Adversarial examples – from discovery to generalization on humans

Weronika Ormaniec[1]

[1]Faculty of Computer Science, Electronics and Telecommunications. AGH University of Science and Technology, Krakow, Poland

In this review, I am going to focus on three papers on the subject of adversarial examples. Those inputs to the ML model are really similar in human perception to the ones that belong to class A but instead, according to the model, they belong to class B. The first two papers "Intriguing Properties of Neural Networks" by Szegedy et al. (1) and "Explaining and Harnessing Adversarial Examples" by Goodfellow et al. (2) are two of the fundamental papers discussing the problem of adversarial examples for computer vision. On the other hand, "Adversarial Examples that Fool both Computer Vision and Time-Limited Humans" by Elsayed et al. (3) is a recent work trying to generalize adversarial examples over the most complicated neural net we know of– human brain.

People have been designing attacks against machine learning models since at least 2004. Firstly, that research was mostly focused on fooling spam detectors. In 2013 Battista Biggio found that you can fool neural networks and around the same time Christian Szegedy discovered that just by an optimization algorithm you can launch an attack against deep neural networks. The paper "Intriguing properties of neural networks" by Szegedy et al. followed up this discovery.

Firstly, the paper argues that it is the entire space of activations rather than particular neurons that contains the majority of semantic information. Specific unit inspection can only confirm some of our intuition about what is really happening, but it will not be enough to understand the whole process of mapping inputs to outputs.

Secondly, it notes the existence of adversarial examples for computer vision. The paper states that the output layer unit is a highly non-linear function of the input. That makes the network encode a non-local generalization over the input space. Therefore, it is possible to classify images that were not originally in the training set. Those images can be far from the data the network was trained on (in pixel space) but they share the label and statistical structure with some of the training inputs. It is proven that a small random change of the input does not influence the correct classification. However, it is also shown that well-prepared small perturbations can cause the input to be classified incorrectly. Moreover, the paper describes how to traverse the data space to find adversarial examples, since it is hard to find them just by picking a random perturbation. Researches define classifier f that maps high-dimensional input space into k classes. They want to find r such as $f(x+r) = l$ where l is an arbitrarily chosen class and $x + r \in [0,1]^m$. Afterwards, they minimize r in L^2 norm.

For each analyzed network (a fully connected network with a few hidden layers and softmax classifier, a classifier trained on the top of an autoencoder, both on MNIST dataset; "AlexNet" on ImageNet dataset; "QuocNet" trained on YouTube images) they were able to generate adversarial examples, impossible to distinguish from normal images, but misclassified by the networks they were trained on. Moreover, while tested on a different model than the one they were trained on, adversarial examples still influenced the classification. Finally, adversarial examples were also hard for models trained on different datasets, but their effectiveness decreased considerably.

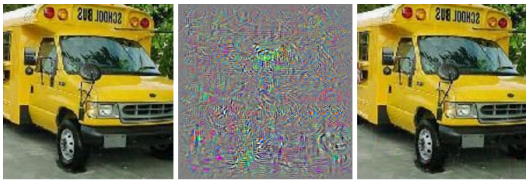

Fig. 1. Adversarial example from (1). Left–correctly predicted sample, center–difference between correct image, and image predicted incorrectly magnified by 10x, right–adversarial example. Right image was predicted to be an "ostrich, Struthio camelus".

In this paper, researchers also suggest that it is possible to train the model with adversarial examples and acquire some kind of immunity to them. This hypothesis was successfully tested only on the MNIST dataset with a two-layer non-convolutional neural network

The major advantage of this paper is the subject it undertakes. Firstly, it may just seem funny that some model classifies a bus as an ostrich, but on second thought one starts to appreciate the danger adversarial examples imply, especially in the future era of autonomous cars, when it will be crucial for the model to recognize a STOP sign as a STOP sign. There are lots of things this work does not fully cover, for example, why adversarial examples generalize or are there any other methods to generate them or make the model immune to them. In 2015, these questions required more research. That is why "Explaining and Harnessing Adversarial Examples" by Goodfellow et al. was published.

Before this paper, some researchers suggested that the extreme non-linearity of the models causes adversarial examples. This work shows that high-dimensional linear transformations are enough to cause adversarial examples. Firstly, in the case of the dot product of weight vector and adversarial examples, researchers show that since L^∞ norm does not have to grow with the increasing number of dimensions, but the change of activation can at the same time grow linearly, really small changes in values of input can make a huge difference in output. This explains why high-dimensional linear models can have adversarial examples.

Secondly, they argue that our non-linear models have activation functions crafted to behave in a linear way because then they are easier to optimize. This suggests that perturbations of linear models should also influence neural nets. They propose a method called "fast gradient sign method". In order to create noise, they compute the sign of the gradient of the cost function of model weights, input the image and desired output and multiply it by chosen epsilon. As shown in experiments,

adding it to the image causes the shallow softmax classifier, maxout network and convolutional maxout network to misclassify the input. The fact that it works also confirms that the existence of adversarial examples is caused by the linear nature of models.

In this paper, researchers also presented a way to counteract adversarial examples. They defined a new cost function with the fast gradient sign method as a regularizer and they used it (combined with dropout) to train their maxout network. It lowered the error rate from 0.94% to 0.84% for the entire test set. Then researchers made some additional changes in the model and after adversarial training, the error rate fell from 89.4% to 17.9% on the same set of adversarial images. However, the confidence of those predictions was still quite high.

Before this paper, It was conjectured that adversarial examples finely tile space like the rational numbers among the reals. However, if it was true, how would it be possible for various non-linear models to classify out-of-distribution point in the same way? Here comes the hypothesis. Adversarial examples occur in contiguous regions of 1-D subspace defined by fast gradient sign method which was tested by shifting epsilon value during computations. And why different classifiers assign the approximately same class to one adversarial example? According to the paper, because they all learn approximately the same weights and are able to generalize. The stability of weights results in stability of predictions-both correct and incorrect.

In the first two reviewed works authors focused on the existence of adversarial examples, ways to generate them and use to create better networks. They also tried to answer the question why do those examples generalize. Those two papers do not exhaust the topic of adversarial images and since their publication, new discoveries in this field have been made, including new techniques of attacks with adversarial examples (4). Therefore, further research is still needed. In this review, I would like to consider one more topic the previous papers have not discussed. The last paper I'm going to mention also focuses on generalizing adversarial examples. This time authors try to generalize them to humans.

Since "Adversarial Examples that Fool both Computer Vision and Time-Limited Humans" by Elsayed et al. was published 3 years after the one I mentioned earlier, there has been some progress in the field of adversarial examples. New discoveries of black box adversarial example construction techniques that create adversarial examples for specific models without access to their parameters and architecture have been made.

Although it has been recently shown that there exist some similarities between the primate visual system and deep CNNs, there are also major differences between machine and human vision. First of all, resolution of the image perceived by a human is not constant and if the change in the image happens in its periphery it may become undetectable to a human eye. Moreover, the human eye is sensitive to spatial features and the brain does not treat what eyes see as one image but it constantly explores the scene. That is why so far, it has been believed that humans are immune to adversarial examples but it has not been properly investigated yet.

This paper underlines the fact that adversarial examples are not designed to be different from human judgment but to cause a mistake during classification by the network. Researchers assign each adversarial example the same label it had at the beginning of the perturbing process and assume it did not influence the true class.

Researchers used 6 images from ImageNet and divided them into 3 categories: pets(cats and dogs), hazard (snakes and spiders), vegetables(broccoli and cabbage). Then for each class pair (A and B) they generated adversarial examples that generalize across different models and make them classify B as A and vice versa.

During the experiment, after looking at a fixation point on the screen, people were shown an image (for $63-71$ ms) and they were supposed to classify it as one of the classes from the defined pairs. The images could have been normal (unperturbed images from ImageNet), adversarial (with specific perturbations added), flipped (adversarial perturbation was vertically flipped before adding to the initial image) and false (presenting objects from neither of two classes person could have chosen) but with adversarial perturbation.

Firstly, adversarial examples were tested on two models they were not generated on Inception V3 and ResNet V2 50. Both models performed well on normal images ($> 75\%$ accuracy) but the attacks with adversarial or false examples were successful between 57% and 89% of the time. Adding flipped perturbations have not changed the result much from the one obtained on the clean dataset which means this flipped dataset was good for validation.

Fake images were perturbed into one of the classes a person could have chosen with equal probability and it occurred that people tended to choose the class the image was perturbed into. Researchers assumed that if perturbation does not influence the decision, people should choose the perturbed class 50% of the time. It was shown that after perturbations this rate was actually between 51.55% and 54.5%.

While testing on images with classes that were actually one of the possible to choose from the effect is also noticeable. The average accuracy of humans dropped almost 10% on adversarial examples, while flipped examples changed it only a few percents.

The paper does not show how adversarial examples specifically work but spots a few patterns. It suggests that top-down and lateral connections used normally by humans while they are not time limited may be the answer to the problem of adversarial examples we are dealing with now. However, it does not eliminate the possibility that adversarial examples for our kind of neural nets also exist.

First two papers focus mostly on the technical part of the existence of adversarial examples. They motivate to think about the security of our models and the safety of using them on a regular basis. The third work however, states more disruptive question. Would it be possible to manipulate humans using adversarial examples crafted for their neural nets? That possibility exists and there is more research needed in both fields adversarial examples for machine learning and adversarial examples for humans.

References

1. Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, Rob Fergus. Intriguing Properties of Neural Networks. *International Conference on Learning Representations (2014)*.
2. Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and Harnessing Adversarial Examples. *International Conference on Learning Representations (2015)*.
3. Gamaleldin Fathy Elsayed, Shreya Shankar, Brian Cheung, Nicolas Papernot, Alex Kurakin, Ian Goodfellow, Jascha Sohl-dickstein. Adversarial Examples that Fool both Computer Vision and Time-Limited Humans. *NeurIPS (2018)*.
4. Anish Athalye, Logan Engstrom, Andrew Ilyas, Kevin Kwok. Synthesizing robust adversarial examples. *arXiv:1707.07397 (2017)*