Project Proposal

Amirhossein Sohrabbeig, Masooma Nazari

March 2022

1 Goal

How can a trained model be quickly adjusted to investigate the achievable trade-offs between its standard and robust accuracies without having to re-train it many times? We try to answer this question during our studies for this project by providing a framework based on [1]. This framework is built on top of adversarial training, but with a novel, model-conditional training approach added on top of it. The weight hyper-parameter of the robust loss term is viewed as a user-specified input in our framework. It samples data points and model instances from the objective family, which are parameterized by distinct loss term weights, during training. As a result, the model learns to base its behavior and output on the hyper-parameter in question. As a result, we could freely switch between standard and robust accuracies in the same model at the testing time by simply swapping the hyper-parameters as inputs.

2 Problem Statement

Recent work has demonstrated that deep neural networks are vulnerable to adversarial attacks—inputs that are almost indistinguishable from natural data and yet classified incorrectly by the network. Adversarial training (AT) has been proposed to solve this problem, which is a powerful strategy for improving model robustness by including adversarial data in model improvement. The problem of Adversarial training can be expressed as a MinMax optimization problem [2]:

$$\mathcal{L}(\theta) = E_{(x,y)\sim D}[\max_{\delta \in S} L(f_\theta(x+\delta), y)] \quad (1)$$

Where f_θ indicates the training model parameterized by θ, and δ is the adversarial perturbation overlaid on the input. Inner maximization seeks adversarial samples that maximize loss within the given permutation range, whereas outer minimization seeks adversarial samples that minimize loss.

While adversarial defense methods are gaining increasing attention and popularity in safety/security- critical applications, their downsides are also noteworthy. To begin with, most adversarial defence approaches, such as adversarial training, come at the expense of standard accuracy [3]. Many investigations have proven both conceptually and experimentally that there is an inherent accuracy-robustness trade-off [4, 5].

In many applications, a significant loss of standard performance on clean data is unacceptable and could result in serious consequences. Instead, model resilience against excessively malicious adversarial attacks is prized more than conventional performance. We want to know how much may model robustness be improved without compromising standard performance.

3 Previous Work

At the moment, adversarial training is largely regarded as the most successful strategy for improving the adversarial robustness of deep learning models in practice [6]. However, adversarial training still has a long way to go before completely handling adversarial attacks. On MNIST, the current adversarial training methods can build a robust model with a worst-case accuracy of roughly 90% [2]. Adversarial training only reaches around 45 percent and 40 percent on SVHN for slightly more complex datasets, such as CIFAR-10, which is far from sufficient. In addition, adversarial training reduces the ability of deep learning models to generalize [7].

The concept of adversarial regularisation was first introduced in [8]. They also included a regularization component in the objective function, which is based on FGSM and is represented as $\mathcal{L}(\theta, x + \epsilon sign(\nabla_x \mathcal{L}(\theta, x, y))$. [9] scaled up this FGSM-based regularization term to ImageNet by adjusting the ratio of adversarial examples in batches. The success of their methods is demonstrated using single-step attacks, as they claim that the linearity of neural networks is due to the presence of adversarial examples [8]. [10]

computed the absolute difference between the adversarial loss and its first-order Taylor expansion, indicating that more robust models have lower local linearity values. For adversarial robustness, they used a Local Linearity Regularization instead of the FGSM-based regularisation.

[4] decomposed the robust error R_{rob} as the sum of natural error R_{nat} and boundary error R_{db}, which was different from earlier methods. When the distance between data and the decision border is sufficiently tiny (less than epsilon), boundary error happens, which is also why adversarial examples exist. As a result, they proposed TRADES as a method to reduce R_{db} by solving the following problem:

$$\min_{f} E\{\mathcal{L}(f(x), y) + \max_{x' \in B(x,\epsilon)} \mathcal{L}(f(x), f(x'))/\lambda)\} \tag{2}$$

where λ is a regularisation coefficient that determines the degree of regularisation. Such decomposition has been shown to be effective, with TRADES surpassing PGD-AT on CIFAR-10 and error rates lowered by 10%. The regularisation term in TRADES is designed to push natural instances and their adversarial equivalents together, regardless of whether natural data is classified correctly or not. [11] looked into the impact of misclassified examples and suggested Misclassification Aware adveRsarial Training (MART), which focuses on misclassified examples with weights of $1 - P_y(x, \theta)$, where $P_y(x, \theta)$ is the likelihood of the ground truth label y.

Imperceptible noises could cause significant changes in feature space due to the amplification of deep models [8]. Some studies look at adversarial training via the lens of representation. Adversarial Logit Pairing (ALP), which encourages logits for pairs of cases to be similar, was proposed by [9]. Unfortunately, ALP is initially ineffective due to incorrect adversarial training aims [12]. [13] also used the popular triplet loss for regularisation, which uses adversarial instances as anchors, to improve the alignment of natural data representations and their adversarial counterparts.

Adversarial regularisation is one of the most important types of adversarial training [14]. Adversarial regularisation is more flexible than the original formulation of adversarial training, and it necessitates a thorough understanding of adversarial robustness. Furthermore, the breakdown of robust error allows unlabeled data to improve adversarial robustness.

4 Approach

Our goal to learn a distribution of DNNs $f(., \lambda; \theta) \sim F$, conditioned on $\lambda \sim P_\lambda$, so that different DNNs sampled from this learned distribution, while sharing the set of parameters θ, could have different accuracy-robustness trade-offs depending on the input λ.

While standard DNN training samples data, OAT proposes to also sample one $f(., \lambda; \theta) \sim F$ per data. Each time, we set f to be conditioned on λ: concretely, it will take a hyperparameter $\lambda \sim P_\lambda$ as the input, while using this same λ to modulate the current Adversarial Training loss function:

$$\mathcal{L}(x, y, \lambda) = E_{(x,y)\sim D, \lambda \sim P_\lambda}[(1 - \lambda)\mathcal{L}_c + \lambda \mathcal{L}_a] \tag{3}$$

The gradient w.r.t. each (x, y) is generated from the loss parameterized by the current λ. Therefore, OAT essentially optimizes a dynamic loss function, with λ varying per sample, and forces weight sharing across iterations.

5 Evaluation

5.1 Datasets and Models

We evaluate our proposed method on WRN-16-8 using SVHN and ResNet34 using CIFAR-10. We also include the STL-10 dataset which has fewer training images but higher resolution using WRN-40-2. All images are normalized to [0, 1].

5.2 Evaluation metrics

Standard Accuracy (SA) refers to the accuracy of categorization on the original clean test set. The (default) accuracy is denoted by SA. Robust Accuracy (RA) is the accuracy with which adversarial images produced from the original test set are classified. The model's robustness is measured by RA. We develop the SA-RA frontier, an empirical Pareto frontier between a model's achievable accuracy and robustness, by measuring the SAs and RAs of the models dedicatedly trained by PGD-AT with varied (fixed) lambda values, to more directly evaluate the trade-off between SA and RA. In the our trained models, we might also change lambda input, and ideally, the resulting SA-RA trade-off curve should be as close to the SA-RA frontier as possible.

5.3 Existing codes

In this project, we will take advantage of the following github repositories:

1. https://github.com/yaodongyu/TRADES
2. https://github.com/p-lambda/robust_tradeoff
3. https://github.com/VITA-Group/Once-for-All-Adversarial-Training

This list might be more comprehensive in the final report.

References

[1] Haotao Wang, Tianlong Chen, Shupeng Gui, TingKuei Hu, Ji Liu, and Zhangyang Wang. Once-for-all adversarial training: In-situ tradeoff between robustness and accuracy for free. *Advances in Neural Information Processing Systems*, 33:7449–7461, 2020.

[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] Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Alexander Turner, and Aleksander Madry. Robustness may be at odds with accuracy. *arXiv preprint arXiv:1805.12152*, 2018.

[4] Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric Xing, Laurent El Ghaoui, and Michael Jordan. Theoretically principled trade-off between robustness and accuracy. In *International conference on machine learning*, pages 7472–7482. PMLR, 2019.

[5] Ludwig Schmidt, Shibani Santurkar, Dimitris Tsipras, Kunal Talwar, and Aleksander Madry. Adversarially robust generalization requires more data. *Advances in neural information processing systems*, 31, 2018.

[6] Anish Athalye and Nicholas Carlini. On the robustness of the cvpr 2018 white-box adversarial example defenses. *arXiv preprint arXiv:1804.03286*, 2018.

[7] Jacob Buckman, Aurko Roy, Colin Raffel, and Ian Goodfellow. Thermometer encoding: One hot way to resist adversarial examples. In *International Conference on Learning Representations*, 2018.

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

[9] Harini Kannan, Alexey Kurakin, and Ian Goodfellow. Adversarial logit pairing. *arXiv preprint arXiv:1803.06373*, 2018.

[10] Chongli Qin, James Martens, Sven Gowal, Dilip Krishnan, Krishnamurthy Dvijotham, Alhussein Fawzi, Soham De, Robert Stanforth, and Pushmeet Kohli. Adversarial robustness through local linearization. *Advances in Neural Information Processing Systems*, 32, 2019.

[11] Yisen Wang, Difan Zou, Jinfeng Yi, James Bailey, Xingjun Ma, and Quanquan Gu. Improving adversarial robustness requires revisiting misclassified examples. In *International Conference on Learning Representations*, 2019.

[12] Logan Engstrom, Andrew Ilyas, and Anish Athalye. Evaluating and understanding the robustness of adversarial logit pairing. *arXiv preprint arXiv:1807.10272*, 2018.

[13] Chengzhi Mao, Ziyuan Zhong, Junfeng Yang, Carl Vondrick, and Baishakhi Ray. Metric learning for adversarial robustness. *Advances in Neural Information Processing Systems*, 32, 2019.

[14] Uri Shaham, Yutaro Yamada, and Sahand Negahban. Understanding adversarial training: Increasing local stability of supervised models through robust optimization. *Neurocomputing*, 307:195–204, 2018.