

07_Generative Models and XAI

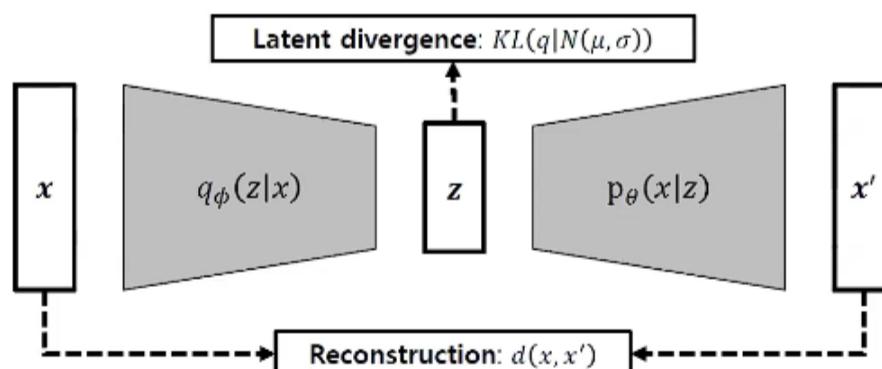
1. Background

A. Generative Models

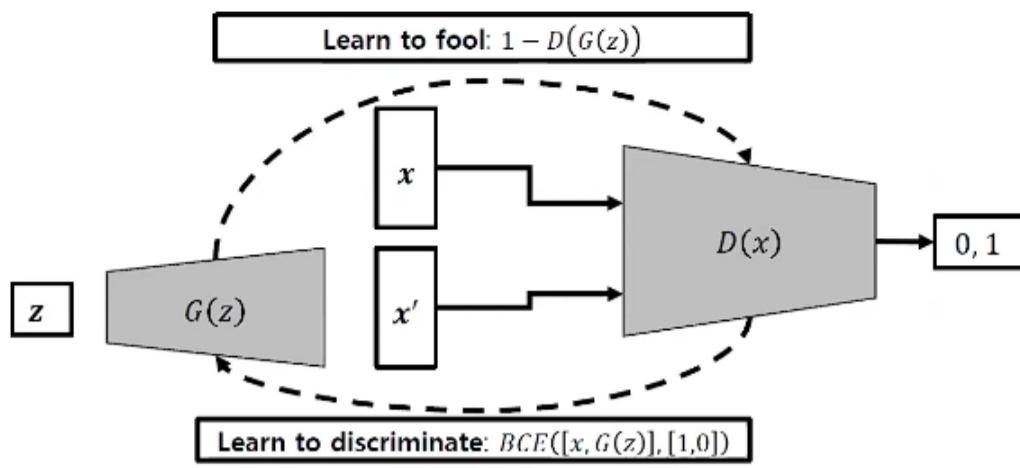
생성형 모델은 특정 데이터 분포를 모방하도록 훈련되었고, 그 데이터 분포에 따르는 새로운 예시들을 생성할 때 사용하는 모델을 뜻한다. 여러 도메인이나 데이터 종류에 적용이 가능하다. 이미지 생성 모델들에 대해서 알아보자.

Types of Generative Models

1. Variational Autoencoder (VAE)

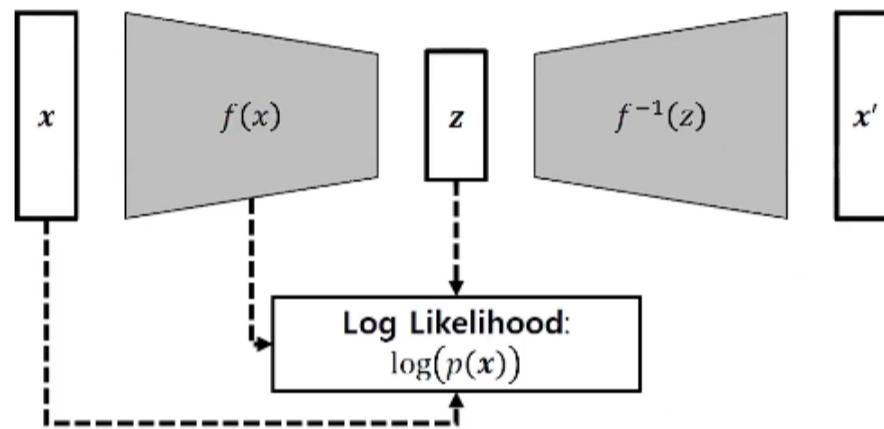


2. Generative Adversarial Network (GAN)



3. Flow-based

- VAE와 GAN과 달리, 확률 분포를 명확하게 정의하고, 역변환을 통해 데이터를 생성한다.
- Encoder와 Decoder가 역함수 관계를 갖는 것이 특징이다.
 - $z = f(x)$
 - $x = f^{-1}(z)$



Strengths & Weaknesses

- VAE:
 - + Easy to use
 - Generates blurry images
- GAN:
 - + Generates sharp images
 - Can suffer mode collapse
- Flow:
 - + Does not lose any information
 - Computationally expensive

B. XAI Methods

XAI

민감한 tasks들에 대하여 설명이 필요한 경우가 많이 존재한다.

- 투자 AI : 왜 그 종목에 투자했는지 설명이 필요
- 이미지 분류 AI : 이미지를 왜 그렇게 분류했는지 설명 필요

[방법]

결과에 대한 입력의 중요도

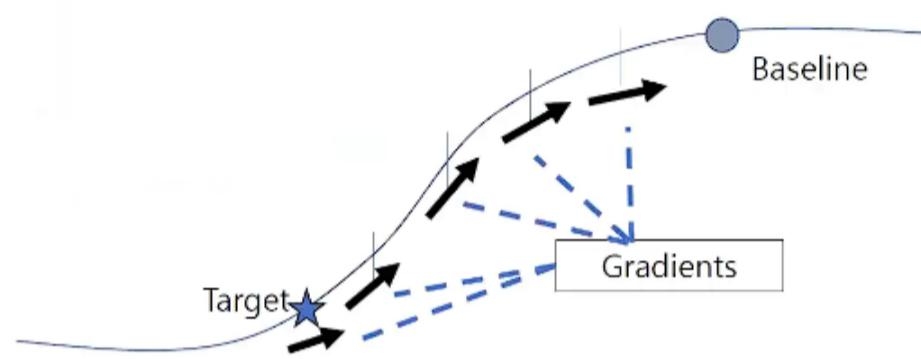
모델 특성을 나타내는 예시를 뽑는 것

모델 내부에 내재된 인간이 이해할 수 있는 컨셉을 추출하는 것

Feature Attribution

특정 입력이 모델의 결과값에 미치는 중요성, 영향을 계산하는 방법.

LIME, KernelSHAP, IG 등의 기법들은 baseline과 target간의 차이를 설명하는 방식들이다.



Example-Based Explanation

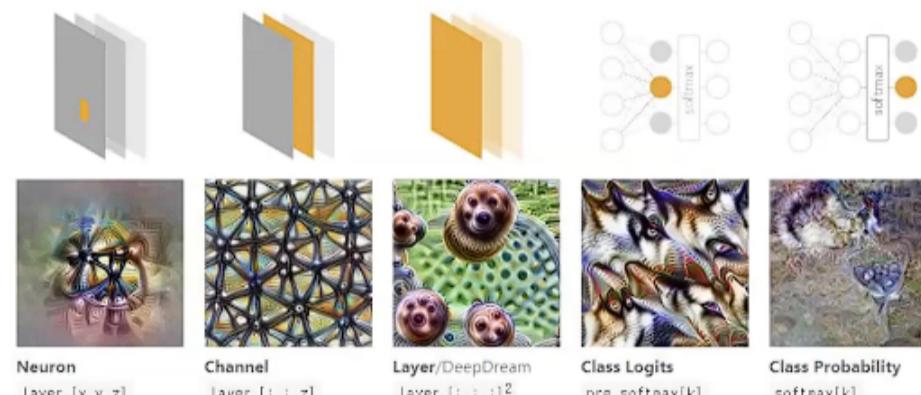
모델의 특성을 반영하는 example을 찾는 방법. 크게 3가지가 사용된다.

- Factual : 분류 모델이 있을 때, 어떤 특정 클래스에 해당하는 예시를 뽑는 것.
- Counterfactual : 특정 target이 있을 때, 이 target으로부터 얼마나 바뀌어야 다른 클래스에 속할 수 있는지를 찾는 것.
- Prototypes : 특정 클래스에 가장 기본이 되는 behavior만 반영하는 즉, 중점적인 예시들을 프로토타입이라고 할 수 있다.

Concept Explanation

모델 내부에 있는 파트들이나 잠재 공간들이 어떤 Concept을 담고 있는지 분석하는 방법.

네트워크 분석이나, 잠재 공간 분석 등이 있다.



C. Generative Models in XAI

그렇다면, 생성 모델과 XAI가 어떤 연관이 있을지 알아보자.

생성 모델은 XAI에서 Explainer(설명자)와 Explainee(설명 대상) 두 가지 역할을 할 수 있다.

1. Explainer 역할

- 생성 모델이 다른 모델을 설명하는 도구로 사용될 수 있다.
- 특정 모델 $f(x)$ 의 예측을 설명하기 위해 생성 모델 $G(z)$ 를 활용.
- 이 과정에서 Counterfactuals, Baseline 설정 등이 가능하다.

[하단 이미지 좌측]

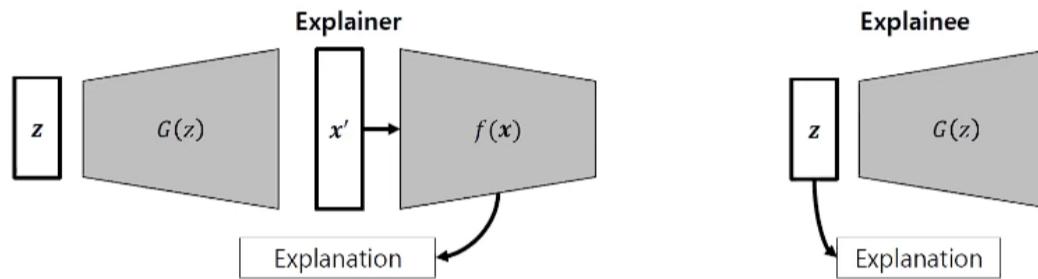
- 잠재 변수 z 에서 생성 모델 $G(z)$ 를 통해 샘플 x' 을 생성.
- 샘플 x' 을 예측 모델 $f(x)$ 에 입력하여 결과를 확인.
- $f(x)$ 의 예측을 설명하는 '설명'을 생성.

2. Explainee 역할

- 생성 모델 자체가 설명되어야 하는 대상이 될 수도 있다.
- 생성 모델 $G(z)$ 내부의 작동 방식과 잠재 공간의 구조를 이해하는 것이 목표.

[하단 이미지 우측]

- 잠재 변수 z 를 생성 모델 $G(z)$ 에 입력.
- 생성된 결과를 분석하여 '설명'을 생성.



2. Generative Models as Explainers

생성 모델을 설명 모델로 사용할 경우, 주로 설명에 필요한 특정 예시를 뽑는데 사용이 된다.

1. Feature Attribution : 모델이 예측을 수행하는 데 어떤 특징이 중요한지 분석.
 - GANMEX (GAN-based Model Explainability)
2. Example-based Explanation : 모델이 결정 경계를 어떻게 정의하는지 설명.
 - Diffeomorphic Counterfactuals

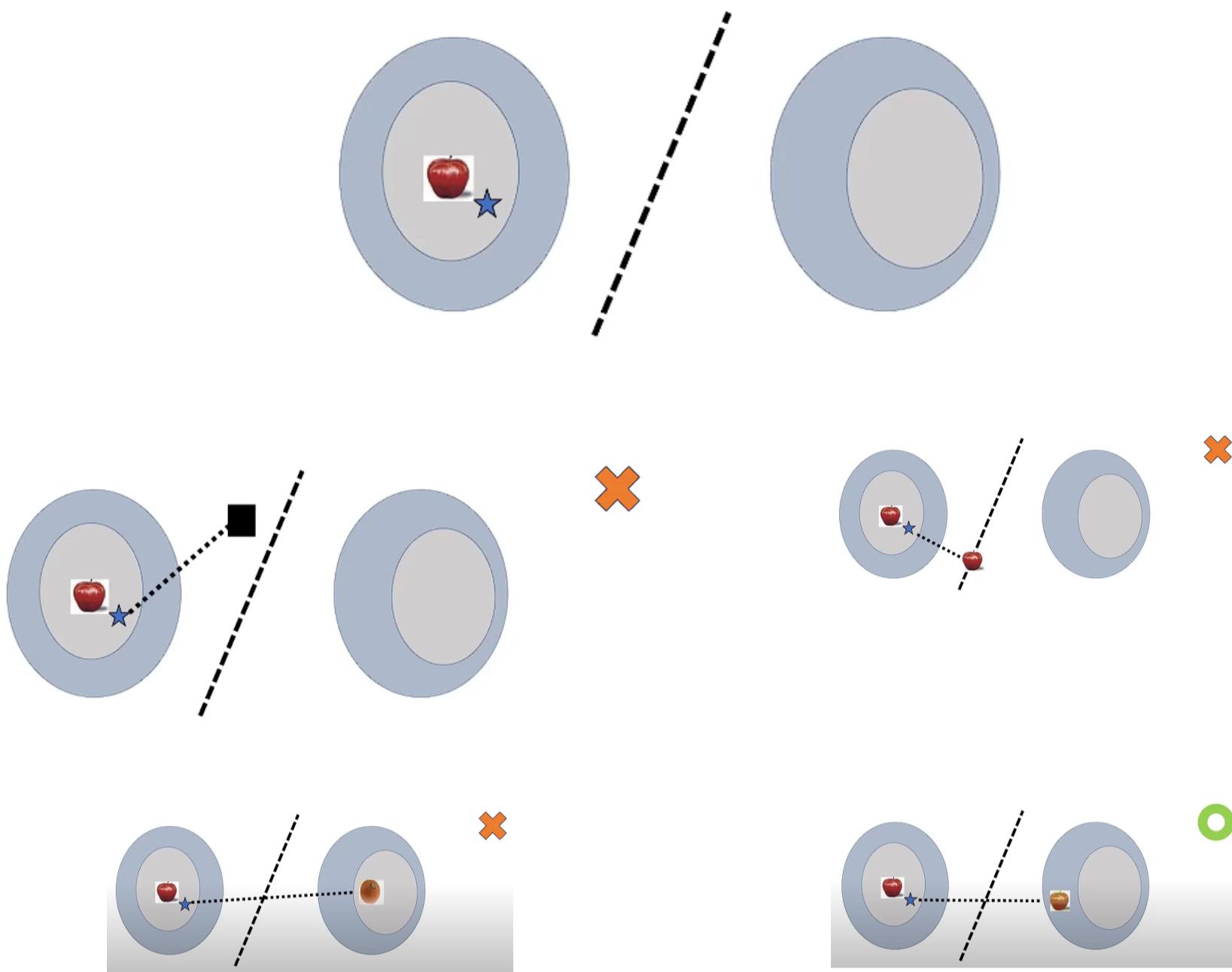
What is a 'good' baseline for feature attribution?

좋은 baseline은 다음과 같은 조건을 만족해야 한다:

1. 목표 클래스(Target Class)에 속해야 함
 - 생성된 샘플이 해당 클래스의 특징을 반영해야 한다.
2. 현실적인 샘플이어야 함(Realistic)
 - 자연스럽고 실제 데이터와 유사해야 한다.
3. 입력과 가까운 값이어야 함(Close to the Input)
 - 원본 데이터와 너무 큰 차이가 없어야 한다.

아래의 그림 예시를 살펴보자.

모델이 예측한 결과가 사과이다. 이 샘플과 비교할 기준점 baseline을 정하는 것이 아주 중요한데, baseline을 정할 때 위의 3가지 조건을 고려하여 어떤 baseline이 적절한지 살펴보자.



첫 번째, zero를 기준점으로 설정하는 것은 현실적인 데이터 분포에서 너무 멀어진다 **×**

두 번째, 결정 경계 근처로 이동한 것은 실제 의미 있는 특징을 반영하지 못한다 **×**

세 번째, 다른 클래스에서 가장 가까운 샘플은 실제 원본 샘플과 너무 다를 가능성이 높다 **×**

마지막 네 번째, 이는 입력 샘플과 가장 가깝고, 현실적이며, 여전히 같은 클래스에 속한다 **✓**

그러므로 네 번째 baseline이 모델 해석에 가장 적절한 기준점이 된다!!

A. GANMEX

적절한 Baseline을 찾는 위의 3가지 조건을 수식화하면 아래와 같다.

$$B_c(x) = \underset{\hat{x} \in R^n}{\operatorname{argmin}} (|x - \hat{x}|_1 - \log R(\hat{x}) - \log S_c(\hat{x}))$$

Close to input
 Realistic
 Target class

[수식 설명]

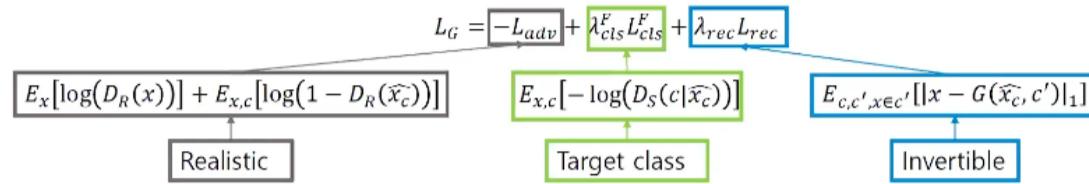
- $|x - \hat{x}|_1$
 - 입력과의 거리 최소화, 생성된 기준점이 원본 입력과 너무 멀어지지 않도록 제약.
- $\log R(\hat{x})$
 - 현실성 유지, 현실적인 샘플이 되도록 GAN의 판별기를 활용하여 조절.
- $\log S_c(\hat{x})$
 - 목표 클래스에 속해야 함, 생성된 샘플이 목표 클래스의 특징을 유지하도록 학습.

StarGAN

StarGAN은 c' 클래스에 속하는 입력 x 를 다른 목표 클래스 c 의 유사한 이미지를 생성하는 모델이다.

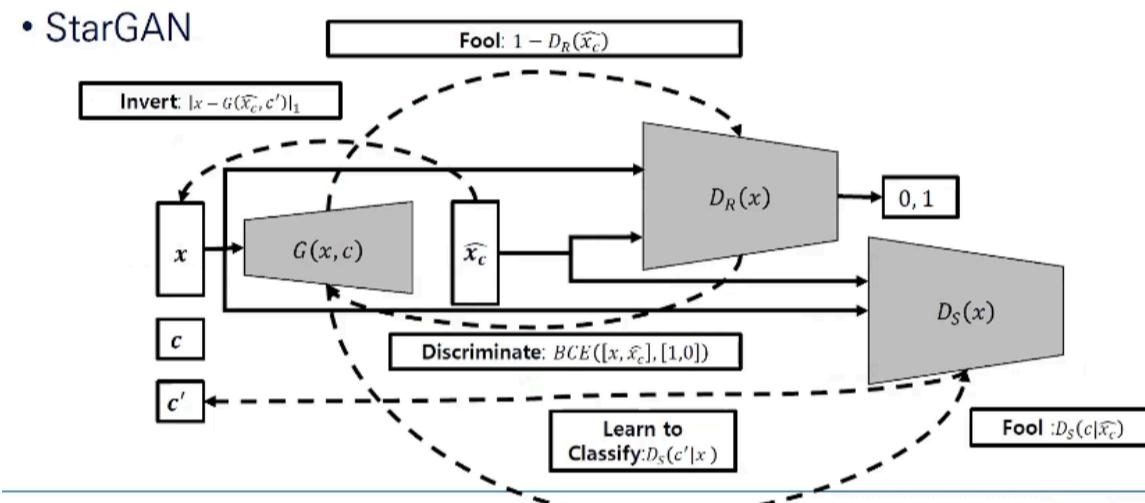
학습 과정에서 다음 3가지 조건을 만족해야 한다.

1. Target Class
2. Realistic
3. Invertibility → 원본 데이터로 되돌릴 수 있어야 함.

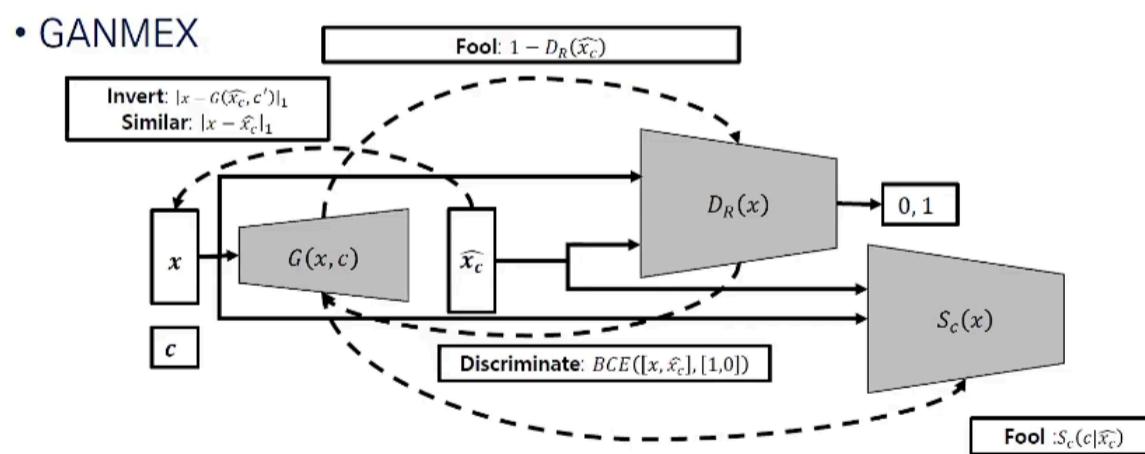


StarGAN의 네트워크 구조는 아래와 같이 구성되어 있다.

- 생성기(G) : 입력 x 를 받아 목표 클래스 c 에 속하는 새로운 샘플을 생성.
- 판별기($D_R(x)$) : 생성된 데이터가 실제 데이터인지 판별.
- 분류기($D_S(x)$) : 생성된 데이터가 올바른 목표 클래스에 속하는지 확인.

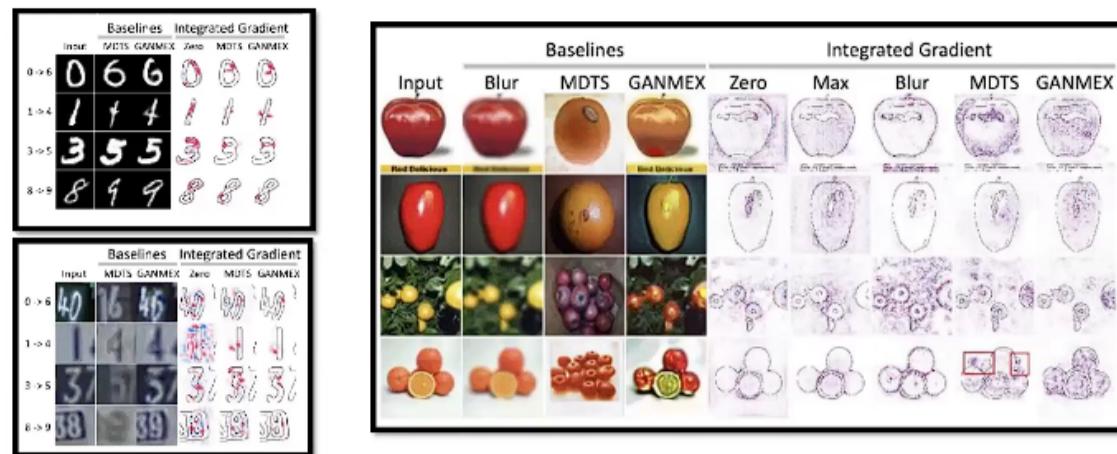


GANMEX는 추가적으로 유사성 제약(Similarity Constraint)을 포함하여 더욱 신뢰할 수 있는 기준점을 찾는다.



Results

이와 같은 방법으로 현실적이면서 다른 클래스에 있는 예시들이 생성됨을 확인할 수 있다.



B. Diffeonmorphic Counterfactuals

아래 이미지가 'brown hair'로 분류되었다.

만약 'blonde hair'로 분류되려면 어떤 작은 변화가 일어나야 할까?

여기서 좋은 Counterfactual은 머리 색만 바꿀 것이다.



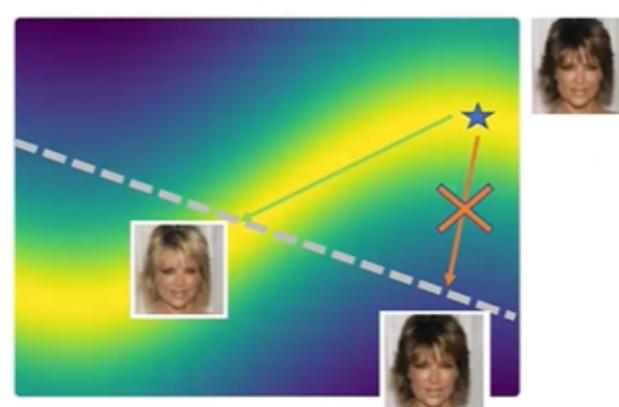
Diffeomorphic Counterfactuals은 모델의 결정 경계를 넘어 반사실 예제를 생성하는 방법 중 하나다.

이 접근 방식은 단순히 결정 경계를 따라 이동하는 것이 아니라

데이터의 구조를 보존하면서(manifold-aware) 반사실을 생성하는 것이 목표이다.

아래 이미지를 살펴보자.

- 점선은 클래스 결정 경계(decision boundary)를 나타낸다.
- 오른쪽의 이미지 (입력 이미지, brown hair)
- 2가지 방법 비교:
 - 직접 결정 경계로 이동(빨간 화살표)
 - 데이터 매니폴드(manifold)를 따라 이동(파란 화살표)



What is Diffeomorphism?

Diffeomorphism은 두 개의 매니폴드 사이의 1:1 변환을 의미한다.

Flow-based 모델을 사용하여 데이터 매니폴드를 유지하며 counterfactual을 생성.

[주요 특징]

- 미분 가능 : 역전파 가능
- 전단사 : 고유한 매팅 보장

- 가역적 : 원래 데이터로 복원 가능

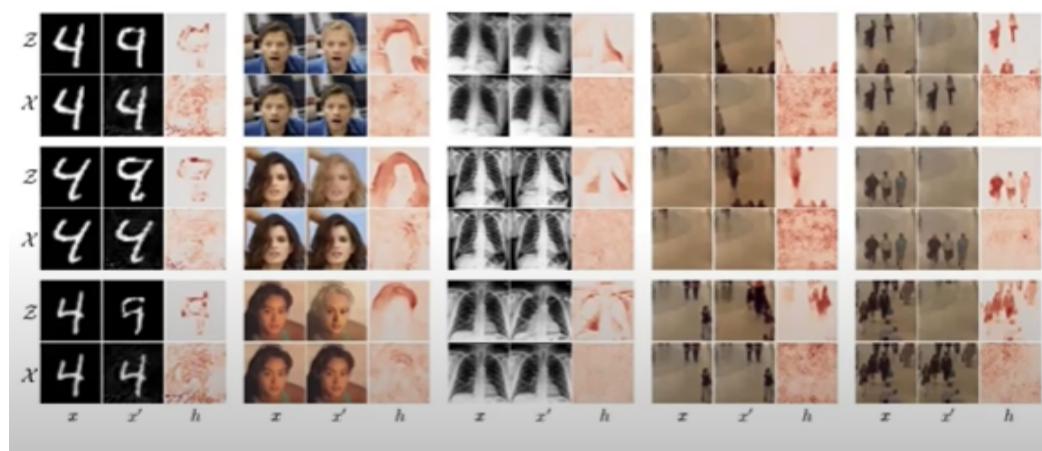
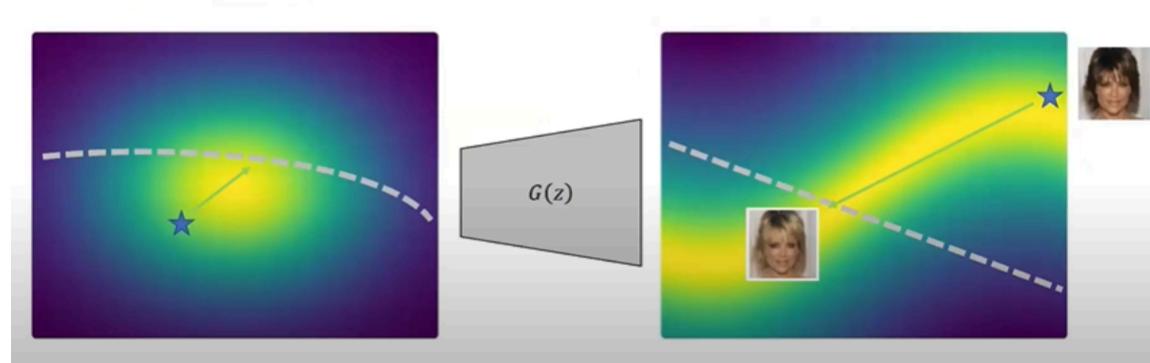
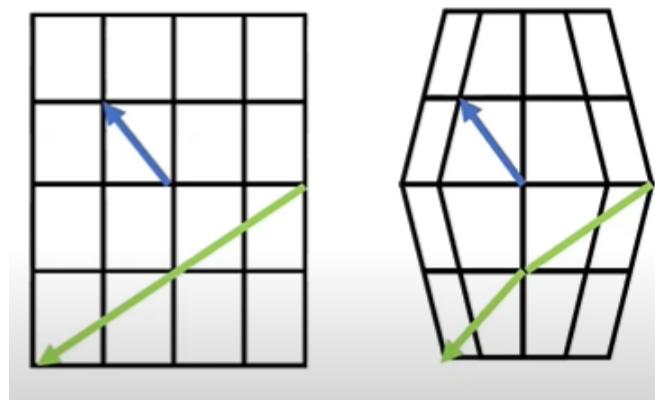
아래 그림은 Diffeomorphism의 개념을 시각적으로 표현한 것이다.

1. 원쪽 그림 (원래 공간)

- 정사각형 그리드로 표현된 좌표 공간을 보여준다.
- 여기서 초록색 화살표는 어떤 방향으로의 변환을 의미하고, 파란색 화살표는 특정한 벡터 이동을 나타낸다.
- 이 상태에서는 공간이 변형되지 않았으며, 기존의 좌표 구조가 유지되고 있다.

2. 오른쪽 그림 (Diffeomorphic 변환 후의 공간)

- 좌표 공간이 비선형적으로 변형됨을 볼 수 있다.
- 기존의 격자 구조가 비틀어졌으며, 선형 변환이 아닌 비선형 변환이 적용된 모습이다.
- 여기서 중요한 점은:
 - 격자가 끊기거나 찢어지지 않고, 연속적으로 변형되었다는 것 → 변환이 **미분 가능**
 - 각 격자가 고유한 위치를 유지하며 역변환이 가능하다 → **가역성**



3. Generative Models as Explainees

이제는 생성 모델(GAN, Flow-based model 등)을 어떻게 설명할 수 있는지 알아보자.

우선 생성 모델을 설명하는 이유가 뭘까?

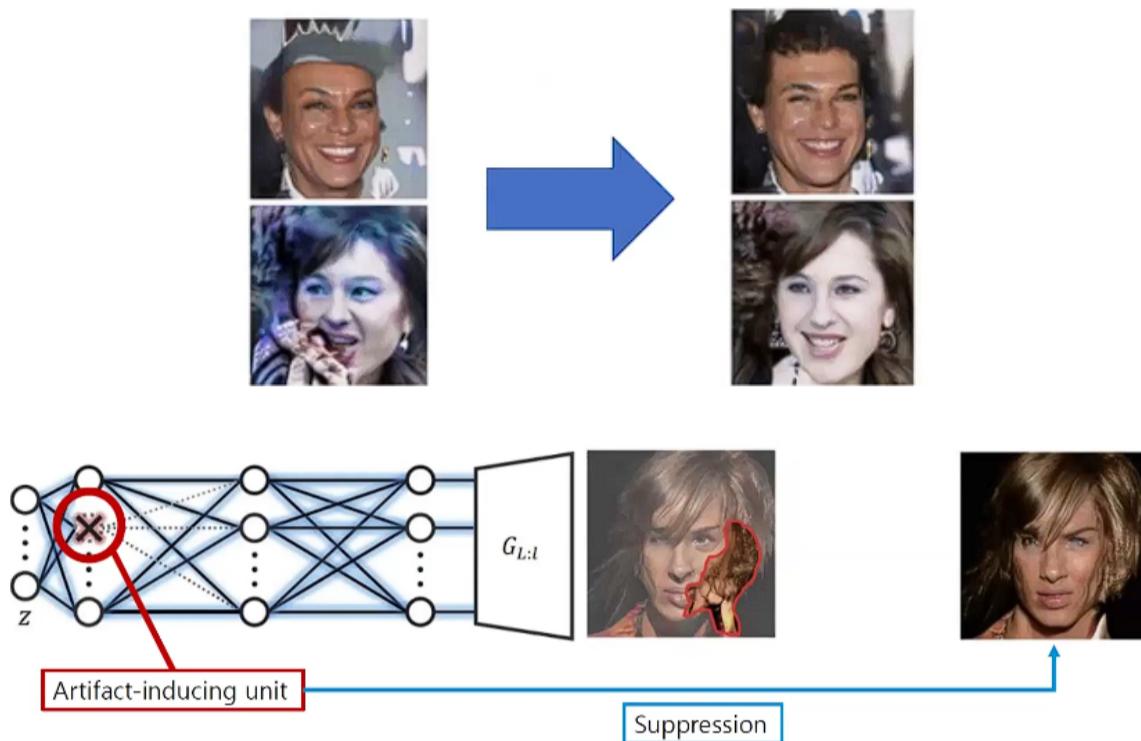
- 생성 모델이 생성한 이미지를 이해하고, 해당 과정에서 어떤 특징이 반영되는지를 파악하는 것이 중요하다.
- 네트워크 내부에서 어떤 개념을 학습했는지 파악하기 위해 네트워크 해부(Network Dissection) 기법을 사용한다.
- 잠재 공간 분석(Latent Space Analysis)을 통해 생성 모델의 각 부분이 어떤 개념을 표현하는지 연구한다.

A. Automatic Correction

생성 모델이 생성한 이미지에서는 결함(Artifacts)이 발생할 수 있다.

이러한 결함을 수정하기 위해 모델을 다시 학습하는 것이 아니라,

결함을 유발하는 특정 뉴런을 찾아 비활성화하는 방식을 사용한다.



💡 핵심 개념

- 네트워크 해부 : 특정 뉴런이 어떤 개념을 담당하는지 확인하는 과정이다.
- 결함을 유발하는 뉴런(Artifact-inducing unit) 식별 : 일부 뉴런이 특정한 결함을 생성하는 역할을 할 수 있다.
- 뉴런 억제(Suppression) : 이러한 뉴런을 찾아 비활성화함으로써 결함을 제거한다.

📌 결함 보정 과정

- 분류기 학습 : 결함이 있는 이미지와 정상 이미지를 분류하는 모델을 학습한다.
- 특징 중요도 분석 : 이미지에서 결함이 발생하는 부분을 찾는다.
- 생성 모델과 결함 매칭 : 생성 모델의 특정 뉴런이 결함을 생성하는지 확인하고 이를 비활성화 한다.

Single-Layer Suppression

단일 계층에서 뉴런을 억제하는 것만으로는 모든 결함을 제거할 수는 없다.

여러 뉴런이 하나의 결함을 생성할 수도 있기 때문이다.

→ 여러 개의 뉴런을 분석하여 더 효과적으로 결함을 제거하는 방법이 필요하다.





하지만, 너무 많은 뉴런을 억제하면 이미지가 망가지는 문제가 발생할 수 있다.

예를 들어, 아래와 같은 건물 이미지에서 20, 40, 60, 80, 98%의 억제를 수행했을 때 억제 비율이 높을수록 이미지 품질이 크게 저하되는 것을 확인할 수 있다.



Conclusion

- 생성 모델의 설명 가능성을 높이기 위해 네트워크 해부(Network Dissection) 및 잠재 공간 분석(Latent Space Analysis) 기법이 사용된다.
- 생성 모델이 만든 결함을 해결하기 위해 뉴런 억제(Suppression) 기법이 도입되었다.
- 하지만, 너무 많은 억제는 오히려 이미지 품질을 저하시킬 수 있기 때문에, 적절한 조정이 필요하다.

B. E-GBAS

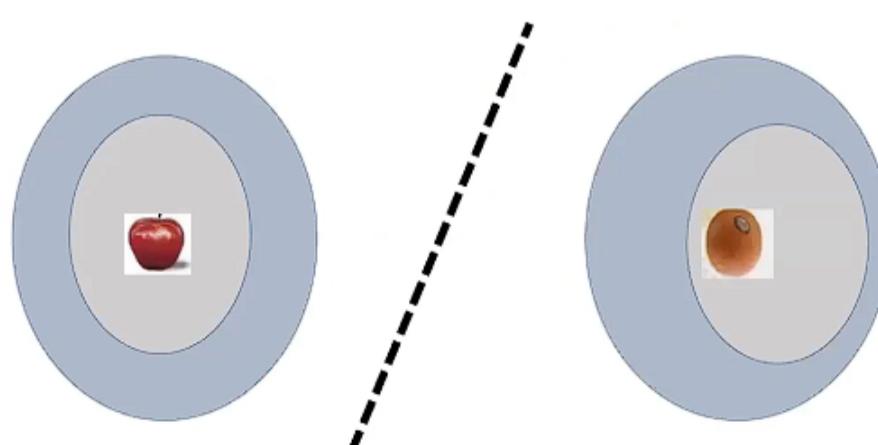
E-GBAS는 생성 모델을 기반으로 한 탐색적 샘플링 기법으로, 생성 경계를 고려하여 더 의미 있는 샘플을 생성하는 방법이다.

이를 통해 생성 모델이 더 일관된 데이터를 학습하고 설명 가능성을 높일 수 있도록 돋는다.

Classification Decision Boundary

아래 이미지는 기존의 분류 결정 경계를 나타낸다.

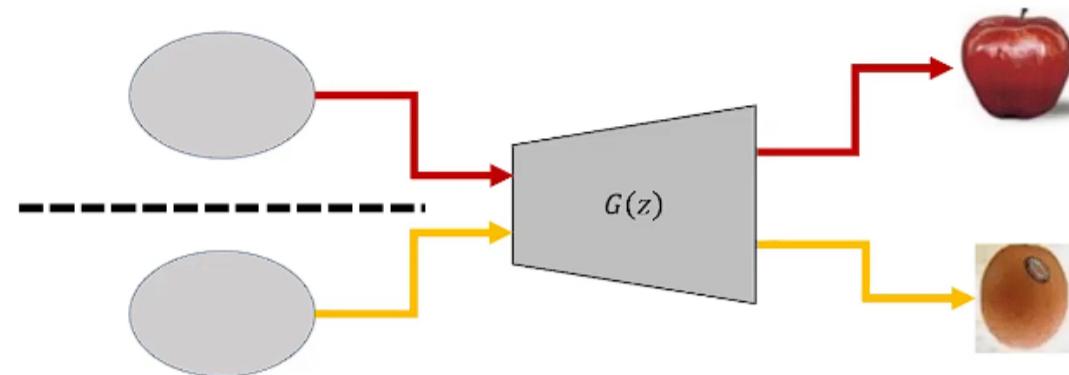
- 사과와 오렌지가 결정 경계의 양쪽에 위치하며, 이는 두 클래스가 구별되는 영역임을 의미한다.
- 즉, 기존 머신러닝 모델에서 입력 데이터가 어느 쪽으로 분류될지 결정하는 기준을 의미한다.



Generative Decision Boundary

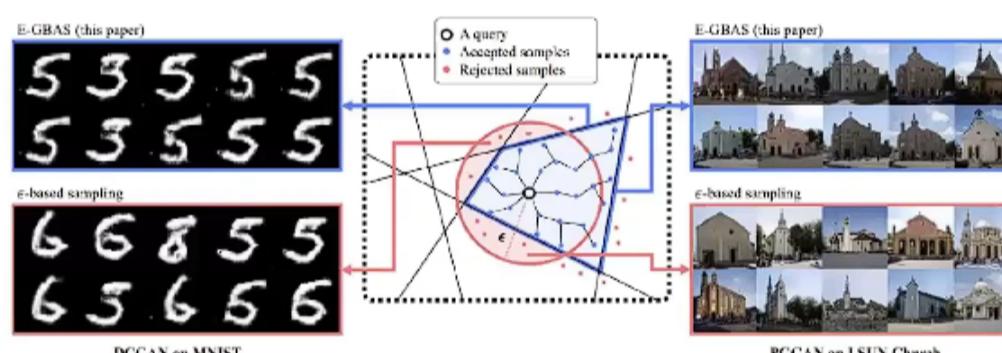
아래 이미지는 생성 모델에서의 결정 경계를 보여준다.

- 생성 모델의 입장에서, 같은 측에 위치한 샘플은 유사한 정보를 가져야 한다.
- 즉, 같은 경계 안에 존재하는 데이터는 비슷한 특성을 공유하는 것이 바람직하다.



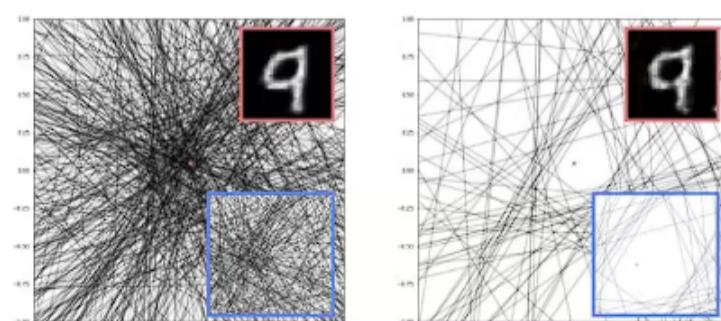
Generative Boundaries & Regions

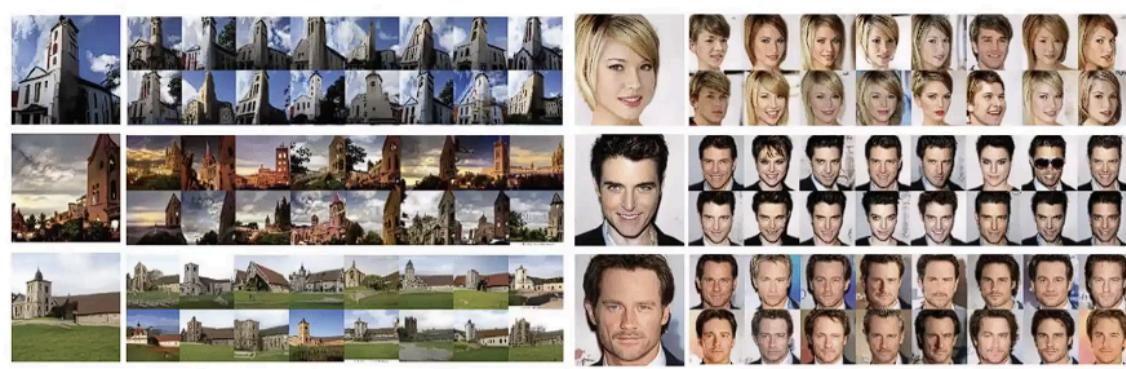
- Generative Boundary : 특정 뉴런의 활성화가 0이 되는 위치에서 정의됨
 - 생성 모델에서 특정 뉴런이 활성화되지 않는다는 것은 해당 뉴런이 특정한 정보를 표현하지 않는다는 것을 의미한다.
- Generative Region : Generative Boundaries에 의해 정의된 여러 영역들의 교차로 형성됨
 - 같은 Generative Region 안에 존재하는 샘플들은 유사한 이미지가 생성된다.
 - 예를 들어, 같은 숫자 "5"를 표현하는 샘플들이 같은 영역에 존재하게 된다.



E-GBAS

- E-GBAS를 적용하면, 같은 생성 경계 안에서 더 일관된 데이터 샘플들이 생성됨.
- 얼굴 생성 예제에서도, 같은 생성 경계 안에서 얼굴의 특징이 유지됨.





[8] Jeon, Gilyoung, Haedong Jeong, and Jaesik Choi. "An efficient explorative sampling considering the generative boundaries of *intelligent*. Vol. 34 No. 04. 2020.

Conclusion

- 생성 모델은 XAI(설명 가능한 인공지능) 분야에서 매우 중요한 역할을 한다.
- 생성 모델은 Explainers (설명 모델)와 Explainees (설명을 받는 대상 모델)로 활용될 수 있음.
 - Explainers: SHAP, GANMEX와 같은 기법으로 특정 예측을 설명하는 용도로 사용됨.
 - Explainees: 생성 모델 자체를 분석하고 설명하는 방법을 연구하는 방향.
- E-GBAS와 같은 기술은 생성 모델이 보다 의미 있는 데이터를 생성할 수 있도록 돋는다.