CS 260 Project Proposal

Wenlong Xiong (204407085)

February 11, 2019

As neural networks become more important for everyday tasks, it is imperative that they are robust against adversarial attacks, in which the model is made to misclassify. In particular, test-time black box attacks (where the model is unknown and fixed and given adversarial examples) form the bulk of real-world situations. Past defenses against these black-box attacks have included distillation to obfuscate gradients that attacks depend on [1], increasing network capacity along with adversarial training [2], using noise addition to create more robust intermediate layers [3], etc.

However, I want to try a different approach - using an denoising autoencoder network as data pre-processing in order to defend against perturbed adversarial examples. Hinton [4] suggests that many models only work well on examples that lie on a thin manifold, and that adversarial examples lie slightly off than manifold. Prior research suggests that denoising autoencoders [5] learn a manifold and project points back onto that manifold during the reconstruction process. In addition, there have been some prior research suggesting that autoencoders are generally difficult to attack [6]. Generic autoencoders as pre-processing have also been shown to be successful in defending against black-box as well [7].

Therefore, it would be interesting to see if using a denoising autoencoder as a preprocessing step to models would defend against black-box attacks. To test out the effectiveness of using a denoising autoencoder, we could compare its effectiveness against Zeroth Order Optimization attacks [8], which is as effective as current state-of-the-art white box attacks. We would use the output of the denoising autoencoder as the input to a pre-trained classifier, and compare the success rate of the attacks without the preprocessing to the success rate of the attacks with preprocessing. In addition, we can compare using a denoising autoencoder as defense against other adversarial defense methods, such as the model described by Madry [2]. I can first start testing my method with the MNIST dataset, and if we achieve a high rate of success, the CIFAR-10 or CIFAR-100 dataset (depending on hardware limitations).

References

[1] Nicolas Papernot, Patrick McDaniel, Xi Wu, Somesh Jha, and Ananthram Swami. Distillation as a defense to adversarial perturbations against deep neural networks, 2015.

[2] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *CoRR*, abs/1706.06083, 2017.

[3] Xuanqing Liu, Minhao Cheng, Huan Zhang, and Cho-Jui Hsieh. Towards robust neural networks via random self-ensemble, 2017.

[4] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples, 2014.

[5] Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. Extracting and composing robust features with denoising autoencoders, 2008.

[6] George Gondim-Ribeiro, Pedro Tabacof, and Eduardo Valle. Adversarial attacks on variational autoencoders, 2018.

[7] Dongyu Meng and Hao Chen. Magnet: a two-pronged defense against adversarial examples, 2017.

[8] Pin-Yu Chen, Huan Zhang, Yash Sharma, Jinfeng Yi, and Cho-Jui Hsieh. Zoo: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models. 2017.