제목 : 강화학습을 이용하여 인공신경망에 대한 adversarial attack을 방어하는 affine transformer에 관한 연구

저자 : 김형진, 이정우

1. 논문 내용

인공지능 모델의 적대적 공격을 방어하는 방법 다룸,

이를 위해 강화학습 알고리즘을 활용하여 affine transformer를 개발

이 transformer는 인공신경망을 보호하고 적대적 공격으로부터 방어 역할

2. 연구 방법

제안된 방법(affine transformer)은 적대적 공격 그래프 생성을 통해 인공지능 모델을 학습하고,

모델이 잘 학습되었는지 loss 값과 reward 값의 변화를 확인하며 공격 그래프를 생성.

이를 통해 기존 방식으로 생성한 공격 그래프와 비교하여 효과적인 방어 메커니즘을 알 수 있었음.

Affine transformer 가 효율적이ㅓㅆ음.