Proposal

Defenses Against Adversarial Attacks

With the development and application of machine learning techniques, more and more applications adapt machine learning algorithms, such as self-driving car, face recognition, and stock prediction. However, the machine learning algorithms has potential and inevitable risk of false negative or false positive, which might cause large loss or risk in commercial applications (autonomous cars). Adversarial machine learning is a technique employed in the field of machine learning which attempts to fool models through malicious input.[1] This technique can be applied for a variety of reasons, the most common being to attack or cause a malfunction in standard machine learning models.

Basically, I would like to cover several aspects of defense of adversarial attacks. For example, defense against Poisoning attacks and defense against adversarial attacks in deep neural networks. There are several ways to defense the adversarial attacks, such as distillation, Fortified networks and PeerNets which can exploit peer wisdom against adversarial attacks. In the proposal, I would like to discuss these ideas in more details

References:

[1]Chang Liu, Bo Li, Yevgeniy Vorobeychik, Robust High-Dimensional Linear Regression.

[2]Robust logistic Regression and classification, Jiashi Feng, Huan xu, Shie Mannor, Shuicheng Yan.

[3]Ali Shafahi, W. Ronnu Huang,Mahyar, Najibi, Octaviam Suciu, Christopher StuderPoison Frogs! Targeted Clearn-Label Poisoning Attacks on Neural Networks.

[4]Ivan Evtimov, Kevin Eykholt, Earlence Fernandes, Tadayoshi Kohno, Bo Li, Atul Prakash, Amir Rahmati, Dawn Song. August, 2017. Robust Physical-World Attacks on Machine Learning Models.

[5] Distillation as a Defense to Adversarial Perturbations against Deep Neural Networks, IEEE Symposium on Security & Privacy

[6] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, Adrian Vladu Towards Deep Learning Models Resistant to Adversarial Attacks

[7] Alex Lamb, Jonathan Binas, Anirudh Goyal, Dmitriy Serdyuk, Sandeep Subramanian, Ioannis Mitliagkas, Yoshua Bengio Fortified Networks: Improving the Robustness of Deep Networks by Modeling the Manifold of Hidden Representations.

[8] Few-shot Learning With graph neural Net-works Victor Garicia, Joan Bruna.

[9] Deformable Convolutional networks, Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong zhang, Han Hu, Yichen Wei.

[10] CommanderSong: A Systematic Approach for Practical Adversarial Voice Recognition.