Robust Binary Network

October 31, 2019

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, 5, 7]. 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.

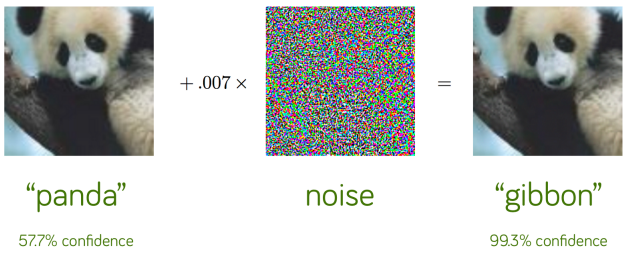

Figure 1: Imperceptible noise cause the state-of-the-art model give wrong predictions with high confidence.

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) [4] is used. Optimize model parameter θ on the adversarial examples $\mathbf{x}'$ found by these methods, we can empirically obtain robust models.

In this project, we will combine model robustness with parameter binarization. We will investigate the robustness of a specific kind of network where all parameters are binary i.e. either $+1$ or -1. Compared with normal network,

binary network is more efficient in terms of computational complexity and memory consumption and is a suitable choice in devices of limited resources.

How to train binary networks in non-adversarial environment is well studied in recent years [1, 2, 6]. Some methods maintain both binary parameters and continuous parameters. They use binary parameters for forward and backward pass but update continuous parameters in each training step, the binary parameters are obtained by projecting continuous parameters into $\{+1, -1\}^d$ space. Other methods put constraints on model parameters $\theta \in \{+1, -1\}^d$ and convert the original ERM problem into a constrained optimization problem. Standard primal-dual based methods like Alternating Direction Method of Multipliers (ADMM) can be used to solve the problem.

2 Plans

This project is suitable for **1 Master Student**. It explores the possibility to combine the method of training binary networks and ones of training robust networks. The planned workload includes:

- Read papers in the reference and get familiar with the topic (3 weeks).
- Implement PGD adversarial training [4] and BinaryConnect [1] (2 weeks).
- Introduce PGD adversarial training into BinaryConnect to train robust binary networks (2 weeks).
- Study optimization-based methods and use them, ADMM for example, to train robust binary networks (3 weeks).
- (Optional) Compare the robustness property of normal network and binary network in both normal training and adversarial training settings.
- Prepare midterm and final presentations, write the report (3 weeks).

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

3 Prerequisites

Minimum:

- Good mathematical foundations: calculus, linear algebra.
- Basic knowledge of optimization: gradient descent.
- Good reading, writing and presentation skills.
- Experience in training deep neural network models and familiar with either TensorFlow or PyTorch. (Pytorch is preferable)

Bonus:

- Familiar with current state-of-the-art deep learning models.
- Familiar with primal-dual method, augmented Lagrangian method, ADMM etc.

4 Contacts

Please contact Chen Liu (chen.liu@epfl.ch) for more details. 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] Matthieu Courbariaux, Itay Hubara, Daniel Soudry, Ran El-Yaniv, and Yoshua Bengio. Binarized neural networks: Training deep neural networks with weights and activations constrained to+ 1 or-1. *arXiv preprint arXiv:1602.02830*, 2016.

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

[4] 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.

[5] 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.

[6] Mohammad Rastegari, Vicente Ordonez, Joseph Redmon, and Ali Farhadi. Xnor-net: Imagenet classification using binary convolutional neural networks. In *European Conference on Computer Vision*, pages 525–542. Springer, 2016.

[7] 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.