2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)

Efficient Adversarial Training with Transferable Adversarial Examples

Haizhong Zheng Ziqi Zhang Juncheng Gu Honglak Lee Atul Prakash
University of Michigan, Ann Arbor
{hzzheng, ziqizh, jcgu, honglak, aprakash}@umich.edu

Abstract

Adversarial training is an effective defense method to protect classification models against adversarial attacks. However, one limitation of this approach is that it can require orders of magnitude additional training time due to high cost of generating strong adversarial examples during training. In this paper, we first show that there is high transferability between models from neighboring epochs in the same training process, i.e., adversarial examples from one epoch continue to be adversarial in subsequent epochs. Leveraging this property, we propose a novel method, Adversarial Training with Transferable Adversarial Examples (ATTA), that can enhance the robustness of trained models and greatly improve the training efficiency by accumulating adversarial perturbations through epochs. Compared to state-of-the-art adversarial training methods, ATTA enhances adversarial accuracy by up to 7.2% *on CIFAR10 and requires* $12 \sim 14\times$ *less training time on MNIST and CIFAR10 datasets with comparable model robustness.*

1. Introduction

State-of-the-art deep learning models for computer vision tasks have been found to be vulnerable to adversarial

$\mathcal{S}$, k-step projected gradient descent method [13, 18] (PGD-k) has been widely adopted to generate adversarial examples. Typically, using more attack iterations (higher value of k) produces stronger adversarial examples [18]. However, each attack iteration needs to compute the gradient on the input, which causes a large computational overhead. As shown in Table 1, the training time of adversarial training can be close to 100 times larger than natural training.

Recent works [17, 19, 20] show that adversarial examples can be transferred between models: adversarial examples generated for one model can still stay adversarial to another model. The key insight in our work, which we experimentally verified, is that, because of high transferability between models (i.e., checkpoints) from neighboring training epochs, attack strength can be accumulated across epochs by repeatedly reusing the adversarial perturbation from the previous epoch.

We take advantage of this insight in coming up with a novel adversarial training method called ATTA (Adversarial Training with Transferable Adversarial examples) that can be significantly faster than state-of-the-art methods while achieving similar model robustness. In traditional adversarial training, when a new epoch begins, the attack algorithm generates adversarial examples from the original input images, which ignores the fact that these perturbations can be