인공지능 성능 향상을 위한 합성데이터 생성 기술 연구 (Synthetic AI in Long-tailed Environments, SAIL)

백승호[1,2], 이상철[1], 조한얼[1], 장효석[1,2], 홍윤기[1,2], 차용준[1,2], 김찬수[1,2,*]

1 한국과학기술연구원 인공지능·정보·추론 연구실,

2 과학기술연합대학원대학교 AI-로봇

*Correspondence should be addressed to eau@ust.ac.kr

On Synthetic AI in Long-tailed Environments for Improving Machine Learning and AI

Seungho Baek[1,2], Sangcheol Lee[1], Haneol Cho[3], Hyo-Seok Jang[1,2], Yoongi Hong[1,2],

Chansoo Kim[1,2,*]

KIST (Korea Institute of Science and Tech.) and UST (Univ. of Sci. and Tech.)

요 약

본 연구는 예외 상황에서 인공지능 예측 모델의 성능 향상을 위한 합성데이터 생성 기술을 기술한다. 합성데이터 생성 기술은 (SAIL, Synthetic AI in Long-tailed environments) 실제 데이터의 특성을 기초로 하여 인공지능 모델 학습에 필요한 데이터를 생성한다. 이를 통해 데이터 부족 문제를 해결하고 모델의 예측력을 강화한다. 이 기술은 GAN, VAE, Transformer 모델들을 결합하고 예외 상황에서 발생하는 다양한 문제를 묘사하기 위해 통계적 측정과 샘플링 기법을 활용한다. 합성데이터 생성 기술 다양한 산업 분야에서 적용되는 AI 모델의 신뢰성과 예측력을 크게 향상시킬 수 있을 것이다.

Ⅰ. 서 론

인공지능(AI, Artificial Intelligence) 모델이 다양한 산업 분야에서 널리 사용됨에 따라, 모자란 데이터를(이를테면 불균형, 불평등 등) 처리하는 능력이 AI 모델의 신뢰성과 성능을 결정하는 중요한 요소로 부상하고 있다. 본 연구는 해당 문제를 해결하기 위한 합성데이터를 생성하는 기술 (SAIL, Synthetic AI in Long-tailed environments)을 다룬다. 이를 통해 데이터 불균형 문제를 해결하고, AI 모델의 신뢰성과 예측력을 높일 수 있는 방법을 제시하고자 한다.

데이터 합성을 위하여 현 상황을 모사하는 다양한 데이터들은 클린징, 시멘틱, 온톨로지, 링크드리스트 등의 기법을 사용하여 멀티모달 형태로 통합된다. 수집된 다양한 데이터들은 메타데이터 검색 기술을 통해 데이터를 효율적으로 관리된다. 특히 시계열 데이터의 경우, 시계열 샘플링 및 조건부 확률 샘플링을 활용하여 현실과 유사한 분포를 유지하도록 구성할 필요가 있다.

Ⅱ. 본론

합성데이터 생성 기술은 AI 모델의 성능과 신뢰성 향상에 중요한 역할을 한다. 본 연구에서 제안하는 SAIL 기술은 다양한 도메인 지식과 빅데이터를 바탕으로 합성데이터를 생성하고 활용하는 방법론이다. 기존의 GAN (Generative Adversarial Network), VAE (Variational Auto-Encoder), 트랜스포머 (Transformer) 모델을 확산 모형과 LoRA (Low-Rank Adaptation) 기술을 통해 효율적으로 예외 데이터를 생성한다. 정보이론과 추론 기반 접근법을 통해 데이터의 구조적 특성을 정확히 모델링한다. DSLM (Domain Specific Language Model) 합성 모듈은 데이터 파이프라인을 자동화하고 최적화한다.

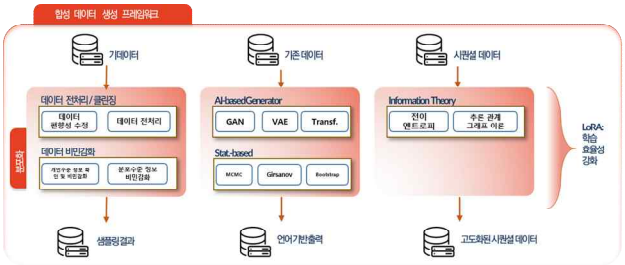
GAN은 랜덤한 잠재 코드 (Latent code)를 입력으로 넣어 새로운 출력 데이터를 얻는 학습 방법이다. 이를 통해 생성기와 판별기 간의 경쟁적 학습을 통해 실제 데이터의 분포를 모방하는 합성데이터를 생성할 수 있다. VAE는 인코더 (encoder)와 디코더 (decoder)를 통해 잠재 공간을 구성하고, 잠재 공간으로부터 우리가 원하는 결과를 decoding 함으로써 합성데이터를 생성한다. 트랜스포머 모델은 디코더가 전부 어텐션 (attention)으로 구현된 모델로서, 입력 데이터를 임베딩하여 다차원 벡터로 변환하여 데이터를 학습한다. 트랜스포머는 학습된 패턴을 기반으로 기존 데이터의 분포와 특성을 반영한 합성데이터를 생성한다. 자연어처리부터 이미지 처리까지 다양한 분야에서 우수한 성능을 보인다. 특히 Stable Diffusion과 같은 기술로 텍스트 기반 이미지 생성에 활용된다.

통계적 측정과 샘플링 기술 또한 합성 데이터 생성에 중요한 역할을 한다. 측도 변환 생성 모듈인 기르사노프 (Girsanov) 변환, 중요도 샘플링 등은 정보이론에 기반한다. 즉, 유의미한 집합에 대해 일종의 크기를 부여하는 수학적 함수인 측도를 사용한다. 한 측도를 다른 측도로 변환하는 기법을 통해 합성데이터를 생성할 수 있다. 부트스트랩 리샘플링 (bootstrap resampling)은 가장 간단하고 효과적인 리샘플링 기법 중 하나이다. 기존 데이터에서 중복을 허용하는 무작위 데이터를 추출한 후, 새로운 표본을 만드는 과정을 거듭 반복하여 새로운 부트스트랩 샘플을 생성하는 방법이다. 마르코프 체인 몬테 카를로(MCMC) 기반 생성 모델은 사용자가 원하는 분포에 맞도록 합성데이터를 생성하는 통계적 데이터 생성 방법이다. 대표적으로 기존 복잡한 확률분포로부터 새로운 샘플을 추출할 수 있도록 제안 분포 (proposal distribution)를 사용하는 메트로폴리스-헤이스팅스 (Metropolis-Hastings) 알고리즘과, 과도감쇄 랑주뱅 역학 (overdamped Langevin dynamics) 을 사용한 랑주뱅 동역학 기반 알고리즘을 기반 등이 대표적이다..

정보이론과 추론 기반 접근법은 데이터 구성 요소 간의 관계를 탐구하여 합성데이터를 생성한다. 전이 엔트로피를 통해 시계열이나 문장과 같은 시퀀셜 데이터의 (Sequential data) 추론 동역학을 탐색한다.

LoRA는 데이터 형태 및 도메인 지식에 기반하여 경량 선형화 신경망 구조를 동적으로 선택하는 기술이다. LoRA를 통해 추가 학습의 효율성을 높이며, 불균형 데이터 보정 기술과 로짓 증강 기법을 통해 예외 데이터의 학습 효율성을 향상시킨다.

DSLM (Domain-specific Language Model)은 특정 도메인이나 주제에 특화된 언어 모델로서, 이를 통해 모델의 명세와 학습 과정을 상세히 정의하고, 효율적인 데이터 파이프라인을 구축할 수 있다. 이 모듈은 데이터 파이프라인의 자동화를 실현하고, 최소한의 인간 개입으로 학습 루틴을 반복할 수 있는 유연한 시스템을 제공한다.



Ⅲ. 결론

본 연구는 인공지능(AI) 모델이 다양한 예외적 상황에서도 높은 성능을 발휘할 수 있도록 합성데이터 생성 기술, SAIL을 약술한다. SAIL 기술은 시뮬레이션을 통해 연쇄적으로 발생하는 상황에 대한 예측을 가능케 할 뿐만 아니라, 예외 상황 데이터를 생성하여 희귀 데이터의 빈도를 증대시킨다.

새롭게 생성된 데이터에 대해 단순 분석에서 나아가 설명성을 확보할 수 있고, 도메인 전문가의 지식을 추가하여 AI 모델의 성능 향상을 기대할 수 있다. 또한 합성데이터 생성을 통해 데이터 수집의 사회적 비용을 크게 절감할 수 있고, 개인정보에 대한 민감성을 줄여 데이터의 자유도와 사회적 합의성을 획득할 수 있다.

ACKNOWLEDGMENT

This research was funded by the grant Nos. 2021-0-02076, 2024-00460980 and 2023-00262155 (IITP) funded by the Korea government (the Ministry of Science and ICT).

참 고 문 헌

[1] Tian, Yonglong, et al. "Stablerep: Synthetic images from text-to-image models make strong visual representation learners." Advances in Neural Information Processing Systems 36 (2024).

[2] Lu, Yingzhou, et al. "Machine learning for synthetic data generation: a review." arXiv preprint arXiv:2302.04062 (2023).

[3] Qi, Di, and Andrew J. Majda. "Using machine learning to predict extreme events in complex systems." Proceedings of the National Academy of Sciences 117.1 (2020): 52-59.

[4] Jiang, Junjie, et al. "Predicting extreme events from data using deep machine learning: When and where." Physical Review Research 4.2 (2022): 023028.

[5] Albeverio, Sergio, Volker Jentsch, and Holger Kantz, eds. Extreme events in nature and society. Springer Science & Business Media, 2006.

[6] Kim, Chansoo, et al. "Hub-Periphery Hierarchy in Bus Transportation Networks: Gini Coefficients and the Seoul Bus System." Sustainability 12.18 (2020): 7297.

[7] S. Lee, Limit theorems for random walk local time, bootstrap percolation and permutation statistics, defended in December 2019.