제목 : 적대적 머신러닝 공격과 방어기법

1. 주요 내용:

- 적대적 공격 분석:

기계 학습 모델의 취약점을 이용하여 공격자가 모델을 속이거나 조작하는 방법을 조사함..

이를 위해 다양한 적대적 공격 유형을 분석하고, 이러한 공격이 기계 학습 시스템의 안정성과 보안에 미치는 영향을 조사했음.

이런 공격- 중독 공격, 회피 공격, 모델 추출 공격, 전도 공격 이 있음을 파악.

- 방어 메커니즘 제안:

논문은 적대적 공격에 대응하여 기계 학습 모델의 견고성을 향상시키는 다양한 방어 메커니즘과 전략을 소개.

Gradient Masking, , Distiliation, Feature Sqeezing 같은 방법이 있음.

이를 통해 안전하고 신뢰할 수 있는 인공지능 모델을 개발하고자 함.

2. 효과 및 중요성:

- 안전한 기계 학습 시스템 개발:

적대적 공격에 대응하여 안전하고 강력한 기계 학습 시스템을 구축하는 게 중요하다고 함.

- 지속적인 보안 강화:

기계 학습 모델의 취약점을 이해하고 방어 기술을 개발함으로써 적대적 위협에 대응하라고 함.