제목 : 인공지능(AI) 적대적 공격 및 적대적 공격에 대한 방어 기술 동향과 군 적용 발전방안

1. 논문 배경:

- 인공지능은 다양한 분야에서 성능을 향상시키고 있지만, 딥러닝과 같은 AI 기술은 보안 취약점을 가지고 있어 딥러닝 보안 문제가 해결될 필요 있음.

2. 주요 내용:

- 적대적 공격 유형:

AI 모델의 신뢰성과 무결성을 위협하는 다양한 적대적 공격 유형이 있음.

중독 공격, 회피 공격, 모델 추출 공격, 전도 공격

- 방어 기술 동향:

적대적 공격에 대응하거나 방어하기 위한 최신 연구 동향이

Gradient Masking, , Distiliation, Feature Sqeezing임.

- 군 적용 발전방안:

군에서 AI를 적극적으로 활용할 때 적대적 공격에 대한 내성을 높이고, AI 기술의 신뢰도를 확보하기 위한 효과적인 방어 전략을 제시