PROPOSAL

DEVISING AN EFFECTIVE BLACK-BOX ADVERSARIAL ATTACK TO EXPOSE VULNERABILITIES OF NEURAL NETWORKS

BACKGROUND WORK

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 1. No. | Title of Paper | Main Contribution | Authors | Citations | Year |
| 1 | Intriguing Properties of Neural Networks | Adversarial Attacks | Christian Szegedy | 1487 | 2013 |
| 2 | Explaining and Harnessing Adversarial Examples | FSGM | Ian Goodfellow | 1346 | 2015 |
| 3 | Practical black-Box Attacks | Black-Box Attacks | Papernot | 256 | 2016 |
| 4 | Boosting Adversarial Attacks with Momentum | BI-FSGM | Dong et. al Tsinghua University, China | 40 | 2018 |
| 5 | Evaluating Robustness of Neural Networks | CLEVER: Extreme value Theory Approach | IBM | 20 | 2018 |

FOCUS AREAS

Neural Networks

CNNs

RNNs

Activation Functions

White-Box and Black-Box Attacks

Adversarial Training

Robustness

TECHNOLOGIES

Python

Tensorflow

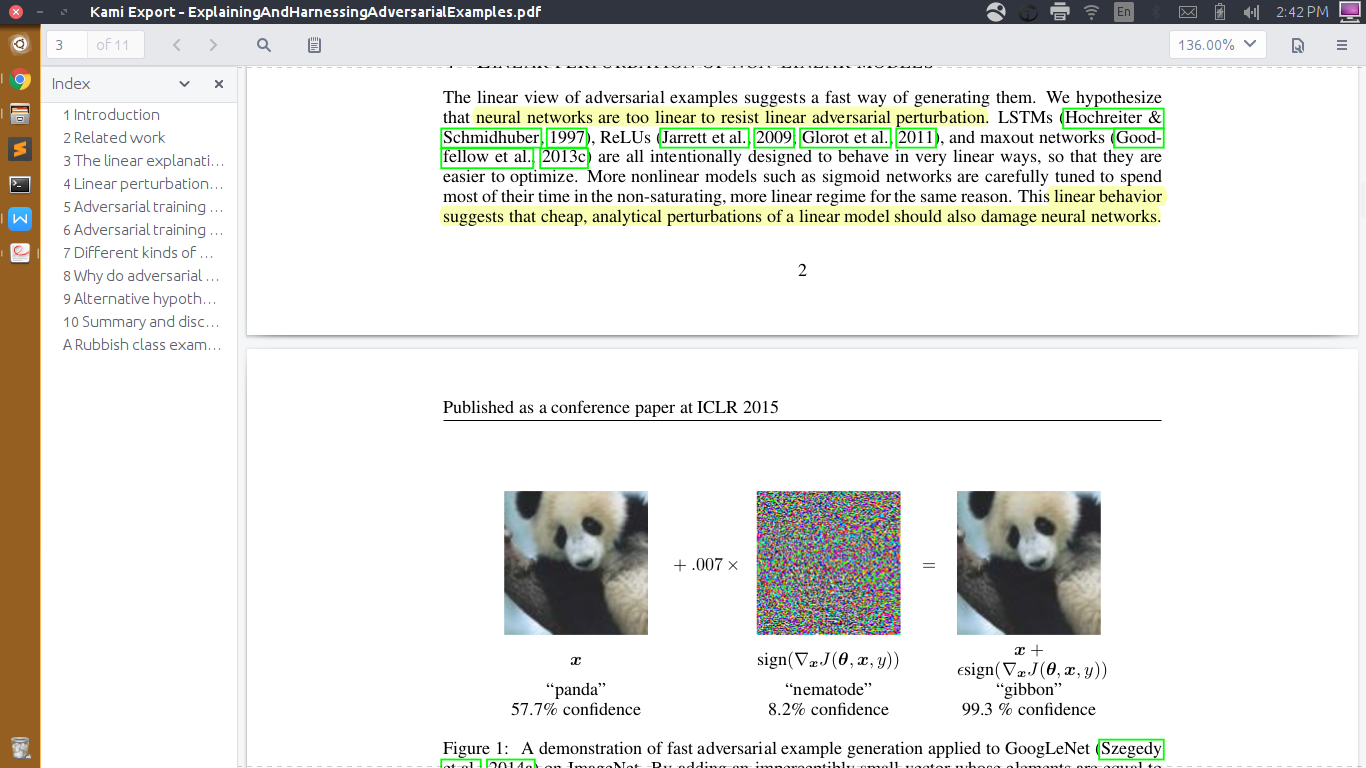
Adversarial Robustness Toolbox by IBM

INTRODUCTION

The work on the robustness of Neural Networks against Adversarial Attacks is an important area of research today with major contributions by top researchers in the field of Deep Learning. If we have a look at the current research focus of Goodfellow, Hinton etc., notably, this topic of Adversarial Attacks and robustness of Neural Nets forms a significant portion of their research today and with good reason.

Several machine learning models, including neural networks, consistently misclassify adversarial examples—inputs formed by applying small but intentionally worst-case perturbations to examples from the dataset, such that the perturbed input results in the model outputting an incorrect answer with high confidence.

Shallow softmax regression models are also vulnerable to adversarial examples.Training on adversarial examples can regularize the model. These results suggest that that classifiers based on modern machine learning techniques, even those that obtain excellent performance on the test set, are not learning the true underlying concepts that determine the correct output label. An example generated for one model is often misclassified by other models, even when they have different architectures or were trained on disjoint training sets. Moreover, when these different models misclassify an adversarial example, they often agree with each other on its class.



Objective: Devising an effective Defense mechanism to enhance the Robustness of Neural Networks against Adversarial Attacks

Project Start: August 2018

Introduction:

Several machine learning models, including neural networks, consistently misclassify adversarial examples—inputs formed by applying small but intentionally worst-case perturbations to examples from the dataset, such that the perturbed input results in the model outputting an incorrect answer with high confidence. Classifiers based on modern machine learning techniques, even those that obtain excellent performance on the test set, are not learning the true underlying concepts that determine the correct output label. An example generated for one model is often misclassified by other models, even when they have different architectures or were trained on disjoint training sets. Moreover, when these different models misclassify an adversarial example, they often agree with each other on its class. In this project, we analyse the representation of image features in neural networks and how it can be exploited by adversarial inputs to produce erroneous classification as the output. The work on the robustness of Neural Networks against Adversarial Attacks is an important area of research today with major contributions by top researchers in the field of Deep Learning. If we have a look at the current research focus of Goodfellow, Hinton etc., notably, this topic of Adversarial Attacks and robustness of Neural Nets forms a significant portion of their research today and with good reason. As we move towards a world that is increasingly being touched by innovations in Computer Vision, the Robustness of neural networks is very central to the efficacy of the models.

Technologies: Convolutional Neural Networks, Capsule Networks, Adversarial Training, Adversarial inputs, FSGM etc.

Languages: Python, Tensorflow etc.

Devising an effective Defense mechanism to increase the Robustness of Neural Networks against Adversarial Attacks Several machine learning models, including neural networks, consistently misclassify adversarial examples—inputs formed by applying small but intentionally worst-case perturbations to examples from the dataset, such that the perturbed input results in the model outputting an incorrect answer with high confidence. In this project, we analyse the representation of image features in neural networks and how it can be improved to prevent exploitation by adversarial inputs. Technologies:Convolutional Neural Networks, Capsule Networks, Adversarial Training, Adversarial inputs, FSGM Languages:Python, Tensorflow