Review “Adversarial ML”

Name: Wenhui Zhang
Email: wuz49@ist.psu.edu

Paper Name：

McDaniel, P., Papernot, N., & Celik, Z. B. (2016). Machine learning in adversarial settings. *IEEE Security & Privacy*, *14*(3), 68-72.

Contribution:

This paper reviews various models and possibilities of malicious inputs designed to fool machine learning models. And it gives an introduction on adversarial examples on reinforcement learning agents. Then it proposed the idea that increase of model resilience could increase security of machine learning algorithms.

Motivation:

Neural networks and machine learning techniques are widely used in industry and have great impact to our life, such as auto-driving. However, neural networks are vulnerable to adversarial examples. Inputs of machine learning techniques that could be intentionally perturbed to remain visually similar to the source input. Adversarial examples are inputs to machine learning models that an attacker has intentionally designed to cause the model to make a mistake.. And this action could cause a misclassification and so on.

Related works:

Various models with different architectures misclassify same adversarial examples, and they are trained on different subsets of the training data. Deep networks are vulnerable to adversarial examples, due to misguided assumption.

Methodology:

This paper claims its own discovery that elasticity and resilience are the key to machine learning models against adversary examples.

Results:

This paper proposed an idea and did not show experiment results. Attacking a machine learning model is easy, while defending is difficult. Adversarial training provides regularization and semi-supervised learning, and the out-of-domain input problem brings up security concern for model-based optimization generally.

Take away:

There are many failed defenses against adversary examples, such as generative pretraining, removing perturbation with an autoencoder, confidence-reducing perturbation at test time, various non-linear units etc. etc. Of all methods, adversarially trained neural nets have the best empirical success rate on adversarial examples.Thus, neural nets might be more secure than other models.