Submission Summary

Conference Name
2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)

Paper ID
8212

Paper Title
DAFAR: Detecting Adversaries by Feedback-Autoencoder Reconstruction

Abstract
Deep learning has shown impressive performance on challenging perceptual tasks. However, researchers found deep neural networks vulnerable to adversarial examples. Since then, many methods are proposed to defend against or detect adversarial examples, but they are either attack-dependent or shown to be ineffective with new attacks.

We propose DAFAR, a feedback framework that allows deep learning models to detect adversarial examples in high accuracy and universality. DAFAR has a relatively simple structure, which contains a target network, a plug-in feedback network and an autoencoder-based detector. The key idea is to capture the high-level features extracted by the target network, and then reconstruct the input using the feedback network. These two parts constitute a feedback autoencoder. It transforms the imperceptible-perturbation attack on the target network directly into obvious reconstruction-error attack on the feedback autoencoder. Finally the detector gives an anomaly score and determines whether the input is adversarial according to the reconstruction errors. Experiments are conducted on MNIST and CIFAR-10 data-sets. Experimental results show that DAFAR is effective against popular and arguably most advanced attacks without losing performance on legitimate samples, with high accuracy and universality across attack methods and parameters.

Created on
11/9/2020, 6:00:22 PM

Last Modified
11/17/2020, 2:34:25 PM

Authors
Haowen Liu (Shanghai Jiao Tong University) < issacliewx@sjtu.edu.cn>
Ping Yi (Shanghai Jiao Tong University) < yiping@sjtu.edu.cn>
Hsiao-Ying Lin (Huawei International) < Lin.hsiao.ying@huawei.com>
Jie Shi (Huawei International) < shi.jie1@huawei.com>

Primary Subject Area
Adversarial Learning, Adversarial Attack and Defense Methods -> Adversarial Attack and Defense Metho

Secondary Subject Areas
Adversarial Learning, Adversarial Attack and Defense Methods -> Adversarial examples

Submission Files
DAFAR__Detecting_Adversaries_by_Feedback_Autoencoder_Reconstruction.pdf (694.5 Kb, 11/17/2020, 2:34:19 PM)