ESANN 2018 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence
and Machine Learning. Bruges (Belgium), 25-27 April 2018, i6doc.com publ., ISBN 978-287587047-6.
Available from http://www.i6doc.com/en/.

# Properties of $\text{adv}^{-1}$ – Adversarials of Adversarials

Nils Worzyk and Oliver Kramer [*]

University of Oldenburg - Dept. of Computing Science
Oldenburg - Germany

**Abstract**. Neural networks are very successful in the domain of image processing, but they are still vulnerable against adversarial images – carefully crafted images to fool the neural network during image classification. There are already some attacks to create those adversarial images, therefore the transition from original images to adversarial images is well understood.

In this paper we apply adversarial attacks on adversarial images. These new images are called $\text{adv}^{-1}$. The goal is to investigate the transition from adversarial images to $\text{adv}^{-1}$ images. This knowledge can be used to 1.) identify adversarial images and 2.) to find the original class of adversarial images.

## 1 Introduction

Neural networks have been shown to perform close to or even better than human beings on image classification tasks. Szegedy et al. [1] for example report a 3.5% top-5 error on the ImageNet dataset. But neural networks are still vulnerable to *adversarial images*, carefully crafted images to fool the classification of an image. This problem was first brought up by Szegedy et al. [2]. For example, Evtimov et al. [3] applied an adversarial attacks to street sign classifiers – they fooled the neural network to classify a Stop sign as a Speed Limit sign – and thereby showed the necessity of robust classifiers in safety-critical systems.

On the opposite side, there has also been research on defending systems against adversarial images. One approach by Goodfellow et al. [4] is called *adversarial training*. They inject previously created adversarial images into the training set, but with the correct label. Kurakin et al. [5] showed, that this technique is not robust against iterative adversarial examples. Other defences like binary classification proposed by Gong et al. [6] or defensive distillation, introduced by Papernot et al. [7], can be fooled by more sophisticated attacks as Carlini and Wagner [8] showed.

Besides several attacks and defences, only few research has been done on the properties of adversarial images. For example, Tabacof and Valle [9] explored the space of *adversarial islands* within the pixel space. To investigate the size of those adversarial islands, they first created adversarial images. Then they perturbed the original images and the adversarial images with random noise and measured how many images stay in the same class or in the case of adversarial

[*] This research is funded by the German Research Foundation through the Research Training Group DFG-GRK 1765: “System Correctness under Adverse Conditions”.

images, how many switch back to the correct original class. They reported that adversarial images appear in larger islands.

In this paper, we will investigate the behavior of adversarial images, if we apply adversarial attacks, instead of random noise. Those new images, the adversarials of adversarials, are called $adv^{-1}$ to distinguish between both types of adversarial images. Since the the transformation of original images to adversarial images is well understood, we investigate, if the transformation of adversarial images to $adv^{-1}$ images is different. That knowledge can be used to answer two questions:

1. Can one identify adversarial images by differences between the transition of original to adversarial images and adversarial to $\text{adv}^{-1}$ images?

2. Do $\text{adv}^{-1}$ images "return" to their original true class in pixel space, or do they "move" to another adversarial class?

The remainder of this paper is organized as follows. In Section 2 we will briefly describe the basic idea of creating adversarial images and introduce the attacks used in this paper. In Section 3 the experiment to answer the two questions above is described, followed by the results and discussion presented in Section 4. In Section 5 we will conclude the paper.

## 2 Creating adversarial images

In general, adversarial attacks try to find the minimal perturbation $\delta$ to an original image $x$, according to some distance metric $\mathcal{D}$, such that the perturbed image $x'$ is classified differently than the original image, and the perturbed image is still in the value range of the trained network, e.g. $[0, 1]$. Formally written:

$$\begin{aligned} \text{minimize} \quad & \mathcal{D}\left(x, x+\delta\right) \\ \text{such that} \quad & \mathcal{C}\left(x+\delta\right) = t \\ & x+\delta \in \left[0,1\right]^n , \end{aligned} \tag{1}$$

where $\mathcal{C}$ denotes the classifier, $t$ is the target class and $t \neq \mathcal{C}\left(x\right)$.

In this paper we use five attacks, provided by the framework Cleverhans [10][1], which solve Equation 1 in different ways, namely:

1. Fast Gradient Sign Method (FGSM), proposed by Goodfellow et al. [4]
2. Basic Iterative Method (BIM), proposed by Kurakin et al. [11]
3. Jacobian-based Saliency Map (JSMA), proposed by Papernot et al. [12]
4. Virtual Adversarial Method (VATM), proposed by Miyato et al. [13]
5. Carlini and Wagner (CW), proposed by Carlini and Wagner [8]

[1] Implementation on `https://github.com/tensorflow/cleverhans`

## 3 Experiments

For our experiments, we use the MNIST dataset [14], the CIFAR-10 dataset [15] and the first 1000 images of the ImageNet dataset [16][2]. The experiment itself is conducted in two stages.

*First stage* In the first stage we create 1000 adversarial images for each of the five attacks named in Section 2. Only FGSM was adapted to find the smallest perturbation, which successfully creates an adversarial image. To achieve this, we iteratively increase the parameter $\epsilon$, which controls the amount of introduced perturbation.

As classifier for MNIST a simple convolution neural network is used, with a classification accuracy of 98.82%. For CIFAR-10, a wide residual network [17][3] with 91.62% accuracy is used. For both datasets, the classifiers operate on the input value range $[0, 1]$. Therefore we standardized the adversarial images to be within that range. For JSMA as attack on CIFAR-10, we set the upper bound of allowed perturbed pixel $\gamma$ to 0.1. The standard parameter setting of $\gamma = \infty$ leads to infinite calculations. All other parameters are set to the standard setting provided by the Cleverhans framework [10].

For the ImageNet dataset, a pretrained Inception V3 network [1][4] with 78% classification accuracy is used. The adversarial images are standardized to the range $[-1, 1]$. For ImageNet, JSMA as attack could not be used at all, due to computational resources. This problem has also been reported by Carlini and Wagner [8].

For all three datasets, an image is accepted as an adversarial, if the classification of the original image is correct and the classification of the adversarial image is different from the correct class. In that case the $\mathrm{L}_2$ difference between the original and the adversarial image is recorded.

*Second stage* In the second stage, we use *each* attack with the same parameter settings and classification networks as in stage one on *all* the previously created adversarial images to create $\mathrm{adv}^{-1}$ images. For each $\mathrm{adv}^{-1}$ image, the attack, which created the adversarial image in the first stage, and the $\mathrm{L}_2$ difference between the adversarial and $\mathrm{adv}^{-1}$ image is recorded – as well as the number of successfully created $\mathrm{adv}^{-1}$ images per attack and the number of $\mathrm{adv}^{-1}$ images, whose classification reverted to the original true class.

## 4 Results

*Identification of adversarials* Exemplary in Figure 1, the distributions of the $\mathrm{L}_2$ differences for the MNIST dataset are shown. In each sub-plot the left most graph, named orig on the x-axis, shows the distribution of the $\mathrm{L}_2$ differences, if

[2] We use only the first 1000 shuffled images due to efficiency reasons.

[3] `https://github.com/titu1994/Wide-Residual-Networks`

[4] `https://github.com/tensorflow/models/tree/master/research/slim`

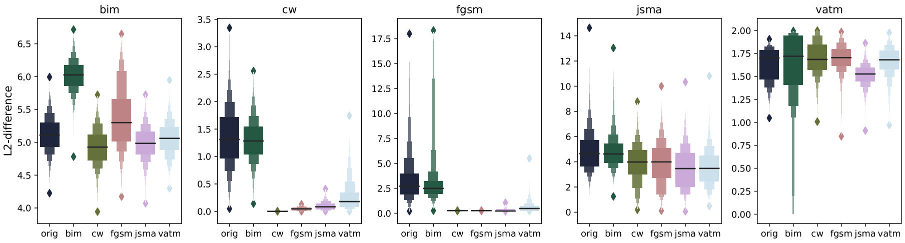

Fig. 1: Distribution of the $L_2$ differences for the MNIST dataset.

we follow the transition *original* $\rightarrow$ *adversarial* (*trans*) for the attack indicated by the title. The other five graphs show the distribution of the $L_2$ differences, if we follow the transition *adversarial* $\rightarrow$ $adv^{-1}$ ($trans^{-1}$), whereby the x-label indicates the attack to produce the adversarial image, and the attack stated as title produces the $\text{adv}^{-1}$ image.

At first sight, especially for CW and FGSM, the mean $L_2$ differences of $trans^{-1}$ are lower for 4 out of 5 attacks in the first stage, than the mean $L_2$ differences of *trans*. To further elaborate this observation we calculate the 99% confidence interval of the mean values for all graphs shown in Figure 1. Then we compare the lower bound of the confidence interval for *trans* with the upper bound of the confidence intervals for $trans^{-1}$. This comparison shows, that the lower bound of *trans* is higher, than the upper bound of $trans^{-1}$, if we use CW, FGSM, or JSMA as second attack, for all attacks, except BIM, in the first stage. These observations indicate, that we can identify a threshold to distinguish between *trans* and $trans^{-1}$ based on the $L_2$ differences. Furthermore this approach of identifying adversarial images seem to be transferable between different attacks. This transferability is an important aspect for a robust defense against adversarial attacks.

For the CIFAR-10 and ImageNet dataset the results are similar. We found, that on CIFAR-10, the lower bound of *trans* is higher than the upper bound of $trans^{-1}$, if we use FGSM in the second stage, for all five attacks used in the first stage. For CW and JSMA as second attacks, this difference can be found for 4 out of 5 attacks. On ImageNet we found this behavior for CW and FGSM as second attacks on 3 out of 4 attacks used in the first stage. Again BIM is the attack, which is difficult to separate.

*Reverting adversarials* In Table 1, the percentage of images, which reverted into their original true class by applying another attack, on MNIST are listed. The columns indicate the attack to create the adversarial images, the rows indicate the attack to produce the $\text{adv}^{-1}$ image. Under certain conditions we can revert up to 99.6% of the adversarial images to their original class, if we simply apply another attack. The best average results is detected, if we use FGSM in the second stage. There we can revert 72.44% of the adversarial images to their initial correct class.

| | BIM | CW | FGSM | JSMA | VATM | mean |
|---|---|---|---|---|---|---|
| BIM | 20.5% | 99.5% | 91.9% | 79.3% | 63.7% | 71% |
| CW | 19.5% | 99.6% | 96.7% | 70.8% | 65.5% | 70.4% |
| FGSM | 22.1% | 99.6% | 97.4% | 74.2% | 68.9% | **72.4%** |
| JSMA | 10.6% | 10.9% | 9.9% | 12% | 10.6% | 10.8% |
| VATM | 5% | 54.3% | 48.1% | 44.6% | 36.8% | 37.8% |
| mean | 15.5% | 72.8% | 68.8% | 56.2% | 49.1% | |

Table 1: Percentage of reverted images on MNIST.

For CIFAR-10 and ImageNet the best average percentage of reverted images is 50.5%, resp. 39.82% if we use FGSM as the second attack. The best single reverting rate is achieved, if we use CW as first and second attack, where 97.7%, resp. 96.6% of the images reverted to their original true class on CIFAR-10, resp. ImageNet.

From another perspective, adversarial images created by BIM have the lowest average reverting rate of 15.54% on MNIST, 0.26% on CIFAR-10 and 0.91% on ImageNet. Whereas adversarial images created by CW have the highest reverting rate with 72.8% on MNIST, 66.5% on CIFAR-10 and 73.5% on ImageNet. This is specially interesting, since Carlini and Wagner [8] reported, that their attack introduces the least $L_2$ perturbation, compared to the other attacks, but yet many of the attacked images can be reverted to their original true class by applying another attack.

## 5 Conclusion

This paper introduced a new perspective on adversarial images, by applying adversarial attacks to adversarial images. Those new images are called $\text{adv}^{-1}$ to distinguish between the different forms of adversarial images. Specifically the transition of original images to adversarial images ($trans$) by an attack, and the transition of adversarial images to $\text{adv}^{-1}$ images ($trans^{-1}$) by another attack, is investigated.

The first question is, if we can distinguish between the two transitions, based on the $L_2$ differences of the corresponding images. In our experiments, we calculated the mean $L_2$ differences, as well as their upper and lower bounds of the 99% confidence intervals, between the images of the two transitions. We found that the mean differences of $trans^{-1}$ are lower than the mean differences of $trans$, for most of the combinations of attacks used in this paper. Furthermore we found, that if we use FGSM as attack to produce the adversarial images, the lower bound of the mean value is higher, than the upper bound of the mean value, if we use FGSM to produce the $\text{adv}^{-1}$ images, for all attacks except BIM (on MNIST and ImageNet) or even all five attacks (on CIFAR-10) used to create the adversarial images. This indicates, that it should be possible to find a threshold to distinguish between original and adversarial images, based on the introduced

ESANN 2018 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2018, i6doc.com publ., ISBN 978-287587047-6. Available from http://www.i6doc.com/en/.

$L_2$ difference.

The second question is, if the adversarial images return to their original correct class if we apply another adversarial attack. Our experiments show, that depending on the attacks used to create the adversarial images, there is a high chance for this behaviour. On MNIST 99.6% of the adversarial images, created by the attack of Carlini and Wagner [8] (CW), revert to their original correct class, if we apply FGSM to create the $\text{adv}^{-1}$ image. For CIFAR-10 97.5% behave that way, and for ImageNet 90%. Moreover, averaged over all applied attacks to create adversarial images, 72.4% revert to their original true class, if we apply FGSM as second attack. For CIFAR-10, on average, 50.5% of the adversarial images revert, if attacked by FGSM, and on ImageNet 39.8% revert on average.

## References

[1] Christian Szegedy, et al. Rethinking the inception architecture for computer vision. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2818–2826, 2016.

[2] Christian Szegedy, et al. Intriguing properties of neural networks. *arXiv:1312.6199*, 2013.

[3] Ivan Evtimov, et al. Robust physical-world attacks on machine learning models. *arXiv:1707.08945*, 2017.

[4] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv:1412.6572*, 2014.

[5] Alexey Kurakin, Ian Goodfellow, and Samy Bengio. Adversarial machine learning at scale. *arXiv:1611.01236*, 2016.

[6] Zhitao Gong, Wenlu Wang, and Wei-Shinn Ku. Adversarial and clean data are not twins. *arXiv:1704.04960*, 2017.

[7] Nicolas Papernot, et al. Distillation as a defense to adversarial perturbations against deep neural networks. In *IEEE Symposium on Security and Privacy (SP)*, pages 582–597. IEEE, 2016.

[8] Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *IEEE Symposium on Security and Privacy (SP)*, pages 39–57. IEEE, 2017.

[9] Pedro Tabacof and Eduardo Valle. Exploring the space of adversarial images. In *International Joint Conference on Neural Networks (IJCNN)*, pages 426–433. IEEE, 2016.

[10] Nicolas Papernot, et al. cleverhans v1.0.0: an adversarial machine learning library. *arXiv:1610.00768*, 2016.

[11] Alexey Kurakin, Ian Goodfellow, and Samy Bengio. Adversarial examples in the physical world. *arXiv:1607.02533*, 2016.

[12] Nicolas Papernot, et al. The limitations of deep learning in adversarial settings. In *IEEE European Symposium on Security and Privacy (EuroS&P)*, pages 372–387. IEEE, 2016.

[13] Takeru Miyato, et al. Distributional smoothing with virtual adversarial training. *arXiv:1507.00677*, 2015.

[14] Yann LeCun. The mnist database of handwritten digits. *http://yann. lecun. com/exdb/mnist/*, 1998.

[15] Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. 2009.

[16] Jia Deng, et al. Imagenet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255. IEEE, 2009.

[17] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. *arXiv:1605.07146*, 2016.