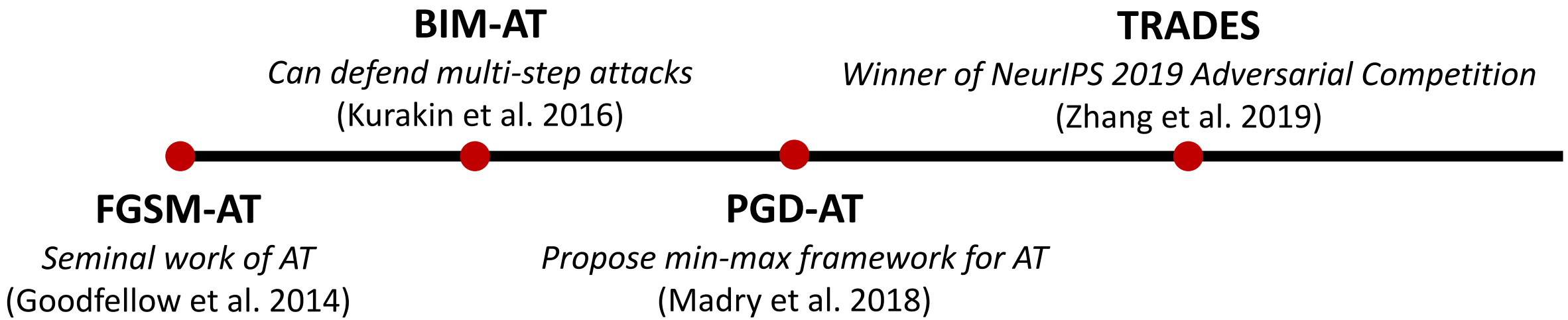
Bag of Tricks for Adversarial Training

Tianyu Pang, Xiao Yang, Yinpeng Dong, Hang Su, Jun Zhu

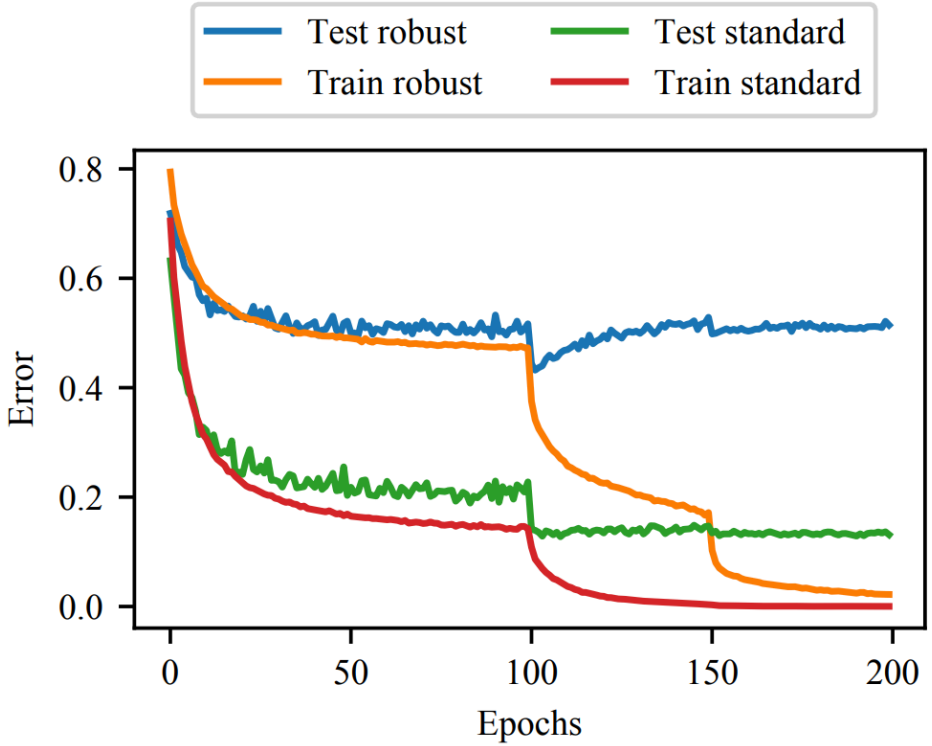
Department of Computer Science and Technology
Tsinghua University

Milestones of adversarial training frameworks (2014-2019)



What happened in 2020?

Rice et al. (ICML 2020) find that simply early stopping the training process of PGD-AT can attain the gains from almost all the previously proposed improvements, including the state-of-the-art TRADES.



(From Rice et al.)

- ***TRADES also applied early stopping by decaying learning rate at 75th epoch and used the checkpoint of 76th epoch.***

What happened in 2020?

Gowal et al. (2020) find that TRADES actually performs better than PGD-AT

Key takeaways. Contrary to the suggestion of Rice et al. (2020) (i.e., *"the original PGD-based adversarial training method can actually achieve the same robust performance as state-of-the-art method"*, see Sec. 2.1), TRADES (when combined with early-stopping – as our setup dictates) is more competitive than classical adversarial training. The results also highlight the importance of strong evaluations beyond PGD^{20} (including evaluations of the validation set used for early stopping).

(From Gowal et al.)

What happened in 2020?

Gowal et al. (2020) find that TRADES actually performs better than PGD-AT

> **Key takeaways.** Contrary to the suggestion of Rice et al. (2020) (i.e., *"the original PGD-based adversarial training method can actually achieve the same robust performance as state-of-the-art method"*, see Sec. 2.1), TRADES (when combined with early-stopping – as our setup dictates) is more competitive than classical adversarial training. The results also highlight the importance of strong evaluations beyond PGD^{20} (including evaluations of the validation set used for early stopping).

(From Gowal et al.)

Zhang et al. (2018): TRADES performs better than PGD-AT
Rice et al. (2020): PGD-AT performs better than TRADES
Gowal et al. (2020): TRADES performs better than PGD-AT

Paradox???

Who is wrong? Nobody

Zhang et al. (2018):
TRADES (weight decay 2×10^{-4})
PGD-AT (weight decay 2×10^{-4})

Rice et al. (2020):
TRADES (weight decay 2×10^{-4})
PGD-AT (weight decay 5×10^{-4})

Gowal et al. (2020):
TRADES (weight decay 5×10^{-4})
PGD-AT (weight decay 5×10^{-4})

Slightly different values of weight decay can lead to largely different conclusions in the adversarial setting!

Overlooked training settings could affect our evaluations on the defenses, especially in public benchmarks.

Training settings in previous work are highly inconsistent

Method	l.r.	Total epoch (l.r. decay)	Batch size	Weight decay	Early stop (train / attack)	Warm-up (l.r. / pertub.)
Madry et al. (2018)	0.1	200 (100, 150)	128	2×10^{-4}	No / No	No / No
Cai et al. (2018)	0.1	300 (150, 250)	200	5×10^{-4}	No / No	No / Yes
Zhang et al. (2019b)	0.1	76 (75)	128	2×10^{-4}	Yes / No	No / No
Wang et al. (2019)	0.01	120 (60, 100)	128	1×10^{-4}	No / Yes	No / No
Qin et al. (2019)	0.1	110 (100, 105)	256	2×10^{-4}	No / No	No / Yes
Mao et al. (2019)	0.1	80 (50, 60)	50	2×10^{-4}	No / No	No / No
Carmon et al. (2019)	0.1	100 (cosine anneal)	256	5×10^{-4}	No / No	No / No
Alayrac et al. (2019)	0.2	64 (38, 46, 51)	128	5×10^{-4}	No / No	No / No
Shafahi et al. (2019b)	0.1	200 (100, 150)	128	2×10^{-4}	No / No	No / No
Zhang et al. (2019a)	0.05	105 (79, 90, 100)	256	5×10^{-4}	No / No	No / No
Zhang & Wang (2019)	0.1	200 (60, 90)	60	2×10^{-4}	No / No	No / No
Atzmon et al. (2019)	0.01	100 (50)	32	1×10^{-4}	No / No	No / No
Wong et al. (2020)	0∼0.2	30 (one cycle)	128	5×10^{-4}	No / No	Yes / No
Rice et al. (2020)	0.1	200 (100, 150)	128	5×10^{-4}	Yes / No	No / No
Ding et al. (2020)	0.3	128 (51, 77, 102)	128	2×10^{-4}	No / No	No / No
Pang et al. (2020a)	0.01	200 (100, 150)	50	1×10^{-4}	No / No	No / No
Zhang et al. (2020)	0.1	120 (60, 90, 110)	128	2×10^{-4}	No / Yes	No / No
Huang et al. (2020)	0.1	200 (cosine anneal)	256	5×10^{-4}	No / No	Yes / No
Cheng et al. (2020)	0.1	200 (80, 140, 180)	128	5×10^{-4}	No / No	No / No
Lee et al. (2020)	0.1	200 (100, 150)	128	2×10^{-4}	No / No	No / No
Xu et al. (2020)	0.1	120 (60, 90)	256	1×10^{-4}	No / No	No / No

What we investigate

- Early stopping adversarial intensity
- Warmup w.r.t. learning rate or perturbation
- Batch size
- Mode for batch normalization when computing PGD
- Label smoothing
- Optimizer
- Weight decay
- Model architecture
- Activation function

Details can be found in our paper

Takeaways

Takeaways:
(i) Slightly different values of weight decay could largely affect the robustness of trained models;
(ii) Moderate label smoothing and linear scaling rule on l.r. for different batch sizes are beneficial;
(iii) Applying eval BN mode to craft training adversarial examples can avoid blurring the distribution;
(iv) Early stopping the adversarial steps or perturbation may degenerate worst-case robustness;
(v) Smooth activation benefits more when the model capacity is not enough for adversarial training.

- **Adversarial training is more sensitive to these usually overlooked hyperparameters, compared to standard training.**

- **Standardize the basic training setting enables fairer benchmarks.**

Thanks

Code: https://github.com/P2333/Bag-of-Tricks-for-AT