제목 : 적대적 공격을 방어하기 위한 StarGAN 기반의 탐지 및 정화 연구

1. 주요 내용:

- 적대적 예제에 대한 방어 방법을 연구
- StarGAN을 활용하여 다양한 적대적 공격을 탐지하고 정화하는 방법 제안.
- StarGAN 모델을 학습시키고, Categorical Entropy loss를 추가하여 훈련

  그 후 다양한 적대적 예제를 생성하여 모델을 훈련시키고, 생성자는 적대적 예제를 정화하고 판별자는 탐지

2. 실험 결과:

- CIFAR-10 데이터셋을 사용하여 실험한 결과, 약 68.77%의 탐지 성능, 약 72.20%의 정화 성능, 그리고 약 93.11%의 방어 성능을 보임.