Towards Deep Learning Models Resistant to Adversarial Attacks

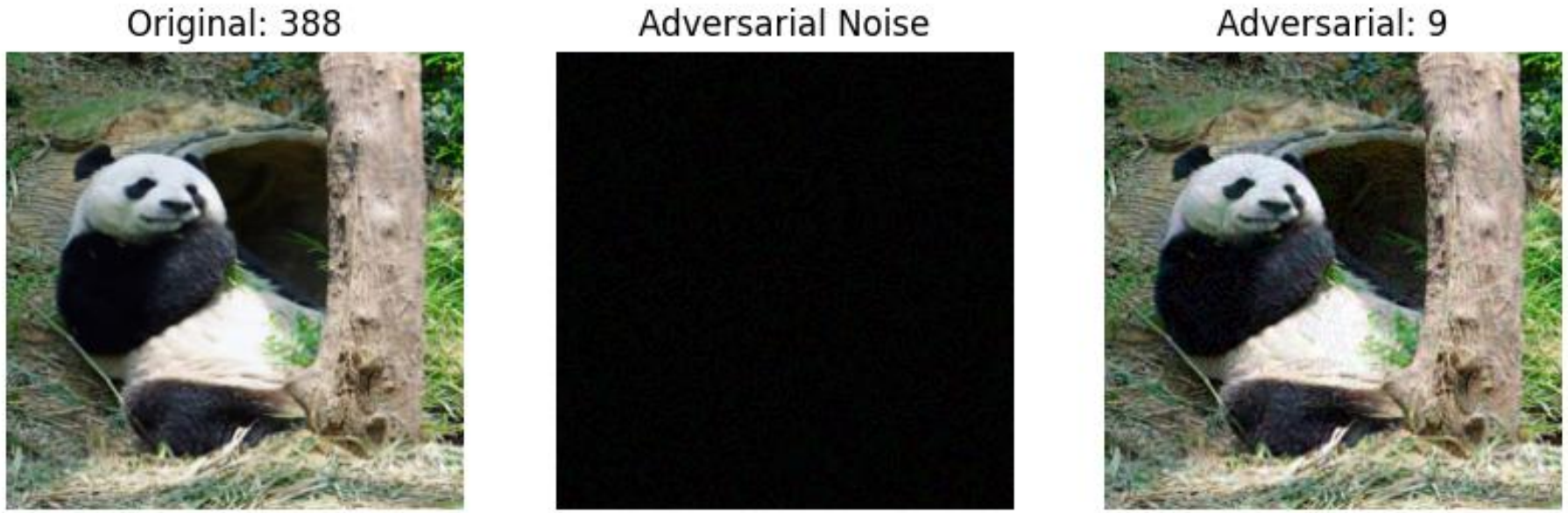
Aleksander Madry외 4인

Sunjun Hwang
RAISE Lab

배경 및 문제 제기

- 딥러닝 모델은 적대적 공격에 취약함.

딥러닝 모델이 작은 변화에 의해 잘못된 출력을 내도록 속이는 공격 기법



배경 및 문제 제기

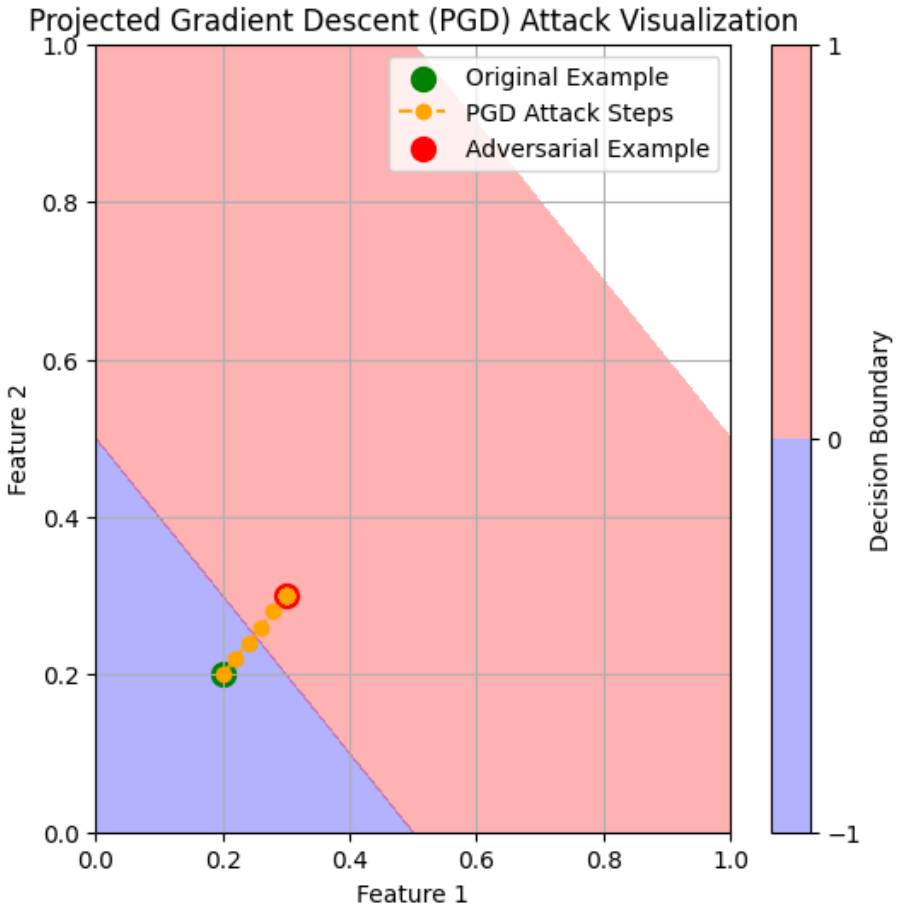
논문의 핵심 질문

1. 어떻게 하면 딥러닝 모델을 **적대적 공격에 강하게 만들 수 있을까?**

2. 기본 방어 기법의 한계점은 무엇인가?

논문의 핵심 기여, 실험 방법

논문의 주요 기여



PGD

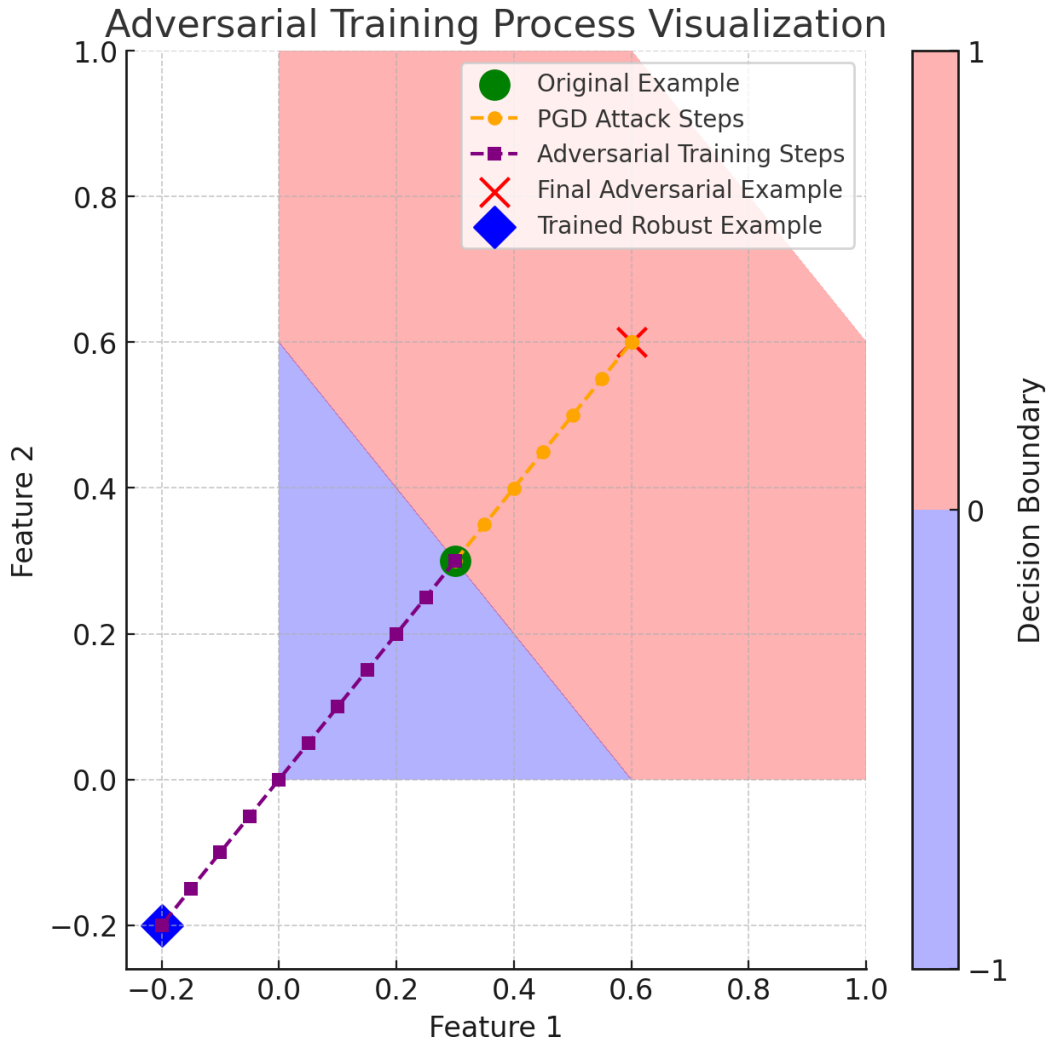
FGSM

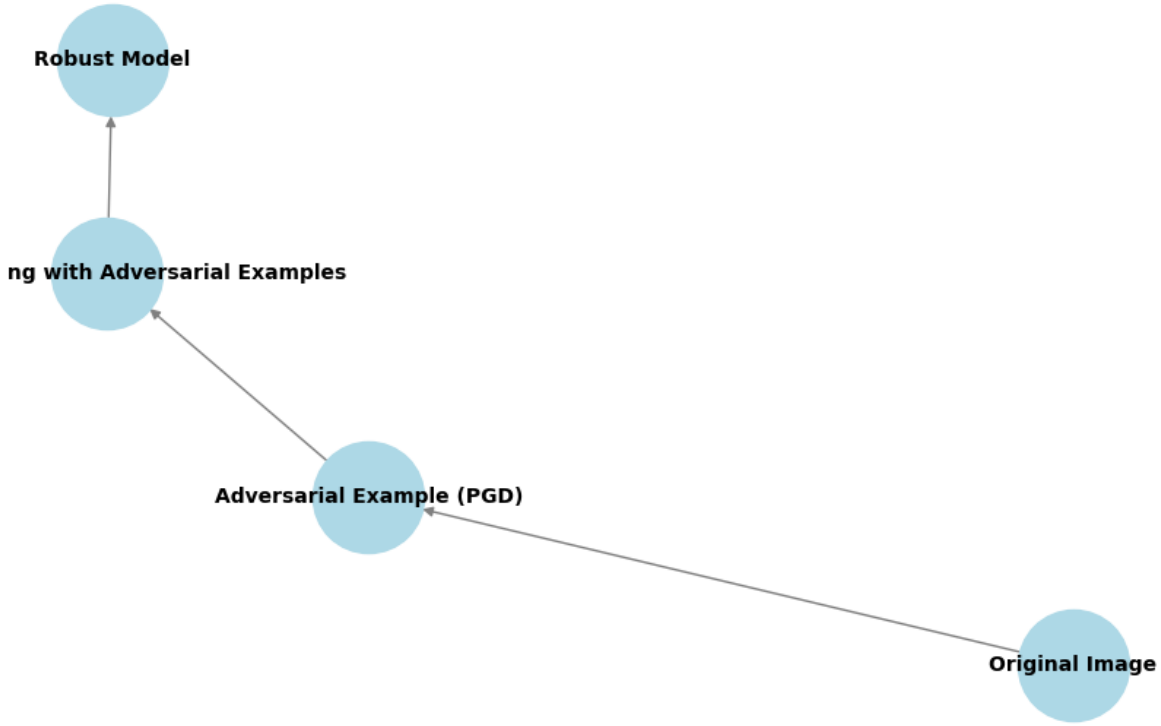
$$x = x + \epsilon sign(\nabla_x J(\theta, x, y))$$

PGD

$$x^{t+1} = \Pi_{x+S}(x^t + \alpha sign(\nabla_x J(\theta, x, y)))$$

Adversarial Training 과정

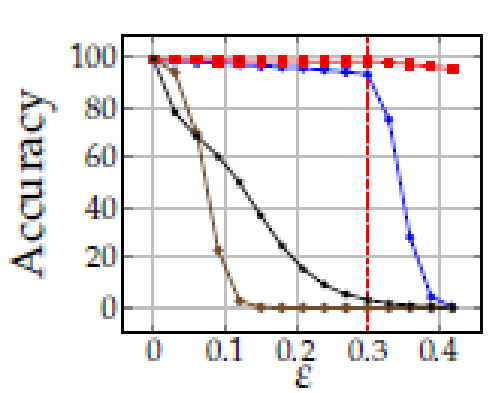




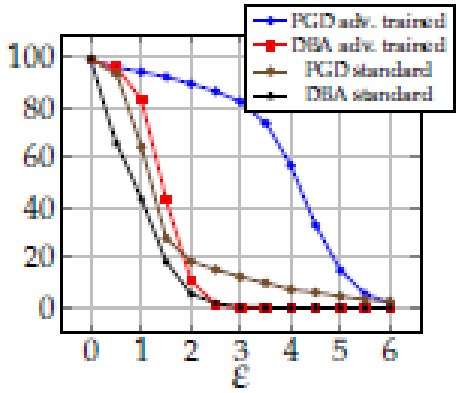
일반 모델 VS Adversarial Training 모델

모델 유형	훈련 방식	적대적 공격에 대한 강건성	특징
일반 모델	일반적인 데이터로 학습	취약	적대적 예제에 쉽게 속음
Adversarial Trained Model	PGD 공격을 포함하여 훈련	향상됨	적대적 예제에 대한 내성이 증가

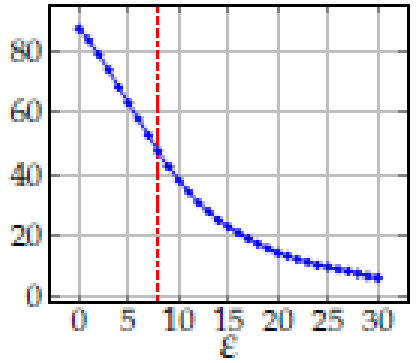
논문의 실험 결과



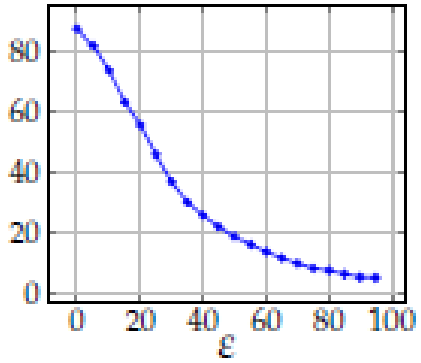
(a) MNIST, ℓ_∞-norm



(b) MNIST, ℓ_2-norm



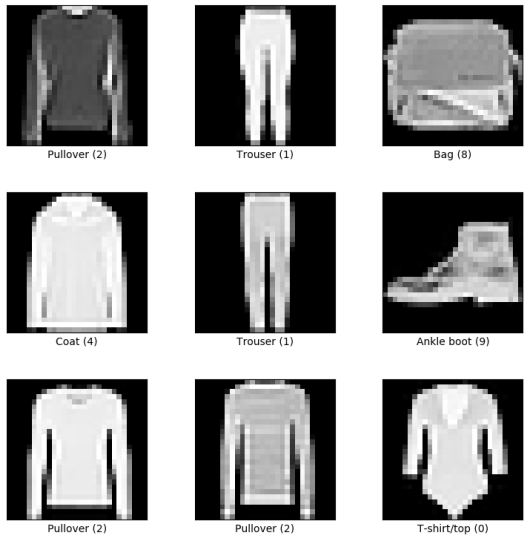
(c) CIFAR10, ℓ_∞-norm



(d) CIFAR10, ℓ_2-norm

내가 한 추가 실험 (My Experiment)

실험 목표: 논문의 방법이 다른 데이터 셋에서도 효과적일까?



새로운 데이터셋 활용: FashionMNIST, SVHN

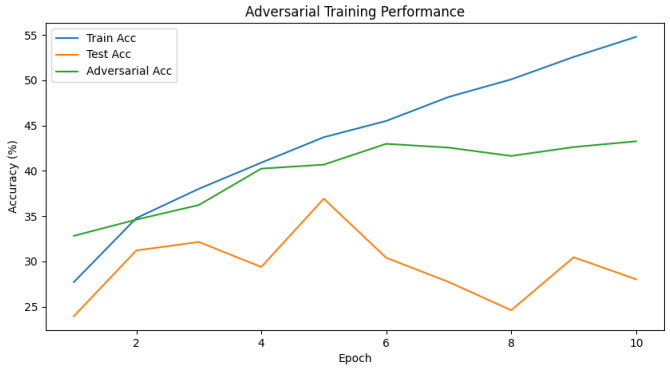
내가 한 추가 실험 (My Experiment)

실험 목표: 논문의 방법이 다른 데이터 셋에서도 효과적일까?

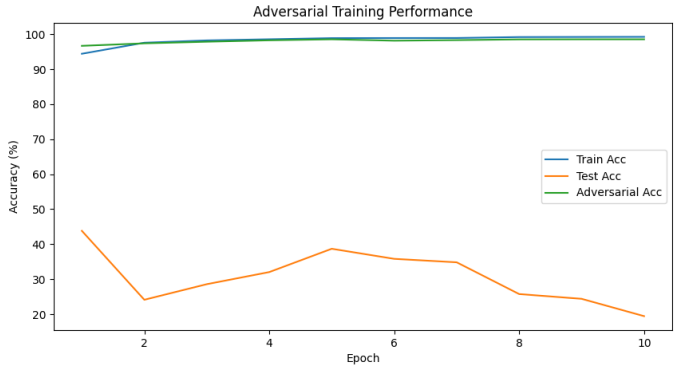
CW 및 AutoAttack 추가 수행

내가 한 추가 실험 (My Experiment)

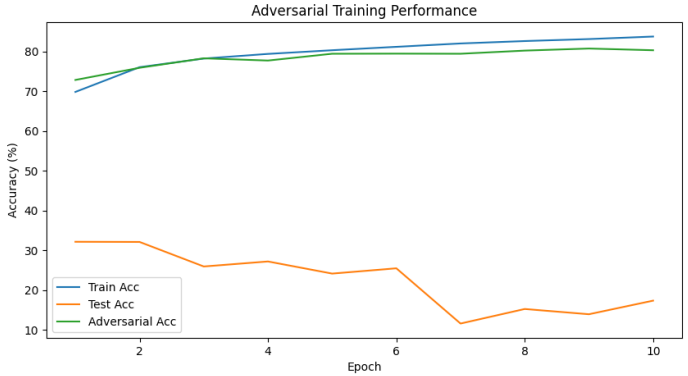
실험 결과



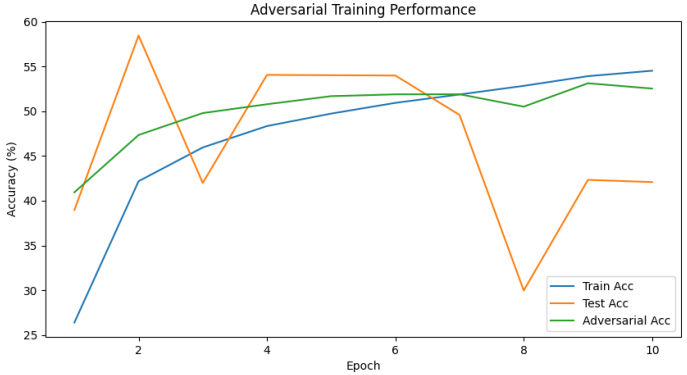
CIFAR-10 Adversarial Training Performance



MNIST Adversarial Training Performance



FashionMNIST Adversarial Training Performance



SVHN Adversarial Training Performance

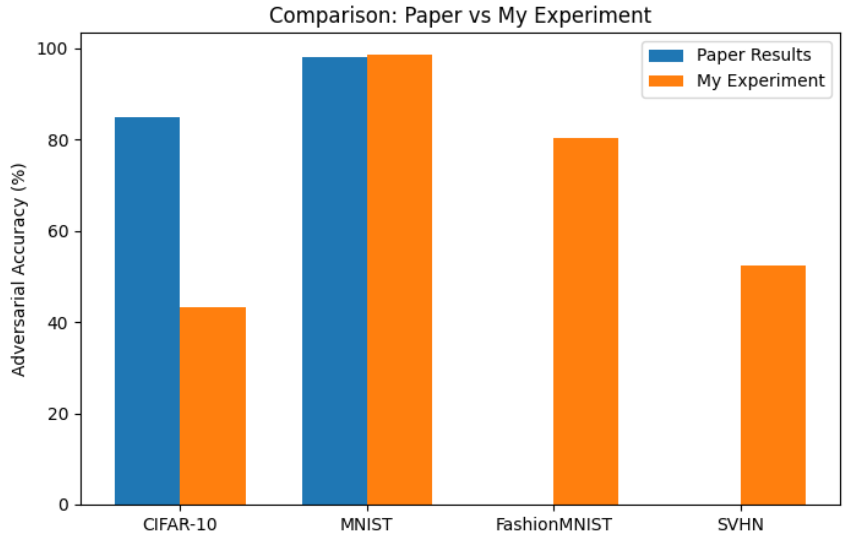
내가 한 추가 실험 (My Experiment)

실험 결과

FashionMNIST와 SVHN은 논문에서 실험하지 않은 데이터셋이므로, Adversarial Training이 동일하게 적용될지 확인하기 위해 추가 실험을 수행했다. 결과적으로 FashionMNIST에서는 높은 강건성이 유지되었으나, SVHN에서는 상대적으로 낮은 Adversarial Accuracy를 보였다.

논문결과와 내 실험 결과 비교

데이터셋	논문의 결과	내 실험 결과
CIFAR-10	PGD 사용 시 강건성 유지	일반 정확도 하락 (최대 43.27%)
MNIST	높은 강건성(98% 이상)	내 실험에서도 98.53% 유지
FashionMNIST		80.34% 유지
SVHN		52.52% 유지



Limitations & Futurework

논문의 한계점

1. PGD 공격에만 초점을 맞춤
2. CIFAR-10에서 MNIST보다 강건성이 낮음

내 실험의 한계

1. CIFAR-10에서 일반 정확도가 크게 하락
2. FashionMNIST는 강건성을 유지했으나, CW&AutoAttack에는 취약
3. SVHN에서는 Adversarial Training 효과가 덜함.

Future work

1. PGD가 아닌 다른 적대적 공격에 강한 모델 개발
2. CIFAR-10의 일반 정확도 저하를 해결할 방법 연구
3. 다양한 데이터셋을 활용한 추가 실험 필요

Conclusion

- "논문의 Adversarial Training 방법은 **MNIST, CIFAR-10에서 강건성을 유지**하지만, 일부 공격(CW, AutoAttack)에는 여전히 취약함."
- "내 실험에서는 **FashionMNIST와 SVHN에서 Adversarial Training을 적용했으나, 데이터셋에 따라 효과가 다르게 나타남.**"
- "CIFAR-10에서 일반 정확도가 하락하는 문제가 있어 **Adversarial Training의 부작용에 대한 연구 필요.**"

Thanks!

Do you have any questions?

sunjun7559012@yonsei.ac.kr
010 -8240-7559 | https://sites.google.com/view/yohanko

연세대학교 YONSEI UNIVERSITY

RAISE Lab
Reliable Artificial Intelligence & System Engineering