On the Claim for the Existence of "Adversarial Examples" in Deep Learning Neural Networks

Costas Neocleous[1] and Christos N. Schizas[2]
[1]*Department of Mechanical Engineering and Materials Science and Engineering, Cyprus University of Technology, 30 Archbishop Kyprianou Str., 3036, Limassol, Cyprus*
[2]*Department of Computer Science, University of Cyprus, 1 University Avenue, 2109, Nicosia, Cyprus*

Keywords: Deep Neural Networks, Adversarial Examples, Feature Distribution in Neural Networks.

Abstract: A recent article in which it is claimed that adversarial examples exist in deep artificial neural networks (ANN) is critically examined. The newly discovered properties of ANNs are critically evaluated. Specifically, we point that adversarial examples can be serious problems in critical applications of pattern recognition. Also, they may stall the further development of artificial neural networks. We challenge the absolute existence of these examples, as this has not been universally proven yet. We also suggest that ANN structures, that correctly recognize adversarial examples, can be developed.

1 INTRODUCTION

In a recent paper that has been presented by a team of researchers from Google, Facebook, New York University and the University of Montreal (Szegedy et al., 2014) at the 2nd International Conference on Learning Representations in April 2014, two debatable and counter-intuitive properties of deep neural networks were claimed. In the authors own words (hereby quoted in italics), these are (bold letters indicate the most important points):

QUOTES

CLAIM #1: On the distribution of semantic information

> *…Generally, it seems that* ***it is the entire space of activations, rather than the individual units, that contains the bulk of the semantic information****…*

CLAIM #2: On the existence of adversarial examples

> *…we find that* ***applying an imperceptible non-random perturbation to a test image, it is possible to arbitrarily change the network's prediction****. These perturbations are found by optimizing the input to maximize the prediction error. We term the so perturbed examples "adversarial examples"…*

and elswhere,

> *… we found that* ***adversarial examples are relatively robust, and are shared by neural networks with varied number of layers, activations or trained on different subsets of the training data****. That is, if we use one neural to generate a set of adversarial examples, we find that these examples are still statistically hard for another neural network even when it was trained with different hyperparemeters or, most surprisingly, when it was trained on a different set of examples.*
>
> *…Our main result is that for deep neural networks, the smoothness assumption that underlies many kernel methods does not hold. Specifically, we show that by using a simple optimization procedure, we are able to find adversarial examples, which are obtained by imperceptibly small perturbations to a correctly classified input image, so that it is no longer classified correctly. This can never occur with smooth classifiers by their definition…*
>
> *…The above observations suggest that* ***adversarial examples are somewhat universal and not just the results of overfitting to a particular model or to the specific selection of the training set****.*

Neocleous C. and N. Schizas C..
On the Claim for the Existence of "Adversarial Examples" in Deep Learning Neural Networks.
DOI: 10.5220/0005152503060309
In *Proceedings of the International Conference on Neural Computation Theory and Applications* (NCTA-2014), pages 306-309
ISBN: 978-989-758-054-3

In this position paper we attempt to evaluate and critically appraise the two claims made by Szegedy et al., henceforth referred to by the abbreviation SZSBEGF2014.

The SZSBEGF2014 researchers used a large, but not exhaustive, range of different artificial neural network (ANN) structures to study the properties they claim that exist. However, all of them were of feedforward (FF) topology. The structures studied ranged from simple to complicated ones, such as the deep NN paradigms (11-layered MLFF network).

They have also done numerous tests on well-known machine learning databases. Namely, on the Mixed National Institute of Standards and Technology database (MNIST) and on the ImageNet database. They also did simulations on image samples from YouTube.

Considering their findings, we will present some supportive and some counter-supportive arguments to either claim.

2 CRITIQUE OF THE CLAIM FOR THE DISTRIBUTION OF SEMANTIC INFORMATION

The SZSBEGF2014 authors claim that, following numerous studies on a diverse set of artificial neural network structures, they had found that *"…it seems that it is the entire space of activations, rather than the individual units, that contains the bulk of the semantic information…"*.

That is, the various semantic factors are encoded in a distributive manner in a multitude of artificial neuronal units rather than in specific single artificial units (neurons).

This is in contrast to the prevailing theory that suggests that the activation levels of individual hidden layer neurons correspond to expressions for a meaningful feature.

In natural/biological neural networks, it is well established that different parts of the brain process different afferent signals. There is a hierarchical structure, and different modules specialize on processing specific tasks. Thus, there exists specificity of function that seems to happen in modular manner, where groups of neurons, operate in concerted manners. They process information in a distributed manner. So, within modules, the information is distributed and we cannot attribute specific semantics to specific single neurons.

Based on the previous comments, we propose that the SZSBEGF2014 claim concerning the semantic distribution, is correct within the context of activations of single artificial neurons belonging to a neural subsystem (module). Different modules however, can specialize in processing certain features that are associated to semantic attributes.

It is however difficult, and possibly impossible to precisely identify the functionality and boundaries of such modules, and hence the specific attributes they process. It could be that the distribution properties that we presume that exist in groups of neurons in a module, also exist for groups of modules in a larger supersystem.

It is appropriate to point that the SZSBEGF2014 researchers have done their simulations only on feed-forward neural structures. But the brain is highly dynamic, having a vast number of local and remote feedbacks.

3 CRITIQUE OF THE CLAIM FOR THE EXISTENCE OF ADVERSARIAL EXAMPLES

The SZSBEGF2014 researchers claim that they found blind spots in the process of generalization of feedforward ANNs. Indeed, they say that by "*…applying an imperceptible non-random perturbation to a test image, it was possible to arbitrarily change the network's prediction…"*. If this is universally true, then we have here a case where the network generalization is deficient.

This hypothesis and the associated claim has been tested by the SZSBEGF2014 researchers through the systematic generation of specific images (cases) that they called "adversarial samples". That is, even though they were very similar to some other samples (having imperceptible differences as seen by a human eye and brain), they were not correctly classified.

Basically, the SZSBEGF2014 researchers developed an optimization algorithm that, starting from a correctly classified image, tries to find a small perturbation of this that drives the output of the network to a wrong classification. The phenomenon is a case where starting from slightly different initial conditions, the network gives a diverse output.

If this is true, we have a serious situation where feedforward network classifications, and more specifically the deep neural network paradigms, fail to generalize. They lead to false classifications, and thus – for crucial applications – may result to severe repercussions, that may even lead to human deaths.

It is pointed here that such phenomena have been well established in engineering and science, as for example in some dynamic-chaotic systems that are highly sensitive to initial conditions. In these, a small perturbation of the initial conditions could drive the system to totally different extremes (behaviour).

As pointed, this second claim is a counter-generalization observation, which may have a profound negative impact on the development of ANNs, especially for critical applications such as in critical medical diagnostic and other systems such as in autonomous/driverless cars and other vehicles, in crucial google glass applications, in salvage operations, in critical/sensitive military applications, etc.

Here are some more specific examples that may make one think twice before relying on feedforward ANNs for decisions:

- A self-driving/autonomous car that uses an ANN (e.g. deep neural network) does not recognize a human standing in front of the car. It may interpret the road as clear, resulting in highly risky and dangerous situations for pedestrians.
- An ANN that is used in a critical medical diagnostic operation that misclassifies as a false positive a specific cancer image or medical signal.
- An ANN that is used in military operations and misclassifies a building as having terrorists that should be bombed!
- A prisoner convicted to death penalty, where the realization of this verdict depends on his/her IQ being above a certain threshold, which had been wrongly established by an ANN (e.g. case of Ted Herring in Florida State, USA).

An interesting issue that comes to mind is whether such "blind spots" also exist in biological neural networks. We know that certain blind spots (static or dynamic) have been observed, e.g. the attentional blink (Marois et al. 2000; Neokleous et al 2009). This occurs in a large number of individuals. That is, it has a high statistical significance, but it is not universal. That is, we may speculate that some biological neural networks express a uniformly blind spot.

Even though, for most people, the brain has an impressive capacity to recognize images in diverse orientations, lighting conditions, deformations, modifications, perturbations etc., may occasionally make wrong classifications, generalizations, interpretations. It can even properly identify words in the well-known Cambridge University observation, popularized by the following extract:

"The phaonmneal pweor of the hmuan mnid, aoccdrnig to a rscheearch at Cmabrigde Uinervtisy, it dseno't mtaetr in waht oerdr the ltteres in a wrod are, the olny iproamtnt tihng is taht the frsit and lsat ltteer be in the rghit pclae. The rset can be a taotl mses and you can sitll raed it whotuit a pboerlm. Tihs is bcuseae the huamn mnid deos not raed ervey lteter by istlef, but the wrod as a wlohe".

If it is a fact that the biological brains may misclassify, misinterpret, miscalculate and misunderstand, then this creates numerous legal, ethical, and philosophical questions.

Concerning the claim for the existence of adversarial examples, we suggest the following criticism.

a) This is a premature claim. The SZSBEGF2014 researchers have tested a large number of ANN structures, but they were all of feedforward topology. They did not say whether they have also tested recurrent structures (dense or sparse). How could we know whether similar behavior occurs in artificial neural structures that have recurrences? We know, for instance that biological neural networks, and more profoundly the human brain, are highly recurrent structures. Thus, more investigation into this issue is needed.

b) Even though one can conduct extensive simulations on diverse networks, there will still be gray areas, unless one manages to prove in a coherent - preferably mathematical formalism - that the blind spots are universal to all network structures. So, here is a new research field for exceptional theoreticians!

c) Considering the cases of blind spots in biological neural recognizing systems such as the human brain, one can observe that these blind spots are not universal. Indeed, one can find human brains that correctly identify images that they were erroneously mislabeled by a large population. Thus, there may be ANNs that can correctly identify adversarial examples. It is rather a matter of finding these networks.

In any case, as it is, it should make us very cautious in building critical applications in which ANNs are embedded, e.g. in medical diagnostic systems for critical diseases.

Indeed, this issue may hold back the development and application of ANNs, analogous to the Minsky and Papert effect that held back developments back in the 1960s.

4 CONCLUSIONS

Adversarial examples, if exist in a universal and absolute manner, can be serious problems in critical applications of pattern recognition. They may also stall the further development of artificial neural networks. However, their absolute existence has not been proven. Nor have they been verified in recurrent neural structures. We believe that appropriate ANN structures that correctly recognize adversarial examples, can be found, developed and applied.

ACKNOWLEDGEMENTS

The Cyprus University of Technology.
The University of Cyprus.

REFERENCES

http://www.image-net.org/ Seen in June 2014.

http://www.nytimes.com/2014/05/31/us/on-death-row-with-low-iq-and-new-hope-for-a-reprieve.html?_r=0 Seen in June 2014.

LeCun Y., Cortes C., Burges C., The MNIST database of handwritten digits. http://yann.lecun.com/exdb/mnist/ Seen in June 2014.

Marois R., Chun M., Gore J., 2000, Neural Correlates of the Attentional Blink, *Neuron*, Vol. 28, pp. 299–308.

Minsky M., Papert S., 1972, *Perceptrons: An Introduction to Computational Geometry*, (2nd edition with corrections, first edition 1969), The MIT Press, Cambridge MA.

Neokleous K., Avraamides M., Neocleous C., Schizas C., 2009. A neural network model of the attentional blink phenomenon. *International Journal of Engineering Intelligent Systems for Electrical Engineering and Communications*, Vol. 17, pp. 115-126.

Szegedy C., Zaremba W., Sutskever I., Bruna J., Erhan D., Goodfellow I., Fergus R., 2014. Intriguing properties of neural networks, *2nd International Conference on Learning Representations*, Rimrock Resort, Canada, April 14-16, 2014.