Adversarial Attacks and Defenses on Neural Networks

Devvrit Prabal Vashisht Surya S. Dwivedi

March 7, 2020

1 PROBLEM STATEMENT

Deep Learning is currently being used extensively in solving problems that were once unattainable. Further, it has become the choice for solving challenging tasks in speech recognition and creating autonomous vehicles. Szegedy et al[1] first discovered that Neural Networks, despite their success, are vulnerable to misclassifying well-designed input called adversarial examples. These adversarial examples are imperceptible to human eyes but can fool a Neural Network with high confidence. Since these models are deployed in safety-critical environment, it becomes important to evaluate various attacks on Neural Networks and investigate further to develop robust defence mechanisms.

2 ACTION PLAN

We will implement and evaluate various state-of-the-art white-box Neural Network attacks. We'll also try to come up with some technique to train Neural Network to make it robust. Further, we will investigate the cause behind the susceptibility of Neural Networks to such attacks. Doing so will help us in devising techniques for defending Neural Networks to such attacks.

3 DATASET

For developing and evaluating image perturbing techniques we will be focusing on datasets with lower resolutions. Therefore, we will be using MNIST and CIFAR-10 dataset. If required, we might even work with the ImageNet dataset.

4 TIMELINE

1. Spring Break - Literature survey and coming up with the idea/algorithm to defend NN to adversarial attacks. We'd aim coming up with training algorithm to make the model robust to adversarial inputs, and at the same time not letting the overall accuracy get affected.
2. March End - Initial results to verify the motivation of our idea.

3. April Mid - More experimentation and an attempt at providing some theoretic guarantees (or at least a strong intuition) behind the algorithm.

4. April End - Comparison with other existing training methods and final results/conclusion of our work.

REFERENCES

[1] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus. Intriguing properties of neural networks. *International Conference on Learning Representations*, 2014.