데이터 분포에 따른 GAN 기반 데이터 증강 평가 및 다양한 수자원 공학 문제에 대한 적용성 연구

정윤상*, 유진경*, 이승연*, 유도근**
*수원대학교 토목공학과
**수원대학교 건설환경에너지공학부
e-mail:dydrkagksror@naver.com

Evaluation of GAN-Based Data Augmentation Depending on Data Distributions and Its Applicability Study to Varioue Water Resources Engineering Problems

Yun-Sang Jeong*, Jin-Gyeong Yoo*, Seung-Yeon Lee*, Do-Guen Yoo**
*Department of Civil Engineering, The University of Suwon
**Department of Civil and Environment Engineering, The University of Suwon

요 약

GAN(Generative Adversarial Network)은 생성자(Generator)와 판별자(Discriminator)가 경쟁적으로 학습하며 실제와 유사한 데이터를 생성하는 딥러닝 기반 모델로, 이미지 생성, 음성 합성, 데이터 증강 등 다양한 분야에서 활발히 활용되고 있다. 특히 수자원 및 수리·수문 분야에서는 관측 기반 자료의 부족, 데이터 불균형 등의 문제로 인해 기계학습 모델의 성능 저하가 빈번히 발생하며, 이에 대한 대응책으로 GAN 기반 데이터 증강이 주목받고 있다. 그러나 GAN의 학습 성능은 단지 모델 구조에 의해서만 결정되는 것이 아니라, 학습에 사용되는 원본 데이터의 확률 분포 특성에도 크게 영향을 받는다. 따라서 이러한 분포 특성이 GAN 학습에 미치는 영향을 정량적이고 체계적으로 분석할 필요가 있다. 본 연구에서는 정규분포, 균등분포, 멀티 모달 분포, 이산등급분포, 편향 분포, 이상치 포함 분포, 시계열 분포 등 다양한 데이터 분포 형태를 정의하고, 각기 다른 분포 형태의 데이터를 입력으로 하여 GAN과 함께 대표적인 데이터 증강 기법인 VAE(Variational Autoencoder)를 각각 학습시켰다. 이후, 생성된 데이터에 대해 분포 유사성, 품질, 학습 안정성 등을 평가하였으며, 이를 위해 손실 함수, KL Divergence, 데이터 분포 유사도 등의 지표를 활용하여 정량적 비교 분석을 수행하였다. 그 결과, 데이터 분포 유형에 따라 증강 기법별 성능이 상이하게 나타났으며, 각 기법의 장단점과 활용 조건을 파악할 수 있었다. 최종적으로 다양한 수자원 분야 적용 가능 문제의 데이터 형태를 살펴보고, 문제 별 데이터 증강에 따른 적용 가능성을 기초적으로 평가하였다. 본 연구는 데이터 분석의 정확성과 활용 범위를 높이기 위해, 데이터를 증강하기 전에 분포 특성을 먼저 이해하고 그에 맞는 증강 방법을 선택하는 것이 필요하다는 점을 강조한다. 이를 통해 향후 수자원 및 물 환경 분야에서 효과적인 데이터 활용과 기계학습 기반 분석을 위한 실질적인 가이드를 제시할 수 있을 것으로 기대된다.

감사의 글

본 결과물은 환경부의 재원으로 한국환경산업기술원의 도시홍수시설의 계획, 운영, 유지관리 최적화 기술개발사업의 지원을 받아 연구되었습니다(RS-2024-00398012).