합성데이터 유용성 측정지표 조사연구

성민경, 이강원, 한주연, *심기창, **김태훈
한국정보통신기술협회, *㈜이지서티, **㈜딥핑소스
[mksung, blong116, hanjy]@tta.or.kr, *gcshim@easycerti.com, **pete.kim@deepingsource.io

Research on usefulness measurement of synthetic data

Sung Min Kyoung, Lee Kangwon, Han Ju Yeun, *Shim Gi Chang, **Tae-hoon Kim
Telecommunications Technology Association, *EASYCERTI Co., Ltd., **Deeping Source Inc.

요 약

인공지능을 효과적으로 활용하기 위해서는 훈련 데이터의 다양성 및 정확성이 확보되어야 하지만 실측 데이터가 부족한 분야, 데이터가 편향된 분야, 개인정보 보호 이슈가 있는 분야 등에서 수집된 데이터는 다양성(충분성, 균등성 등)부족 및 편향성 존재로 인해 데이터 활용이 어려운 상황이다. 이에 따라 최근 합성데이터를 활용하여 데이터 부족과 개인정보 이슈를 해결하려는 사례가 늘어나고 있으며 합성데이터의 유용성 측정에 관한 연구도 활발히 이루어지고 있다. 본 논문에서는 합성데이터의 유용성을 측정할 수 있는 기존 지표조사 및 한국정보통신기술협회(TTA)에서 활용하고 있는 합성데이터 유용성 측정 지표를 알아본다.

Ⅰ. 서 론

인공지능 훈련을 위해서는 양질의 데이터가 대량으로 필요하지만 측정 대상이 되는 사건의 발생 빈도가 낮거나, 데이터가 편향되어 있는 경우 인공지능 훈련에 필요한 데이터를 충분히 확보하기가 어려우며, 충분한 양의 실데이터를 확보할 수 있는 분야라 할지라도 인공지능 모델의 거대화에 따라 데이터 획득 및 정제, 라벨링에 필요한 비용이 기하급수적으로 늘어난다. 또한, 의료 데이터와 같이 개인정보를 침해할 소지가 있는 분야에서는 데이터 수집 과정에서부터 제한이 있으며 활용 또한 쉽지 않다. 이에 따라 인공지능 훈련을 위한 새로운 데이터 확보 방안에 대한 요구가 증대되고 있다.

합성데이터는 생성 모델 등을 통해 원본데이터로부터 생성된 데이터를 의미하며, 적대적 생성 신경망(GAN) 모델의 출현 이래[1], 이를 이용한 데이터 증강 및 합성을 통한 데이터 확보 방안이 주목받고 있다. 합성데이터 기법을 통해 수집 과정에서 확보하기 어려운 데이터를 합성할 수 있어 자동차 자율주행, 희귀 질환, 소수 언어, 결함 탐지, 사이버 보안(침입자 탐지 IDS intruder detection system) 등 실데이터가 부족한 환경에서 활발히 활용된다. 또한, 개인정보보호위원회의 가명정보 처리 가이드라인[2]은 합성데이터가 익명정보 처리기법 중 하나가 될 수 있다고 소개하고 있다. 가트너의 2021년 보고서에 따르면, 2030년에는 인공지능 모델을 훈련시키기 위해 사용되는 데이터 중 합성데이터의 비중이 실데이터의 비중을 상회하게 될 것으로 예측된다.

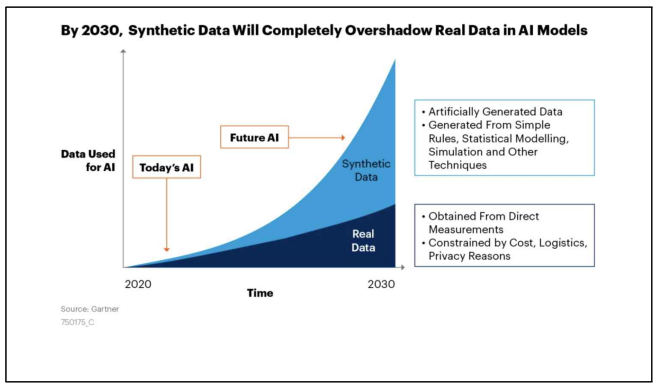

그림 1. 인공지능 훈련용 합성데이터 활용 추세 예측(출처: 가트너 2021)

Ⅱ. 본 론

합성데이터의 유용성 측정은 합성데이터의 형태에 따라 정형데이터 측정과 비정형데이터 측정으로 구분된다. 정형데이터에 대한 유용성 측정은 1차원 분포 유사도, 2차원 관계 유사도, 모형 성능 유사도 및 성향점수 등이 있다. 1차원 분포 유사도는 원본데이터와 합성데이터에 대해 그림2와 같이 각 컬럼의 분포 모양을 비교하여 측정할 수 있다. Kolmogorov-Smirnov 통계량[3], 카이제곱 통계량[4]이 1차원 분포 유사도 측정의 대표적인 방법이다.

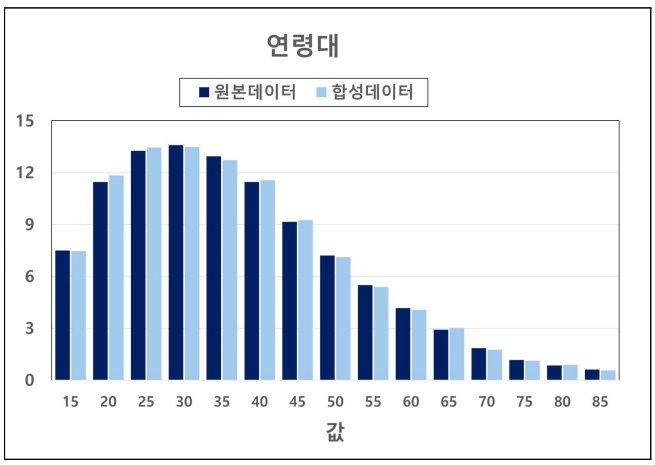

그림 2. 1차원 분포 유사도 측정

2차원 관계 유사도는 컬럼 분포 이외에 각 컬럼의 상관관계의 유사 정도를 지표로 측정한다. 피어슨 상관계수와 Cramer's V 상관계수가 널리 활용된다.

모형성능 유사도 지표는 회귀모형, 로지스틱 회귀 모형으로 원본데이터와 합성데이터를 각각 모형화하고 모형 계수에 대한 신뢰구간을 비교하여 측정한다. 성향점수는 원본데이터와 합성데이터를 일정 비율로 혼합하여 머신러닝 모형이 두 데이터를 구분하는 정확도가 50:50에 가까운지를 측정한다.

비정형데이터에 대한 유용성 측도는 PSNR[5], SSIM[6], MSE[7], Perceptual Loss[8] 등이 있다(표1).

표1. 비정형데이터 주요 측정지표

지표명	지표설명
Peak Signal-to-Noise Ratio (PSNR)	원래 이미지와 비교하여 복원된 이미지의 품질을 비교 측정
Structural Similarity Index Measure (SSIM)	구조 정보를 기반으로 두 이미지 간의 유사성을 측정
Mean Squared Error (MSE)	두 이미지 간의 평균 제곱 차이를 측정
Perceptual Loss	픽셀 수준의 차이가 아닌 인산 수준의 지각 차이를 기반으로 두 이미지 간의 유사성을 측정

한편, 인공지능 학습용 데이터 구축사업에서 비정형 합성데이터에 대한 유용성 검증은 현재 한국정보통신기술협회(TTA)에서 수행하고 있으며[9], FID(Frechet Inception Distance)[10], VTT(Visual Turing Test)[11], 유사도를 주요 지표로 활용한다(표2).

FID는 영상 집합 사이의 거리를 나타내며, 생성된 영상의 품질(화질)을 평가하는데 사용된다. 이 지표는 거리가 가까울수록(값이 작을수록) 품질이 좋은 영상으로 판단한다. FID의 경우 영상/이미지의 흑백/컬러 여부에 따라 기준이 달라지므로 데이터의 상황에 적합한 기준선정이 중요하다.

VTT는 컴퓨터상 이미지가 실제 사물의 모습과 유사한지를 판별하는 과정이다. 실제 이미지와 합성 이미지의 무작위 혼합 샘플을 해당 분야 전문가가 직접 구분하는 시험으로, 정확도 50%가 이상적이며, 너무 높은 정확도는 합성 이미지 불량, 너무 낮은 정확도는 실제 이미지 불량으로 해석할 수 있다. VTT는 해당 분야 전문가의 숙련도 및 성향에 따라 결과가 달라질 수 있으므로 교차검증을 위한 충분한 전문가의 수 확보가 필요하다.

유사도는 합성데이터 기반 인공지능 모델과 원본데이터 기반 인공지능 모델의 정확도 또는 AUC(Area Under the Curve) 차이를 비교한다. 합성데이터와 원본데이터로 각각 인공지능 모델을 훈련한 후 원본데이터의 일부를 각 모델의 테스트 데이터로 평가한다. 평가결과의 차이가 작을수록 합성데이터의 품질이 높은 것으로 판단한다. 본 항목에서 유의할 점은 합성데이터로 훈련한 모델의 평가결과가 높은 것이 중요한 것이 아니라 얼마나 원본데이터로 학습한 모델의 평가결과가 유사한지가 높은 품질의 기준이라는 것이다.

표2. 학습용 데이터 구축사업에서 활용 중인 비정형 합성데이터 유용성 측정지표

지표명	지표설명
FID (Frechet Inception Distance)	- 영상 집합 사이의 거리를 나타내며, 생성된 영상의 품질을 평가하는데 사용 - 합성 이미지와 실제 이미지 두 그룹 간의 다변량 정규분포의 거리를 계산 - 거리가 가까울수록 좋은 영상으로 판단
VTT (Visual Turing Test)	- 컴퓨터상 이미지가 실제 사물에 필적하는지를 판별하는 과정 - 실제 이미지와 합성 이미지의 무작위 혼합 샘플을 해당 분야 전문가가 직접 구분하는 시험 - 정확도 50%가 이상적이며, 너무 높으면 합성 이미지 불량, 너무 낮으면 실제 이미지 불량
유사도	- 합성데이터 기반 AI 모델과 원본데이터 기반 AI 모델의 정확도 또는 AUC 차이 비교 - 합성데이터 및 원본데이터로 각각 훈련한 후(Training Set), 원본데이터의(Testing Set)로 평가 - 각 성능 간 차이가 낮을수록 합성데이터의 품질이 높은 것으로 판단

Ⅲ. 결론

본 논문에서는 생성된 합성데이터의 유용성을 측정하기 위해 정형 합성데이터의 유용성 측정지표와 비정형 합성데이터의 유용성 측정지표를 알아보았으며, 실제 인공지능 학습용 데이터 구축사업에서 활용되고 있는 TTA의 유용성 측정지표를 알아보았다. 향후 연구로는 TTA에서 수행한 유용성 측정지표 관련 실제 케이스 연구 및 측정 결과를 바탕으로 유용성을 향상시키기 위한 방안에 대해 연구할 예정이다.

ACKNOWLEDGMENT

이 논문은 2021년 과학기술정보통신부의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임(No.2021-0-00634 '대용량 정형 데이터 대상 개인정보 가명.익명처리 자동화 및 안정성 검증 기술개발', No.2021-0-00393 '영상 등 멀티미디어 데이터의 온전한 AI 학습활용을 보장하는 복원 불가형 개인식별정보 익명처리 핵심 기술 개발')

참 고 문 헌

[1] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengjo, "Generative adversarial networks", in Communications of the ACM, 2020.

[2] 개인정보보호위원회, "가명정보 처리 가이드라인," 2022.

[3] V. W. Berger, and Y. Zhou, "Kolmogorov-Smirnov Test: Overview", in Encyclopedia of Statistics in Behavioral Science, 2014.

[4] M. L. McHuge, "The Chi-square test of independence", in Biochemia Medica, 2013.

[5] J. Korhonen, and J. You, "Peak signal-to-noise ratio revisited: Is simple beautiful?", n Proceedings of Fourth International Workshop on Quality of Multimedia Experience, 2012.

[6] D. Brunet, E. R. Vrscay, and Z. Wang, "On the mathematical properties of the structural similarity index", in IEEE Trasactions on Image Processing, 2012.

[7] Z. Wang, and A. C. Bovik, "Mean squared error: Love it or leave it? A new look at signal fidelity measures", in IEEE Signal Processing Magazine, 2009.

[8] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution", in Computer Vision-ECCV, 2016.

[9] 인공지능 학습용 데이터 품질관리 가이드라인 및 구축 안내서 v3.0, 과학기술정보통신부, 한국지능정보사회진흥원, 한국정보통신기술협회, 2023.

[10] Y. Yu, W. Zhang, and Yun Deng, "Frechet inception distance for evaluating GANs", in arXiv:1603.08155, 2021.

[11] D. Geman, S. Geman, N. Hallonquist, and Laurent Younes, "Visual turing test for computer vision systems", in Proceedings of the National Academy of Sciences, 2014