Writing Session - eXplainable AI

생성형 모델
eXplainable AI

20212549 김채원 / 20192780 유광열 / 20212568 이서연

CONTENTS

01 연구 배경

02 연구 주제 및 개요

03 사전 연구

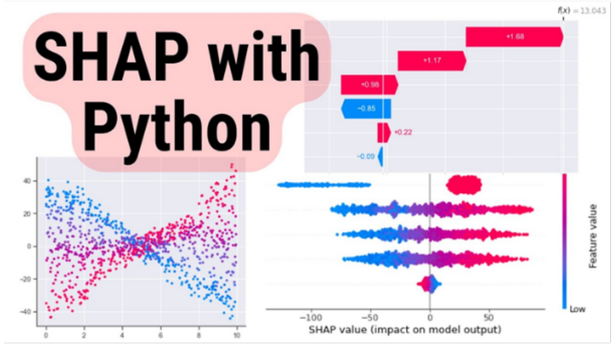
04 진행 상황

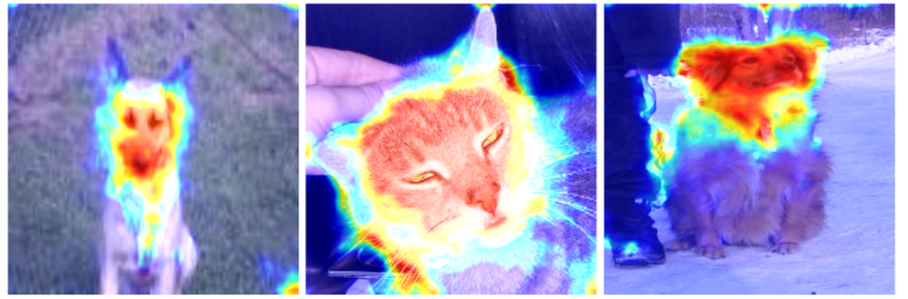
05 연구 계획

eXplainable AI?

AI 모델을 설명하고 해석할 수 있게 하는 기술과 프레임워크를 의미

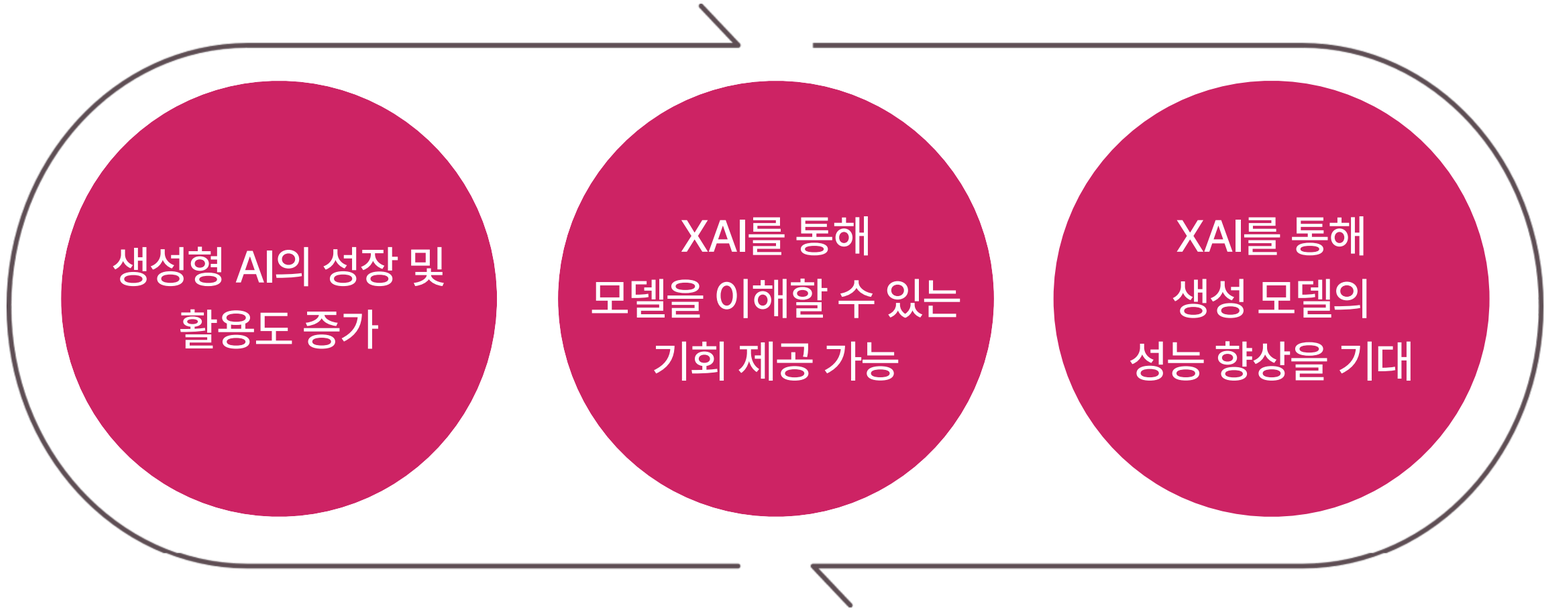
블랙박스 형태의 복잡한 AI 모델을 투명하고 해석 가능한 방식으로 만들어,
AI를 사용하는 사람들에게 모델의 작동 방식을 이해할 수 있는 기회를 제공





연구 배경

“ 생성형 AI + eXplainable ”



연구 주제 및 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 주제 및 개요

어떤 부분으로 인해
Real / Fake라고 판별했는지
이해 가능

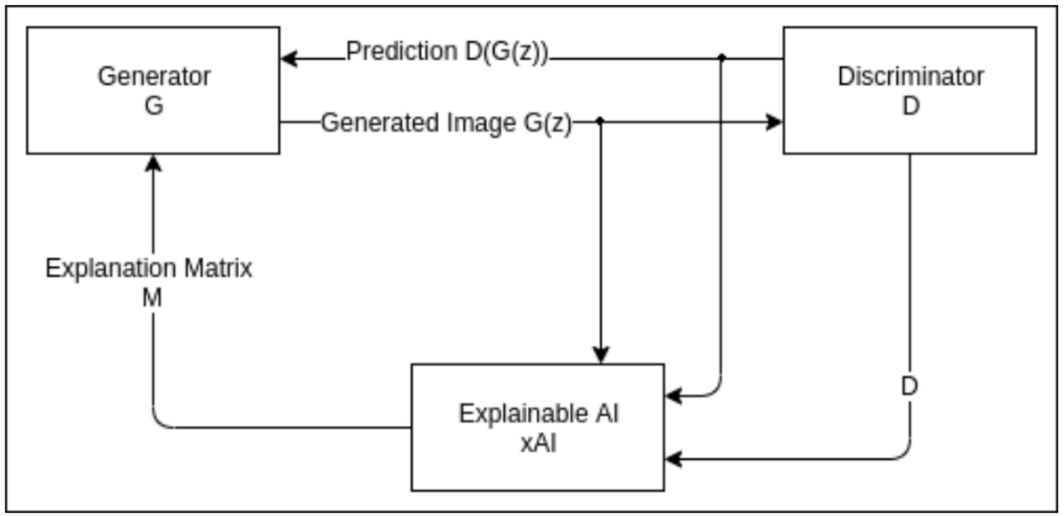
판별 역량 강화

생성형AI
성능 향상

사전 연구 : xAI - GAN

**Discriminator가 분류를 수행한 이유를 구체화하는 xAI system을 이용하여,
Generator에게 하나의 loss값이 아닌 더 풍부한 feedback을 제공**

- Saliency map과 같은 xAI를 이용하여, input image의 모든 feature에 대해 Discriminator 결과에 영향을 미친 정도에 따라 0~1 범위의 값 할당 (행렬 M의 형태)
  - explanation matrix M을 계산한 다음, 경사하강법 알고리즘에서 Generator output의 gradient와 element-wise 곱을 계산한 값으로 업데이트
    - 가장 중요한 feature에 gradient를 집중시키고 덜 중요한 feature에는 gradient를 제한하는 역할을 수행하게 됨

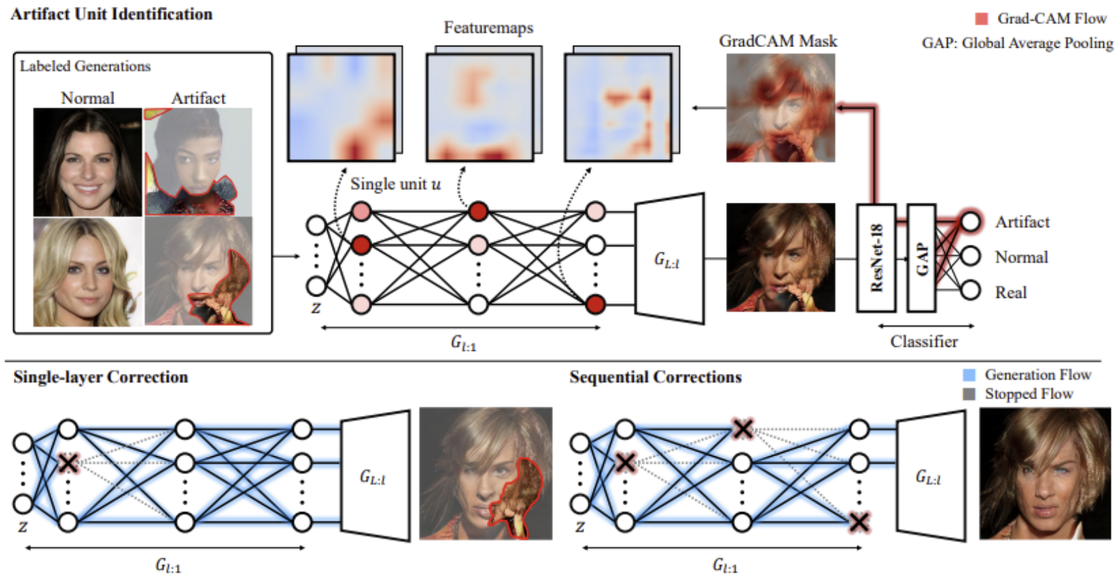


Automatic Correction of Internal Units in Generative Neural Networks (2021)

사전 연구 : Error Correction Techniques for GANs

GradCAM을 활용하여 각 레이어의 artifact unit을 식별하고, 이를 통해 생성 모델의 오류를 보정

- GradCAM을 활용하여 레이어별로 오류를 일으키는 내부 unit을 감지하고, 이와 같은 unit들의 활성화를 방지하여 생성 모델의 오류를 수정
  - 각 레이어에서 발생한 artifact(결함) unit을 찾고, 이를 마스크로 변환
  - 마스크와 생성기 내부 유닛 간의 IoU를 계산하고, 여러 샘플에 대한 IoU의 평균을 해당 레이어의 결함 점수로 사용

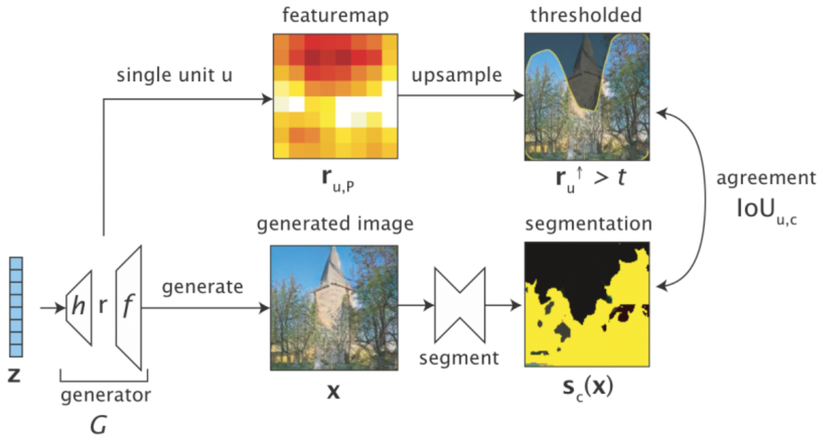


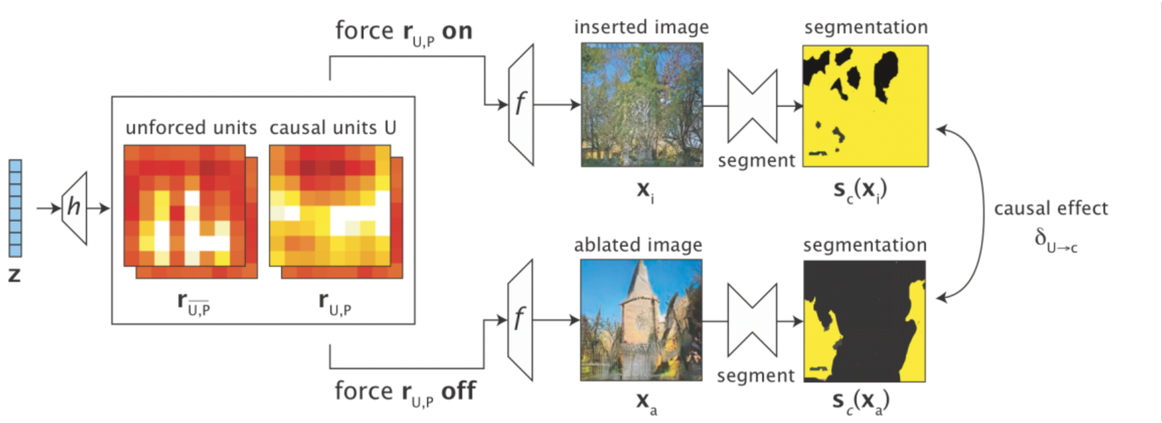
Automatic Correction of Internal Units in Generative Neural Networks (2021)

사전 연구 : GAN DISSECTION

FeatureMAP과 segmentation의 IOU차이와 turnON/OFF를 통한 Casual effect를 구하여 GAN의 영향력 설명

- generator를 사용하여 생성된 r중 single unit u (featuremap의 one channel)를 추출, segmentation 결과와 픽셀에서 얼마나 일치하는지 확인
- r에 대하여 turn on.off로 나누어 진행, 두개의 segmentation결과로 casual effect 계산
- 최종적으로 r과 class의 관계 파악, casual effect 확인

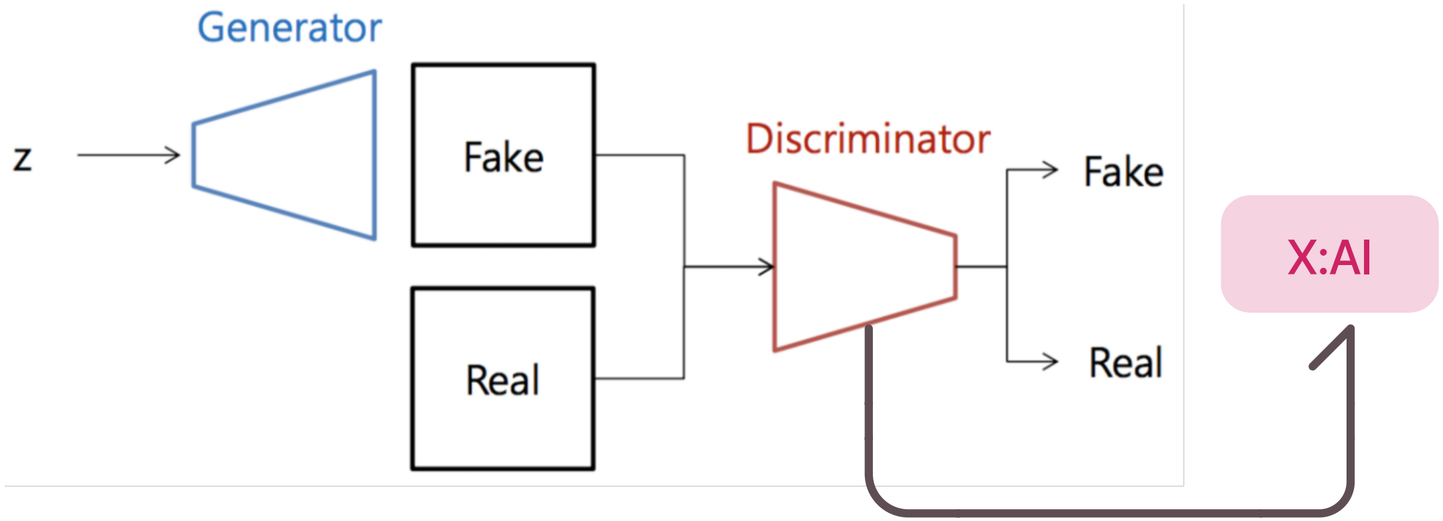




GAN Dissection: GAN Dissection Visualizing and Understanding Generative Adversarial Networks (2019)

연구 진행 상황

Generative Adversarial Networks Discriminator에 Explainable 진행



Discriminator X:AI

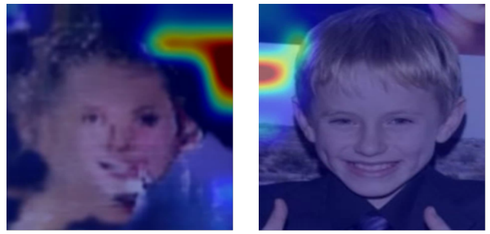
Fake와 Real을 구분하기 위해 Discriminator가
이미지의 어떤 부분을 주로 고려하는지 설명하는 X:AI

연구 진행 상황

DCGAN Discriminator에 다양한 CAM 적용

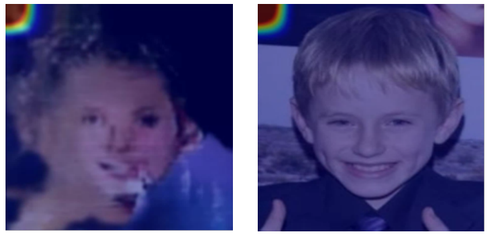
[GradCAMpp]

[ScoreCAM]



[LayerCAM]

[SmoothGradCAMpp]



[XGradCAM]

DCGAN Discriminator에 다양한 CAM을 적용한 결과, 인간이 인식하기 어려운 설명력을 보인다는 한계

연구 계획

Diffusion 모델에 eXplainbe AI 적용

**가중치 업데이트 시,
기존 Gradient에
중요도를 곱해주는 방식**

- xAI-GAN의 방법론을 Diffusion에 적용
- 이미지의 feature에 대한 Discriminator 결과의 Gradient를 사용하여 중요도를 평가 (Saliency map)
- 이 값을 기존 loss에 대한 Gradient에 곱하여 업데이트하여, 가장 영향력 있는 feature에 학습 process를 집중

**중요도 결과에 따라
noise를 다르게 적용**

- Saliency map으로 이미지 feature에 대해 Discriminator 결과에 대한 중요도를 도출
- 중요도가 높은 영역에는 이미지에 노이즈를 적용하지 않고, 중요도가 낮은 영역에는 큰 노이즈를 적용하여 Diffusion 모델이 중요한 부분에 집중하여 학습

Writing Session - eXplainable AI

Q & A

2024. 01. 18