합성 데이터의 성능 평가를 위한 통계적 데이터 분석 방법 고찰

연구목적

연구목적 1

데이터 부족 문제를 해결하기 위해 합성데이터의 **수요 증가**

연구목적 2

합성 데이터의 품질 평가에 대한 **지표 부족**

이를 보완하기 위해 통계적 방법을 사용하여 합성 데이터와 실제 데이터의 분포를 분석하는 연구가 필요하다.

사전 연구

01 합성데이터란 무엇인가?

: 실제 데이터가 아니라 실제 데이터에서 생성되어 **실제 데이터와 통계 속성이 동일한 데이터**.

02 GAN이란 무엇인가?

: **Generator(생성기)**와 **Discriminator(판별기)**라는 서로 다른 두 개의 네트워크를 적대적으로 학습시키며 **실제 데이터와 비슷한 데이터를 생성**해내는 모델

03 t-SNE란 무엇인가?

: 높은 차원의 복잡한 데이터를 **2차원에 차원 축소하는 방법**으로 낮은 차원 공간의 시각화에 주로 사용하며 차원 축소할 때는 비슷한 구조끼리 데이터를 정리한 상태이므로 데이터 구조를 이해하는데 도움을 준다.

연구방법

01 GAN을 통해 합성데이터 생성

GAN의 Generator(생성기)를 이용해 실제 데이터 분포에 가까운 합성 데이터 분포를 생성한다.

02 t-SNE로 3차원의 이미지 데이터를 2차원으로 축소

앞서 GAN으로 생성한 데이터를 t-SNE를 통해 차원 축소하기 위해 Gradient Descent를 사용하여 KL divergence를 최소화하며, 이를 시각적으로 표현한 그래프를 통해 데이터의 패턴을 발견한다.

03 커널 밀도 추정으로 두 분포간 유사도 확인

커널 밀도 추정을 사용하여 두 분포를 시각화하고 분포가 유사한지를 확인한다.

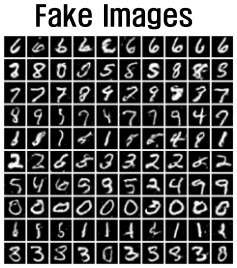
04 스미르노프 검정을 통한 적합도 검정

스미르노프 검정을 실시하여 두 분포가 같은지 여부를 검증한다.

연구결과

01 GAN을 통해 합성데이터 생성

Fake Images



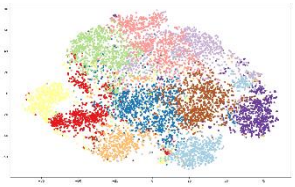
Real Images

실제 데이터 세트를 기반으로 GAN의 Generator(생성기)를 사용하여 합성 데이터 세트를 생성하였다.

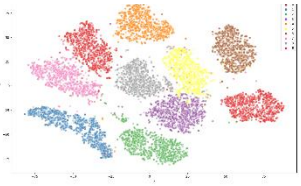
02 t-SNE로 3차원의 이미지 데이터를 2차원으로 축소

앞서 GAN으로 생성한 3차원 이미지 데이터인 합성 데이터와 실제 데이터를 t-SNE를 이용해 2차원으로 시각화한 결과는 다음과 같다.

tSNE_fake_with_kmeans



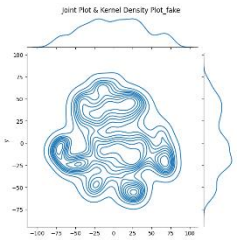
tSNE_real_with_kmeans



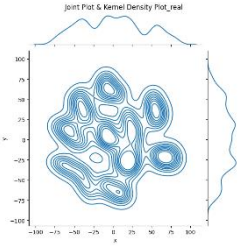
두 데이터 간의 분포를 비교해보면 확실히 GAN으로 생성한 데이터가 잘 분류되지 않고 boundary 근처에서 애매하게 존재하는 것을 볼 수 있다.

03 커널 밀도 추정으로 두 분포간 유사도 확인

Fake Images



Real Images



합성데이터에서도 실제 데이터와 유사한 위치 분포를 보였지만 경계가 부드러워 (5, 8 등) 생성된 데이터에서 일부 숫자(7, 9 등)는 애매한 모양을 가진 것으로 확인된다.

04 스미르노프 검정을 통한 적합도 검정

정규성 검정 결과 합성 데이터와 실제 데이터는 정규분포를 따르지 않는다. 따라서 스미르노프 검정을 사용해 분포 모양이 얼마나 유사한지 검정할 수 있다.

H_0 : **모든** 합성 데이터의 분포는 실제 데이터 분포와 **동일**
H_1 : **적어도 하나의** 합성 데이터가 실제 데이터와 **다르다**

스미르노프 검정 결과 검정 통계량이 약 0.55이고 유의확률이 0에 가깝기 때문에 귀무가설(H_0)을 기각하였고, 합성 데이터와 실제 데이터는 일치하다는 충분한 근거가 없다는 결론을 얻을 수 있었다.

결론 및 제언

01 GAN의 Discriminator는 Generator가 생성한 합성 데이터에서 실제 데이터와 유사한 것만을 선택한다. 이외에도, 더 실제 데이터와 더 유사한 합성 데이터를 만들기 위해 이들 간의 유사도를 검정하였다.

02 xai 기법을 활용하여 유효한 feature를 추출하여 2차원으로 이미지 데이터를 매핑해 보면 더 정확한 검정이 가능할 것으로 예상된다.

03 합성 데이터의 경계가 모호할 때, 해당 데이터를 학습하여 딥러닝 모델이 값을 부드럽게 구분하는 경계를 생성하도록 함으로써, 추후 복잡한 데이터를 학습하고 예측할 때 성능을 향상시킬 수 있다.