240228-시도한 방법들

시도한 방법들

- 시연에서 기본 baseline 비교 (WRN16-8 + CIFAR-10)

model	clean accuracy(%)	robust accuracy(%)
SGD	96.67	8.79
Adam	77.00	6.83
SAM	95.20	19.17
AT	69.6	60.32
TRADES	65.18	59.79

- SAM+AT (optimizing twice)
- SAM+AT (bi-level optimization?)
- Label smoothing
- Suggesting loss at loss landscape view

비교 논문 및 baseline

- Marc Khoury et al., (2018). On the Geometry of Adversarial Examples. Arxiv
- SAM
  - Pierre Foret et al., (2021). Sharpnesss-Aware Minimization for Efficiently Improving Generalization. ICLR
  - Zeming Wei et al., (2023). Sharpness-Aware Minimization Alopne can Improve Adversarial Robustness. ICML workshop
- SWAD
  - Junbum Cha et al., (2021). SWAD: Domain Generalization by Seeking Flat Minima. NeurIPS
- SGD with large step
  - Maksym Andriushchenko et al., (2023). SGD with Large Step Size Learns Sparse Features. ICML
- AWP
  - Dongxian Wu et al., (2020). Adversarial Weight Perturbation Helps Robust Generalization. NeurIPS
- Trade off
  - Dimitris Tsipras et al., (2019). Robustness May Be at Odds with Accuracy. ICLR
- Defense survey
  - Zhuang Qian et al., (2022). A Survey of Robust Adversarial Training in Pattern Recognition: Fundamental, Theory, and Methodologies. ScienceDirect
  - FGSM-AT
    - Christian Szegedy et al., (2014). Intriguing Properties of Neural Networks. Arxiv
  - PGD-AT
    - Aleksander Madry et al., (2018). Towards Deep Learning Models Resistant to Adversarial Attacks. Arxiv
  - BAT
    - Jianyu Wang et al., (2019). Bilateral Adversarial Training: Towards Fast Training of More Robust Models Against Adversarial Attacks

	Clean accuracy	Adversarial accuracy with PGD	Memory	FLOPS/Training time
SGD				
Adam				
SAM				
FGSM-AT				
PGD-AT				
TRADES				
BAT				
AWP				
ours				

SAMAT에서 SAM을 사용한 이유와, 그냥 AT보다 좋은 이유

- Natrual network for CIFAR10
  - Clean accuracy: 92.20%
  - Robust accuracy: 0.00%
- PGD-7(ϵ=8)-AT
  - Clean accuracy: 79.57%
  - Adversarial accuracy: 41.93%
- SAM: model의 generalization 성능을 높여준다.
- AT: model의 robustness 성능을 높여주지만, clean accuracy가 낮아진다는 trade-off가 있다.
- 그냥 AT 보다 좋은지 확인이 먼저 필요..