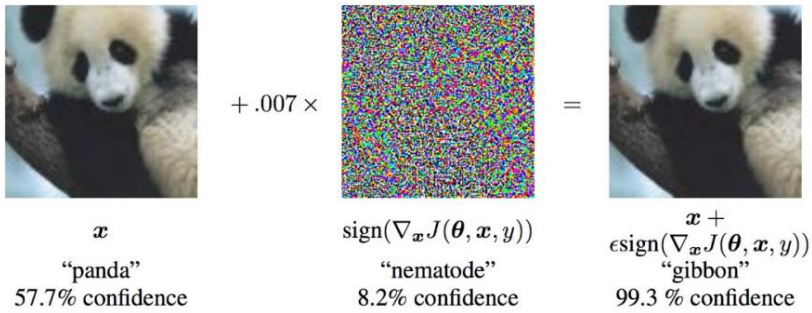
Data augmentation with adversarial examples에 대한 모델의 정확도 향상

이연우	정민기	한대일
인공지능응용학과	인공지능응용학과	인공지능응용학과
21102376	21102385	21102398

1. 문제 제기



본 프로젝트 주제는 ML2 Week 4 에서 다룬 Data augmentation with adversarial examples이라는 주제에 대한 의구심을 바탕으로 한다. 즉, 이 연구의 문제 의식은 왜곡된 이미지 데이터를 인식할 때, 모델의 인식률이 사람에 비해 기이할 정도로 뒤떨어진다는 것이다. 따라서 이 연구의 목적은 모델에게 data augmentation의 존재를 좀더 명확하게 학습시키는 방법을 고찰하여 가장 효율적인 해결책을 제시하는 것이다.

2. 연구 계획

- CIFAR, MNIST, ImageNet 등 잘 알려진 공용 데이터셋을 활용한다.
- Data augmentation에 대한 다양한 학습 방법을 개발하고 비교 및 실험한다. 예를 들면, data augmentation이 전혀 적용되지 않은 일반적인 CNN 모델과 오로지 data augmentation 여부를 분류하는 supervised learning 모델을 ensemble 하여 새로운 모델을 개발하고, 이에 대한 성능을 평가한다. Augmented data에 대하여 어떻게 labeling하는 것이 효율적일지 연구한다.
- 이를 위해 최대한 다양한 data augmentation에 대해 공부하고 이를 구현하려고 노력한다. 다른 연구자들이 관련된 문제를 어떻게 해결하려고 하였는지 찾아본다. 효율적인 Ensemble 방식과 그 구현에 대하여 학습한다.

3. 평가

일반적인 CNN 모델, data augmentation을 학습한 CNN 모델, 새롭게 개발한 모델의 accuracy를 각각 평가하여 비교한다.