**Critical Review of “Explaining and Harnessing Adversarial Examples”**

Ian J. Goodfellow, Jonathon Shiens & Christian Szegedy

*Google Inc. Mountain View CA*

Rafael Carvalhaes Possas, University of Sydney, Australia

1. Introduction

This paper can be considered a seminal publication in the field of adversarial examples on machine learning. It shows that generalization of neural networks depends mainly on their architecture and training set and the main cause of this vulnerability is their linear nature. Even though, this linear nature is explained, there is not enough information that proves this statement in the paper. One should already understand what a linear behavior in high-dimensional spaces looks like. By following previous works this paper shows that machine learning models can misclassify images with high confidence by only causing slightly small perturbations to every pixel in the input image. The quantitative results are enough to prove the efficiency of the Fast Gradient Sign Method and all its variants.

1. Summary

The text provides a brief and good explanation on why neural networks can behave linearly in high dimensional spaces. The method proposed is clearly explained and derived mathematically so readers can repeat the experiment. A good visual representation of what the model is doing is given through real images and representations of what the noise being added to generate the adversarial looks like.

Neural Networks are proven to be too linear to resist linear adversarial perturbation. For instance, Sigmoid networks are tuned to spend most of their time in the non-saturating linear regime so that they are easier to optimize. The methods then, explores this scenario by using gradient information plus random noise so that images are perturbed yielding wrong label outputs when submitted to state of the art DNN algorithms.

1. Critique

This work was mainly based into the paper from Szegedy et al. 2014 where he discusses intriguing properties of neural networks and other related models. It does a really good job on summarizing previous work that supports his claims and builds up foundation knowledge to the paper current readers.

Most of the experiments are well designed and results are presented accordingly so the hypothesis is clearly proven. For instance, the work uses state of the art Neural Networks along with the popular ImageNet dataset. These are the baseline sources of every computer vision tasks nowadays as they have the higher performance on ILSCV competitions. The seminal contribution of this paper is that networks with stacked layers are not non-linear as everyone thinks, but models with linear behavior in high-dimensional spaces. This is well explained throughout the entire paper and good comparisons between different architectures of DNNs are made.

1. Conclusion

The method proposed in this work is highly cited in several papers in this field. Most of the work developed has this as a baseline technique and different variations are applied in order to explore the characteristics of adversarial examples in machine learning. The technique is simple but efficient in generating examples with high confidence, which shows that Deep Learning systems are not as robust as most would think since they are easily fooled.