2023. 12. 29

2023 Poster Session 중간 발표

eXplainable AI

김채원 유광열 이서연

CONTENTS

이전까지의 연구

XAI for Regression Problem

XAI 기술을 활용하여
회귀 문제에 대해 모델의 설명가능성을 확인

기존 설명 모델(Grad-CAM)에
Regression 예측을 위한 Layer 추가

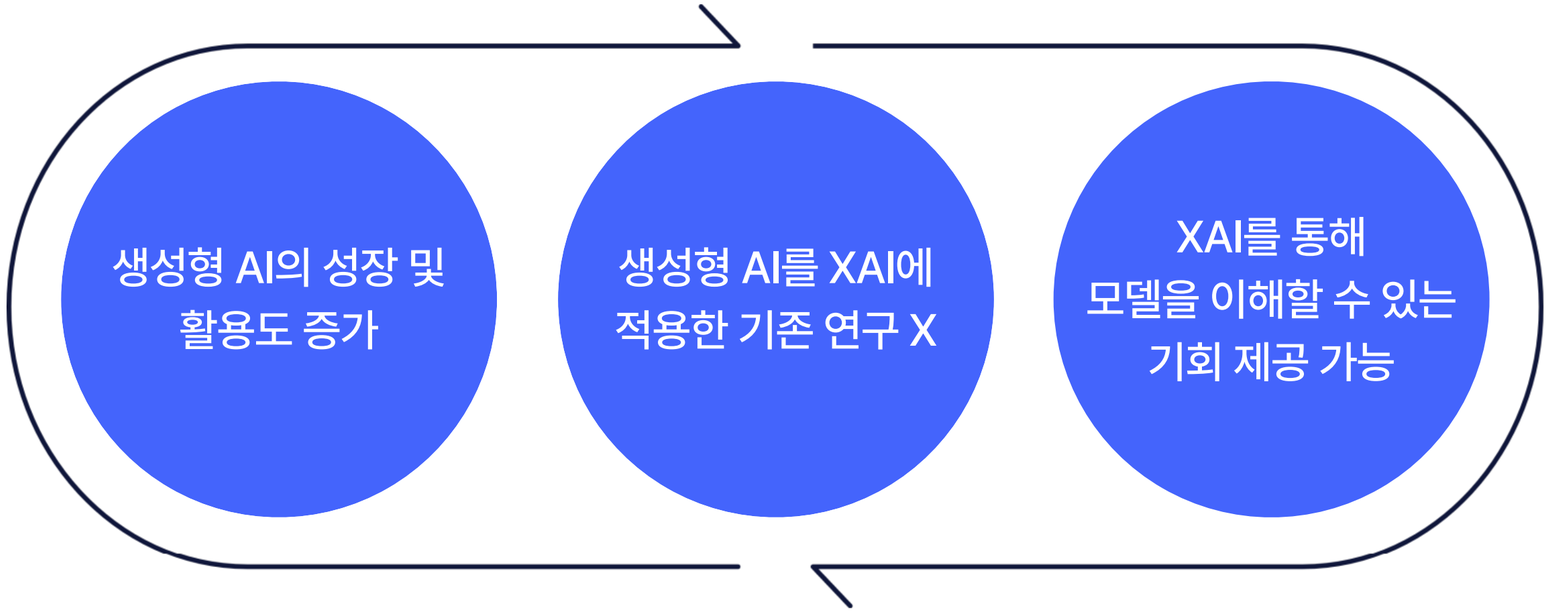
XAI를 통한 성능 향상

이미지에서 중요도를 추출한 후
데이터 변형을 통한 성능 향상

Class Activation Map의 결과로 나온 Heatmap
(이미지에서 어떤 부분을 집중하는지)을 기존 이미지에 더하는
데이터 변형을 거친 후 모델 성능 비교

연구 주제 및 배경

“ 생성형 AI + eXplainable ”



연구 개요

Generative AI + Explainable AI (XAI)

생성형 AI의 " Discriminator " 에 XAI 기술을 적용

GAN, Diffusion(+ Discriminator)과 같은 생성형 AI 모델에 대하여,
모델의 Discriminator가 내린 의사 결정을 이해하고자 함

연구 개요

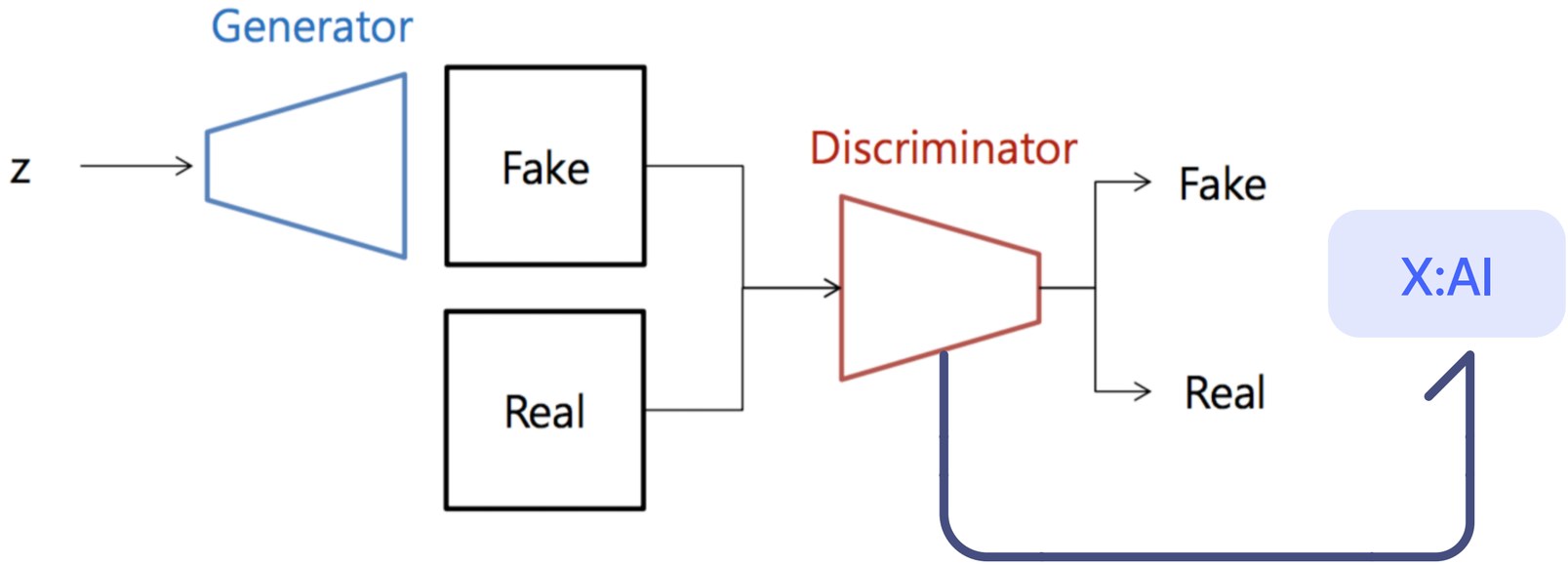
판별 역량 강화

어떤 부분으로 인해
Real / Fake라고 판별했는지
이해 가능

생성형AI
성능 향상

연구 진행 상황

Generative Adversarial Networks Discriminator에 Explainable 진행



Discriminator X:AI

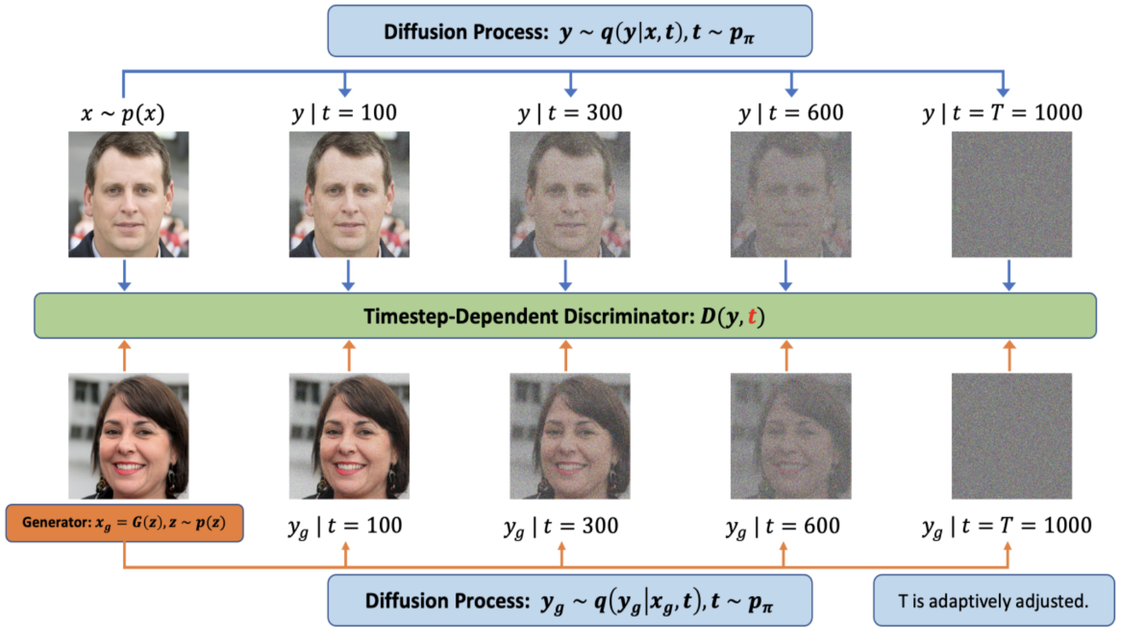
Fake와 Real을 구분하기 위해 Discriminator가
이미지의 어떤 부분을 주로 고려하는지 설명하는 X:AI

연구 진행 상황

Diffusion 기반 생성형 모델에 Discriminator 추가

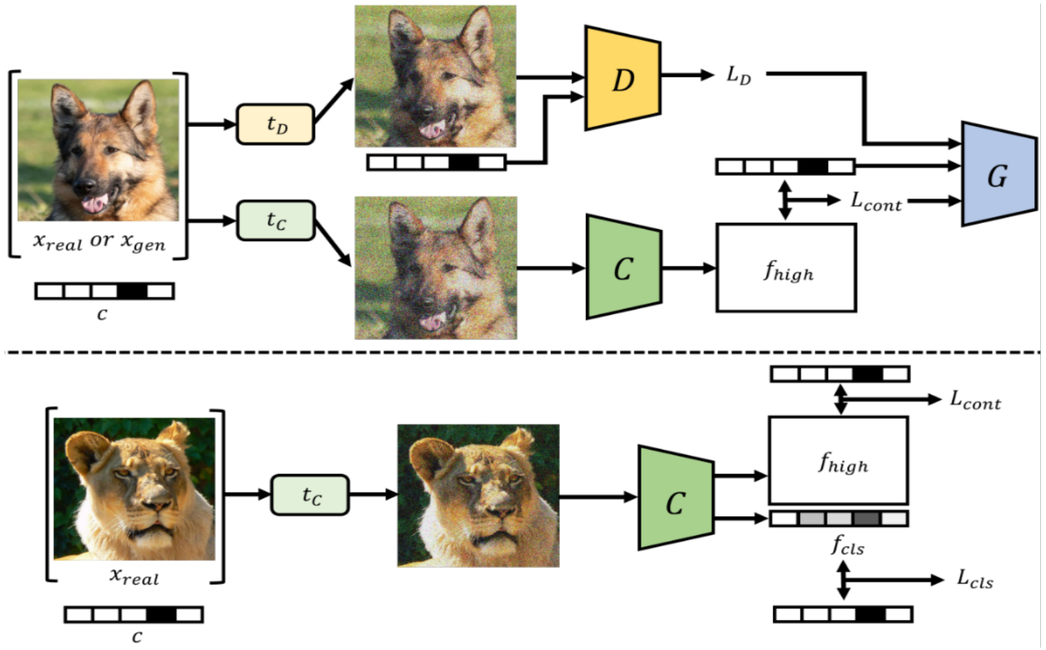
Diffusion_GAN

: Diffusion을 활용하여 Discriminator에 들어가는 input 다양화



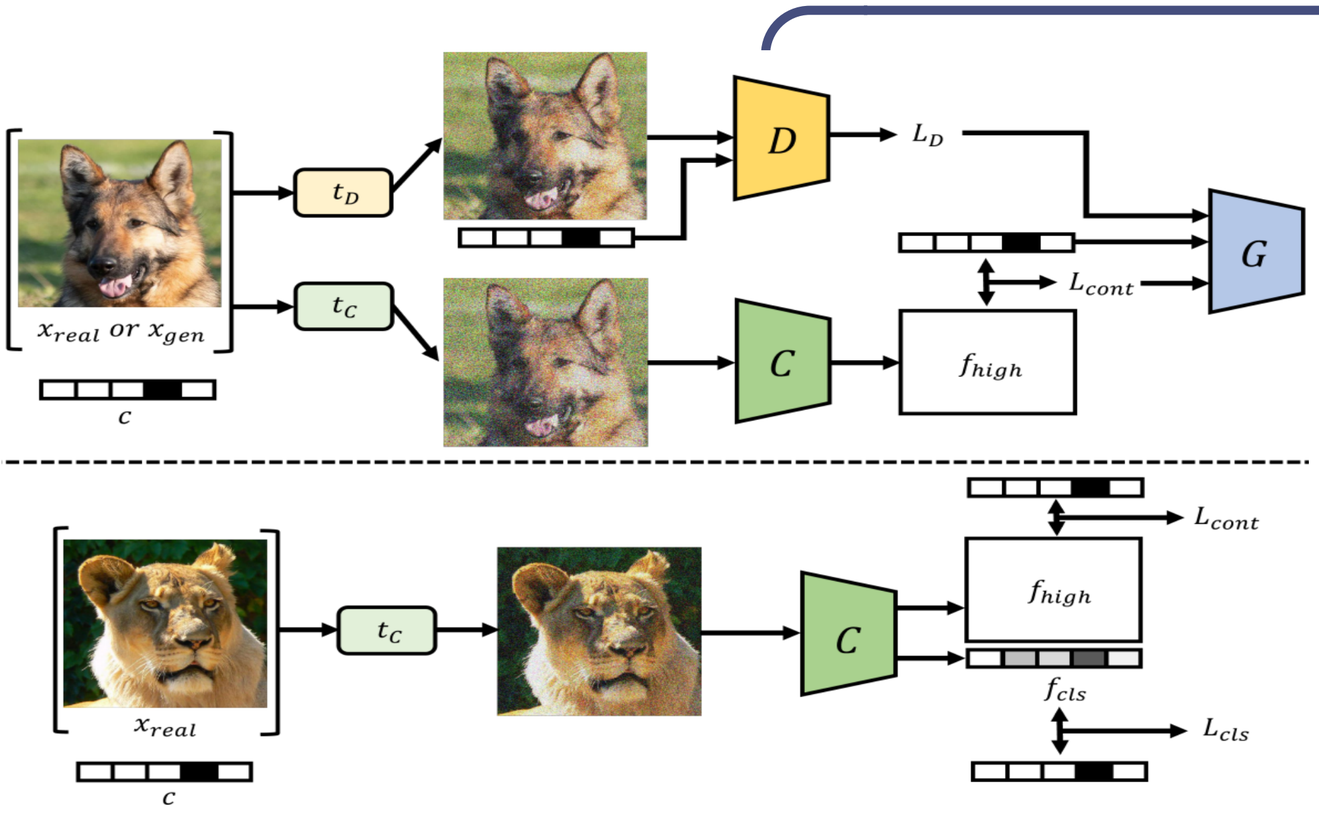
DuDGAN

: 기존 GAN에 Generative를 diffusion을 대체, Classifier 추가



연구 진행 상황

DuDGAN 모델에 존재하는 Discriminator Explainable 진행



X:AI

Discriminator X:AI

이미지에 대해 Discriminator가 어느 부분을 보고 Fake와 Real로 판단하는지 X:AI

연구 계획

eXplainable Generative AI

1) GAN + Discriminator X:AI 구현
→ GAN을 통한 Discriminator X:AI 구현 마무리

2) DuDGAN+ Discriminator X:AI 구현 (additional)
→ 최신 생성형 AI 트렌드를 반영하여 Diffusion based model에도 해당 방법론 적용

생성형 모델 개선

Discriminator X:AI를 통해 얻은 결과를
side information으로 활용하여 모델 성능 향상 정도 확인

2023 Poster Session 중간 발표

Q & A