**REVIEW FORM**

Paper Title: Explaining and Harnessing Adversarial Examples

List of Authors: Ian J Goodfellow, Jonathon Shiens & Christian Szegedy

Publication Year/Venue: 2014

ERA/Core ranking for the publication venue: AAA

Number of citation the paper received so far: 131

Reviewer SID#: 450645880

Reviewer Full Name: Rafael Carvalhaes Possas

Have you contacted with your supervisor to confirm reviewing this paper (Y/N): Yes

PLEASE EVALUATE THE MANUSCRIPT BASED ON THE FOLLOWING SCALE:

1 = Poor 2 = Below average 3 = Average 4 = Good 5 = Excellent x = Not applicable

Poor Excellent N/a

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 1. This paper is based on rigorous academic standards. | 1 | 2 | 3 | 4 | 5 | x |
| 1. This paper is presented in a format which is accessible by practitioners. It focuses on justification, results and implementation; has readable style | 1 | 2 | 3 | 4 | X | x |
| 1. The paper has clarity of presentation. It is well organized, clearly written. It is easy to read and free from grammatical or spelling errors. | 1 | 2 | 3 | X | 5 | x |
| 1. The paper makes a significant contribution to the body of knowledge. It is highly significant, breaks new ground, and provides a foundation for future research. | 1 | 2 | 3 | 4 | X | x |
| 1. The topic of this paper is relevant, timely, and of interest to the audience of the target venue. | 1 | 2 | 3 | 4 | X | x |
| 1. The rationale for the paper is well grounded. It is based on a known theory or on an interesting issue. | 1 | 2 | 3 | 4 | X | x |
| 1. The research methodology for the study is appropriate and applied properly. | 1 | 2 | 3 | X | 5 | x |
| 1. The content of this paper is technically accurate and sound. | 1 | 2 | 3 | 4 | X | x |
| 1. The supporting evidence in this paper is strongly reliable and properly validated. | 1 | 2 | 3 | 4 | X | x |

Your comments: (Please present specific details using the following categories.)

Summary of the paper:

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.

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.

Strengths:

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.

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

Minor/Major weaknesses and suggestions for further improvements:

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. The explanation of the technique could be deeper and more clear so readers could replicate the experiment easily. The overhead also is a minor issue as it requires the use of pre-trained and specific architectures of deep nets to be replicated. As a further improvement the author could make the code available with the examples used in the paper on his website.