**초록**

# **제목**

|  |  |  |
| --- | --- | --- |
| 이연우 | 정민기 | 한대일 |
| 인공지능응용학과 | 인공지능응용학과 | 인공지능응용학과 |
|  |  | 21102398 |

사람의 육안은 쉽게 구별할 수 있지만, 딥러닝 모델에게는 구별이 쉽지 않은 데이터를 생성하는 기술을 적대적 공격이라고 한다. 이 프로젝트에선 적대적 공격을 구현하여, 이미지 분류 모델에 대한 적대적 공격의 성능을 측정한다. 그리고 적대적 공격에 대하여 데이터 증강 및 모델의 층 수 증대 등의 효과적인 방어책을 모색하여 결과를 도출한다. 실험 결과, 가장 효율적인 방어책은 ~~~으로 나타났다.

1. **서론**

본 프로젝트는 머신러닝2 강의 4주차에서 다루었던 적대적 예제에 대한 의구심을 토대로 한다. 이미지 분류 문제에서 적대적 예제란, 적대적 공격으로 생성한 이미지 데이터를 의미한다. 적대적 공격이란 전통적인 가중치 업데이트 방식과 반대되는 방향으로 가중치를 업데이트하여 DNN 이미지 분류 모델이 다루는 데이터에 대하여 최악에 가까운 노이즈 이미지를 생성하는 이론이다.

본 프로젝트는 상기의 적대적 공격에 대하여 효과적인 방어 방법이 무엇인지 찾아내는 것을 그 골자로 한다. 구체적으로 몇 가지 데이터 증강 이론과 모델의 레이어 깊이에 대한 변화를 통해서 효과적인 방어 방법에 대하여 고찰해 볼 것이다.

데이터셋으로는 MNIST 데이터셋과 CIFAR10 데이터셋을 사용한다. 두 데이터셋에 대하여 같은 실험을 반복하여, 실험의 일관성을 확인할 것이다. DNN 모델은 LeNet-5모델을 수정하여 사용할 것이다.

적대적 공격 모델로는 Fast Gradient Sign Method(FGSM)을 사용할 것이다. FGSM은 신경망 모델이 적대적 변동 (adversarial perturbation)을 견디기에는 지나치게 선형적이라는 가정하에, 이미지 분류의 결정 경계를 넘을 수 있는 최소한의 노이즈 분포를 찾는 방법론이다.

데이터 증강 방법으로는 Random Horizontal Flip, Random Rotation, Random Affine, Color Jitter, Gaussian Blur를 각각 사용하여 모델을 학습시킨 뒤, 각각의 모델에 대하여 적대적 공격을 감행하여, epsilon에 따른 accuracy 감소량을 확인할 것이다.

- 본론과 실험결과에 대한 간략한 설명

1. **본론**

* 1 ~ 2쪽 분량
* 프로젝트의 결과물 (조사 내용, 시스템, 알고리즘 등)에 대한 자세한 설명
* 주제와 관련된 배경 지식, 관련 기술에 대한 간략한 소개 및 프로젝트와의 관련성
* 본론과 실험결과에 대한 간략한 설명

1. **실험 결과**

* 1 ~ 2쪽 분량
* 실험이 원래 목표로 했던 주제를 잘 해결할 수 있는지에 대한 검증
* 다양한 그림과 표를 사용, 해당 실험 결과들을 제시/분석

1. **결론**

* 초록과 유사, 프로젝트에 대한 요약과 의의를 1문단

1. **평가 기준**

최종 보고서 초안 (6점)

프로젝트 주제 설정부터 해당 주제를 해결하기 위해 수행했던 관련 기술 조사, 프로젝트에서 제시하고 있는 방법, 실험 설계 및 검증이 어떤 식으로 진행되었는지 등을 평가할 계획입니다.

최종 보고서 수정안 (6점)

수정안 평가에서는 최종 보고서 자체에 대한 평가를 진행할 계획입니다.

프로젝트 결과물 (코드) (2점)

코드 동작여부, readme 파일 작성 여부 등을 평가할 계획입니다.

**참고 문헌**

[1] ~