Journal of the KIMST, Vol. 26, No. 5, pp. 422-430, 2023
ISSN 1598-9127(Print) · 2636-0640(Online)
DOI https://doi.org/10.9766/KIMST.2023.26.5.422

Research Paper 화생방·환경 부문

수학적 변환과 심층 생성 모델을 활용한 DMMP와 2-CEES의 모의 라만 분광 생성

박성원[1] · 정보성[1] · 김홍중*,[1]

[1] 고려대학교 수학과

Generating Synthetic Raman Spectra of DMMP and 2-CEES by Mathematical Transforms and Deep Generative Models

Sungwon Park[1] · Boseong Jeong[1] · Hongjoong Kim*,[1]

[1] *Department of Mathematics, Korea University, Korea*

(Received 10 March 2023 / Revised 15 October 2023 / Accepted 17 November 2023)

Abstract

To build an automated system detecting toxic chemicals from Raman spectra, we have to obtain sufficient data of toxic chemicals. However, it usually costs high to gather Raman spectra of toxic chemicals in diverse situations. Tackling this problem, we develop methods to generate synthetic Raman spectra of DMMP and 2-CEES without actual experiments. First, we propose certain mathematical transforms to augment few original Raman spectra. Then, we train deep generative models to generate more realistic and diverse data. Analyzing synthetic Raman spectra of toxic chemicals generated by our methods through visualization, we qualitatively verify that the data are sufficiently similar to original data and diverse. For conclusion, we obtain a synthetic dataset of DMMP and 2-CEES with the proposed algorithm.

Key Words : Raman Spectrum(라만 분광), Toxic Chemical Detection(독성 화학 탐지), Synthetic Data Generation(모의 데이터 생성), Variational Auto Encoder(변분적 오토 인코더), Generative Adversarial Networks(적대적 생성 신경망)

1. 서 론

라만 분광은 물질의 고유 특징을 나타내는 스펙트럼으로써 화학 물질 탐지에 유용하게 활용된다. 특히 라만 분광기는 접촉 없이도 물질 분석이 가능하기에 작전 상황에서 화학 오염 여부를 탐지하는 데 사용할 수 있고, 이에 따라 지표면에서 비접촉으로 화학 물질을 탐지하기 위한 라만 분광법에 대한 연구[1-5]가 활발히 진행되고 있다. 이때, 실제 화학 오염 탐지를 위해서는 측정된 라만 분광이 어떤 독성 물질의 분광과 동일한지 분별해내는 과정이 필요하고, 기계학습 모델

* Corresponding author, E-mail: hongjoong@korea.ac.kr
Copyright ⓒ The Korea Institute of Military Science and Technology

이 이 일을 할 수 있도록 학습된다면 독성 화학 탐지를 신속하고 정확하게 수행할 수 있을 것이다.

최근 딥러닝이 다방면으로 사람에 필적하고, 때때로는 넘어서는 결과들을 보여주고 있다[6]. 이러한 점을 미루어 볼 때, 라만 분광에 딥러닝 기반의 모델을 적용하면 라만 분석 과정의 자동화가 가능할 것으로 기대할 수 있다[7]. 딥러닝은 많은 데이터를 사용할수록 더 높은 정확도의 모델을 얻을 수 있지만[8] 라만 분광은 다양한 환경에서 데이터를 수집하는 것이 어려운 상황이 많고, 이것이 딥러닝을 적용할 때의 제약이 된다. 예를 들어, 라만 분광기를 활용해 지표면에서 독성 물질을 검출하는 자동화된 딥러닝 모델을 얻고자 하는 경우, 각기 다른 환경에서 독성 물질의 라만 분광을 여러 번 측정해야 하는데 이를 위해선 큰 비용이 들어가게 된다.

따라서 본 논문에서는 이러한 문제를 해결하기 위해, 각각의 물질에 해당하는 라만 분광 데이터가 충분하지 못한 상황에서 물질별로 그와 유사한 모의 신호를 생성하는 알고리즘을 제안한다. 우선 라만 분광의 특성을 유지시키는 4가지 수학적 변환을 설계하였고, 오토 인코더(autoencoder), 변분적 오토 인코더(VAE; Variational Autoencoder)[9], 적대적 생성 신경망(GAN; Generative Adversarial Networks)[10]를 학습한 후, 이를 토대로 독성 물질 Dimethyl Methyl Phosphonate(DMMP)와 2-Chloroehtyl Ethyl Sulfide(2-CEES)에 대한 모의 라만 분광 데이터셋을 구축하였다.

본 논문이 기여하는 바는 다음과 같다. 첫째, 직접적인 실험 없이 많은 양의 모의 독성 물질 데이터를 생성하였다. 이는 실험을 위한 비용 절감과 더불어 강력한 기계학습 모델 학습에 도움이 된다. 둘째, 디노이징에 사용되는 기법을 적절히 활용하여 라만 분광 데이터에서 피크 위치를 유지한 채로 다양한 변형을 가할 수학적 변환들이 설계되었다. 단순히 신호 데이터에 노이즈를 더하는 것이 아닌 라만 분광의 성질을 보존하는 것에 특화된 변환을 제안하였다. 셋째, 생성한 데이터로 심층 생성 모델을 학습시켜 더욱 다양한 데이터셋을 구축하였다. 2장에서는 관련 연구를 간략히 요약하고, 3장에서 모의 데이터 생성 방법을 설명한다. 4장에서는 생성된 데이터를 시각화해 분석하고, 이를 기반으로 모의 데이터셋을 개선한다.

2. 기술현황 분석

본 연구에서는 라만 분광의 모의 데이터 생성을 다루고자 하며 이번 장에서는 다양한 측면의 선행 연구들을 요약한다.

2.1 라만 분광과 딥러닝

라만 분광이 어떤 물질을 나타내는지 분류하거나, 주어진 라만 분광에서 노이즈를 제거하는 등 다양한 의도의 알고리즘, 딥러닝 모델이 연구되고 있다. 노이즈를 제거하기 위해 웨이블릿 변환을 활용하거나 디노이징 오토인코더, GAN을 학습시키는 연구가 진행중이다[11-13]. 또한 라만 분광 분류를 위한 CNN모델을 학습시키거나 GAN의 판별자를 활용하는 접근도 이루어지고 있다[14,15]. 라만 분광에 딥러닝 모델을 학습시키는 시도들이 성공적인 결과를 보여주는 가운데, 적은 개수의 라만 분광만을 가지고 충분한 양의 모의 데이터를 생성할 수 있는 알고리즘은 독성 화학 탐지 등 다양한 문제에 활용될 수 있을 것이다.

2.2 신호 데이터 증강 알고리즘

데이터 증강은 강력한 딥러닝 모델을 학습시키기 위한 중요한 요소로써 증강과 관련한 많은 연구가 이루어지고 있다. 시계열 데이터를 시간 도메인 또는 주파수 도메인에서 노이즈를 더하는 등의 방식으로 증강하거나, 심층 생성 모델을 활용하는 방법들이 연구되었다[16]. 특히 GAN과 같은 딥러닝 기반의 방법을 활용해 모의 데이터를 생성하는 시도가 많았다[17,18]. 특정 도메인의 데이터에 대해서는 데이터의 성질을 유지하는 증강 방법들이 성공적인 결과를 얻어냈고, 웨어러블 센서 데이터, 뇌파 데이터 등에 관한 연구가 이루어졌다[19,20].

특정한 센서 데이터들에 대한 많은 증강 기법이 연구된 데 반하여 라만 분광 데이터에 관해서는 데이터의 성질을 유지하는 변환에 관한 연구가 이뤄지지 않았다. 기존의 신호 증강 기법들은 라만 분광에 적합하지 않기에, 본 연구에서는 라만 분광의 특색을 반영한 데이터 증강 기법을 제안한다.

3. 연구 방법

이번 장에서는 라만 분광의 증강 알고리즘을 제안

한다. 3.1절에서는 독성 물질 데이터에 관해 설명하고, 3.2절에서는 수학적 변환으로 증강된 데이터를 생성하는 방법을, 3.3절에서는 해당 변환들로 생성된 데이터를 이용한 기계학습 모델의 학습 방법을 소개한다.

3.1 라만 분광 데이터 획득

모의 데이터셋 구축을 위해 독성 물질 DMMP와 2-CEES를 사용한다. 레이저 파장 532 nm, 측정시간 5초, 레이저 지름 300 μm를 사용해 물질별로 하나의 분광 데이터를 측정했고, Fig. 1은 측정된 데이터를 보여준다. 웨이브 넘버 400 cm^{-1}부터 1500 cm^{-1}까지 0.5 cm^{-1} 간격인 길이 2200의 데이터를 입력 신호로 사용하였다. 본 연구에서 사용된 데이터에는 노이즈와 베이스라인이 관측되지 않아 별도의 전처리 과정이 없이 측정된 분광 데이터를 직접 연구에 활용하였다. 만일 측정된 라만 분광에 노이즈와 베이스라인이 관측될 경우 디노이징과 베이스라인 제거가 사전에 필요할 것이다.

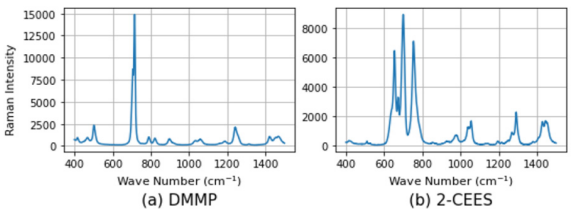

Fig. 1. Raman spectra of DMMP and 2-CEES

3.2 수학적 변환을 활용한 신호 변형

라만 분광 데이터가 균등한 간격의 인덱스를 가지며 길이가 N인 일차원 신호 $F[0],\dots,F[N-1]$와 같이 주어진 경우, 라만 분광으로부터 중요치 않은 신호를 필터링해 변환을 가한다. 이를 위해 기저 분해 방식인 이산 푸리에 변환과 이산 웨이블릿 변환을 활용하는 변환 2가지를 제안한다. 더불어 무작위성으로 라만 분광의 피크를 변환하는 변환 2가지를 제안한다.

3.2.1 이산 푸리에 변환

이산 푸리에 변환은 주어진 신호로부터 특정 주파수에 해당되는 노이즈를 제거하는 데에 활용된다. 본 논문에서는 여기에 기반을 두어 큰 주기만을 사용해 주어진 신호의 피크의 추세를 유지하면서 변화를 가할 알고리즘을 제안한다. 입력 신호에 대해 푸리에 변환은 다음과 같이 주어진다.

$$f[k]=\sum_{n=0}^{N-1}F[n]e^{-\frac{2\pi i}{N}kn}\quad(k=0,\dots,N-1)$$

이때 역변환은 다음과 같이 나타낼 수 있다.

$$F[n]=\frac{1}{n}\sum_{k=0}^{N/2+1}\left(f[k]e^{\frac{2\pi i}{N}kn}+f[N-k]e^{\frac{2\pi i}{N}(N-k)n}\right)$$

여기서 일부 K개의 항만을 이용해 역변환하여 피크의 정보를 유지시키며 입력 신호를 변환하고자 한다. 입력 신호가 실수라면 모든 k에 대해 $f[k]=\overline{f[N-k]}$가 성립하고 $f[0]$도 실수이므로 다음은 실수 신호가 된다.

$$\hat{F}_{DFT}[n]=\frac{1}{n}\sum_{k=0}^{K}\left(f[k]e^{\frac{2\pi i}{N}kn}+f[N-k]e^{\frac{2\pi i}{N}(N-k)n}\right)$$

주기가 큰 일부만으로 신호를 복원하게 되면 원본에서 피크가 없거나 상대적으로 낮은 피크가 있는 위치에서는 과도한 진동이 발생한다. 이로부터 새로운 피크가 만들어지는 일을 막기 위해 임계값 파라미터 θ를 사용해 그보다 큰 값만 변환한다.

$$F_{DFT}[n]=\begin{cases}\hat{F}_{DFT}[n] & \text{if}\quad F[n]>\theta\\ F[n] & \text{otherwise}\end{cases}\tag{1}$$

3.2.2 이산 웨이블릿 변환

이산 웨이블릿 변환은 웨이블릿 족의 기저로 분해하는 변환이다. 이를 신호의 피크의 추세를 유지하는 변환으로 활용한다. 이산 웨이블릿 변환과 역 이산 웨이블릿 변환은 [21]에, 이를 활용한 디노이징은 [22]에 자세히 정리되어 있다. 입력 신호에 웨이블릿 족 w로 L 레벨만큼 이산 웨이블릿 변환을 적용하면 근사계수 a_L과 디테일 계수 $d_1,\dots,d_L$을 얻는다. 여기서 디테일 계수의 값만을 바꾸어 피크 정보를 유지한 채로 신호에 변형을 가한다. d_L로 노이즈 임계값을 추정하고 이로부터 디테일 계수들을 $\hat{d}_1,\dots,\hat{d}_L$로 임계화한다. 그 후 근사계수 a_L과 임계화된 디테일 계수 $\hat{d}_1,\dots,\hat{d}_L$에 w로 역 이산 웨이블릿 변환을 적용한 것을 $\hat{F}_{DWT}$로 둔다. 필터의 모양에 의존해 피크의 모양이 변형된다. 푸리에 변환에서와 같은 이유로 원본에서 임계값

파라미터 θ보다 큰 부분만 변환해준다.

$$F_{DWT}[n]=\begin{cases}\hat{F}_{DWT}[n] & \text{if } F[n]>\theta \\ F[n] & \text{otherwise}\end{cases} \quad (2)$$

3.2.3 랜덤 초이스와 랜덤 스케일링

앞서 제안한 두 변환을 사용한다면, 각 피크의 상대적 높낮이와 폭에 변화가 작다는 점이 한계가 된다. 또한 두 변환 모두 θ보다 작은 부분에는 전혀 변화가 일어나지 않는다. 그러므로 이러한 문제점을 해결함과 동시에 생성하는 데이터에 무작위성을 더하기 위해 랜덤 초이스와 랜덤 스케일링 방법을 제안한다.

랜덤 초이스 알고리즘은 입력 신호의 인덱스를 랜덤으로 선택된 새로운 인덱스를 대체한다. 입력 신호의 인덱스 $0,\dots,N-1$를 균등하지 않은 간격의 실수 인덱스 $I_0,\dots,I_{N-1}$로 대체할 것이다. 0에서 $N-1$사이의 임의의 실수를 N개 샘플링해 모아둔 집합을 X라 하자. 이때 샘플링된 점들이 특정 구간에 몰리는 것을 방지하기 위해 전체 인덱스를 적당히 분할하고, 분할된 구간별로 균등분포를 이용해 샘플링한다. I_n을 X에서 $n+1$번째로 작은 원소로 두자. 이를 이용해 랜덤 초이스 변환을 정의한다.

$$F_C[n]=F[I_n] \quad (3)$$

I_n은 실수값을 가지므로 실수 인덱스에 대한 신호의 계산이 필요한데, 이를 위해 선형 보간을 이용한다.

랜덤 스케일링 알고리즘은 각 피크의 높낮이를 임의로 변형한다. 스케일링 펙터 $S[n]$를 다음과 같이 정의한다. $\gamma=\{x_1,\dots,x_k\}$가 입력 신호의 피크들의 집합이라 하자. $n\in\gamma$인 경우,

$$S[n]\sim Uniform(\alpha,\beta)$$

와 같이 샘플링해 결정하며, $n\notin\gamma$인 경우 $S[n]$은 선형 보간을 이용해 계산한다. 이로부터 랜덤 스케일링 변환을 다음과 같이 정의한다.

$$F_S[n]=S[n]F[n] \quad (4)$$

Fig. 2에 이번 절에서 제안된 수학적 변환들이 주어진 신호를 어떻게 변환하는지 그림으로 표현되어 있다. 회색선이 변환 전의 신호, 점선은 변환 후의 신호이다. (1)은 주어진 신호를 주기가 큰 함수만으로 표현하기 때문에 각 피크를 가지는 곳에서 피크의 모양을 더욱 완만한 형태로 변형시키고, (2)는 사용되는 웨이블릿 w의 필터 모양에 따라서 피크의 모양을 변화시킨다. (3)은 각 피크의 폭과 상승, 하락하는 모양을 변형하고, (4)는 피크별 높낮이를 임의로 늘리거나 줄인다.

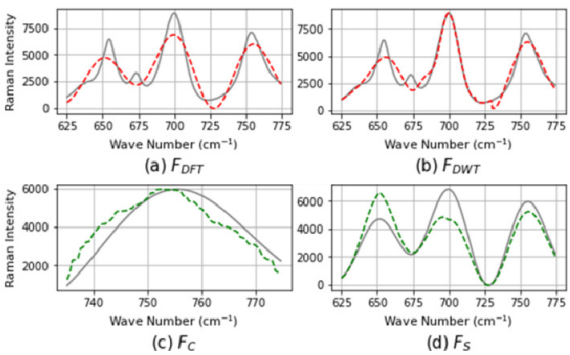

Fig. 2. Original signals(solid line) and signals transformed by mathematical transforms (dashed line)

3.3 생성 모델을 활용한 모의 분광 생성

더욱 다양한 모의 신호를 생성하기 위해 딥러닝 기반의 생성 모델을 학습시켰다. 데이터의 길이 N을 2200으로 사용하면 모델 학습의 안정성이 떨어져 선형 보간을 이용해 그보다 작은 값인 256으로 조정해 사용하였다.

3.3.1 변분적 오토 인코더

수학적 변환으로 생성된 데이터의 복잡한 패턴을 잠재변수로 인코딩하고 효과적으로 디코딩해 복원하는 오토인코더를 학습시키고자 한다. 다음과 같이 복원 손실과 정규화 손실에 대한 항으로 구성된 손실함수를 사용한다.

$$L_{VAE}(x)=-E_{q(z|x)}[\log p(x|z)]+D_{KL}(q(z|x),q(z)) \quad (5)$$

디코더는 Fig. 3과 같은 구조의 합성곱 블록(CB; Convolutional Block)을 활용하고, 두 번의 일차원 합성곱과 배치정규화(BN; Batch Normalization), 선형 업샘플링으로 구성된다. C와 L은 각각 입력 채널의 개수와 길이이다.

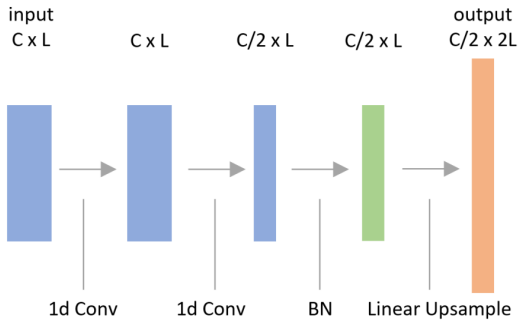

Fig. 3. Convolutional block architecture

Table 1. Encoder and decoder architecture

인코더 구조	디코더 구조
C1d, (64, 4, 2)	FC, 2048
C1d with BN, (128, 4, 2)	FC, 4096
FC, 1024	FC, 128×64
FC, 32 and FC, 32	CB, (64, 5)
	CB, (32, 5)
	C1d, (32, 5, 1)
	C1d, (1, 5, 1)

변분적 오토인코더의 학습에 사용할 인코더와 디코더의 구조는 Table 1에 정리되어 있다. 각 행이 하나의 계층에 대응된다. 일차원 컨볼루션(C1d; 1d Convolution)의 파라미터는 순서대로 필터의 개수, 커널 길이, 스트라이드를 의미한다. 완전연결(FC; Fully Connected) 계층의 파라미터는 노드의 개수다. 합성곱 블록의 파라미터는 순서대로 출력 채널의 개수와 모든 일차원 컨볼루션의 커널 길이다. 배치 정규화는 모두 해당 계층 활성화 함수 이전에 적용된다. 모든 활성화 함수로 음수부분에 0.2가 곱해진 Leaky ReLU를, 디코더의 최종 층에만 시그모이드를 사용했다.

3.3.2 적대적 생성 신경망

다음으로 생성 모델로써 가장 활발히 연구되는 GAN을 활용해 모의 신호를 생성하였다. GAN은 실제 데이터와 유사한 샘플을 생성하는 생성자를 학습하기 위한 모델로 생성자와 판별자의 최대최소게임으로 학습된다. 본 논문에서는 GAN의 학습에는 최대최소게임의 수렴을 돕는 Wasserstein GAN에 기울기 패널티를 주는 손실함수[23]를 사용하였다.

$$L_{GAN} = E_{\tilde{x}\sim P_g}[D(\tilde{x})] - E_{x\sim P_r}[D(x)] + \lambda E_{P_{\hat{x}}}[(\| \nabla_{\hat{x}} D(\hat{x}) \| - 1)^2] \quad (6)$$

실제적인 모의 신호의 생성을 위해 GAN의 생성자의 초깃값으로 사전학습된 VAE의 디코더를 사용한다. 판별자의 구조는 최종 층만 제외하고는 Table 1의 인코더 구조와 같고, 최종 층은 하나의 노드에 시그모이드 함수가 활성화 함수로 사용된다.

제안한 방법들로 모의 신호를 생성하는 과정이 Fig. 4에 표현되어 있다. 붉은 화살표가 생성 과정, 푸른 화살표가 학습 과정을 의미한다. 먼저 원본 신호를 각각 (1)과 (2)를 이용해 변환한 후 순차적으로 (3)과 (4)로 변환해준다. 그 결과로 수학적 변환으로 생성된 모의 데이터를 얻는다. 그 후 해당 모의 데이터를 이용해 (5)를 이용해 VAE를 학습한다. 나아가 학습된 VAE의 디코더를 사전학습된 생성자로 초기화하여 (6)을 이용해 GAN을 학습한다. 학습된 VAE와 GAN으로 기계학습으로 생성된 모의 데이터를 얻는다.

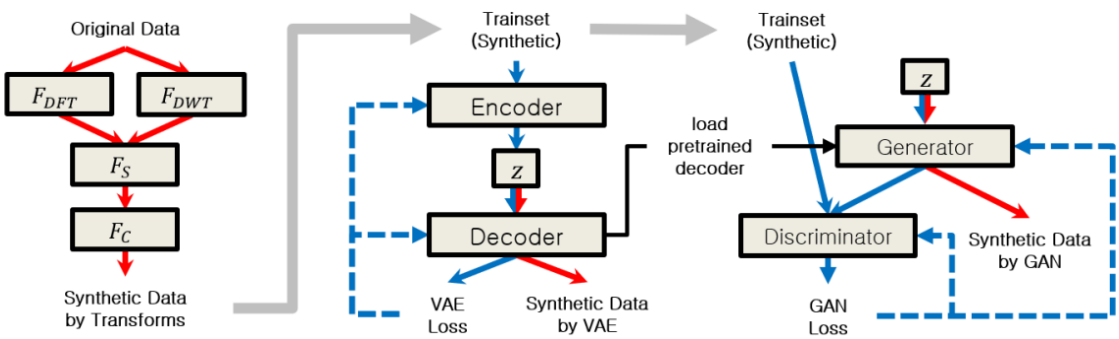

Fig. 4. Procedure for generating synthetic signal data

4. 실험 결과

3장에서 제안한 방법들로 주어진 DMMP와 2-CEES의 모의 라만 분광 데이터를 직접 생성하고, 그 결과를 분석하고자 한다.

4.1 수학적 변환의 파라미터

수학적 변환에 필요한 파라미터들은 물질별 명확한 차이를 유지하면서 다양성을 확보하기 위한 값들로 본 연구에서는 임의로 지정했다. F_{DFT}의 K와 F_{DWT}의 L은 Fig. 5에서 나타난 바와 같이 특정 값까지는 원본과의 평균 절대 편차가 작은 값으로 유지되다가 급격히 증가하는 추세를 보이기 때문에, 변화율이 가장 커지는 위치에서 값을 택했다. K는 Table 2의 값을, L은 5를 사용한다. 웨이블릿 w는 결정된 L 값에 대해 분해가 가능한 필터 중에서 Table 2과 같이 다양하게 선택했다. 또한 F_{DFT}와 F_{DWT}의 임계치 θ는 각 원본 신호에 대한 최대값과 최소값의 차의 5 %가 되도록 결정했고, F_{DWT}의 디테일 계수 임계값 추정에는 전역 임계값을 사용했다. F_S의 α, β는 각 물질이 확실하게 구분되는 값으로 각각 0.8, 1.2로 택했으며, F_C의 분할로 100 cm^{-1} 간격의 균등분할을 사용했다.

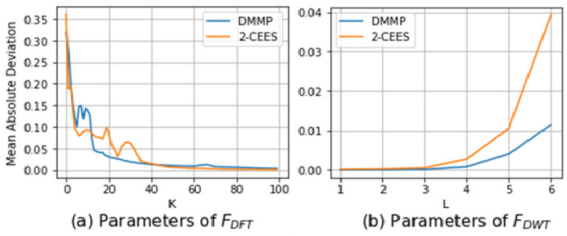

Fig. 5. Mean absolute deviation between the original signal and signals transformed by mathematical transforms with respect to parameters

Table 2. Parameters for mathematical transforms by discrete fourier transform and discrete wavelet transform

변수명	값
K	20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30
w	bior2.2, bior3.1, coif2, coif7, coif11, db2, db3, db5, db9, db12, db13, db19, db30, rbio2.4, rbio3.9, sym6, sym12

4.2 수학적 변환을 이용한 데이터 생성

우선 수학적 변환을 통해 데이터를 생성하고 결과를 분석한다. VAE와 GAN 학습에 일정한 범위의 데이터를 사용하기 위해 생성된 데이터에는 0과 1사이의 값으로 최대최소 스케일링을 적용했다. Fig. 6은 생성된 모의 데이터와 그것의 평균(진한 선)과 최솟값, 최댓값의 범위(연한 범위)를 표현한 것이다. 모의 데이터가 이상치 없이 트랜드를 따라가면서도 변동폭을 가진다. 더불어 피크가 있는 부분에서의 변동폭이 피크가 없는 곳에 비해 더 큰데, 이는 피크에 다양한 변형을 가하고자 했던 의도를 반영하는 결과다. Fig. 1과 비교해보면 생성된 신호의 평균이 원본 신호와 같은 위치에서 피크를 가지고, 이로부터 라만 분광의 특성이 유지되었음을 확인할 수 있다.

Fig. 7은 변환으로 생성된 데이터를 PCA와 t-SNE로 시각화한 결과다. 검은 점은 원본 데이터를 의미한다. 물질별로 변환으로 생성된 데이터는 원본 데이터와 충분히 가까운 위치에 모여 있으며 명확하게 구분되면서도 적당한 분산을 가져 흩어져 있음을 확인할 수 있다.

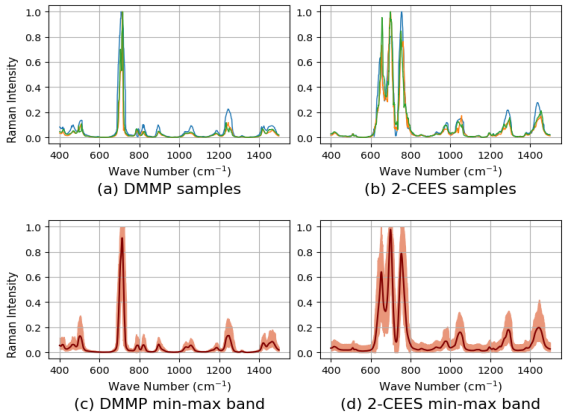

Fig. 6. Synthetic raman spectra generated by mathematical transforms

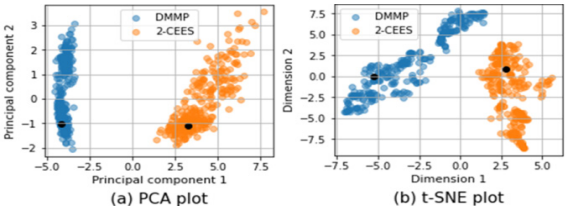

Fig. 7. Visualization of synthetic raman spectra generated by mathematical transforms

4.3 생성 모델을 이용한 데이터 생성

다음으로 변환으로 생성한 데이터를 훈련집합으로 생성 모델을 학습시켜 얻는 새로운 데이터를 확인한다. VAE는 물질별로 0.001 학습율의 Adam으로 1000 에포크씩, GAN은 생성자는 0.00005 판별자는 0.0001의 학습율로 RMSprop을 활용해 물질별로 1000 에포크씩 학습시켰다.

Fig. 8, 9, 10에서 수학적 변환과 생성 모델로 생성된 데이터의 분포를 상세하게 비교하고 있다. 각 생성 모델의 훈련데이터와 생성된 데이터를 같은 개수만큼 가지고 t-SNE를 수행한다. 이때, 생성모델이 생성한 데이터는 기존 데이터 길이인 2,200으로 선형보간하여 비교하였다. Fig. 8은 VAE의 잠재공간에서 샘플링으로 생성된 데이터과 훈련데이터를 시각화한 결과다. VAE로 생성된 데이터의 분포들 모두 훈련데이터의 분포와 충분히 겹쳐지고, 이는 VAE가 훈련데이터에 내재된 분포를 잘 모방하게끔 학습되었다는 것을 의미한다.

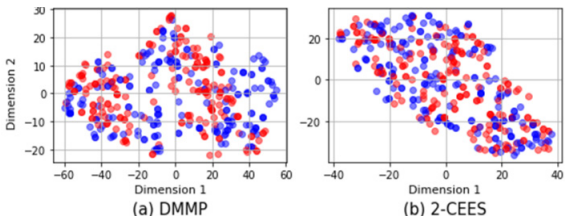

Fig. 8. t-SNE plots of raman spectra generated by mathematical transforms(red) and VAE(blue)

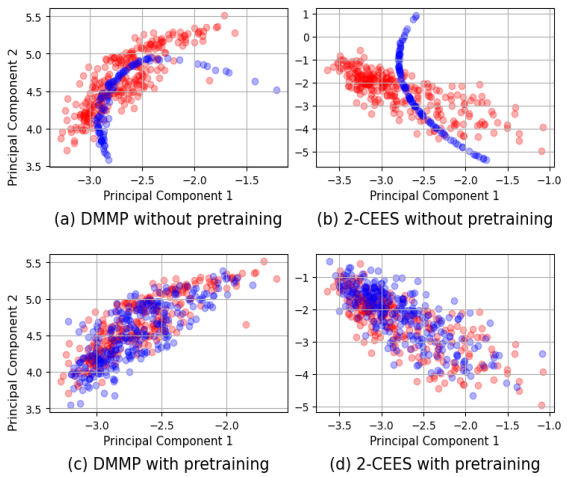

Fig. 9. PCA plots of raman spectra generated by mathematical transforms(red) and GAN(blue)

Fig. 9는 학습된 GAN이 생성한 데이터를 시각화한 결과다. 사전학습 없이 학습을 진행했을 때, 생성된 데이터가 충분히 다양하지 못하며, 개선을 위해 VAE의 디코더를 생성자의 초깃값으로 초기화하여 GAN을 학습시킨 결과, 수학적 변환 데이터와 충분히 유사하며 다양한 데이터가 생성되었다.

Fig. 10에는 최종적으로 GAN이 생성한 모의 데이터와 훈련데이터의 t-SNE 시각화가 나타나 있다. Fig. 8과 비교해보았을 때도 GAN이 생성한 데이터의 분포가 훈련데이터의 분포와 더욱 잘 겹쳐지는 것을 확인할 수 있다. 마찬가지로 GAN이 훈련데이터를 잘 모방한다는 것을 알 수 있다.

생성된 데이터와 원본과의 차이로 각 방법으로 생성한 데이터들의 분포가 유사한지 알아볼 것이다. Table 3은 각 방법으로 생성된 데이터와 원본 데이터 사이의 상관계수 나타낸 표이다. 표의 값을 통해 생성된 모의데이터가 원본과 높은 상관도를 가지도록 생성되었음을 확인할 수 있다.

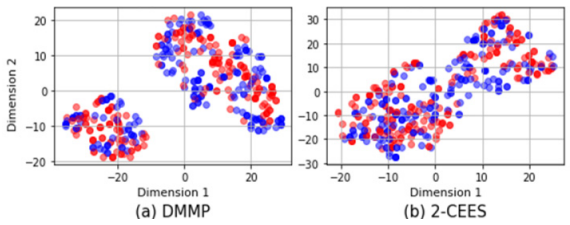

Fig. 10. t-SNE plots of raman spectra generated by mathematical transforms(red) and GAN(blue)

Table 3. Correlation cofficients between generated and original raman spectra

물질	변환	VAE	GAN
DMMP	0.9423	0.9459	0.9589
2-CEES	0.9684	0.9712	0.9749

5 결 론

본 논문은 적은 수의 데이터만을 가지고 그와 유사한 모의 라만 분광을 생성하는 알고리즘을 제안하고, 이를 통해 어떤 물질의 라만 분광이 주어지면 수학적 변환과 생성모델의 학습을 거쳐 모의 라만 분광 데이

터셋을 구축한다. 독성 물질인 DMMP와 2-CEES의 라만 분광 데이터로 모의 데이터를 생성한 후 해당 물질에 대한 충분한 크기의 모의 라만 분광 데이터셋을 구축하였다. 향후 본 연구에서 제안한 알고리즘으로 생성한 모의 라만 분광을 이용하여 다양한 환경에서 독성 물질을 검출하는 탐지 모델의 개발을 진행할 것이다.

후 기

이 성과는 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(No. 2021R1F1A1054766).

References

[1] Young Jin Koh, “The Design and Test of the Stand-off Surface Chemical Contaminant Detection System based on Raman Spectroscopy,” Journal of the Korea Institute of Military Science and Technology, 22.3, 433-440, 2019.

[2] Sun-Kyung Choi, et al., “Deep UV Raman spectroscopic study for the standoff detection of chemical warfare agents from the agent-contaminated ground surface,” Journal of the Korea Institute of Military Science and Technology, 18.5, 612-620, 2015.

[3] S. K. Choi, et al., “Analysis of Raman Spectral Characteristics of Chemical Warfare Agents by using a 248 nm UV Raman Spectroscopy,” Bulletin of the Korean Chemical Society, 40(3): 279-284, 2019.

[4] S. K. Choi, et al., “Detection of toxic chemicals on the surface by a stand-off Raman spectroscopy,” Bulletin of the Korean Chemical Society, 40(6): 483-484, 2019.

[5] J. H. Lee, et al., “Detection of hazardous chemical using dual-wavelength Raman spectroscopy in the ultraviolet region,” Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy, 287: 122061, 2023.

[6] LeCun, Yann, Yoshua Bengio, and Geoffrey Hinton, “Deep learning,” nature, 521.7553, 436-444, 2015.

[7] Luo, Ruihao, Juergen Popp, and Thomas Bocklitz, “Deep Learning for Raman Spectroscopy: A Review,” Analytica 3.3, 287-301, 2022.

[8] Goodfellow, Ian, Yoshua Bengio, and Aaron Courville, Deep learning, MIT press, 2016.

[9] Kingma, Diederik P., and Max Welling, “Auto-encoding variational Bayes,” International Conference on Learning Representations, 2014.

[10] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David WardeFarley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio, “Generative adversarial nets,” Advances in Neural Information Processing Systems, 2014.

[11] Hao Chen, et al., “An adaptive denoising method for Raman spectroscopy based on lifting wavelet transform,” Journal of Raman Spectroscopy, 49.9, 1529-1539, 2018.

[12] Chang Sik Lee, et al., “Denoising Autoencoder based Noise Reduction Technique for Raman Spectrometers for Standoff Detection of Chemical Warfare Agents,” Journal of the Korea Institute of Military Science and Technology, 24.4, 374-381, 2021.

[13] Xiangyun Ma, et al., “Conditional Generative Adversarial Network for Spectral Recovery to Accelerate Single-Cell Raman Spectroscopic Analysis,” Analytical Chemistry, 94.2, 577-582, 2022.

[14] Jae-Hyeon Park, et al., “CNN based Raman Spectroscopy Algorithm That is Robust to Noise and Spectral Shift,” Journal of the Korea Institute of Military Science and Technology, 24.3, 264-271, 2021.

[15] Yu, Shixiang, et al., “Classification of pathogens by Raman spectroscopy combined with generative adversarial networks,” Science of The Total Environment, 726, 138477, 2020.

[16] Wen, Qingsong, et al., “Time series data augmentation for deep learning: A survey,” International Joint Conference on Artificial Intelligence, 2021.

[17] Yoon, Jinsung, Daniel Jarrett, and Mihaela Van der Schaar, “Time-series generative adversarial networks,”

Advances in neural information processing systems, 2019.

[18] Xu, Tianlin, et al., “Cot-gan: Generating sequential data via causal optimal transport,” Advances in Neural Information Processing Systems, 2020.

[19] T. T. Um et al., “Data augmentation of wearable sensor data for Parkinson's disease monitoring using convolutional neural networks,” Proceedings of the 19th ACM ICMI, 2017.

[20] Lotte, Fabien, “Signal processing approaches to minimize or suppress calibration time in oscillatory activity-based brain-computer interfaces,” Proceedings of the IEEE 103.6, 871-890, 2015.

[21] Mallat, Stéphane, “A wavelet tour of signal processing,” Elsevier, 1999.

[22] Guomin Luo and Daming Zhang, “Wavelet Denoising,” Advances in Wavelet Theory and Their Applications in Engineering, Physics and Technology, 2012.

[23] Ishaan Gulrajani, et al., “Improved training of wasserstein gans,” Advances in neural information processing systems, 2017.