Sharpness-Aware Minimization Alone can Improve Adversarial Robustness

Introduction

- SAM은 모델의 generalization을 높이기 위한 학습 방법이다.

$$\max_{||\epsilon|| \leq \rho} L(w+\epsilon) + \lambda ||w||_2^2$$

- Adversarial attack은 input에 작은 perturbatioin을 줌으로써 모델이 부정확한 예측을 하게 만드는 방법이다.
- 본 연구에서는 SAM이 adversarial examples를 어떻게 대응하는지를 보며 SAM의 robustness를 바라본다.
- 본 연구는 SAM만을 적용한 모델이 robustness를 높인다는 발견을 했으며, AT보다 계산적인 이득을 가지면서도 정확성을 잃지 않는 모습을 보여주었다.
- 본 논문에서는 두 가지의 질문을 제시한다. (Research Questions)
  - RQ1. SAM이 왜 adversarial robustness를 증가시키는가?
  - RQ2. SAM이 AT(adversarial training)의 경량의 대체제로 쓰일 수 있는가?
- 본 논문의 contribution은 다음과 같다.
  1. SAM을 사용하는 것은 clean accuracy의 손실 없이 adversarial robustness를 높일 수 있다.
  2. Non-robust features(NRF)를 제거하는 의의를 가진 SAM과 AT 사이의 관계를 논의한다. 둘은 natural accuracy와 robust accuracy 사이의 trade-off인 perturbation strength에서 차이가 있다.
  3. SAM이 특정 요구사하에서 AT의 경량 대체제가 될 수 있음을 제안한다.

Backgrounds

2.1. Sharpness awareness minimization

- (1997. Hochreiter & Schmidhuber et al. "Simplifying neural nets by discovering flat minima") loss에서의 flat minima가 model의 generalization을 확보한다는 제안 이후로 entropy SGD, stochastic weight averaging(SWA) 등 여러 시도가 있어 왔다.
- SAM의 좋은 generalization 능력은 다음 식이 갖는 제약에 의해 보장된다.

$$L_{\mathscr{D}}(w) <= \max_{||\epsilon|| \leq \rho} L(w+\epsilon) + h(\frac{\lambda ||w||_2^2}{\rho^2})$$

h: strictly increasing function

- Language model, fluid dynamics 등 많은 application이 존재한다.
- Adaptive SAM(ASAM)< Efficient SAM(ESAM), LookSAM, Sparse SAM(SSAM), Fisher SAM(FSAM) 등 많은 개선된 알고리즘이 존재한다.

Backgrounds

2.2. Adversarial robustness

- (2014. Goodfellow et al. "Explaining and harnessing adversarial examples") adversarial example의 발견 이후로 많은 관심이 있어 왔다.
- 다양한 adversarial attack 기법과 defense 접근법이 제안되었다.
- Adversarial training(AT)의 방법은 다음과 같은 식으로 표현된다.

$$\min_{w} E_{(x,y)\sim\mathcal{D}} \max_{||\delta||\leq\epsilon} L(w; x+\delta, y)$$

- 식의 max를 구할 땐 보통 PGD로 계산된다.

$$x^{t+1} = \Pi_{\mathcal{B}(x,\epsilon)}(x^t + \alpha \cdot \text{sign}(\nabla_x \ell(\theta; x^t, y)))$$

- AT는 computational overhead나 class-wise fairness, natural accuracy 감소 등을 주된 문제로 가지고 있다.
- AT는 robustness accuracy와 natural accuracy 간의 intrinsic trade-off를 가짐이 알려져 있다.
- 가장 대표적인 flat loss를 이용한 adversarial training은 AWP(Adversarial Weight Perturbation)이 있다.
- AWP는 input이나 feature space에 perturbation을 통해 AT에서의 SAM을 구현한다.
- AWP도 natural accuracy의 손실이 있다는 단점이 있다. 또 flat loss landscape가 좋은 robustness를 가진다는 점이 잘 설명되지 못했다.

Empirical understanding

- AT explicitly adds perturbations to input examples
- SAM perturbates the parameters to implicit data augmentation

Experiments

Table 1. Natural and Robust Accuracy evaluation on **CIFAR-100** dataset.

Method	Natural Accuracy	ℓ_∞-Robust Accuracy		ℓ_2-Robust Accuracy	
		$\epsilon = 1/255$	$\epsilon = 2/255$	$\epsilon = 16/255$	$\epsilon = 32/255$
ST	76.9	13.6	1.7	44.5	21.2
SAM ($\rho = 0.1$)	78.0	19.6	3.0	51.5	27.2
SAM ($\rho = 0.2$)	78.5	23.1	4.2	54.2	31.3
SAM ($\rho = 0.4$)	**78.7**	**28.3**	**6.5**	**57.0**	**36.2**
AT (ℓ_∞-$\epsilon = 1/255$)	**73.1**	**60.4**	46.6	**67.4**	**61.5**
AT (ℓ_∞-$\epsilon = 2/255$)	70.1	60.3	50.6	65.7	60.6
AT (ℓ_∞-$\epsilon = 4/255$)	66.2	59.3	**52.0**	62.8	58.9
AT (ℓ_∞-$\epsilon = 8/255$)	60.4	55.1	50.4	57.0	54.3
AT (ℓ_2-$\epsilon = 16/255$)	**74.8**	52.8	31.4	66.3	57.7
AT (ℓ_2-$\epsilon = 32/255$)	73.2	57.4	40.6	**67.1**	**61.0**
AT (ℓ_2-$\epsilon = 64/255$)	70.7	58.1	45.9	66.1	60.7
AT (ℓ_2-$\epsilon = 128/255$)	67.4	**58.2**	**48.7**	63.9	60.4