Network Pruning in Adversarial Training

November 3, 2020

1 Introduction

Modern deep neural networks are the state-of-the-art techniques for many applications such as computer vision and natural language processing, but they are vulnerable to adversarial attacks [3, 8, 11]. As Figure 1 shows, small but well-designed noise makes the state-of-the-art model predict wrong label with very high confidence. The existence of adversarial examples reveals some unsatisfying properties of modern deep learning model and poses a threat to safety-critic applications.

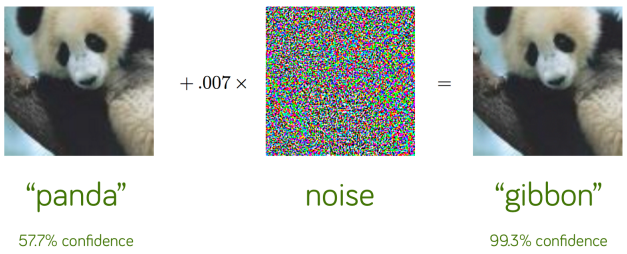

Figure 1: Imperceptible noise cause the state-of-the-art model give wrong predictions with high confidence.

In order to obtain robust models against adversarial attack, the following problem is studied in place of traditional empirical risk minimization (ERM). We define $\mathcal{S}_\epsilon(\mathbf{x}) = \{\mathbf{x}' | \|\mathbf{x}' - \mathbf{x}\| < \epsilon\}$ as the adversarial budget.

$$\min_{\theta} \max_{\mathbf{x}' \in \mathcal{S}_\epsilon(\mathbf{x})} \ell(\mathbf{x}', \theta) \tag{1}$$

Due to the non-convexity of the loss function $\ell(\mathbf{x}', \theta)$, it is difficult to solve the inner maximization problem exactly. Instead, gradient-based method such as Fast Gradient Sign Method (FGSM) [3] and Projected Gradient Descent (PGD) [7] is used. Optimize model parameter θ on the adversarial examples $\mathbf{x}'$ found by these methods, we can empirically obtain robust models.

On the other hand, modern deep neural networks are over-parameterized. Millions of, even billions of, parameters make it different to be deployed on memory-deficient devices, like mobile phones. To compress the neural network

without sacrificing the performance too much, many methods have been proposed, such as pruning [6], quantization [1] and weight coding [5].

In this project, we will combine model robustness with compression, especially model pruning. There are already lots of works studying either aspect and a few recently [12, 4, 10] for both. The first part of the project is to reproduce the state-of-the-art pruning methods for adversarially robust neural networks. Based on that, we will explore the methods to either improve the performance or implement the network pruning under more difficult settings. For example, related to popular *Lottery Ticket Hypothesis* [2], some recent work finds that randomly weighted networks contains subnetworks of competitive performance even without any training [9]. Despite many interesting phenomena in network pruning, most of them are under vanilla settings, i.e., no adversarial attack. We would like to explore whether or not these phenomena hold when we consider adversarial attacks.

This project is suitable for **1 Master Student**. The supervisor will provide basic work for adversarial training (in PyTorch). The student is expected to do experiments based on that.

2 Workloads

The planned workload includes:

Week 1-3 Read papers about adversarial training [3, 7] and its combination with network pruning [4, 10]. Get familiar with computational environment and run the code provided.

Week 4-5 Reproduce the results of paper [4, 10].

Week 6-8 Read papers about Lottery Hypothesis [2] and more classic works about network pruning [5, 9]. Prepare the mid-term presentation. Easter break.

Week 9 - 13 Explore network pruning in adversarial trainings. Check if the phenomena in vanilla settings still hold under adversarial settings.

Week 14 - 15 Summarize the project, writing the report and prepare for the final presentation.

Optional Compare the properties of the original and the pruned network, such as weight distribution, network topology.

The schedule might vary based on the actual progress. If the results are good, we seek some machine learning conference for paper submission.

3 Evaluation

The grade will based on the quality of the results, reports. There will be one midterm and one final presentations in the lab.

4 Prerequisites

Minimum:

- Good mathematical foundations: calculus, linear algebra, probability.
- Basic knowledge of optimization.

- Good reading, writing and presentation skills.
- Good coding skills in Python and familiar in PyTorch.

Bonus:

- Knowledge or project experience about network pruning or adversarial robustness.
- Experience in training state-of-the-art deep learning models on a cluster.

5 Contacts

Please contact Chen Liu (chen.liu@epfl.ch) for more details, with your transcripts and CV attached. It is preferred if you can also provide your github homepage, showing your previous projects. More projects in IVRL lab is available on https://ivrl.epfl.ch/available-projects/.

References

[1] Matthieu Courbariaux, Yoshua Bengio, and Jean-Pierre David. Binaryconnect: Training deep neural networks with binary weights during propagations. In *Advances in neural information processing systems*, pages 3123–3131, 2015.

[2] Jonathan Frankle and Michael Carbin. The lottery ticket hypothesis: Finding sparse, trainable neural networks. In *International Conference on Learning Representations*, 2019.

[3] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

[4] Shupeng Gui, Haotao N Wang, Haichuan Yang, Chen Yu, Zhangyang Wang, and Ji Liu. Model compression with adversarial robustness: A unified optimization framework. In *Advances in Neural Information Processing Systems*, pages 1285–1296, 2019.

[5] Song Han, Huizi Mao, and William J Dally. Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding. *arXiv preprint arXiv:1510.00149*, 2015.

[6] Song Han, Jeff Pool, John Tran, and William Dally. Learning both weights and connections for efficient neural network. In *Advances in neural information processing systems*, pages 1135–1143, 2015.

[7] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

[8] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, Omar Fawzi, and Pascal Frossard. Universal adversarial perturbations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1765–1773, 2017.

[9] Vivek Ramanujan, Mitchell Wortsman, Aniruddha Kembhavi, Ali Farhadi, and Mohammad Rastegari. What's hidden in a randomly weighted neural network? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11893–11902, 2020.

[10] Vikash Sehwag, Shiqi Wang, Prateek Mittal, and Suman Jana. On pruning adversarially robust neural networks. *arXiv preprint arXiv:2002.10509*, 2020.

[11] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

[12] Shaokai Ye, Kaidi Xu, Sijia Liu, Hao Cheng, Jan-Henrik Lambrechts, Huan Zhang, Aojun Zhou, Kaisheng Ma, Yanzhi Wang, and Xue Lin. Second rethinking of network pruning in the adversarial setting. *arXiv preprint arXiv:1903.12561*, 2019.