# Reproduce the results of the article "Explaining and Harnessing Adversarial Examples"

**JIANG Chenao 1** CHENAO.JIANG@ETU.SORBONNE-UNIVERSITE.FR
*Master Ingénierie des Systèmes Intelligents*
*Sorbonne Université*
*Paris, France*

**LIU Hanlin 2** HANLIN.LIU@ETU.SORBONNE-UNIVERSITE.FR
*Master Systèmes Avancés et Robotiques*
*Sorbonne Université*
*Paris, France*

**Chen Yuwang 3** YUWANG.CHEN@ETU.SORBONNE-UNIVERSITE.FR
*Master Ingénierie des Systèmes Intelligents*
*Sorbonne Université*
*Paris, France*

**WANG Haoyu 4** HAOYU.WANG@ETU.SORBONNE-UNIVERSITE.FR
*Master Systèmes Avancés et Robotiques*
*Sorbonne Université*
*Paris, France*

**Editor:** Machine Learning Avancé (2021-2022)

## Abstract

In the era of data computing driven by deep learning algorithms, it is crucial to ensure the security and robustness of the algorithms : some machine learning models consistently misclassify adversarial examples. Szegedy et al. (2014b) argue that the primary cause of neural networks' vulnerability to adversarial perturbation is their linear nature, and uses this view to generate a simple and fast method of generating adversarial examples - FGSM. In this project, we focus on generating adversarial samples and confirming their impact on neural networks using the FGSM method, and implementing adversarial training of linear models as well as deep networks. We work on generating adversarial samples on different neural networks and observe how the samples interfere with the classification of the neural network. After the adversarial training is completed, we compare the robustness of the neural network to the adversarial interference before and after training and confirm the effectiveness of the adversarial training.We will discuss the ability of different architectures of neural networks to resist interference. In addition,we will discuss the ability of different architectures of neural networks to resist interference. And, we will also study...

**Keywords:** Adversarial examples, FGSM, Adversarial trainin, GoogLeNet, Maxout, DNN,etc...

## 1. Introduction

Szegedy et al. (2014b) argue that the vulnerability of machine learning models to interference from adversarial examples is due to the extreme non-linearity of deep neural networks. In line with this view, Szegedy et al.devise a fast method for generating adversarial examples, *FGSM*, and show that adversarial training can indeed greatly improve the robustness and accuracy of neural networks, and can provide additional regularisation benefits. In their study, they first explain the existence of adversarial examples for linear models, and explain the principle of *FGSM*. Afterwards, Szegedy et al. implement adversarial training on a variety of linear models and deep networks, and demonstrate the effectiveness and practicality of the training. The resistance of different architectures of neural networks to adversarial examples is also discussed, and a deeper insight into adversarial examples is provided.

The aim of this project is to re-implement the algorithms in the paper and to reproduce all the experimental results obtained. We have approached the reproduction of the paper in three main directions : firstly, the existence of adversarial samples. Second, the impact of adversarial attacks. Third, the practicality of adversarial training.

To do this, we first need to understand the adversarial example and its existence : we study its rationale and its generation method, *FGSM*, and generate an adversarial example based on a neural network model. Secondly, we look at the effect of this adversarial example on linear neural network models, in particular logistic regression networks (*simple linear neural networks, softmax, maxout, etc.*), and implement the adversarial example training in these networks to see the effectiveness of the adversarial training. Thirdly, we will focus on adversarial training of deep networks and try to demonstrate that deep networks are more robust to adversarial examples than simple linear neural networks. In this section, we will work on generating adversarial samples and implementing adversarial training on *GoogLeNet* and *CNN*, etc. Fourthly, we will compare the robustness of the adversarial samples and the results of the adversarial training on the above different architectures and hope to try more different approaches based on the adversarial training, such as *early stopping* and expanding the model to improve the accuracy of the model.

Finally, we will extend the study with an adversarial sample and discuss as well as compare the results obtained from the above experiments ...... to complete ......

## 2. Présentation de l'algorithme

Researchers have identified a serious security concern with existing neural network models : an attacker can easily fool a neural network by adding specific noise to benign samples, often undetected. The attacker uses perturbations that are not perceptible to human vision/audition, which are sufficient to cause a normally trained model to output false predictions with high confidence, a phenomenon that researchers call adversarial attacks.

Existing adversarial attacks can be classified as white-box, grey-box and black-box attacks based on the threat model. The difference between these three models lies in the information known to the attacker, and the FGSM approach is a white-box attack in which the threat model assumes that the attacker has complete knowledge of his target model, including the model architecture and parameters. The attacker can therefore create an adversarial sample directly on the target model by any means. The attacker can therefore create an adversarial sample directly on the target model by any means.

## 3. Données

**Présentation des bases de données utilisées. Précisez et justifiez les éventuelles différences avec l'article de référence.**

# 4. Evaluation expérimentale

## 4.1 Description de l'expérience

**Description de l'expérience réalisée, méthodologie et métriques d'évaluation.**

## 4.2 Résultats

**Présentation des résultats expérimentaux obtenus et comparaison par rapport à ceux de l'article de réference.**

## 4.3 Discussion

**Discussion critique à partir des résultats obtenus**

# 5. Conclusion

**Conclusion : un résumé synthétique des principaux résultat obtenus et présentation des principales pistes d'amélioration possibles.**

# 6. Bibliographie

**Bibliographie : une liste complète des principaux articles de l'état de l'art ou ayant inspiré la démarche, et qui seront référencés de manière pertinente dans le rapport.**

Par exemple (Goodfellow et al., 2016) et (Some and Other, 2016)

# Références

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. `http://www.deeplearningbook.org`.

Author Some and Author Other. Test references. *A Journal*, 2016.