2022년도 대한전기학회 하계학술대회 논문집 2022. 7. 13 ~ 16

소량의 신호데이터 증량을 위해 구축된 생성 네트워크 구조의 설계: 비교연구

정병국*, 김진율**, 오성권***
수원대학교

Design on Generative Network Structures Constructed for Small Signal Data Augmentation: Comparative Studies

Byeong-Kuk Jeong*, Jin-Yul Kim**, Sung-Kwun Oh***
The University of Suwon

Abstract - 기계학습 및 딥러닝 알고리즘을 사용하여 문제를 해결하기 위해서 충분한 양의 데이터를 확보하는 것은 매우 중요한 조건 중 하나이다. 다수의 알고리즘에서 학습데이터의 수가 데이터가 갖는 입력변수의 수보다 적은 경우, 알고리즘의 성능이 저하되는 과적합 현상이 발생할 수 있다. 본 연구에서는 소량의 신호 데이터를 알고리즘에 적용하고, 과적합 현상을 방지하기 위해 다양한 생성 네트워크를 이용하여 가짜 신호 데이터를 생성한다. 생성된 가짜 신호 데이터로 학습한 알고리즘의 일반화 성능을 비교하는 것으로 각 생성 네트워크의 성능을 간접적으로 분석하고, 생성 네트워크를 통해 알고리즘의 일반화 성능을 개선할 수 있음을 입증한다.

1. 서 론

현재 다양한 분야에서 주어진 소량의 신호데이터를 통해 문제를 해결하기 위해 다양한 기계학습 및 딥러닝 알고리즘이 사용되고 있다[1]. 이러한 알고리즘은 데이터의 표본이 매우 적거나 획득하기 어려운 경우 매우 제한적으로 적용할 수밖에 없는 문제를 내포하고 있다. 이러한 문제를 위해 기존에는 잡음을 추가한 데이터를 전체 데이터 집합에 추가하거나, 데이터를 분할하여 k-겹 교차 검증을 사용하는 등의 부분적으로 해결하는 방법들을 사용하였다. 본 논문에서는 소량의 신호데이터의 학습과 알고리즘의 일반화 성능을 개선하기 위하여 생성 네트워크들을 구축하고, 실제 데이터의 정보를 부분적으로 포함하는 가짜 데이터를 생성하여 데이터 집합을 증량하고자 한다. 본 연구를 통해 다양한 생성 네트워크를 통해 실제 데이터의 정보를 부분적으로 포함하는 가짜 데이터를 생성할 수 있고, 가짜 데이터를 통해 증량된 데이터 집합이 알고리즘의 일반화 성능을 개선할 수 있음을 연구 결과를 통해 입증한다.

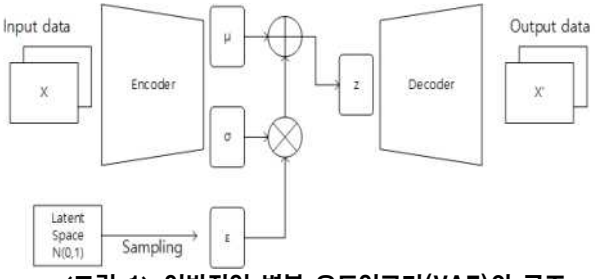
2. 본 론

2.1 생성 네트워크의 구축

생성 네트워크는 어떠한 입력 데이터 **X**를 가장 잘 표현할 수 있는 특징을 잠재 변수 z에 반영하여 입력 데이터 **X**와 동일하지 않은 유사한 데이터 **X'**를 출력하는 네트워크이다. 본 연구에서 소량의 신호데이터 증량을 위해 고려된 생성 네트워크는 변분 오토인코더(Variational Auto-Enocder, VAE)와 심층곱 적대적 생성 신경망(Deep Convolutional Generative Adversarial Networks, DCGAN)을 사용하였다[1-2]. 두 개의 생성 네트워크의 입력으로 소량의 부분방전 데이터를 사용하여 가짜 부분방전 데이터를 생성하고, 생성된 가짜 부분방전 데이터를 합성곱 신경망(Convolutional Neural Networks, CNN) 분류기의 학습데이터로 활용하여 분류기의 일반화 성능을 비교한다[3].

2.1.1 변분 오토인코더의 구축

변분 오토인코더(VAE)의 구조는 오토인코더(Auto-Encoder, AE)와 같이 내부 신경망으로 구성된 인코더와 디코더로 이루어진다[1]. 인코더는 입력 데이터 X의 확률분포를 정규분포로 가정하고 평균 μ와 표준편차 σ를 추정하여 입력 X의 확률밀도를 추정한다. 디코더의 입력인 잠재 변수 **z**는 인코더에서 추정된 **μ**와 **σ**와 잠재공간에서 샘플링한 분포 ε로 구성되며 이때, ε는 디코더에서 데이터를 재생성할 때, 입력 데이터 **X**와 완전히 동일한 데이터가 생성되지 않도록 하기 위한 분포이다. 디코더는 잠재 변수 **z**를 입력으로 원 데이터 **X**와 같은 크기의 데이터 **X'**를 생성한다.

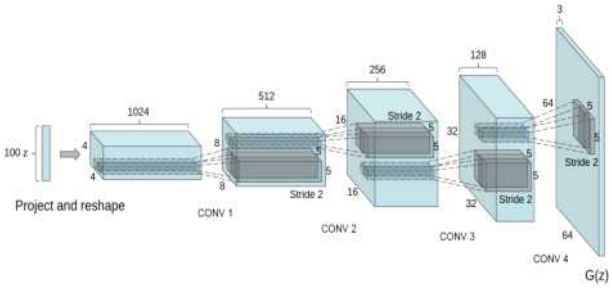


〈그림 1〉 일반적인 변분 오토인코더(VAE)의 구조

2.1.2 심층곱 적대적 생성 신경망의 구축

심층곱 적대적 생성 신경망(DCGAN)은 기존 생성적 적대 신경망(GAN)의 내부 신경망 구조를 합성곱 신경망 구조로 변환한 생성 네트워크의 일종으로, 생성 모델 G와 판별 모델 D의 상호 경쟁 학습을 통해 데이터를 생성한다[2]. 생성 모델 G는 잠재공간에서 추출된 잠재변수 z를 데이터 공간으로 변환하여 가짜 데이터를 생성한다. 판별 모델 D는 입력 데이터 X를 사전에 학습하고, 생성 모델의 출력 G(z)를 입력받아 이를 실제 데이터인지 가짜 데이터인지 이진 분류를 통해 구분한다. 그림 2는 일반적인 DCGAN 생성 모델의 구조를 나타내며, 식 (1)는 일반적인 GAN의 목적함수이다.

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



〈그림 2〉 일반적인 DCGAN 생성 모델의 구조

2.2 시뮬레이션 및 결과 고찰

본 실험에서 사용된 신호데이터는 부분방전 신호 데이터는 정상상태 200개, 자유도체 방전 150개, 절연체 방전 150개, 코로나

방전 150개, 부유 방전 150개로 전체 800개, 5가지 부류로 구성되어있다. 각 데이터는 [600×128]의 크기를 갖는 이미지 데이터로 생성 네트워크의 입력으로 사용되었다[3-4]. 두 생성 네트워크를 통해 구축된 학습데이터 집합의 수는 총 1200개로, 실데이터의 다섯 가지 부류의 비율에 맞추어 생성하였다.

<표 1> 실험에 사용된 부분방전 데이터의 정보

Data set	Classes	Number of patterns	Data Size
Partial Discharge	Normal	200	[600×128]
	Particle	150	
	Void	150	
	Corona	150	
	Floating	150	
Generated Fake Partial Discharge	Normal	300	[600×128]
	Particle	225	
	Void	225	
	Corona	225	
	Flaoting	225	

VAE를 사용하여 생성된 가짜 데이터를 추가한 데이터 집합과 DCGAN을 사용하여 생성된 가짜 데이터를 추가한 데이터 집합을 CNN 패턴 분류기를 통해 분류한 결과를 표 2로 나타내었다.

<표 2> 실데이터로 학습한 결과와 가짜 데이터로 학습한 결과

Training data set	**Classes**	**Each Accuracy**	**Total Accuracy**
Original Partial Discharge	Normal	58.75%	84.815%
	Particle	66.66%	
	Void	100%	
	Corona	98.66%	
	Floating	100%	
Generated Fake data through VAE	Normal	61%	84.872%
	Particle	70.66%	
	Void	97.33%	
	Corona	95.36%	
	Flaoting	100%	
Generated Fake data through DCGAN	Normal	63.00%	**85.666%**
	Particle	69.33%	
	Void	97.33%	
	Corona	98.66%	
	Flaoting	100%	

실 부분방전 데이터만을 이용하여 학습 및 분류한 결과 전체 분류 정확도는 84.815%로 전체 부류 중 정상상태 및 자유도체 방전의 분류 정확도는 각각 58.75%와 66.66%로 매우 낮은 정확도를 보였다. VAE를 통해 생성된 가짜 부분방전 데이터로 학습한 분류기를 실데이터로 검증한 결과 실제 데이터만 사용한 결과보다 부분적으로 2~3%의 정확도가 개선되었지만, 전체 분류 정확도 개선에 있어 0,06%의 저조한 결과를 얻었다. 반면 DCGAN을 통해 생성된 가짜 부분방전 데이터로 학습된 분류기는 VAE를 사용하였을 때 보다 전체 정확도에서 약 0.8% 정도 개선된 정확도를 확인하였으나 추가적인 개선 방안이 필요할 것으로 보인다.

3. 결 론

본 논문에서는 VAE와 DCGAN을 사용하여 가짜 부분방전 데이터를 생성하고 CNN 패턴 분류기의 학습데이터로 사용하였다. 실험 결과 실 부분방전 데이터만을 사용한 결과보다 두 가지 부류의 분류 정확도 개선이 부분적으로 이루어졌지만 다른 부류의 정확도가 오히려 감소하는 모습을 확인할 수 있었다. 이러한 원인에 있어 생성 네트워크가 실제 데이터가 갖는 분포 및 특징을 학습하는 과정에 있어 추가적인 개선 및 제약이 필요한 것으로 판단된다. 향후 생성 네트워크의 신뢰성을 높이기 위한 연구를 수행하고자 한다.

감사의 글

본 연구는 2021년도 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임(NRF-2021R1F1A1056102)

[참 고 문 헌]

[1] D. P. Kingma, M. Welling, “An Introduction to Variational Autoencoders,”, *Foundations and Trends in Machine Leraning*, Vol. 12, No. 4, pp. 307-392, Dec., 2019.

[2] Radford, Alec, Luke Metz, and Soumith Chintala. "Unsupervised representation learning with deep convolutional generative adversarial networks," *arXiv:1511.06434 [cs.LG]*, 2016.

[3] 박준용, 오성권. "부분방전 데이터 처리 방법에 따른 CNN 기반 패턴 분류기의 비교 연구," *전기학회논문지* Vol. 70, No. 3, pp. 515-525, 2021.

[4] 박준용, 김영일, 오성권., "DCGAN을 응용한 가상의 실험 데이터 구축 및 부분방전 패턴 분류기 모델 설계," *대한전기학회 학술대회 논문집*, pp.1773-1774. 2021.