

생성적 적대 신경망(GAN)을 이용한 시계열 데이터 증강 기법 적용 및 성능 평가에 관한 연구

이창현, 신동훈, 전호석, 정진우, 권기원*, 임태호

호서대학교, 한국전자기술연구원*

lygyg6322@naver.com, kwonkw@keti.re.kr*

A Study on Application of Time Series Data Enhancement Technique and Performance Evaluation using Generative Adversarial Neural Network (GAN)

ChangHyun Lee, DongHun Shin, HoSeok Jeon, JinWoo Jung, KiWon Kwon*, TaeHo Im

Hoseo Univ, Korea Electronics Technology Institute*

요 약

본 논문은 시계열 데이터의 수집이 원활하게 이루어지지 못하는 것을 증강기법을 통해 해결하고자 제안된 방법이다. 특히 시계열 데이터 증강은 적은 양의 데이터로 높은 정확성을 나오게 하는 것이 중요하다. 데이터 증강을 인공지능을 사용해서 진행하기 위해 생성적 적대 신경망(GAN)을 이용하여 시계열 데이터 증강을 진행하였고, Test 데이터 세트로 확인을 했을 때 평균적으로 높은 정확성을 보여주었기에 시계열 데이터가 부족한 상황에서 생성적 적대 신경망(GAN)을 통해 데이터 증강을 진행했을 때 충분히 사용 가능하다는 것을 증명하였다.

I. 서 론

최근 시계열 데이터는 초연결 사회, 4차 산업혁명 등 기존 사회에 획기적인 변화를 가져오는 기술의 진보를 위한 기반으로 수요가 증가하고 있지만, 시계열 데이터의 수요에 비해 데이터의 수집은 원활하게 이루어지지 못하고 있다. 이러한 문제를 해결하기 위해 다양한 데이터 증강 기법들이 나오고 있으며, 특히 기존의 데이터 증강 방식은 적은 시계열 데이터에서 적용하려면 복잡한 계산과정을 거쳐야 한다는 단점이 존재하므로 인공지능을 이용해서 데이터 증강을 하려는 시도가 늘어나고 있고, 적은 시계열 데이터로도 데이터 증강이 가능하기에 인공지능 데이터 증강 기법에 대한 연구가 자연스럽게 주목받고 있다. 인공지능의 경우 복잡한 계산식을 주입할 필요 없이 원본 시계열 데이터의 특징을 찾아 원본 시계열 데이터와 유사한 데이터를 생성하는 것이 가능하므로 적은 시계열 데이터로 데이터 증강을 할 수 있는 학습 기반 방법을 응용한 인공지능 모델을 사용하였을 때 실제로 사용한 원본 데이터를 기반으로 증강이 되어 나오는 데이터를 사용할 수 있는지 확인을 해볼 필요가 있다. 본 논문에서는 생성적 적대 신경망(GAN)에서 나오는 증강 데이터의 평가에 대해 다루었다.

II. 선행 연구

기존의 데이터 증강 기법은 사람이 직접 값을 조작해야 하는 기본 데이터 확장 방법과 고급 데이터 확장 방법으로 분류된다(그림 1). 시간 영역과 주파수 영역에서 노이즈 패턴을 주입하는 것과 원래의 입력 시계열에서 값을 직접 조작하는 방식인 기본 데이터 확장 방법이 있고, 분해 기반 방법, 통계 생성 모델, 학습 기반 방법으로 분류되는 고급 데이터 확장 방법이 있다. 분해 기반 방법의 경우 시계열 데이터의 특정 부분을 자동 회귀 모델과 같은 확률 계산으로 증강하는 방식이다. 통계 생성 모델의 경우 기존의 통계 모델을 사용하여 시계열 데이터를 모델링 하는 방법이다. 학습 기반의 경우 실제 데이터의 특성을 모방하여 진짜인 데이터처럼 비슷하게 만드는 방법이다 [1].

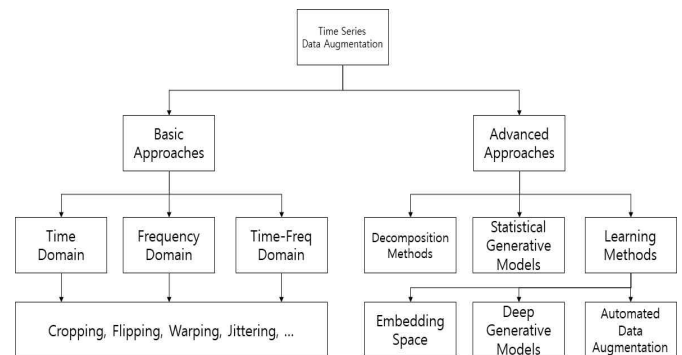


그림 1. 시계열 데이터 증강 방법 분류

III. GAN 모델 설계

전체적인 GAN 모델은 그림 2와 같다. Keras의 Sequential model을 이용하여 만들어진 모델인 1D Generative Adversarial Network라는 모델을 응용하여 설계된 모델이다 [2]. GAN은 원본 시계열 데이터를 입력받고, 원본 시계열 데이터를 기반으로 생성자가 임의의 데이터를 생성하고, 판별자가 원본 시계열 데이터와 생성자가 만들어 낸 임의의 데이터를 비교하고 정확도를 산출합니다. 그 후 산출된 정확도를 생성자가 다음 데이터를 만들 때 참고하게 된다.

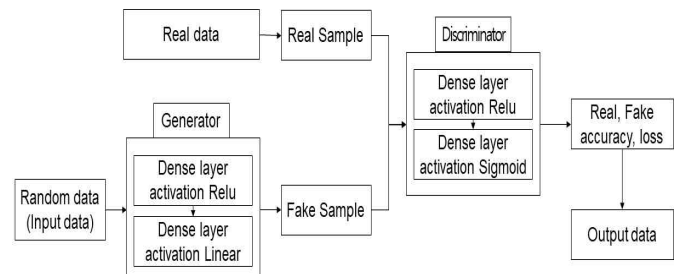


그림 2. 생성적 적대 신경망(GAN) 모델 구조

IV. 실험 과정

실험은 UCR 데이터에서 로스앤젤레스의 101 North 고속도로 진입로에 설치된 루프 센서로 수집되는 교통 데이터인 DogerLoopGame 라는 시계열 데이터를 이용하여 진행되었으며, DogerLoopGame의 경우 Class는 2개로 이루어져 있고 Class_1은 게임이 없을 때이며 Class_2는 게임이 있을 때를 의미한다. 실험에 사용한 원본 데이터 세트는 Class당 10개씩 총 20개이며, Test 데이터 세트는 Class당 69개씩 총 138개에서 임의로 20개를 선정해서 진행했다. 생성적 적대 신경망(GAN)으로 데이터를 원본 데이터와의 유사성이 20%~100%, 40%~100%, 60%~100%, 80%~100%인 것으로 각각 100개씩 Class마다 증강을 진행하였고, 총 800개의 증강 데이터를 얻었다. 시계열 데이터를 m 차원의 공간 궤적에 그리고, 각 공간 궤적에 위치한 점 간의 거리를 이용하여, 이미지화를 시키는 Plot인 Recurrence Plots 알고리즘으로 증강 데이터를 이미지화를 진행하였다 [3]. 이미지화를 진행한 이유는 동일한 데이터 세트에 대해 Recurrence Plots라는 알고리즘으로 이미지화를 한 후 CNN 모델로 평가를 진행했을 때, 같은 데이터 세트에 대해 기존 혹은 최신 시계열 데이터 분류 알고리즘과 비교 시 대부분의 데이터 세트에서 CNN 모델을 사용했을 때 오차율이 많이 개선되었다는 결과를 적용하기 위해 이미지화를 진행했다 [3]. 그림 3과 같이 생성적 적대 신경망(GAN)으로 증강된 데이터와 원본 데이터와 유사한 것을 알 수 있다.

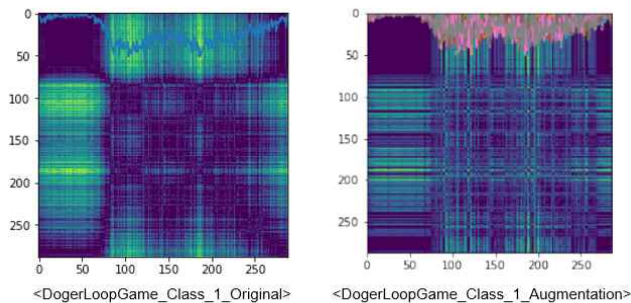


그림 3. 원본 데이터와 증강 데이터 비교

이미지화된 증강 데이터 세트를 그림 4와 같이 각각의 Class 별로 YOLOV4 라는 CNN 모델을 이용하여 이미지학습을 진행 후 UCR에서 제공하는 각각의 Class마다 Test 데이터 세트에서 10개의 데이터 세트를 Recurrence Plots 알고리즘으로 이미지화했다. 총 20개의 변환된 이미지 데이터를 이용해서 YOLOV4 모델로 정확성을 평가하는 실험을 진행했다. YOLOV4 모델을 사용한 이유는 Object Detection에서 영상이나 이미지에서 뛰어난 인식률을 보이는 모델이므로 본 실험에서 증강 데이터의 이미지들을 인식하는 용도로 사용했다.

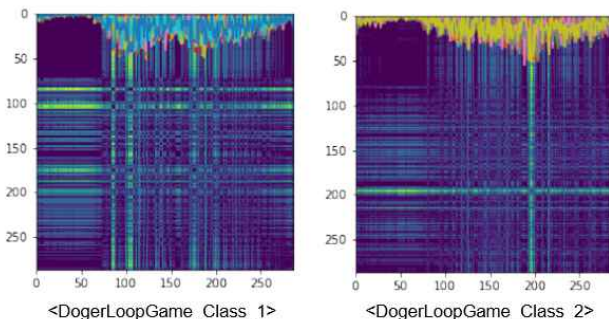


그림 4. DogerLoopGame Class 훈련 데이터 이미지

V. 실험 결과

본 논문에서 실험은 PC에서 YOLOV4 모델로 훈련되어 나온 Weights 파일을 이용해서 훈련된 증강 데이터가 Test 데이터 세트로 평가를 진행했을 때 정확성이 얼마나 나오는지 20%~100%, 40%~100%, 60%~100%, 80%~100% 원본 데이터와의 유사성으로 분류된 Class 별 데이터의 정확성 평균을 측정했다. 표 1과 같이 결과값이 나왔으며, Class_1의 경우 최소 77%, 최대 89.8%라는 높은 정확성이 나왔으며, Class_2도 최소 78.9%, 최대 80% 정도의 높은 정확성을 얻을 수 있었다.

표 1. DogerLoopGame 실험 결과(Class 별 평균 정확도)

	20%~100%	40%~100%	60%~100%	80%~100%
class_1	77%	78.3%	87.9%	89.8%
class_2	78.9%	79.9%	80%	80%

VI. 결론

본 논문은 시계열 데이터의 수요에 비해 데이터의 수집은 원활하게 이루어지지 못하고 있다는 점에서 데이터의 수집을 원활하게 하고자 데이터 증강을 통해 해결하는 방법을 제안했다. GAN을 통해 증강된 데이터를 CNN 모델로 정확성 평가를 그림 5와 같이 확인을 했을 때 평균 77% 이상의 원본 데이터와의 정확성을 보여줬으며, 원본 데이터와 유사성이 80%~100%인 데이터만 했을 때는 최대 90% 정도 정확성을 보여줬기 때문에 시계열 데이터가 부족한 경우 충분히 GAN을 이용한 데이터 증강을 충분히 사용할 수 있을 것으로 판단이 된다.

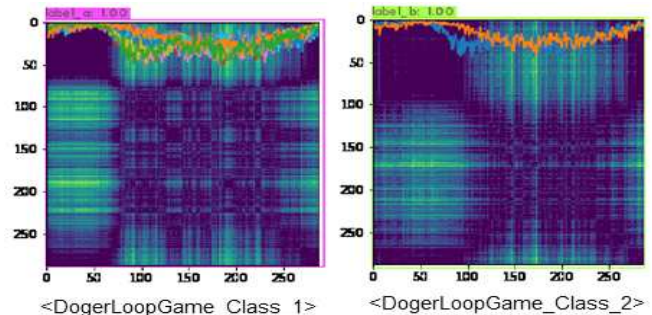


그림 5. Test 데이터 정확성 확인

ACKNOWLEDGMENT

“이 논문은 2021년도 해양수산부의 재원으로 해양수산과학기술진흥원의 지원을 받아 수행된 연구임”(No.20200611, 컨테이너 위험화물 자동검색 및 복합탐지 시스템 개발)

“본 연구는 과학기술정보통신부 및 정보통신기획평가원의 대학ICT연구센터지원사업의 연구결과로 수행되었음”(IITP-2021-2018-0-01417)

참 고 문 헌

- [1] Qingsong Wen, Liang Sun, Fan Yang, Xiaomin Song, Jingkun Gao, Xue Wang, Huan Xu "Time Series Data Augmentation for Deep Learning: A Survey", arXiv preprint arXiv:2002.12478, 2020.
- [2] L. Cui, P. Zhao, K. Wang, J. Yang, X. Bu, "A Kind of Arbitrary Signal Generator Based on 1D Generative Adversarial Network", IEEE 8th Data Driven Control and Learning Systems Conference, 2019.
- [3] Hatami N, Gavet Y, Debayle J. "Classification of time-series images using deep convolutional neural networks" In: International conference on machine vision, 2017.