Adversarial Training and Loss Landscape

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 [1, 3, 4]. 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.

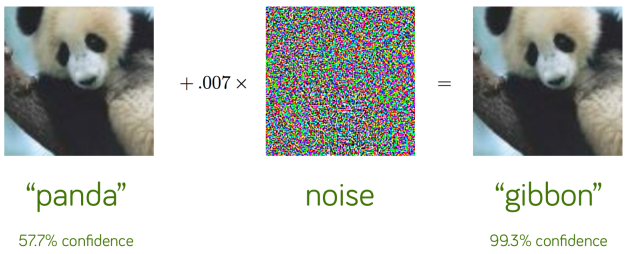

Figure 1: Imperceptible noise cause the state-of-the-art model give wrong predictions with high confidence.

In order to obtain robust models against adversarial attack, the following problem is studied in place of traditional empirical risk minimization (ERM). We define $\mathcal{S}_\epsilon(\mathbf{x}) = \{\mathbf{x}' | \|\mathbf{x}' - \mathbf{x}\|_\infty < \epsilon\}$ 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) [1] and Projected Gradient Descent (PGD) [2] is used. Optimize model parameter θ on the adversarial examples $\mathbf{x}'$ found by these methods, we can empirically obtain robust models.

FGSM [1] perturbs the input $\mathbf{x}$ by the sign of its gradient: $\mathbf{x}' = \mathbf{x} + sign\left(\nabla_{\mathbf{x}} \ell(\mathbf{x}, \theta)\right)$. FGSM is one-step gradient ascent and corresponds to the maximization of the first-order Taylor extension of $\ell(\mathbf{x}', \theta)$: $\ell(\mathbf{x}', \theta) \sim \ell(\mathbf{x}, \theta) +$

$\langle \mathbf{x}' - \mathbf{x}, \nabla_{\mathbf{x}} \ell(\mathbf{x}, \theta) \rangle$. PGD [2] extends FGSM by running gradient ascent iteratively in multiple times. It is further strengthened by adding a random noise to the initial clean input.

In this project, we first study the validity and strength of FGSM-based and PGD-based adversarial training. We will confirm that FGSM-based training can be broken by PGD attack. Furthermore, we study the properties of neural networks by different training methods, including normal training, FGSM-based training, PGD-based training. Since the input has high dimensions, so we will use dimension-reduction techniques such as Principle Component Analysis (PCA) to visualize the loss function of different input perturbations.

2 Plans

This project is suitable for **1 Senior Bachelor Student**. The planned workload includes:

- Read papers in the reference and get familiar with the topic (3 weeks).
- Implement FGSM and PGD, use them to attack or adversarially train networks. (2 weeks)
- Study techniques to visualize the high dimensional functions. (2 weeks)
- Visualize the loss function of networks trained in different ways. (3 weeks)
- (Optional) Investigate why FGSM-based adversarial training can be broken by PGD training.
- Prepare midterm and final presentations, write the report (3 weeks).

The schedule might vary based on the actual progress.

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 dimension-reduction techniques such as PCA.

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

[2] 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.

[3] 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.

[4] 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.