적대적 생성 신경망을 이용한 시계열 데이터 생성기법 조사

서은성, 석준희*
고려대학교, *고려대학교

seoun4612@korea.ac.kr, *jseok14@korea.ac.kr

Survey on Time series data Generation Using Generative Adversarial Networks

Seo Eun Seong, Seok Jun Hee*
Korea Univ., *Korea Univ.

요 약

최근 딥 러닝의 발달로 인하여 이미지 분류나 자연어 처리 등 많은 분야에서 딥 러닝을 적용하는 것이 좋은 성과를 보였다. 이러한 딥러닝 모델은 다량의 데이터를 이용해 학습하는 방법을 통하여 성능을 높여왔다. 하지만 딥러닝 모델의 학습을 위한 충분한 데이터를 획득하는 것이 제한되는 분야가 다수 존재한다. 본 논문에서는 이러한 데이터의 부족 문제를 해결하기 위하여 적대적 생성 신경망을 이용한 시계열 데이터 생성 기법에 대해 소개하고 특징을 살펴보았다.

Ⅰ. 서 론

딥러닝(Deep Learning)은 여러 비선형 변환기법의 조합을 통해 높은 수준의 추상화를 시도하는 기계학습 알고리즘의 집합으로, 현재 학계에서 여러 분야에서 훌륭한 성과를 보인다. 특히 이미지 처리나 자연어 처리 분야에서 빅 데이터의 활용은 기존의 모델에 비해 성능을 크게 높이는 데 기여하였다. 또한, 시계열 데이터의 분류, 이상 탐지, 예측의 영역서도 큰 활약을 보여주었다. [1] 하지만 모든 분야에서 인공지능 모델의 학습을 위한 빅데이터를 획득하는 것은 불가능에 가깝다. 따라서 데이터의 양의 부족한 분야에서는 학습을 위한 빅데이터의 형성을 위해 다양한 데이터 증강기법을 도입하였다. 이러한 증강된 데이터를 학습에 사용하는 것으로 시계열 데이터에서도 성능의 향상을 꾀할 수 있음이 이전 연구에서 확인되었다.[3] 다양한 데이터 증강 기법 중, 2014년에 공개된 적대적 생성 신경망(Generative Adversarial Networks, GAN) 기법은 원본 데이터의 분포와 유사한 데이터를 생성할 수 있다는 점에서 많은 관심을 받았다. 본 논문에서는 GAN을 이용하여 시계열 데이터를 생성하는 대표적인 2가지 기법들에 대해서 소개하고 어떠한 특징이 있는지 비교해 보았다.

Ⅱ. 관련 연구

적대적 생성신경망은 2014년 제안된 딥러닝 모델로, 판별기(Discriminator, D)와 생성기(Generator, G)라는 두 개의 신경망이 적대적 관계 학습을 통해 실제로 학습 데이터와 분포가 유사한 데이터를 생성하는 기법이다. G는 노이즈로부터 실제 데이터의 분포를 모사해서 D가 구별하지 못하는 방향으로 학습하고, D는 G가 생성한 데이터와 실제 데이터를 구분하는 방향으로 학습한다. 이와 같은 GAN의 목적 함수는 식(1)과 같이 표현된다.

$$\min_G \max_D V(D,G)$$
$$= E_{x\sim P_{data}(x)}[\log D(x)] + E_{x\sim P_z(z)}[\log(1-D(G(z)))]$$

식1. GAN의 손실함수

위 식(1)에서 P_{data}는 원본데이터를 의미하고 P_z는 D가 P_{data}의 확률분포와 유사하게 생성한 데이터이다. 이러한 GAN 모델의 학습 구조도는 그림 1과 같이 표현할 수 있다.

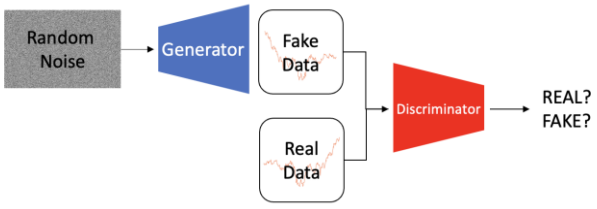

그림1. GAN의 학습구조도

Ⅲ. RNN 구조를 가진 GAN: RGAN, RCGAN

[Esteban et al., 2017] 에서 저자는 실제와 같은 다변량 시계열 데이터를 생성할 수 있는 Recurrent GAN(RGAN)과 Recurrent Conditional GAN (RCGAN)를 제안했다. [4]

RGAN은 생성기와 판별기에 RNN구조인 LSTM(Long Term-Short Memory)를 채택하였다. RCGAN은 RGAN의 발전된 형태로 일부 조건부 입력 C_n이 추가된 형태의 모델로 조건부 입력을 통하여 생성되는 데이터를 조절할 수 있는 GAN 모델이다. 두 모델의 판별기는 매시간당 예측과 실제 시계열 간의 Average Negative Cross-Entropy를 최소화하도록 훈련하고 생성기는 합성

시계열 데이터가 실제 시계열 데이터와 유사한 분포를 갖도록 하는 방향으로 학습된다. 이러한 RCGAN 모델의 학습 구조도는 그림 2와 같이 표현할 수 있다.

해당 논문은 시계열 데이터 생성 연구 분야에서의 TSTR(Train on Synthetic, Test on Real)이라는 새로운 평가지표를 제시했다. GAN을 통해 생성된 데이터가 얼마나 진짜와 같은지 검증하는 것은 여전히 어려운 문제로 남아있는데, 특히, GAN을 통해 생성된 시계열 데이터의 유사성은 육안으로 쉽게 확인하기 어려워 더욱 해결하기 힘든 문제이다. 본 논문의 저자가 제시한 새로운 평가지표인 TSTR은 합성된 데이터를 통해 학습하고 실제 데이터로 테스트하여 원본 데이터의 특성을 얼마나 잘 보존하고 있는지 확인하는 방법으로 생성된 데이터의 유효성을 평가하는 지표로 이상적이다. 이러한 지표를 통하여 우리는 GAN을 통해 생성된 합성데이터가 얼마나 훌륭하게 실제 학습 데이터를 대체할 수 있는지를 확인할 수 있다.

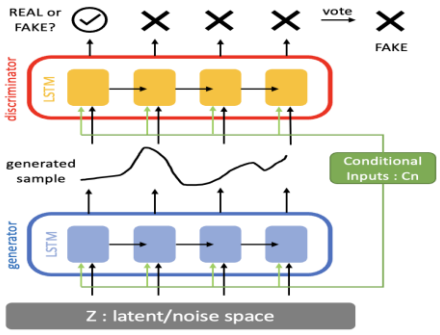

그림2. RCGAN의 학습구조도

Ⅳ. 자기회귀 모델과 GAN의 융합: TimeGAN

본 논문[yoon et al,. 2019]은 GAN의 비지도 방식과 자기회귀(Auto Regressive, AR)모델의 지도방식을 결합한 시계열 데이터 생성 모델인 TimeGAN을 제시하였다.[5] TimeGAN이란 차원이 축소된 임베딩 공간에서 GAN의 비지도 손실과 AR모델의 지도손실을 동시에 최소화하여 학습하는 GAN 모델을 지칭한다. 이러한 TimeGAN 모델의 학습 구조도는 그림 3과 같이 표현된다. TimeGAN은 기존의 시계열 데이터 생성기법에 비해 두 가지의 장점을 보였다.

첫째, 기존의 GAN 모델은 특정 도메인에서만 시계열 데이터를 생성할 수 있다는 단점이 존재하였다. 이에 비해 TimeGAN은 비지도학습과 지도학습을 결합하는 방법으로 다양한 영역에서 적용할 수 있도록 기존의 모델을 개선하였다. 기존의 시계열 생성 GAN은 시간적 역학에 따른 상관관계를 고려하지 못하고 적대적 피드백에만 의존한다는 단점을 가지고 있었다. 동시에, AR 모델과 같은 지도 학습은 시계열 데이터 고유의 시간적 역학을 세밀하게 조정할 수 있지만 훈련 데이터에 의존적인 한정된 분포라는 단점이 존재했다. 시간적 역학을 보존하면서 시계열 데이터를 생성시키는 새로운 메커니즘을 제안했다. 제시된 방법론을 통해 생성된 시계열 데이터는 기존의 방법을 통해 생성된 데이터에 비해 여러 영역에서 강점을 보였다.

둘째, 특징과 잠재 표현 간의 가역성 매핑을 제공하는 임베딩 네트워크를 도입하여 적대적 학습 공간의 차원을 축소하였다. 이러한 방법론은 복잡한 시스템의 시간적 역학이 더 낮은 차원의 변동요인에 바뀐다는 점으로부터 착안하여 제시되었다. 또한, 지도손실은 임베딩 네트워크와 생성기의 네트워크에서 동시에 훈련되면서 손실을 최소화하는 것을 목표로 하는데, 이때 저차원의 잠재 공간이 시간적 관계의 학습 효율을 증가시킨다.

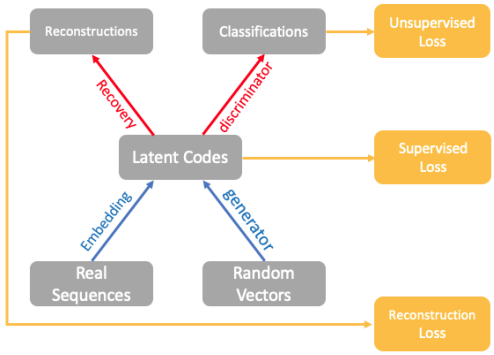

그림3. TimeGAN의 구조도

Ⅴ. 결론

본 논문에서는 GAN을 이용한 대표적인 두 개의시계열 데이터 생성 기법에 대하여 탐구하였다. RGAN은 판별기와 생성기에 RNN구조를 채택한 GAN 모델로서 매 시점에서의 실제와 비슷한 데이터 분포의 생성하는 GAN 모델이다. TimeGAN은 지도 학습과 비지도 학습을 상호보완 하여 시간적 역학을 보존하고 다양한 시계열 데이터를 생성하는데 강점을 보인 GAN 모델이다. 이러한 다양한 GAN 모델 들의 도입을 통하여 시계열 데이터의 분류, 예측, 이상탐지와 같은 다양한 영역에서 인공지능 모델들의 발전이 기대된다.

ACKNOWLEDGMENT

이 논문은 한국연구재단의 지원을 받아 수행된 연구임 (NRF-2022R1A2C2004003).

참 고 문 헌

[1] Wen, Qingsong, et al. "Time series data augmentation for deep learning: A survey." arXiv preprint arXiv:2002.12478 (2020).

[2] Goodfellow, Ian, et al. "Generative adversarial nets." Advances in neural information processing systems 27 (2014).

[3] Iwana, Brian Kenji, and Seiichi Uchida. "An empirical survey of data augmentation for time series classification with neural networks." Plos one 16.7 (2021): e0254841.

[4] Esteban, Cristóbal, Stephanie L. Hyland, and Gunnar Rätsch. "Real-valued (medical) time series generation with recurrent conditional gans." arXiv preprint arXiv:1706.02633 (2017).

[5] Yoon, Jinsung, Daniel Jarrett, and Mihaela Van der Schaar. "Time-series generative adversarial networks." Advances in Neural Information Processing Systems 32 (2019).