Recent Advances in Adversarial Training for Adversarial Robustness

Tao Bai [1] Jinqi Luo [1] Jun Zhao [1] Bihan Wen [1] Qian Wang [2]

[1]Nanyang Technological University [2]Wuhan University

In this paper, we give a detailed overview of adversarial training in image classification, which has attracted extensive attention of researchers. We believe that this survey can provide up-to-date findings and developments happening on adversarial training. Notably, we carefully review and analyze adversarial training with a novel taxonomy, uniquely discuss the poor generalization ability from different perspectives, and present future research directions.

Definition of Adversarial Training

Adversarial training [Madry *et al.*, 2018] is a min-max problem: the inner maximization problem is finding the worst-case samples for the given model, and the outer minimization problem is to train a model robust to adversarial examples. Formally it is defined as:

$$\min_{\theta} \mathbb{E}_{(x,y)\sim\mathcal{D}} \left[\max_{\delta\in B(x,\varepsilon)} \mathcal{L}_{ce}(\theta, x+\delta, y) \right], \tag{1}$$

where $(x, y) \sim \mathcal{D}$ represents training data sampled from distribution $\mathcal{D}$ and $B(x, \varepsilon)$ is the allowed perturbation set, expressed as $B(x, \varepsilon) := \{x + \delta \in \mathbb{R}^{h\times w\times c} \mid \|\delta\|_p \leq \varepsilon\}$.

One popular method for solving this optimization problem is using Projected Gradient Descent (PGD) [Madry *et al.*, 2018], which is shown below:

$$x^{t+1} := \mathrm{Proj}_{x+B(x,\varepsilon)} \left(x^t + \alpha \,\mathrm{sign}\left(\nabla_{x^t} \mathcal{L}_{ce}\left(\theta, x^t, y\right)\right)\right), \tag{2}$$

where t is the current step and α is the step size. Madry *et al.* (2018) investigated the inner maximization problem from the landscape of adversarial examples and gave both theoretical and empirical proofs of local maxima's tractability with PGD. Thus, PGD-based adversarial training became a critical benchmark and is regarded as the standard way to do adversarial training in practice.

Taxonomy of Adversarial Training for Adversarial Robustness

A summary of selected adversarial training methods is provided in Table 1.

- **Adversarial Regularization** uses both clean and adversarial data for training.
- **Curriculum-based Adversarial Training** utilizes weak adversarial examples.
- **Ensemble Adversarial Training** augments training data with adversarial examples generated from multiple target models.
- **Adversarial Training with Adaptive ϵ** generates proper adversarial examples with the intrinsic adversarial robustness of individual samples taken into consideration.
- **Adversarial Training with Semi/Unsupervised Learning** utilizes larger datasets for adversarial training.
- **Efficient Adversarial Training** promotes the efficiency of adversarial training.
- **Other Variants**

Generalization Problems in Adversarial Training

Standard Generalization

Adversarial training is observed that hurts standard accuracy badly [Madry *et al.*, 2018]. **On the one hand,** one popular viewpoint is the trade-off between adversarial robustness and standard accuracy. **On the other hand,** it is confirmed the existence of adversarial examples on the manifold of natural data, adversarial robustness on which is equivalent to generalization.

Adversarially Robust Generalization

The phenomenon that adversarially trained models do not perform well on adversarially perturbed test data is firstly observed in [Madry *et al.*, 2018]. In other words, there is a large gap between the training accuracy and test accuracy on adversarial data.

Generalization on Unseen Attacks

Adversarially trained models, which are robust to a specific attack, *e.g.*, l_∞ adversarial examples, can be circumvented easily by different types of attacks, *e.g.*, other l_p norms, or larger ϵ, or different target models [Kang *et al.*, 2019].

Future Direction

- **Min-Max Optimization in Adversarial Training.** The robustness of adversarially trained models is not guaranteed [Kang *et al.*, 2019].
- **Overfitting in Adversarial Training.** The generalization gap between adversarial training accuracy and testing accuracy is very large.
- **Beyond Adversarial Training.** Though many theories have been proposed for improving adversarial training, it is undeniable that these improvements are less effective than claimed [Pang *et al.*, 2021].
- **Adversarial Training in Other Domains.** Adversarial training has been successfully applied to texts [Miyato *et al.*, 2017], graphs [Dai *et al.*, 2019], audios [Pandey and Wang, 2018], and reinforcement learning [Pattanaik *et al.*, 2017].

Taxonomy	Publication	Model Architecture	Attack	ϵ	Dataset	Accuracy
Adversarial Regularization	[Qin *et al.*, 2019]	ResNet-152	PGD_{50}	4/255	ImageNet	47.00%
	[Zhang *et al.*, 2019b]	Wide ResNet	CW_{10}	0.031/1	CIFAR-10	84.03%
	[Wang *et al.*, 2020]	ResNet-18	PGD_{20}	8/255	CIFAR-10	55.45%
	[Kannan *et al.*, 2018]	InceptionV3	PGD_{10}	16/255	ImageNet	27.90%
	[Mao *et al.*, 2019]	Wide ResNet	PGD_{20}	8/255	CIFAR-10	50.03%
Curriculum	[Zhang *et al.*, 2020]	Wide ResNet	PGD_{20}	16/255	CIFAR-10	49.86%
	[Cai *et al.*, 2018]	DenseNet-161	PGD_{7}	8/255	CIFAR-10	69.27%
	[Wang *et al.*, 2019]	8-Layer ConvNet	PGD_{20}	8/255	CIFAR-10	42.40%
Ensemble	[Pang *et al.*, 2019]	Wide ResNet	PGD_{10}	0.005	CIFAR-100	32.10%
	[Kariyappa and Qureshi, 2019]	ResNet-20	PGD_{30}	0.09/1	CIFAR-10	46.30%
	[Yang *et al.*, 2020]	ResNet-20	PGD_{20}	0.01/1	CIFAR-10	52.40%
Adaptive ϵ	[Balaji *et al.*, 2019]	ResNet-152	PGD_{1000}	8/255	ImageNet	59.28%
	[Ding *et al.*, 2020]	Wide ResNet	PGD_{100}	8/255	CIFAR-10	47.18%
	[Cheng *et al.*, 2020]	Wide ResNet	PGD_{20}	8/255	CIFAR-10	73.38%
Semi-Unsupervised	[Alayrac *et al.*, 2019]	Wide ResNet	FGSM	8/255	CIFAR-10	62.18%
	[Carmon *et al.*, 2019]	Wide ResNet	PGD_{10}	8/255	CIFAR-10	63.10%
	[Zhai *et al.*, 2019]	Customized ResNet	PGD_{7}	8/255	CIFAR-10	42.48%
	[Hendrycks *et al.*, 2019]	Wide ResNet	PGD_{20}	0.3/1	ImageNet	50.40%
Efficient	[Shafahi *et al.*, 2019]	Wide ResNet	PGD_{100}	8/255	CIFAR-10	46.19%
	[Wong *et al.*, 2020]	ResNet-50	PGD_{40}	2/255	ImageNet	43.43%
	[Andriushchenko and Flammarion, 2020]	ResNet-50	PGD_{50}	2/255	ImageNet	41.40%
	[Kim *et al.*, 2021]	PreActResNet-18	FGSM	8/255	CIFAR-10	50.50%
	[S. and Babu, 2020]	Wide ResNet	PGD_{40}	8/255	MNIST	88.51%
	[Song *et al.*, 2019]	Customized ConvNet	PGD_{20}	4/255	CIFAR-10	58.10%
	[Vivek and Babu, 2020]	Wide ResNet	PGD_{100}	0.3/1	MNIST	90.03%
	[Huang *et al.*, 2020]	Wide ResNet	PGD_{20}	8/255	CIFAR-10	45.80%
	[Zhang *et al.*, 2019a]	Wide ResNet	PGD_{20}	8/255	CIFAR-10	47.98%
Others	[Dong *et al.*, 2020]	Wide ResNet	PGD_{20}	8/255	CIFAR-100	29.40%
	[Wang and Zhang, 2019]	Wide ResNet	CW_{200}	4/255	CIFAR-10	60.30%
	[Zhang and Wang, 2019]	Wide ResNet	PGD_{20}	8/255	CIFAR-100	47.20%
	[Pang *et al.*, 2020]	Wide ResNet	PGD_{500}	8/255	CIFAR-10	60.75%
	[Lee *et al.*, 2020]	PreActResNet-18	PGD_{20}	8/255	Tiny ImageNet	20.31%
Benchmark	[Madry *et al.*, 2018]	ResNet-50	PGD_{20}	8/255	CIFAR-10	45.80%

Table 1. A summary of experimental results for various adversarial training methods. All the attacks are under l_∞ norm.

Conclusion

In this paper, we present recent advances of adversarial training methods for adversarial robustness. To our best knowledge, for the first time, we review adversarial training with a novel taxonomy and discuss the generalization problem in adversarial training. We also summarize the benchmarks and provide performance comparisons of different methods. Despite extensive efforts, the vulnerability of deep learning models to adversarial examples has not been completely solved by adversarial training and several open problems remain yet to solve.

Acknowledgements

This paper is supported by 1) Singapore Ministry of Education Academic Research Fund Tier 1 RG128/18, Tier 1 RG115/19, Tier 1 RT07/19, Tier 1 RT01/19, Tier 1 RG24/20, and Tier 2 MOE2019-T2-1-176, 2) NTU-WASP Joint Project, 3) Singapore NRF National Satellite of Excellence, Design Science and Technology for Secure Critical Infrastructure NSoE DeST-SCI2019-0012, 4) AI Singapore (AISG) 100 Experiments (100E) programme, and 5) NTU Project for Large Vertical Take-Off & Landing (VTOL) Research Platform. Bihan Wen was supported in part by the National Research Foundation (NRF), Singapore, through the Singapore Cybersecurity Consortium (SGCSC) Grant Office, under SGCSC_Grant_2019-S01. Qian Wang's work was partially supported by the NSFC under Grants U20B2049 and 61822207. We would like to express our very great appreciation to Prof Bo Li and Mr. Qi Bi for their valuable and constructive suggestions of this research work.