Training Robust Neural Networks

Clément Preaut, Iwan Quemada, Nicolas Durat

Quick-Attacks

8 November 2022

Introduction

- Adversarial ML is a method that trick models by providing deceptive input. It includes both the generation and detection of adversarial examples created to deceive classifiers
- An adversarial attack is a method to generate adversarial examples. It is an input to a ML model that is purposely designed to cause a model to make a mistake in its predictions
- In whitebox attack, attacker has a complete target model access (i.e parameters)
- In blackbox attack, attacker has no model access and only observe targeted model outputs
- Adversarial training defends models by augmenting training data with adversarial cases. By training on both data types, it attempts to reduce the adversarial examples risk

Naive Classifier

- Dataset: CIFAR-10 - 60000 32x32 images (10 classes)
- Basic Architecture: 54 % accuracy (training set: 50000 images)
- Optimization: NLLLoss, SGD (learning_rate = 0.001, momentum = 0.9)

```
==========================================================================================
Layer (type:depth-idx)                   Output Shape              Param #
==========================================================================================
Net                                      [32, 10]                  --
├─Conv2d: 1-1                            [32, 6, 28, 28]           456
├─MaxPool2d: 1-2                         [32, 6, 14, 14]           --
├─Conv2d: 1-3                            [32, 16, 10, 10]          2,416
├─MaxPool2d: 1-4                         [32, 16, 5, 5]            --
├─Linear: 1-5                            [32, 120]                 48,120
├─Linear: 1-6                            [32, 84]                  10,164
├─Linear: 1-7                            [32, 10]                  850
==========================================================================================
Total params: 62,006
Trainable params: 62,006
Non-trainable params: 0
Total mult-adds (M): 21.06
==========================================================================================
Input size (MB): 0.39
Forward/backward pass size (MB): 1.67
Params size (MB): 0.25
Estimated Total Size (MB): 2.31
==========================================================================================
```

Figure: Naive Classifier Summary

Whitebox Attack Mechanisms

FGSM (Fast Gradient Sign Method, Goodfellow et al., 2015)

$$x^{adv} = x + \epsilon sign(\nabla_x L(\theta, x, y)) \tag{1}$$

- Comparably efficient computing times
- Perturbations are added to every feature in a single step

PGD (Projected Gradient Descent, Madry et al., 2019)

$$x^{t+1} = \Pi_{B(x_0,\epsilon)}(x^t + \alpha sign(\nabla_x L(\theta, x, y))) \quad where \quad x_0 = x \tag{2}$$

- Simple, flexible and very efficient
- Perturbations applied many times with a small step size

Preliminary results

- PGD : Imperceptible changes until 0.02

Figure: Influence of ϵ on picture visualization

Preliminary results

- FGSM : Fast attack
- PGD : Best performing attack

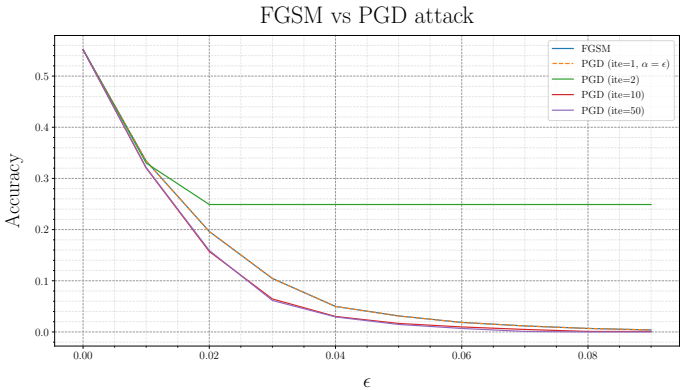

Figure: Influence of ϵ on Accuracy

Preliminary results

- Necessary Condition : $\alpha \times ite \geq \epsilon$

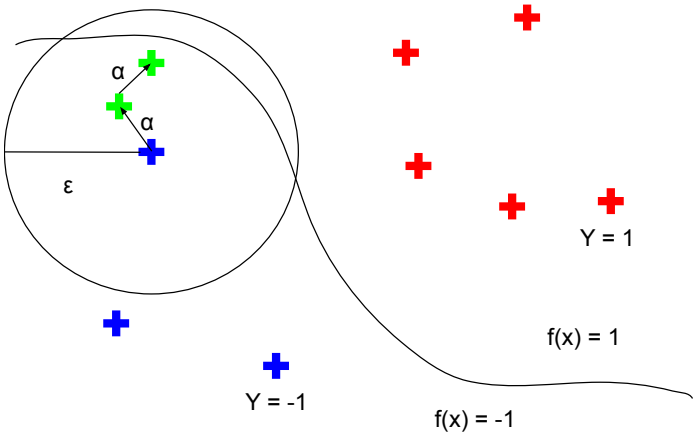

Figure: Not enough disturbed picture

Adversarial Training

- Provides a defense against a particular set of attacks
- Minimizing the worst case error when the data is perturbed by an adversary
- Form of active learning, where the model is able to request labels on new points
- Create and incorporate adversarial examples into the training process
- Inner maximization problem is approximated with FGSM or PGD attack

Adversarial Training (Defense mechanism, Goodfellow et al., 2015)

$$\hat{L}(\theta, x, y) = \alpha L(\theta, x, y) + (1 - \alpha) L(\theta, x + \epsilon sign(\nabla_x L(\theta, x, y))) \tag{3}$$

- Effective on a specific attack in practice
- Lack of adaptability, generalization of robustness difficult

Adversarial Training

- W/O: Best performing on unattacked images but very attacks sensitive
- With: Less performances without attack. More attack resistant

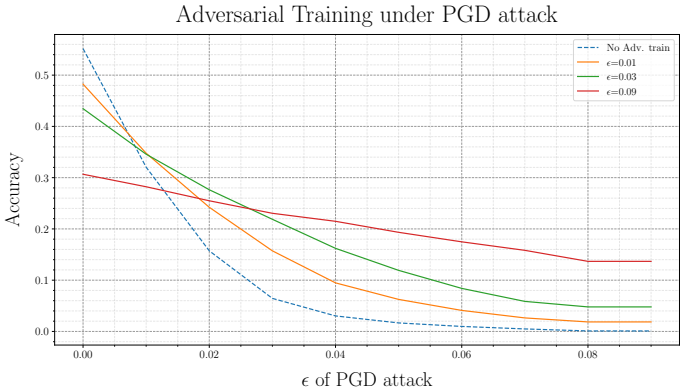

Figure: Impact of the ϵ in Adversarial Training ($\alpha = 0.5$) on Accuracy, under PGD attacks

References I

Ian Goodfellow, Jonathon Shlens, and Christian Szegedy. 2015. Explaining and harnessing adversarial examples.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. 2019. Towards deep learning models resistant to adversarial attacks.