capability to recognise the classes involved. For this reason we focus on what we call *cognitively adversarial examples*, which are those obtained as variations of samples that the classifier can recognise correctly. On the other hand, while other underlying reasons have been proposed in the literature for the existence of adversarial examples, in this paper we argue that the phenomenon of adversarial examples is rooted in the inescapable trade-off that exists in machine learning (including DL) between fitting and generalization. This hypothesis is supported by experiments carried out in which the robustness to adversarial examples is measured with respect to the degree of fitting to the training samples. The goal of this paper is not to introduce a novel method, but to advance our knowledge about the phenomenon, its root causes and implications.

This paper is organised as follows. In Section 2 we summarize previous work that focused on the nature of adversarial examples. Section 3 describes the methods and datasets used. Section 4 shows the experimental results. The paper concludes with a broader discussion (Section 5) and the main conclusions.

2 Previous Work

Since adversarial examples (henceforth AEs) first drawn attention of researchers, the two major lines of associated research have been: 1) generating AEs and 2)

defending against AEs. In this work we will not cover either, and the reader is referred to recent surveys. In parallel to those two lines, however, a significant body of work has been carried out to delve into the root causes of AEs and their implications.

In early work, the high nonlinearity of deep neural networks was suspected as a possible reason explaining the existence of adversarial examples [2]. On the other hand, later in [4] it is argued that high-dimensional linearities cause the adversarial pockets in the classification space. This suggests that generalization (as implied by the less complex linear discrimination boundaries) have a detrimental effect that produces AEs. In the same line, in [5] it is stated: "*Unlike the initial belief that adversarial examples are caused by the high non-linearity of neural networks, our results suggest instead that this phenomenon is due to the low flexibility of classifiers*".

In [2] the authors had suggested a preliminary explanation for the phenomenon, arguing that low-probability adversarial pockets are densely distributed in input space. In later work [6] the authors probed the space of adversarial images using noise of varying intensity and distribution. They showed that adversarial images appear in large regions in the pixel space instead.

In [7] the authors review the existing literature on the topic and conclude that up to 8 different explanations have been given for AEs. The prevailing trend, however, seems to focus on the linear/non-linear and in general in the overfitting problems of the classifier. Under two interpretations (the boundary tilting hypothesis [8] and in [9]) the authors argue that the phenomenon of AEs is essentially due to overfitting and can be alleviated through regularisation or smoothing of the classification boundary.

In the recent work [10] the authors perform experiments on a contrived synthetic dataset and conclude that low (test) error classification and AEs are intrinsically associated. They contend that this does not imply that defending against adversarial examples is impossible, only that success in doing so would require improved model generalization. Thus, they argue that the only way to defend against AEs is to massively reduce that error. However, we note that this would be in apparent contradiction with the main finding in that paper (that AEs appear with low classification error). Thus, while generalization may help reduce error in general, without additional considerations it would not necessarily remove AEs.

In [11] the authors point out that an implicit assumption underlying most of the related work is that the same training dataset that enables good *standard* accuracy also suffices to train a robust model. The authors argue that the assumption may be invalid and suggest that, for high-dimensional problems, adversarial robustness can require a significantly larger number of samples. Similar conclusions are drawn in [12], where it is stated that adversarial vulnerability increases with input dimension.

One interesting work that indirectly relates to ours is [13], where a *Deep K-Nearest Neighbors* (DkNN) method is proposed. DkNN inspects the internals of a deep neural network (DNN) at test time to provide 3 outputs: prediction,

confidence and credibility. The DkNN algorithm uses a standard already-trained deep neural network. During inference, for a test input it compares each layer's outputs with those of the nearest neighbors used to train the model (including the network's output layer). The authors show that for adversarial examples the number of different nearest-neighbor labels obtained from the layers is often high. While there are practical issues with the method (specifically: a) it is not clear how one would use the three outputs from DkNN in practice and b) the algoritm requires of a calibration set -a holdout set that does not overlap with the training or test sets-), the connection made with nearest neighbors is intriguing.

In the related literature some findings have transpired that appear to be particularly interesting:

- The phenomenon does not seem to be tied to specific architectures or particular subsets of the training data ([14,15]). In particular, it is possible to transfer adversarial examples from models with known parameters and architecture to other models with unknown parameters and architecture
- The phenomenon is not exclusive of deep learning ([5,16])
- Despite the fact that methods have been proposed to increase robustness to AEs, the phenomenon appears to be essentially unavoidable (see for example [17,18,19])

In summary, despite substantial research, the exact cause of the phenomenon is still poorly understood and remains unsolved. It is not clear, for example, whether the phenomenon is due to overfitting or, on the contrary, to a lack of expressive power. Neither it is clear if the phenomenon is due to the (limited) amount of training samples that are available or due to the large input dimension (or the relationship between these two). On the other hand, even though there appears to be consensus that AEs are unavoidable, it is not clear why this has to be so.

3 Datasets and Methods

In existing research work, AEs are typically generated as variations from the test samples, irrespective of the classifier's decisions on those. This means that some AEs may be generated from test samples that are themselves not correctly classified. In contrast to this, as humans the cognitive dissonance we experience with AEs occurs because we can perceive the classes involved. Thus, for a classifier it should not come as a surprise that AEs are missclassified when they are tiny variations of samples that are missclassified. For these reasons in the following we focus on AEs generated from samples from the test set that are classified correctly (i.e. *cognitively adversarial examples*).

In an adversarial attack, an imperceptible change on an image from an *original* class makes the classifier confuse it with the so-called *target* class. When we think of the actual cause of error in this adversarial example we may ask, which class manifold is not being correctly modelled? that of original class? that of the target class? (see Figure 2). The fact is that the AE actually implies a bad

modelling of both. AEs are consistent with an incorrect modeling of the classes whereby manifolds span excess or defect space.

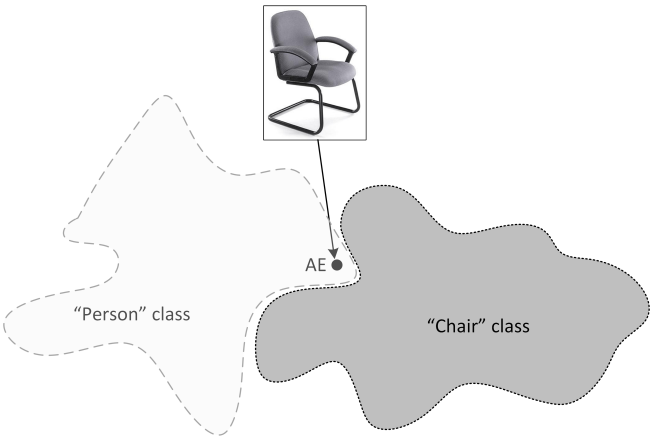

Fig. 2. Adversarial example *vs* trivial example. Is the error shown here due to a bad modelling of the Person class manifold or a bad modelling of the Chair class manifold?

In general, we can always analyze our approximations to the class manifolds in terms of bias and variance. Bias measures how far off the estimated class sample positions are from the true positions, while variance refers to the spread of the estimated positions. Because of the so-called bias-variance dilemma, reducing one quantity automatically increases the other. On the other hand, AEs are practically co-located with (i.e. extremely close to) samples that are correctly handled by the classifier, i.e. they are practically situated on true class positions. This means that reducing model bias (and therefore increasing model variance) should reduce error in those samples (albeit error may increase for other samples), see Figure 3. That is exactly the hypothesis that we address below in the experiments. The bias and variance errors are in general controlled by the classifier's trade-off between fitting and generalization. Our aim is to test if such change in the fitting-generalization trade-off point reflects in the robustness to AEs.

In the experiments we used the MNIST [20] and CIFAR-10 [21] datasets, arguably the two most common datasets used in research on the nature of adversarial examples (used in 39 of the 48 papers reviewed in [7]). MNIST is a dataset of handwritten grayscale 28x28 images representing the digits 0 to 9. Typically, 60000 images are used for training and 10000 for testing. The CIFAR-10 dataset consists of 60000 32x32 colour images in 10 classes[1], with 6000 images per class (50000 training images and 10000 test images).

[1] airplane, automobile, bird, cat, deer, dog, frog, horse, ship and truck

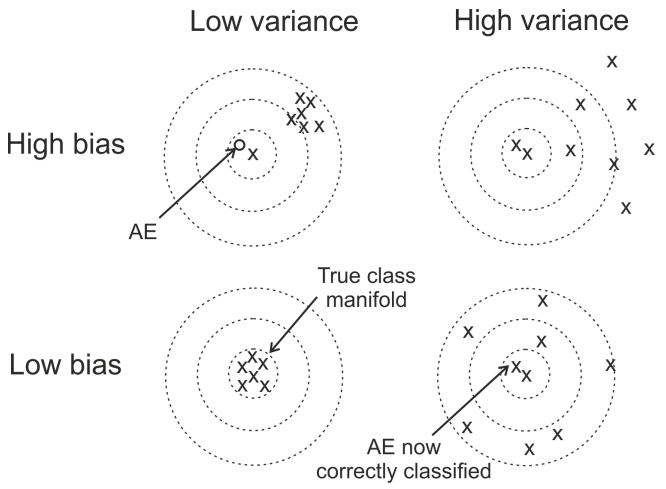

Fig. 3. Bias-variance dilemma.

To validate our hypothesis and show that accuracy in the AE set is linked to the fitting capability, we need a classifier working under various points of the fitting-generalization regime. There are multiple factors, some of which interrelated, that affect the fitting-generalization trade-off point in a deep network: dimensionality of the input space, number of samples available, number of layers, number of epochs, etc. In order to have a direct control of the trade-off point based on a single parameter we decided to use a K-Nearest Neighbor Classifier instead, for K values equal and greater than 1. The K-NN classifier is a natural choice here. It is widely known that large values of K are used to have better generalization, while lower values (down to K=1) may produce overfitting.

Given the dimensionalities involved, an efficient KD-tree-based implementation was used for the K-NN classifier.

The AEs were obtained using two methods: the Fast Gradient Sign Method (a so-called white-box targeted attack, introduced in [4]) and DeepFool [22], targeting all classes and keeping the maximum perturbation threshold at $\epsilon = 0.1$. For both methods we used the LENET-5 network architecture. Figure 4 shows some examples of the AEs generated.

4 Experimental Results

As mentioned above, the AEs were obtained from test samples that were correctly classified by the K-NN, for a $K = Z$ (for a fixed $Z > 1$, Z odd). The attack method (FGSM or DeepFool) was then used to generate a set of AEs from those. This set was filtered to discard samples that were correctly classified by the Z-NN. This leaves a set of samples that was generated from test samples correctly classified by Z-NN but which were not themselves correctly classified

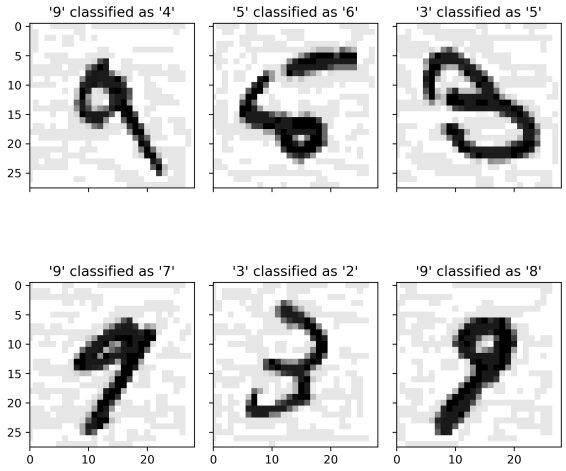

Fig. 4. Sample AEs generated with FGSM for the MNIST dataset.

by Z-NN (thus being adversarial examples). Therefore Z-NN accuracy in this set is 0%.

Then we measured the K-NN classifier accuracy on this set, for values of K smaller than Z, down to K=1. Again, our hypothesis is that accuracy in this set should increase as K gets smaller. In order to discard the possibility of this being a general trend with lower values of K, we also obtained the accuracy for the whole test set and for the subset of test samples in which the classifier gave a wrong decision. The latter is therefore the set of samples for which the classifier gave a wrong decision for K=Z. Note that, by definition, this set and the adversarial set have both 0% accuracy for K=Z.

We repeated the experiment a number of times, each run performing a stratified shuffling of the dataset between the training and test sets (always leaving 60000 samples for training and 10000 samples for test for MNIST, and 50000 samples for training with 10000 samples for test in CIFAR-10). The results are shown in Figures 5 and 6 .

The results show that accuracy in the adversarial set has the highest increase rate as K gets smaller. The accuracy values obtained in the whole test set are always very stable (and very close 100% in the case of MNIST) which makes it difficult to establish a trend in that case. For the other two sets, in order to check if there was a statistically significant difference of trends we applied hypothesis testing in the following way. Let A_i be the accuracy obtained using $K = i$. We calculated the *slope* between accuracies for succesive values of K in the following way: $S_j = A_j/A_{j+2}$, for $j = \{1, 3, 5\}$. Then we used the slopes as the random variables to perform a paired Welch's t-test. In Figure 7 we show the p-values of the test.

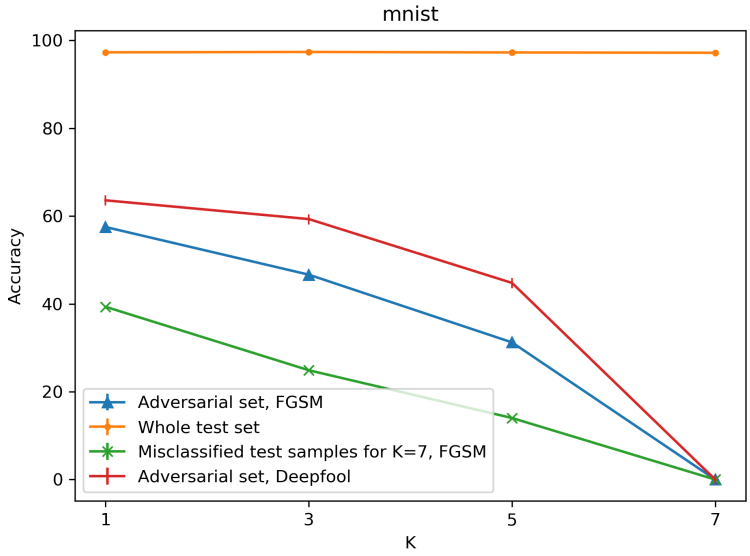

Fig. 5. Left: Accuracy values for the MNIST dataset, using Z=7. Averages of 12 runs.

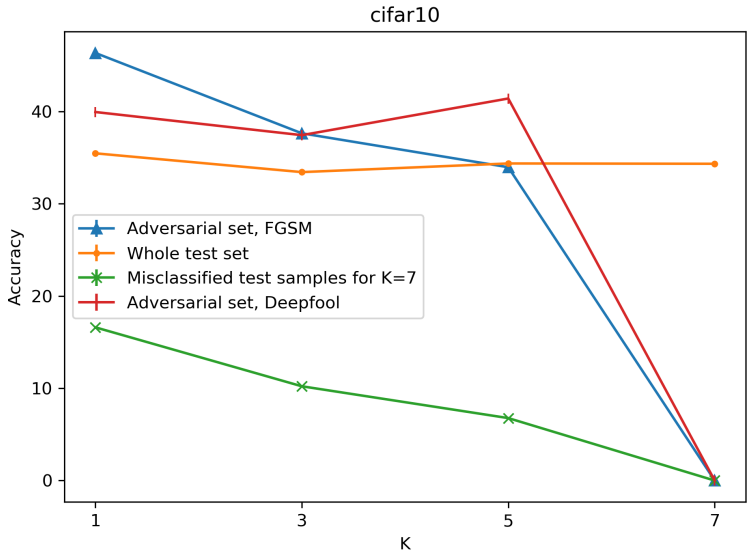

Fig. 6. Left: Accuracy values for the CIFAR-10 dataset, using Z=7. Averages of 8 runs.

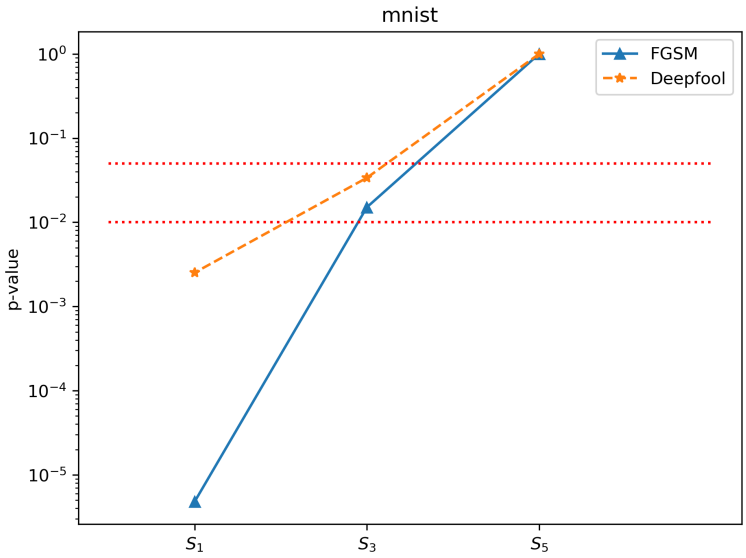

Fig. 7. p-values (represented in logarithmic scale) obtained by the paired Welch's t-test between results in the Adversarial set and those in the misclassified test samples. The dashed horizontal lines represent the 95% and 99% confidence thresholds.

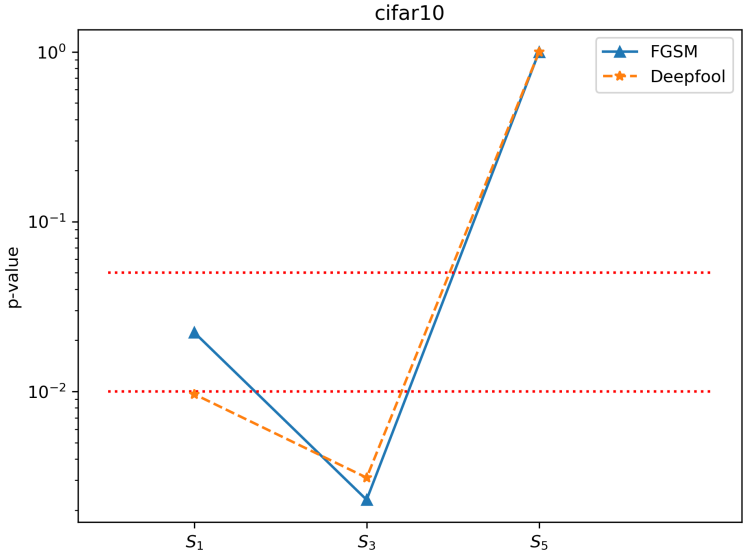

Fig. 8. p-values (represented in logarithmic scale) obtained by the paired Welch's t-test between results in the Adversarial set and those in the Misclassified test samples. The dashed horizontal lines represent the 95% and 99% confidence thresholds.

The values in Figures 7 and 8 show that the trends in the two sets are statistically different. Note also that for S_5 the value is not meaningful since A_7 is 0 in both cases so the slope is actually infinite.

5 Discussion

Our hypothesis that AEs are intrinsic to the bias-variance dilemma has been supported by experiments in which a classifier moving towards the variance extremum showed increasing robustness to the AEs. This increase was, with statistical significance, higher that in the test set, meaning that the increased robustness to AEs did not entail a higher accuracy in general. Overall this is essentially the expected behavior for the well-know bias-variance dilemma: good generalization and robustness to AEs cannot be achieved simultaneously.

In the light of the results, we postulate that the existence of AEs do not reflect a problem of either overfitting or lack of expressive power, as suggested by previous work. Rather, AEs exist in practice because our models lack *both* aspects simultaneously. Rather than being an impossibility statement, this actually calls for methods that have more flexibility to reflect both aspects. While practically all machine learning methods already incorporate some form of trade-off between generalization and fitting, we hypothesize that such trade-offs may be fundamentally different than any such trade-off used by human perceptual learning (since the latter presumably allows for both good generalization and robustness to AEs simultaneously).

Broadly speaking, the goal of machine learning is to find the boundary that best separates two sets of samples. Based on our work, we argue that this goal can be achieved but not necessarily with the best separation between each individual training sample and other samples. We contend that the existence of AEs is a manifestation of the implicit trade-off between fitting and generalization. While the emphasis in machine learning is typically focused on improving generalization, here we argue that the generalization-fitting trade-off is also important. Thus, while the classifier must be flexible enough to avoid overfitting, it must be *flexible* enough to accomodate the *good* effects of overfitting.

6 Conclusions

Based on the observation that adversarial examples indicate that class manifolds are not being modelled accurately, we have argued that the phenomenon is rooted in the inescapable trade-off that exists in machine learning (including DL) between fitting and generalization. This hypothesis is supported by experiments carried out in which the robustness to adversarial examples is measured with respect to the degree of fitting to the training samples, as measured by the K value of a nearest neighbor classifier. As far as the authors know, this is the first time that such reason is proposed as the underlying cause for AEs. The hypothesis should in any case receive additional support through future work in which deep networks are used instead of a K-NN classifier.

While the bias-variance dilemma is posited as the root cause, that should not be considered an impossibility statement. Rather, this would actually call for methods that have more flexibility to reflect both aspects. Current trade-offs between bias and variance or equivalently between fitting and generalization would seem to be themselves biased towards generalization. The cost of that is precisely a lack of robustness to cases such as AEs. If human learning uses (or can be modelled with) any such trade-off, then it should be fundamentally different because presumably it allows for both good generalization and robustness to AEs simultaneously.

Acknowledgments

This work was partially funded by projects TIN2017-82113-C2-2-R by the Spanish Ministry of Economy and Business and SBPLY/17/180501/000543 by the Autonomous Government of Castilla-La Mancha and the ERDF.

References

1. A. L. Yuille, C. Liu, Deep nets: What have they ever done for vision?, CoRR abs/1805.04025 (2018). arXiv:1805.04025.
   URL http://arxiv.org/abs/1805.04025
2. C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. J. Goodfellow, R. Fergus, Intriguing properties of neural networks., CoRR abs/1312.6199 (2013).
   URL http://dblp.uni-trier.de/db/journals/corr/corr1312.html#SzegedyZSBEGF13
3. A. Athalye, L. Engstrom, A. Ilyas, K. Kwok, Synthesizing robust adversarial examples, CoRR abs/1707.07397 (2017). arXiv:1707.07397.
   URL http://arxiv.org/abs/1707.07397
4. I. J. Goodfellow, J. Shlens, C. Szegedy, Explaining and harnessing adversarial examples, arXiv preprint arXiv:1412.6572 (2014).
5. A. Fawzi, O. Fawzi, P. Frossard, Fundamental limits on adversarial robustness, Proceedings of ICML, Workshop on Deep Learning (2015).
   URL http://infoscience.epfl.ch/record/214923
6. P. Tabacof, E. Valle, Exploring the space of adversarial images, 2016 International Joint Conference on Neural Networks (IJCNN) (2016) 426–433 (2016).
7. A. C. Serban, E. Poll, Adversarial examples: A complete characterisation of the phenomenon, CoRR abs/1810.01185 (2018). arXiv:1810.01185.
   URL http://arxiv.org/abs/1810.01185
8. T. Tanay, L. D. Griffin, A boundary tilting persepective on the phenomenon of adversarial examples, CoRR abs/1608.07690 (2016). arXiv:1608.07690.
   URL http://arxiv.org/abs/1608.07690
9. A. Fawzi, S. Moosavi-Dezfooli, P. Frossard, Robustness of classifiers: from adversarial to random noise, CoRR abs/1608.08967 (2016). arXiv:1608.08967.
   URL http://arxiv.org/abs/1608.08967

10. J. Gilmer, L. Metz, F. Faghri, S. S. Schoenholz, M. Raghu, M. Wattenberg, I. J. Goodfellow, Adversarial spheres, CoRR abs/1801.02774 (2018). arXiv:1801.02774.
URL http://arxiv.org/abs/1801.02774
11. L. Schmidt, S. Santurkar, D. Tsipras, K. Talwar, A. Madry, Adversarially robust generalization requires more data, CoRR abs/1804.11285 (2018). arXiv:1804.11285.
URL http://arxiv.org/abs/1804.11285
12. C.-J. Simon-Gabriel, Y. Ollivier, B. Schölkopf, L. Bottou, D. Lopez-Paz, Adversarial vulnerability of neural networks increases with input dimension, CoRR abs/1802.01421 (2018).
13. N. Papernot, P. D. McDaniel, Deep k-nearest neighbors: Towards confident, interpretable and robust deep learning, CoRR abs/1803.04765 (2018). arXiv:1803.04765.
URL http://arxiv.org/abs/1803.04765
14. N. Papernot, P. McDaniel, I. Goodfellow, Transferability in machine learning: from phenomena to black-box attacks using adversarial samples, arXiv preprint arXiv:1605.07277 (2016).
15. Z. B. Charles, H. Rosenberg, D. S. Papailiopoulos, A geometric perspective on the transferability of adversarial directions, CoRR abs/1811.03531 (2018).
16. Y. Wang, S. Jha, K. Chaudhuri, Analyzing the robustness of nearest neighbors to adversarial examples, in: ICML, 2018 (2018).
17. L. Bortolussi, G. Sanguinetti, Intrinsic geometric vulnerability of high-dimensional artificial intelligence, CoRR abs/1811.03571 (2018). arXiv:1811.03571.
URL http://arxiv.org/abs/1811.03571
18. D. Tsipras, S. Santurkar, L. Engstrom, A. Turner, A. Madry, Robustness may be at odds with accuracy, in: International Conference on Learning Representations, 2019 (2019).
URL https://openreview.net/forum?id=SyxAb30cY7
19. A. Shamir, I. Safran, E. Ronen, O. Dunkelman, A simple explanation for the existence of adversarial examples with small hamming distance, CoRR abs/1901.10861 (2019). arXiv:1901.10861.
URL http://arxiv.org/abs/1901.10861
20. Y. LeCun, C. Cortes, MNIST handwritten digit database (2010) [cited 2016-01-14 14:24:11].
URL http://yann.lecun.com/exdb/mnist/
21. A. Krizhevsky, V. Nair, G. Hinton, CIFAR-10 (Canadian Institute for Advanced Research).
URL http://www.cs.toronto.edu/~kriz/cifar.html
22. S. Moosavi-Dezfooli, A. Fawzi, P. Frossard, Deepfool: a simple and accurate method to fool deep neural networks, CoRR abs/1511.04599 (2015). arXiv:1511.04599.
URL http://arxiv.org/abs/1511.04599