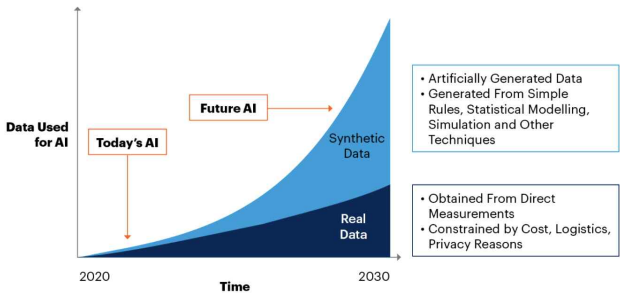
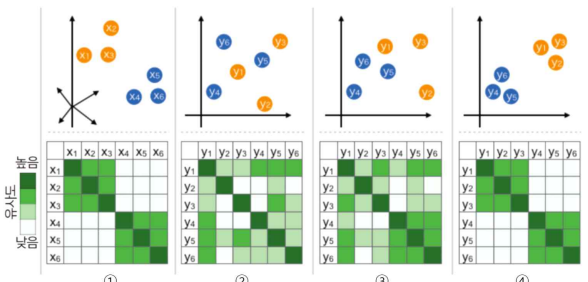


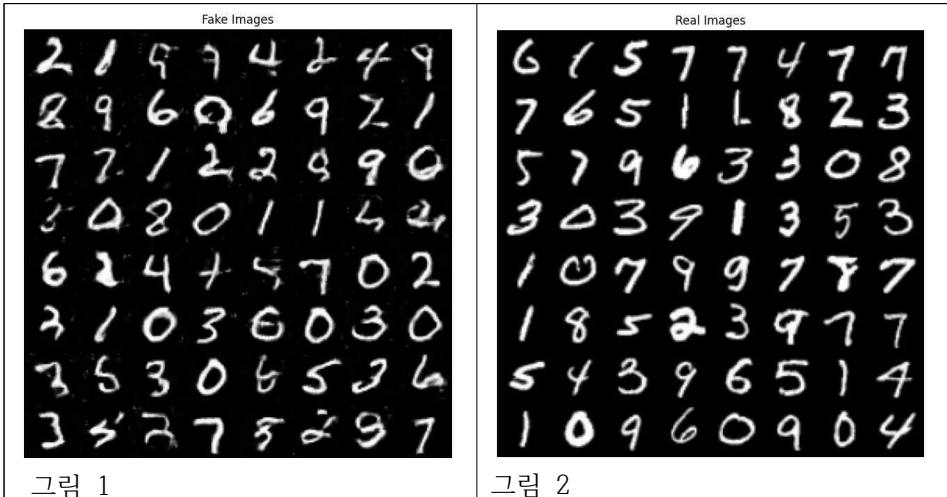
2023 경영경제대학 학술제 참가보고서

제 목	합성 데이터의 성능 평가를 위한 통계적 데이터 분석 방법 고찰				
신청팀 인적사항	성명	학번	생년월일	연락처	E-mail
	김성연	20190146	19980401	01023986822	ksy9744@cau.ac.kr
	정서현	20190247	19990324	01046624794	jsh1021902@naver.com
선정주제	합성데이터의 성능평가를 위한 데이터 분포 분석				
주제 선정 이유 (사회적 현상과 관련지어 서술)	<p>합성 데이터는 데이터 부족 문제를 해결하고 비용 및 시간을 절약할 수 있으며, 가트너의 2021년 보고서에 따르면 합성 데이터가 전체 데이터의 대부분을 차지할 것으로 예측되고 있습니다.</p> <p>By 2030, Synthetic Data Will Completely Overshadow Real Data in AI Models</p>  <p>그러나 현재 합성 데이터의 품질 평가에 대한 지표는 부족하며, 이를 보완하기 위해 통계적 방법을 사용하여 합성 데이터와 실제 데이터의 분포를 분석하는 연구가 필요합니다. 이를 통해 합성 데이터의 품질을 보다 정확하게 평가할 수 있으며, 이를 활용하여 인공지능 분야 등 다양한 분야에서 보다 효과적인 결과를 얻을 수 있습니다. 이러한 연구가 현재 합성 데이터를 다루는 다양한 분야에서 중요하게 다뤄지고 있으며, 이를 통해 실제적인 문제 해결에 큰 도움을 줄 수 있습니다.</p>				
주제 분석 및 연구	<p>I. 사전 연구</p> <p>① 합성 데이터 : 합성 데이터는 실제 데이터가 아니라 실제 데이터에서 생성되어 실제 데이터와 통계 속성이 동일한 데이터를 말합니다.</p> <p>② GAN : GAN은 Generator(생성기)와Discriminator(판별기)라는 서로 다른 두 개의 네트워크를 적대적으로 학습시키며 실제 데이터와 비슷한 데이터를 생성해내는 모델을 일컫습니다.</p> <p>③ t-SNE : t-SNE는 비선형적인 방법의 차원 축소 기법으로 고차원 데이터셋을 시각화하는 데 성능이 좋습니다.</p>  <p>t-SNE 알고리즘은 고차원 공간에서의 점들의 유사성과 그에 해당하는 저차원 공간에서의 점들의 유사성을 계산합니다. 이를 위해 점들의 유사도를 조건부 확률을 이용하여 계산하고 저차원 공간에서 데이터 요소를 완벽하게 표현하기 위해 고차원 및 저차원 공간에서 이러한 조건부 확률 (또는 유사점) 간의 차이를 최소화하려고 시도합니다. 이러한 방식으로, t-SNE는 다차원 데이터를 보다</p>				

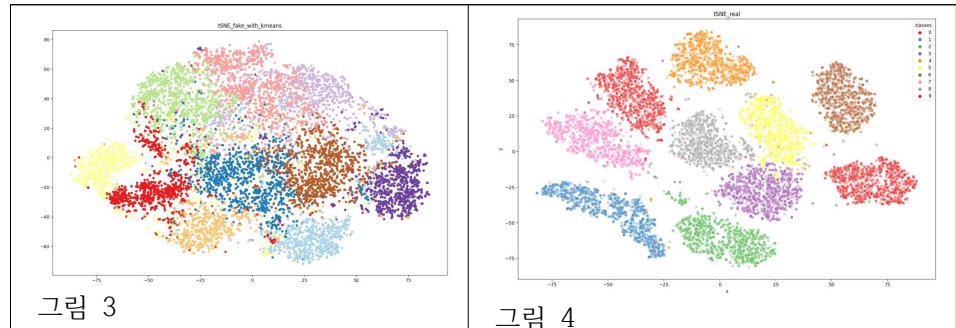
낮은 차원 공간으로 매핑하고, 다수의 특징을 갖는 데이터 포인트의 유사성을 기반으로 점들의 클러스터를 식별함으로써 데이터에서 패턴을 발견할 수 있습니다.

II. 연구 방법 및 결과

① DCGAN을 이용한 합성 데이터 생성 : 60000장의 mnist 학습데이터 세트를 기반으로 DCGAN을 이용해 10000장의 합성 데이터 세트를 생성하였습니다. 아래 왼쪽 데이터(그림 1)가 합성데이터의 결과물입니다.



② t-SNE를 통해 3차원인 이미지를 2차원으로 차원 축소 후 시각화 : 앞서 생성한 3차원 이미지 데이터인 합성 데이터(그림 3)와 실제 데이터를 t-SNE를 이용해 2차원으로 시각화한 결과(그림 4)는 다음과 같습니다.



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

- 이때 합성 데이터는 무작위로 생성하였기에 label이 존재하지 않아 K-Means를 통해 10개의 레이블로 나누어 주었고, 이를 시각화하였습니다.

(from sklearn.cluster import KMeans)

[그림 5는 K-Means를 통해 레이블을 나눈 합성 데이터를 시각화한 것입니다.]

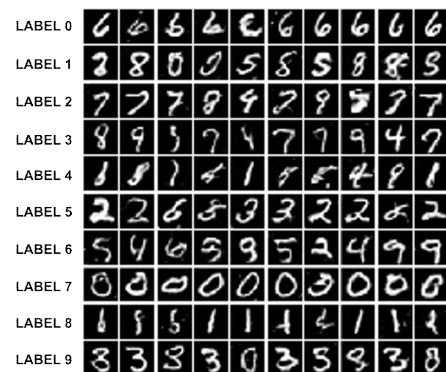


그림 5

이를 통해 kmeans만으로도 충분히 유의미한 분류가 가능함을 볼 수 있습니다.

③ 커널 밀도 추정으로 두 분포간 유사도 확인

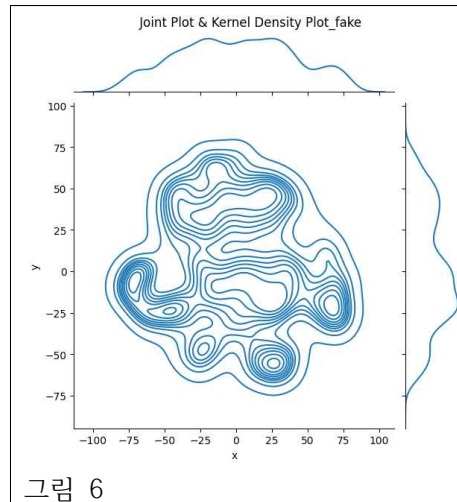


그림 6

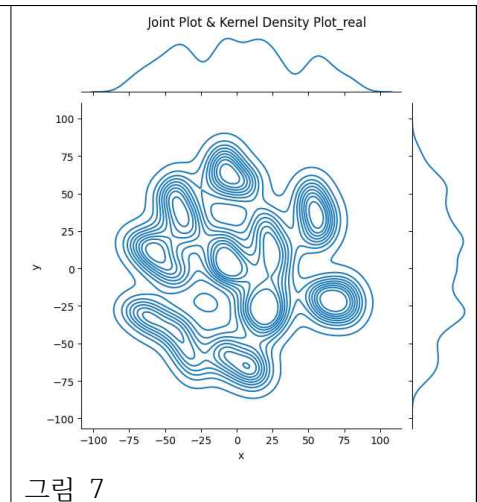


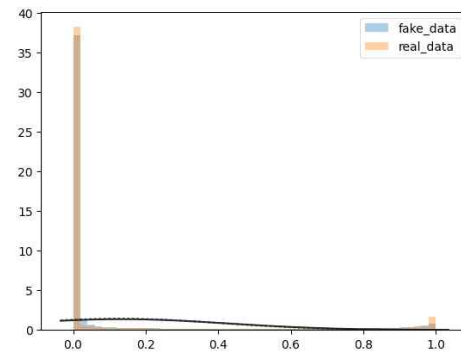
그림 7

seaborn 패키지의 joint plot을 통해 밀도를 파악해 본 결과 글씨가 나타나는 부분의 분포를 파악할 수 있었고, 유사한 위치에서 글자를 생성해내는 것을 볼 수 있었습니다.

차이점은 실제 데이터보다 경계가 부드럽게 (5와 8처럼 유사한 글자의 경계가 뚜렷하지 않게) 데이터를 생성한다는 것이었습니다.

또한 이를 바탕으로 그림 5(합성데이터 샘플)를 다시 보면 7,9 등의 유사한 모양을 가진 숫자에서 7과 9의 중간정도의 느낌으로 보이는 애매한 모양을 만들어내는 것을 볼 수 있습니다.

④ 스미르노프 검정을 통한 적합도 검정



이미지 특성상 배경에 치우치는 경향이 있어 분포가 정규성을 띄지 않습니다. 이와 같이 정규성을 띄지 않는 경우에는 정규화를 하거나 비모수적인 검정을 사용해야 합니다.

하지만 정규화를 하기에는 어려움이 있을 것으로 판단되어 비모수적인 검정을 사용하였습니다.

이를 확인하기 위한 방법으로 카이제곱 검정, 콜모고로프-스미르노프 검정, 윌콕슨-맨-휘트니 검정, 크루스칼-왈리스 검

정, 맨-휘트니 U 검정 등이 있습니다.

이는 간단하게 scipy.stats 패키지를 통해 확인할 수 있습니다.

콜모고로프-스미르노프 검정

```
import scipy.stats as stats
stats.kstest(real_array, fake_array)
```

귀무가설을 “모든 합성 데이터의 분포는 실제 데이터의 분포와 동일하다.” 라고 하고 대립가설을 “적어도 하나의 합성데이터가 실제 데이터의 분포와 동일하지 않다.” 라고 할 때, 스미르노프 검정 결과 검정 통계량이 0.55이고 유의확률이 0에 가깝기 때문에 유의확률 0.05에서 귀무가설을 기각합니다. 이는 합성 데이터와 실제 데이터는 일치한다는 충분한 근거가 없다는 결론을 얻을 수 있습니다.

위와 같은 방식으로 맨-휘트니 검정(from scipy.stats import mannwhitneyu)도 수행해본 결과 같은 결론이 나왔습니다.

III. 결론 및 제언

GAN의 Discriminator는 합성 데이터가 실제 데이터와의 분포가 유사해지도록

		<p>학습하게 하는 역할을 합니다. 결국 이는 현재 우리가 수행한 목표와 일치합니다. 이 외에도 합성 데이터와 실제 데이터를 비교하는 다양한 방법이 존재하고 있습니다. 현재는 시간의 한계로 생성된 합성 데이터와 실제 데이터를 통계학과 강의에서 배운 분포의 개념과 비모수통계학의 지식을 이용하여 커널 밀도 추정과 스미르노프 검정등의 방식을 통해 시각적, 수치적으로 비교하고 있습니다.</p> <p>추후에는 위에서 말한 다른 검정방법으로도 비교해보고 딥러닝 모델을 통해 중요한 변수를 추출하여 차원을 축소한 뒤에 분포를 비교해볼 예정입니다.</p> <p>또한, 합성 데이터의 경계가 모호한 것을 확인하였는데, 이는 해당 데이터를 학습하여 딥러닝 모델이 값을 부드럽게 구분하는 경계를 생성하도록 함으로써, 오버피팅을 막고 추후 새로운 데이터를 예측할 때 더 좋은 성능을 낼 수 있게 하는 역할을 할 것으로 기대됩니다.</p>
현실성 및 지속가능성	주제의 현실 적용 가능성	<p>현재 인공지능 분야에서는 ChatGPT와 같은 생성형 AI를 학습시키기 위한 데이터에 대한 수요가 증가하고 있습니다. 이에 따라, 개인정보와 저작권 등에 대한 문제를 해결할 수 있는 합성데이터의 활용도 더욱 중요시되고 있습니다. 그러나, 이러한 AI 학습 데이터의 생성과 활용에 대한 규제와 갈등도 또한 증가하고 있습니다.</p> <p>특히, 최근 G7 회의에서는 AI를 위한 위험 기반 규제에 대한 합의가 이뤄졌으며, 민주주의와 인권 등을 강조하는 원칙도 수립되었습니다. 이러한 규제와 원칙은 AI 학습 데이터의 생성과 활용에 있어서 중요한 가이드라인이 될 것입니다. 또한, EC가 AI법 초안에 합의하여 AI시스템을 위험도에 따라 규제하며, 생성형 AI 개발에 쓰인 원데이터의 저작권을 모두 공개하도록 하는 조항을 새롭게 포함하였습니다. 이러한 규제와 조치는 AI 학습 데이터의 생성과 활용에 대한 투명성과 공정성을 높이는 데 기여할 것으로 기대됩니다.</p> <p>그리고 합성데이터는 실제 데이터를 비식별화하여 생성되는 데이터로, 개인정보와 저작권 등에 대한 문제를 해결할 수 있습니다. 따라서, 합성데이터의 활용도 더욱 중요시되고 있으며, 합성데이터 생성과 활용에 대한 규제와 지침이 필요합니다. 이를 통해 인공지능 분야에서의 규제와 갈등을 예방하고, 안전하고 공정한 AI 학습 데이터의 생성과 활용을 실현할 수 있을 것으로 기대됩니다. 실제로 저희가 확인해 본 과정들은 시각화하기 편리하며, 딥러닝 모델들에 비해 쉽게 이해할 수 있다는 장점이 있습니다.</p>
	주제의 발전 가능성 및 타당성	<p>합성데이터의 성능평가를 위한 데이터 분포 분석을 연구하는 것은, 합성데이터 생성에 대한 요즘 핫한 토픽에 대해 더욱 명확한 방향성을 제시할 수 있는 연구 주제입니다. 이 연구를 통해, 합성데이터 생성에 대한 지표를 제시하고 향후 합성데이터가 인공지능 분야 등에서 효과적으로 활용될 수 있도록 지원할 수 있습니다. 예를 들어, 이러한 연구 결과를 바탕으로 합성데이터 생성 방법을 개선하거나, 합성데이터의 품질 평가 지표를 제시하여 합성데이터의 품질을 보다 정확하게 평가할 수 있도록 도움을 줄 수 있습니다. 또한, 이러한 연구는 합성데이터 생성 및 활용에 대한 규제와 갈등을 예방하고, 안전하고 공정한 AI 학습 데이터의 생성과 활용을 실현하는 데 기여할 수 있습니다. 이러한 의의로 미래에도 이 연구 분야는 꾸준한 발전이 이루어질 것으로 기대됩니다.</p>
창의성 및 전문성	주제에 대한 분석의 차별성 및 독창성	<p>데이터 평가는 인공지능 분야에서 매우 중요한 연구분야 중 하나입니다. 현재까지는 GAN discriminator를 기반으로 한 평가 메트릭이 가장 많이 사용되어 왔지만, 이 외에도 매우 다양한 평가 방법이 연구되고 있습니다. 이러한 평가 방법은 단순히 생성된 데이터가 실제 데이터와 유사한지를 판단하는 것뿐만 아니라, 생성된 데이터의 분포와 실제 데이터의 분포간의 유사성을 평가하거나, 단순한 데이터 복제(mimicking)를 방지하는 등의 목적으로 다양한 방법이 연구되고 있습니다.</p> <p>특히, 조건부 기반 혹은 지식 기반으로 생성된 데이터의 충실도와 다양성을 평가하는 연구는 매우 중요합니다. 이는 ChatGPT와 같은 생성형 AI가 거짓 정보를 그럴싸하게 늘어놓는 것을 방지하고, 실제로 유용한 정보를 생성할 수 있</p>

		<p>도록 함으로써, 인공지능 분야에서 더욱 발전해 나갈 수 있도록 기여할 것입니다.</p>
	<p>의견 도출의 논리 및 과정</p>	<p>합성데이터의 수요가 많은 의료데이터등의 정보보안이나 개인정보보호가 필요한 데이터들의 경우에는 예측력이 아무리 좋은 모델이라 하여도 설명력이 있는 모델과 데이터가 요구됩니다. 이에 따라 합성 데이터의 품질 평가가 매우 중요한 요소가 됩니다.</p> <p>이에 따라 설명력이 낮은 딥러닝 모델들에 선행해서 기존에 사용했던 통계적 이론들을 이용해서 합성데이터의 설명력을 높일 수 있을지에 대해 연구해보고자 하였습니다.</p> <p>저희는 '좋은 합성데이터'는 실제 데이터셋과 유사한 분포를 가지는 가짜 데이터라 가정하였고 이를 확인하기 위한 방법을 찾고자 하였습니다. 하지만 이를 수행하기 위해서는 이미지 데이터를 낮은 차원으로 축소하는 과정이 필수적이었고, 이를 수행하기 위한 과정이 상당히 복잡하다는 것을 깨달았습니다.</p> <p>t-SNE, PCA, LDA 등의 방법을 찾았고, t-SNE를 활용하여 차원을 2차원으로 축소하여 시각화를 할 수 있었습니다. 또한 모든 픽셀데이터를 min-max scaling을 한 뒤에 합성데이터를 잘 만들었는지 확인하기 위한 작업으로 kmeans 클러스터링을 통해 유사한 데이터끼리 label을 묶는 작업을 수행하였습니다.</p> <p>으며, 이를 통해 보다 효과적인 결과를 얻을 수 있습니다. 이후에는 차원을 축소하고 전처리한 데이터들을 바탕으로 비모수 통계학에서 배운 검정 방법(스미르노프, 맨 휘트니 검정)들을 수행해 볼 수 있었습니다.</p>
	<p>전공융합 적절성</p>	<p>이 주제에 대한 연구를 진행하기 위해, “이미지 데이터 분석을 위한 딥러닝”이라는 전공 수업에서 배운 GAN 모델을 이용하여 합성데이터를 생성하였습니다. 이 합성데이터를 분석하기 위해 차원 축소 방법 중 하나인 t-SNE를 이용하여 이미지 데이터를 2차원으로 축소하였습니다.</p> <p>그리고 이렇게 생성된 합성데이터와 실제 데이터의 분포를 비교하기 위해 “기초 통계학”에서 배운 분포와 밀도를 확인해보고자 하였고, 데이터가 정규분포를 따르지 않았기에 “비모수통계학”에서 배운 검정방법을 활용하고자 하였습니다.</p> <p>스미르노프 검정을 이용하여 실제 데이터와 합성 데이터 간의 분포의 유사도를 비교하고, 이를 통해 합성데이터가 실제 데이터와 얼마나 유사한지를 분석할 수 있었습니다. 추후 이미지 데이터의 feature를 줄이고 추출하는 과정에서 회귀분석 등의 시간에 배운 PCA, LDA 등의 방법도 활용할 수 있을 것입니다.</p>
	<p>의의 및 기대효과</p>	<p>합성데이터는 인공지능 분야에서 매우 중요한 데이터로서 사용됩니다. 합성데이터의 품질 평가는 이러한 데이터를 사용하는 분야에서 매우 중요한 문제입니다. 이러한 문제를 해결하기 위해 통계적 방법을 사용하여 합성 데이터와 실제 데이터의 분포를 분석하는 연구가 활발히 진행되고 있습니다.</p> <p>이러한 연구 결과는 합성 데이터의 성능 개선과 더 나은 응용 분야를 개발하는 데 활용될 수 있습니다. 또한, 합성 데이터의 품질 평가는 머신 러닝 및 인공지능 분야에서 매우 중요합니다. 따라서 이러한 주제의 발전 가능성과 타당성은 매우 높다고 볼 수 있습니다.</p> <p>또한, 합성데이터의 성능 평가가 보다 정확하고 신뢰성 높은 방식으로 이루어질 수 있게 됩니다. 이는 인공지능 분야에서 더욱 정확하고 유용한 모델링 및 예측을 가능하게 하여, 다양한 분야에서의 응용 가능성이 확장될 수 있음을 의미합니다. 또한, 이러한 연구는 합성데이터 생성 및 활용에 대한 규제와 갈등을 예방하고, 안전하고 공정한 AI 학습 데이터의 생성과 활용을 실현하는 데 기여할 수 있습니다. 이러한 의의와 기대효과로 미래에도 이 연구 분야는 꾸준한 발전이 이루어질 것으로 기대됩니다.</p>