# 2023 암호분석경진대회

## 1번 문제

최근 인공지능 기술은 수년 전 Transformer 기술의 출현과 이를 기반으로 하는 초거대 언어모델(LLM)의 등장, 그리고 이를 활용한 ChatGPT의 출현으로 인해 우리 사회와 산업에 큰 충격과 변화를 가져다주고 있다. 하지만, 이러한 AI 기술은 만능이 아니며, AI가 가지고 있는 본연의 특성으로 인해 다양한 예기치 못한 결과도 낼 수 있다.

예를 들어, 아래 그림에 사람이 보기에 동일한 두 개의 이미지가 있다. 왼쪽 이미지에 적절한 노이즈를 추가하여 오른쪽 이미지를 생성했다고 가정하자. 이 때, 객체 인식용 AI는 이 두 이미지를 서로 다르게 인식할 수 있다. 즉, 객체 인식용 AI에서는 하나는 팬더(왼쪽 사진)로 인식하고 다른 하나는 긴팔원숭이(오른쪽 사진)로 오인식할 수 있게 된다. 여기서 사용한 노이즈(Perturbation)를 추가하는 공격을 적대적 AI 공격(Adversarial AI attack)이라고 하며, 아래의 Perturbation attack 기법 외에도, 현실적으로 심각한 위협의 될 수 있는 Active한 공격(Patch 공격 등) 기법이 존재한다.

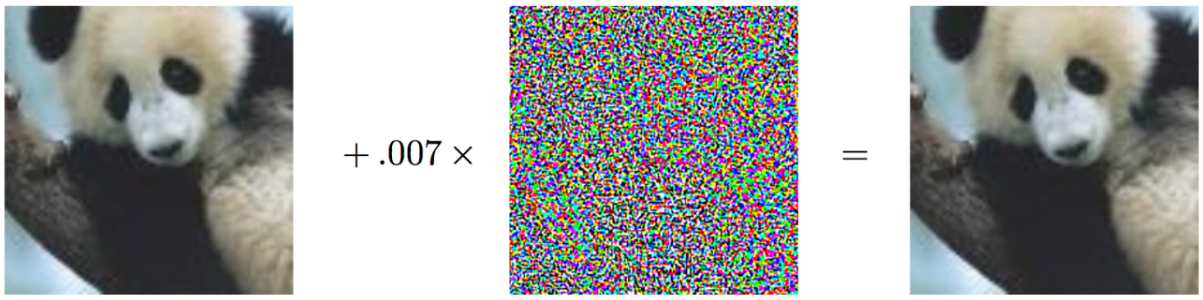


그림 1) 왼쪽은 객체인식용 AI에서 Panda로 인식하지만, 노이즈(Perturbation)가 추가된 오른쪽 사진은 객체인식용 AI에서 긴팔원숭이로 인식하는 공격용 기법이 존재함

본 문제는 이러한 적대적 AI 기술에 대한 이해와 실제 AI를 활용하며, AI에 대한 공격 가능성에 대한 문제다. 쉽게 주어진 문제를 해결하기 위해, 다양한 참고 자료 및 참고 코드를 활용하여 문제를 제시한다.

**문제**

**(문제 1)** 참고자료 (2)의 링크에 나와 있는 코드를 참고하여, CIFAR10 데이터 혹은 ImageNet 데이터에 있는 이미지 중에서 적당한 이미지에 대한 PGD(Projected Gradient Descent) 기반의 Perturbation 공격을 수행하라.

성공적인 AI 모델에 대한 공격이 이뤄졌음을 증명하고, 이때, 사람의 육안으로는 공격 전 이미지와 공격 후 이미지를 (거의) 구분할 수 없음을 보여라.

**(문제 2)** 현실적으로 위협적인 공격 기법으로는 Patch Attack이 존재한다. 참고자료 3의 Figure 1에 제시된 그림을 참고하여, 교통표지판(STOP 신호, 속도제한 신호 표지판 등)에 대한 Patch Attack 사례를 제시하라. 실제 오인식됨을 증명하라.

**주의사항**

(1) 본 실험을 수행하는데 있어서 반드시 PyTorch를 사용해야 한다(Tensorflow 사용 불가)

(2) 공격 대상이 되는 객체 분류용 AI 모델과 데이터세트에 대한 정보, 환경 설정 등에 대한 정보는 다음 사이트의 코드 및 관련 자료를 참고할 수 있다.

https://github.com/Harry24k/adversarial-attacks-pytorch/blob/master/demo/White-box%20Attack%20on%20ImageNet.ipynb

- 본 코드에서는 ResNet-18을 사용하고 있다. 이처럼 ResNet-18을 사용할 수도 있다. 하지만, 만약 기존 시중에 공개된 CNN 계열의 객체분류 모델 중에서, 공격 성능이 더 우수한 모델이 있다면 이를 사용하여 그 결과를 보일 수도 있다. 한편, 문제 풀이자는 새로운 객체 인식 모델을 제시할 수도 있지만, 이 경우에는 최소한 100개 클래스 이상의 객체를 인식할 수 있도록 제시된 모델이 이미 학습되었음을 증명해야 한다.

(3) 실험을 위해 CPU 혹은 GPU 어떤 것을 사용할 수 있다.

**참고자료**

(1) Strengthening Deep Neural Netwroks, https://github.com/katywarr/strengthening-dnns

(2) https://github.com/Harry24k/adversarial-attacks-pytorch

(3) Adversarial Patch by om B. Brown, Dandelion Mané, Aurko Roy, Martín Abadi, Justin Gilmer