FREE VS FAST ADVERSARIAL TRAINING

Giuseppe Capaldi
University of Rome "La Sapienza"
capaldi.1699498@studenti.uniroma1.it

Gianluca Capozzi
University of Rome "La Sapienza"
capozzi.1693255@studenti.uniroma1.it

September 22, 2020

ABSTRACT

In this work we propose a comparison between two methods for adversarial training: Free and Fast. The code is available at https://github.com/not-a-genius/neuralNetworkExam.

1 Introduction

Adversarial training is a method for training robust deep neural networks, training on adversarial examples as a defense against adversarial attacks. This is typically considered to be more expensive than standard training due to the necessity of constructing adversarial examples via a first-order method like projected gradient descent (PGD). The papers presented show that it is possible to train empirically robust models using methods no more costly than standard training in practice. These methods show results comparable to train a model against PGD but at lower cost in terms of computational time. The goal of adversarial training is to learn a model which is not only accurate on the data (accuracy on train and test data) but also accurate on adversarially perturbed versions of the data (i.e. accuracy on images perturbed using the PGD attack). This leads to particular attention at the results we obtained after adversarial training experiments, aware of the presence of an acceptable compromise in accuracy, given a large increase in robustness due to the tradeoff between robustness and generalization [1, 2, 3].

1.1 Related Works

Our work is based on “Adversarial training for free!” paper [4], that presented an algorithm able to eliminate the overhead cost of generating adversarial examples by recycling the gradient information computed when updating the model parameters. While researching related work in the field we discovered a more recent paper, called "Fast is better than free: Revisiting adversarial training" [5] presenting an already known method called FGSM, previously considered ineffective due to what the paper calls “catastrophic overfitting”. This failure condition is preventable with the use of random initialization points. Moreover the same paper reported a further acceleration in training even for free algorithm thanks to standard techniques for efficient training, including cyclic learning rate and mixed-precision arithmetic. This caught our interest and we decided to try to replicate the reported results and compare them with our implementation, in order to confirm or deny these thesis.

1.2 Dataset and architecture

We decided to choose CIFAR-10 as the dataset on which our experiments have been conducted to adapt to the available hardware (our GPU is an Nvidia RTX 2070 super). Both WRN32 and Resnet50 have been tried as models to train on but their lack of accuracy even at high numbers of epochs suggested us to switch to shallower models, so we used PreActResnet18 which is based on Resnet18.

2 Adversarial Machine Learning

Adversarial machine learning is a machine learning technique that attempts to fool models by supplying deceptive input (called also adversarial example). Adversarial examples exploit the way artificial intelligence algorithms work to

disrupt the behavior of artificial intelligence algorithms. In the past few years, adversarial machine learning has become an active area of research as the role of AI continues to grow in many of the applications we use. There's growing concern that vulnerabilities in machine learning systems can be exploited for malicious purposes.

3 PGD attack

4 Free Adversarial training for free!

5 Fast is better than free

6 Conclusions

References

[1] Tsipras et al., 2018: Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Alexander Turner, and Aleksander Madry. Robustness may be at odds with accuracy. ICLR, 1050:11, 2018

[2] Zhang et al., 2019a: Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric P Xing, Laurent El Ghaoui, and Michael I Jordan. Theoretically principled trade-off between robustness and accuracy. ICML, 2019a.

[3] Shafahi et al., 2019a Ali Shafahi, W Ronny Huang, Christoph Studer, Soheil Feizi, and Tom Goldstein. Are adversarial examples inevitable? ICLR, 2019a.

[4] Shafahi et al., 2019b Ali Shafahi, Mahyar Najibi, Amin Ghiasi, Zheng Xu, John Dickerson, Christoph Studer, Larry S Davis, Gavin Taylor, and Tom Goldstein. Adversarial training for free! arXiv preprint arXiv:1904.12843, 2019b.

[5] Wong et al., 2020 Eric Wong and Leslie Rice and J. Zico Kolter, arxiv:2001.03994, 2020.