Bridging Adversarial Robustness and Semi/Self/Un-supervised Learning

Dinghuai Zhang
Peking University
zhangdinghuai@pku.edu.cn

1 Introduction

Deep learning has achieved state-of-the-art performance on many pattern recognition tasks [4]. Nonetheless, a series of recent work show that deep neural networks are typically vulnerable to adversarial perturbations in a relative small scale [8], which means a human-imperceptible perturbation can violate the prediction of modern machine learning models easily. As a natural result, great concerns are posed for researcher to find defensive algorithms for robust and stable models. One of the most successful methods, known as adversarial training [6], use strong adversarial examples as data augmentation to solve a minimax game:

$$\min_{\theta} \mathbb{E}_{(x,y)\sim\mathcal{D}} \max_{\|\eta\|\leq\epsilon} \ell(\theta; x+\eta, y). \tag{1}$$

where ℓ is the cross entropy loss, $\mathcal{D}$ is training dataset and θ is the parameters of model.

On the other hand, many learning settings have emerged with the fast development of machine learning, including semi/self/un-supervised learning which are under different scenarios according to different label information provided. However, it turns out that we there is subtle but important connection between them and adversarial robustness. The most crucial point is that

Clean data and adversarial data can be seen as two different data source domains.

With this insight, many training methods can be proposed naturally to boost the robustness of deep learning models.

2 Bridging Adversarial Robustness and Existing Learning Methods

2.1 Semi-supervised Learning

Virtual adversarial training (VAT) [7], is one of the most classical semi-supervised learning algorithms, it combines two loss term:

$$\mathbb{E}_{(x,y)\sim\mathcal{D}^l} \ell(\theta; x, y) + \lambda \mathbb{E}_{x\sim\mathcal{D}^l \bigcup \mathcal{D}^{ul}} \mathbb{D}\{p(y|x)||p(y|x')\} \tag{2}$$

to smooth the prediction of classifier on both labeled data $\mathcal{D}^l$ unlabeled data $\mathcal{D}^{ul}$, where x' is a perturbed version of x. With our previously mentioned insight, if we take advantage of this technique into adversarial training, then we have:

$$\min_{\theta} \mathbb{E}_{(x,y)\sim\mathcal{D}} \ell(\theta; x, y) + \lambda \cdot \mathbb{E}_{x\sim\mathcal{D}^l \bigcup \mathcal{D}^{ul}} \mathbb{D}\{p(y|x)||p(y|x')\} \tag{3}$$

where x' is the *adversarial examples* of original input x. This has been adopted by [2] [12] [10].

Preprint. Under review.

2.2 Self-supervised Learning

In [3], treating the data as unlabeled data, an auxiliary loss proposed in self-supervised learning is used in adversarial training:

$$\ell_{\mathrm{SS}} = \frac{1}{4}\left[\sum_{r \in \{0^\circ, 90^\circ, 180^\circ, 270^\circ\}} \ell_{\mathrm{CE}}\left(\theta; R_r(x), r\right)\right] \tag{4}$$

where $R_r(x)$ rotates input x for r degree and a 4-way auxiliary classifier is used to predict the rotation degree r. Combine this with previous adversarial training loss can boost the robustness performance.

2.3 Un-supervised Learning

Here we mainly focus on the application of un-supervised domain adaptation algorithms in adversarial robustness. [9] treat the adversarial examples and clean input data like they are from two different source data domain, trying to make the feature extractor to extract the same feature from these two distributions via additional regularization term from domain adaption field. For example, deep CORAL loss [11], Multi-kernel MMD loss [5] [1] can be used with existing adversarial training technique.

References

[1] Karsten M Borgwardt, Arthur Gretton, Malte J Rasch, Hans-Peter Kriegel, Bernhard Schölkopf, and Alex J Smola. Integrating structured biological data by kernel maximum mean discrepancy. *Bioinformatics*, 22(14):e49–e57, 2006.

[2] Yair Carmon, Aditi Raghunathan, Ludwig Schmidt, Percy Liang, and John C Duchi. Unlabeled data improves adversarial robustness. *arXiv preprint arXiv:1905.13736*, 2019.

[3] Dan Hendrycks, Mantas Mazeika, Saurav Kadavath, and Dawn Song. Using self-supervised learning can improve model robustness and uncertainty. *arXiv preprint arXiv:1906.12340*, 2019.

[4] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436, 2015.

[5] Mingsheng Long, Yue Cao, Jianmin Wang, and Michael I Jordan. Learning transferable features with deep adaptation networks. *arXiv preprint arXiv:1502.02791*, 2015.

[6] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018.

[7] Takeru Miyato, Shin-ichi Maeda, Masanori Koyama, and Shin Ishii. Virtual adversarial training: a regularization method for supervised and semi-supervised learning. *IEEE transactions on pattern analysis and machine intelligence*, 41(8):1979–1993, 2018.

[8] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard. Deepfool: a simple and accurate method to fool deep neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2574–2582, 2016.

[9] Chuanbiao Song, Kun He, Liwei Wang, and John E Hopcroft. Improving the generalization of adversarial training with domain adaptation. *arXiv preprint arXiv:1810.00740*, 2018.

[10] Robert Stanforth, Alhussein Fawzi, Pushmeet Kohli, et al. Are labels required for improving adversarial robustness? *arXiv preprint arXiv:1905.13725*, 2019.

[11] Baochen Sun and Kate Saenko. Deep coral: Correlation alignment for deep domain adaptation. In *European Conference on Computer Vision*, pages 443–450. Springer, 2016.

[12] Runtian Zhai, Tianle Cai, Di He, Chen Dan, Kun He, John Hopcroft, and Liwei Wang. Adversarially robust generalization just requires more unlabeled data. *arXiv preprint arXiv:1906.00555*, 2019.