Adversarial attack에서의 Robust Feature 학습: 모델 강건성 향상을 위한 접근법

2024-2 AJOU SOFTCON

이 름 공종혁 지도교수 황원준

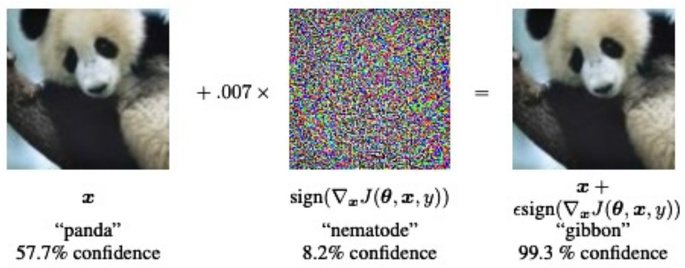
연구배경

Adversarial attack(적대적 공격)은 이미지에 사람이 눈으로는 식별하기 어려운 perturbation(노이즈)을 추가하여 CNN 모델에 치명적인 예측 오류를 유발하는 연구 분야이다. 이러한 공격 기법이 발전함에 따라, adversarial example(적대적 예제)을 활용해 모델의 robustness(강건성)를 향상시키는 adversarial training(적대적 학습) 연구도 함께 발전하였다. 기존 연구들은 사람의 눈에는 보이지 않지만 모델에 영향을 미치는 perturbation의 존재를 정의하고, 이를 이미지와 결합하여 adversarial example을 생성하였다. 또한, 이미지에 추가될 perturbation을 보다 빠르고 효율적으로 원하는 방향으로 유도하기 위해 다양한 adversarial attack 방법론이 개발되었다. 이러한 공격 방법론의 발전에 발맞춰, 모델의 robustness를 강화하기 위한 adversarial training 방법론 역시 지속적으로 발전해 왔다. 기존 adversarial training에서는 모델 학습에 adversarial example로 구성된 데이터셋을 활용하여 모델의 robustness를 향상시켰다. 그러나 이는 원본 데이터에 대한 성능의 저하를 동반한다. *Adversarial Examples Are Not Bugs, They Are Features* [1] 논문에서는 adversarial example의 생성 원인을 이미지의 robust feature와 non-robust feature의 존재를 통해 설명하였다. 논문은 non-robust feature가 인간에게는 무의미해 보일 수 있지만, 모델에는 예측에 유의미한 신호로 작용한다고 주장하였다. 본 연구에서는 이를 기반으로 robust feature와 non-robust feature가 model에 있어 실제로 어떤 영향을 끼치는지와 adversarial exmaple에 대해 효과적인 방어기법이 될 수 있는지에 대해 알아보고자 하였다.

관련 선행 연구

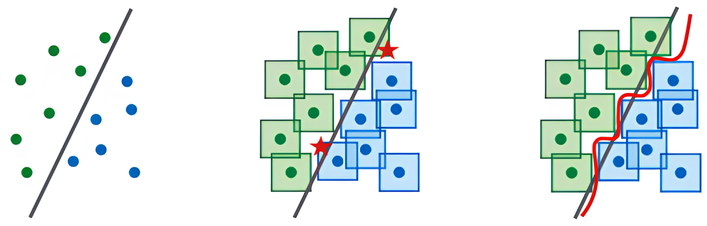
1. FGSM(Fast Gradient Sign Method) [2]

기존에는 모델의 비선형성, 과적합, 모델의 복잡성, 학습 데이터의 특성 등 다양한 요인이 적대적 예제의 원인으로 제시되었다. 하지만 본 논문에서는 고차원 특징 공간에서의 데이터 분포와 모델의 결정 경계의 특성이 적대적 공격에 영향을 미침을 제시하였다. 또한, 본 논문에서는 적대적 예제를 생성하는 알고리즘인 FGSM($\eta = \epsilon * sign(\nabla_x J(\theta, x, y))$)을 제안하였다.



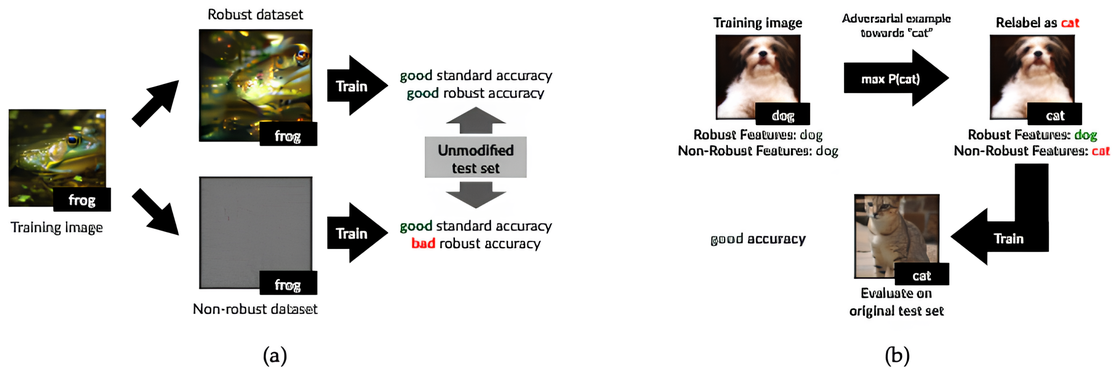
2. PGD(Projected Gradient Descent) [3]

FGSM의 one-step 방식을 확장하여 multi-step 방법을 제안하고, 이를 바탕으로 adversarial example을 생성하였다. 이 방법을 통해 FGSM과 달리 원하는 target에 대한 공격도 가능하며, PGD(Projected Gradient Descent)를 통해 생성된 adversarial example 데이터셋을 학습에 활용할 경우 FGSM보다 더 높은 성능을 보였다.



3.Robust & Non-robust feature

본 논문에서는 이미지에 robust feature와 non-robust feature가 존재함을 정의하였다. robust feature는 shape을 의미하며, non-robust feature는 texture를 의미한다. 사람은 shape을 기반으로 객체를 판단하는 반면, CNN은 convolution을 사용하기 때문에 간극이 긴 shape 보다 texture을 중심으로 학습한다. 이때 adversarial attack이 사람보다 model에 민감한 non-robust feature를 활용한다고 제안하였다.



연구 과정 및 결과

Adversarial attack이 모델 결과에 미치는 영향을 확인하기 위해, 원본 이미지와 perturbation이 적용된 이미지를 다양한 모델에서 비교하였다. epsilon은 5/255로 설정하였다. 실험의 결과는 아래 표와 같다.

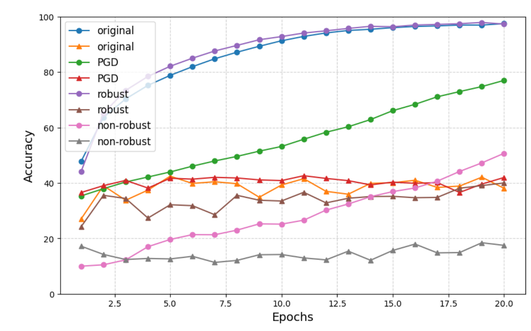
model	AlexNet		GoogleNet		resnet50	
Top1	388	95.63%	388	55.93%	388	95.12%
Top2	805	1.23%	384	34.37%	384	2.51%
Top3	258	0.42%	378	1.0%	387	1.24%
Top4	850	0.37%	380	0.52%	368	0.53%
Top5	267	0.34%	164	0.47%	375	0.13%

AlexNet	FGSM		DeepFool		PGD	
Top1	258	36.49%	267	29.49%	368	94.98%
Top2	435	6.04%	388	29.47%	260	1.43%
Top3	876	4.33%	266	10.21%	248	0.23%
Top4	794	2.87%	257	4.06%	270	0.23%
Top5	153	2.24%	805	3.88%	258	0.22%

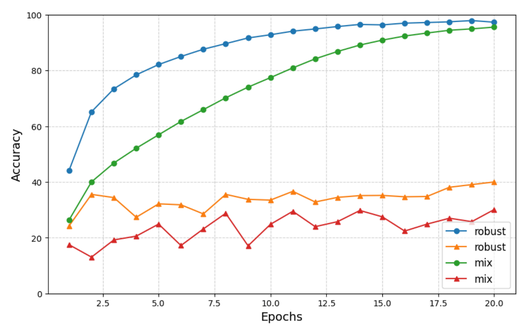
Googlenet	FGSM		DeepFool		PGD	
Top1	384	17.09%	384	50.72%	368	100.0%
Top2	382	9.59%	388	39.14%	369	0.0%
Top3	380	8.53%	378	1.19%	374	0.0%
Top4	368	7.88%	380	0.71%	384	0.0%
Top5	378	5.9%	368	0.51%	378	0.0%

resnet50	FGSM		DeepFool		PGD	
Top1	368	39.94%	384	44.71%	368	100.0%
Top2	382	19.25%	388	44.68%	369	0.0%
Top3	375	8.72%	368	3.92%	384	0.0%
Top4	384	5.92%	387	1.62%	296	0.0%
Top5	850	4.71%	375	0.98%	375	0.0%

기존 ResNet-34 모델을 CIFAR-10 데이터셋을 활용하여 네 가지 방법(origin, PGD example, robust feature, non-robust feature)으로 학습시킨 후, PGD로 증강된 데이터셋으로 평가하였다.



원본 데이터로 학습된 모델과 비교했을 때, PGD를 활용한 학습은 훈련 과정에서 비교적 완만한 정확도의 증가를 보였다. 반면, robust feature로 구성된 데이터셋은 훈련 데이터에서 더 빠르고 높은 정확도의 성능 향상을 나타냈다. 또한, PGD 및 robust feature를 기반으로 학습된 모델은 PGD 공격을 받은 테스트 데이터에서도 우수한 성능을 보였다. 이는 non-robust feature를 제거하고 robust feature를 활용하여 학습을 진행할 경우, 모델이 adversarial example에 대해 보다 robust해질 수 있음을 시사한다. Non-robust feautre의 양에 따른 성능 차이를 알기 위해 일정 비율 혼합하여 학습을 진행하였다.



다양한 비율 조합을 사용해 비교해 보았지만, non-robust feature 데이터가 추가될수록 train과 test의 accuracy는 점점 감소했다.

즉 image의 non-robust feautre를 제거하고 모델을 학습시키면 보다 좋은 성능을 얻을 수 있다.

참고 문헌

1. Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Logan Engstrom, Brandon Tran, Aleksander Madry, “Adversarial Examples Are Not Bugs, They Are Features”, neurips, 2019, 1035
2. Ian J. Goodfellow, Jonathon Shlens, Christian Szegedy, “Explaining and harnessing adversarial examples”, ICLR, 2015, 22848
3. Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, Adrian Vladu, “Towards Deep Learning Models Resistant to Adversarial”, ICLR, 2017, 13570