New Physics: Sae Mulli,
Vol. 70, No. 11, November 2020, pp. 920∼927
http://dx.doi.org/10.3938/NPSM.70.920

Deep Learning Applied to Peak Fitting of Spectroscopic Data in Frequency Domain

Hyeong Seon Park · Seong-Heum Park · Hyunbok Lee · Heung-Sik Kim*
Department of Physics, Kangwon National University, Chuncheon 24341, Korea

(Received 31 August 2020 : revised 17 September 2020 : accepted 21 September 2020)

A data-driven study of material properties and functional materials design based on it requires high-throughput and comparative analyses of the results of experimental spectroscopy with those from first-principles electronic structure calculations. Hence, an efficient machine-learning-based computational tool to extract electronic structure information from experimental data without human intervention is in high demand. Here, we test the capability of deep neural network models to fit photoemission spectroscopy (PES) data in the frequency domain with unknown PES peak positions, numbers, and widths. A one-dimensional convolution neural network (CNN) was employed in combination with fully connected layers (FCL), and the trained model was applied to photoemission spectra for the sulfur $2p$ states in poly(3-hexylthiophene) (P3HT) molecules and oxygen $1s$ states in indium tin oxide (ITO). We conclude by further discussing potential ways to improve the performance of the model.

Keywords: Photoemission spectroscopy, Machine learning, Deep neural network

진동수 영역의 분광학 스펙트럼 분석을 위한 심층 기계학습 모델 적용

박형선 · 박성흠 · 이현복 · 김흥식*
강원대학교 물리학과, 춘천 24341, 대한민국

(2020년 8월 31일 받음, 2020년 9월 17일 수정본 받음, 2020년 9월 21일 게재 확정)

최근 들어 학계 및 산업계에서는 데이터 사이언스를 이용한 물질 분석 및 신물질 디자인이 중요한 연구 주제로 떠오르고 있으며, 이를 위한 필수 요소 중 하나는 실험적 분광학 결과 및 제일원리 전자구조계산 결과에 대한 고속대량 데이터 스크리닝 작업이다. 이러한 작업에는 대량의 실험적 분광학 데이터와 제일원리 전자구조계산 결과들로부터 유용한 정보들을 최소한의 인간의 개입만으로 추출할 수 있는 기계학습 기반 방법론이 필수적이다. 이를 위하여, 본 연구에서는 1차원 진동수 영역에서의 광전자 분광학 (photoemission spectroscopy, PES) 실험 결과들을 입력받아, 이로부터 전자의 여기 에너지, 여기 상태의 수 및 각 PES 피크의 에너지 폭을 얻어 내는 심층 신경망 모델을 만들고 훈련시켜 보았다. 본 모델에서는 1차원 합성곱 신경망(convolution neural network, CNN)을 완전연결 신경망(fully-connected layers, FCL)과 조합하여 사용하였으며, 훈련된 모델은 Poly(3-hexylthiophene) (P3HT) 분자 내 황의 $2p$ 상태 및 인듐 주석 산화물 내 산소의 $1s$ 상태로부터의 PES 스펙트럼을 분석하는데 사용되었다. 마지막으로 현재의 모델을 보다 개선하기 위한 방법에 대한 논의를 덧붙인다.

This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/3.0) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

Keywords: 광전자 분광학, 기계 학습, 심층 신경망

I. 서 론

컴퓨터 하드웨어, 소프트웨어 및 통신 기술 발전에 따라 인류가 축적한 정보의 양은 날이 갈 수록 급격하게 증가하고 있으며, 이에 따라 인류가 지금까지 생성 및 축적해 온 막대한 양의 데이터로부터 과학 및 산업계의 문제들에 대한 해결책의 실마리를 찾고자 하는 '데이터 과학 (data science)'적 접근 방법이 최근 들어 큰 관심을 받고 있다 [1]. 이러한 막대한 양의 데이터를 효율적으로 처리 및 분석하기 위한 '기계 학습 (machine learning)' 방법론 또한 오랜 세월에 걸쳐 발전해 왔으며, 최근에는 생물의 두뇌 신경망 구조에 영감을 받은 '인공 신경망 모델(artificial neural network)' 및 '심층 신경망 모델 (deep neural network, DNN)' 이 이 분야의 핵심 키워드로서 떠오르고 있다 [2,3]. 2010년대 이후의 그래픽 처리 장치 등의 하드웨어의 급격한 발전 및 알고리즘 분야의 꾸준한 연구에 힘입어, 현재 인공 신경망 모델에 기반한 연구들은 다종 다양한 분야에서 기존의 인공지능 분야에서 이루어 온 업적들을 압도하는 결과물들을 내놓고 있으며 [4,5], 물리학 연구에서도 이는 예외가 아니어서 이미 기계학습 방법론을 다체계 양자역학 및 통계물리학, 전산재료과학 등에 적용한 주목할만한 연구 결과들이 발표되고 있는 상황이다. [6–10]

이러한 데이터 과학 및 기계학습 방법론을 응집물질물리학 및 전산재료과학에 적용하여 얻을 수 있는 궁극적인 목표 중 하나로서, 과학 및 산업 응용에 유리한 성질을 갖는 물질을 추론 및 디자인 할 수 있는 '역방향 신물질 디자인 (inverse design)'을 들 수 있다 [11,12]. 다양한 조성 및 구조를 갖는 물질들의 물성 및 실험 결과들에 대한 데이터베이스를 구축하며, 이로부터 연구자가 원하는 성질을 가지기에 유리한 조성 및 구조의 패턴을 데이터베이스 내 정보들을 통해 얻어낼 수 있는 기계학습 모델을 고안 및 학습할 수 있다는 것이 이러한 역방향 디자인의 기본적인 전제조건이라 할 수 있다. 이를 위해서는, 다종 다양한 물질들에 대한 각종 1차 결과들, 즉 각종 분광학적 실험 결과들의 미 가공 데이터 및 해당 물질들에 대한 제일원리 전자구조계산으로부터 얻은 파동함수나 전하 밀도 등의 데이터들로부터 유용한 정보들을 인간의 개입 없이 (또는 최소한의 개입 만으로) 추출해 낼 수 있는 효율적인 기계학습 모델의 개발이 필수적이라 할 수 있다.

본 연구에서는 이를 위한 시작 단계로서, 운동량 의존성을 갖지 않는 에너지 영역에서의 광전자 분광학 (photoemission spectroscopy, PES) 스펙트럼과 같은 1차원 영역에서의 데이터에 대한 비선형회귀를 자동으로 수행할 수 있는 인공신경망 모델을 고안 및 훈련하고, 이를 실제 2가지 물질의 PES 스펙트럼에 적용하여 모델의 성능을 평가하는 작업을 수행하였다. 이를 위해 1차원 합성곱 신경망 (convolution neural network, CNN) 모델 및 완전 연결 신경망 (fully-connected layer, FCL) 구조를 조합한 간단한 모델을 고안 및 훈련하였다. 훈련을 위한 인공 데이터들은 실제 실험결과 분석에 사용되는 유사 Voigt 함수들의 무작위 선형 결합을 통해 생성되었으며, 이를 통해 훈련된 모델을 Poly(3-hexylthiophene) (P3HT) 분자 내 황의 $2p$ 상태 및 인듐 주석 산화물 내 산소의 $1s$ 상태로부터의 PES 스펙트럼에 적용 [13], 도출된 결과들을 스펙트럼 분석툴로부터 얻어진 피팅 결과와 비교하였다.

본 논문에서는 먼저 심층신경망 학습을 위해 사용된 방법론 및 플랫폼을 소개하며, 이어서 신경망 훈련을 위해 만들어진 인공 데이터들 및 신경망 모델에 관해 간략히 설명한다. 그리고 이를 P3HT 및 ITO 스펙트럼에 적용하여 피팅을 수행한 결과를 이전에 알려진 피팅 결과와 비교하며, 신경망 모델 성능의 개선을 위해 필요한 부분들에 및 앞으로의 연구 진행 방향에 대한 논의로서 논문을 마무리하고자 한다.

II. 본론: 계산 방법론, 결과 및 논의

1. 계산 방법론 및 도구

본 연구에서는 신경망 모델 설정 및 학습을 위한 오픈 소스 Keras 프레임워크를 사용하였다 [14]. Keras는 고차원 텐서 연산을 그래픽 처리장치 또는 텐서 연산장치를 사용하여 수행하는 TensorFlow 백엔드 [15] 위에서 작동하며, 파이썬 프로그래밍 언어 및 NumPy 모듈에 대한 기초적인 지식만으로도 신경망 모델 생성 및 학습을 수행할 수 있게 해준다. 초기 단계에서의 연구를 위해서 Google 에서 무료로 제공하는 Colab[2] 환경을 이용하였으며, 마지막 단계 모델 학습을 위해서는 연구실에서 자체적으로 보유한 Jupyter notebook 서버 환경 및 계산 자원을 사용하였다. ITO 및

*E-mail: heungsikim@kangwon.ac.kr

[2] https://colab.research.google.com

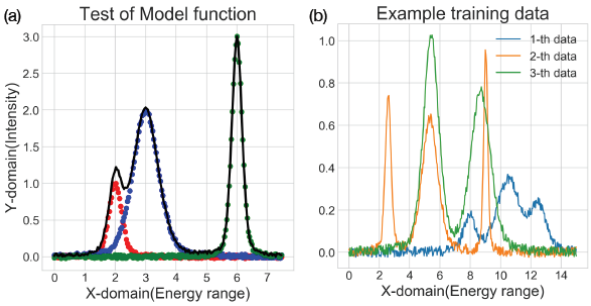

Fig. 1. (Color online) (a) An example of generated training dataset consisting of three pseudo-Voigt peaks; red, blue, green dots show each of three pseudo-Voigt peaks generated by Eq. (1) with noise. Black solid curve is sum of all three peaks. (b) Plots of three generated training data.

P3HT의 X선 PES 실험을 위해서는 K-Alpha+ (Thermo Fisher Scientific Co.) 시스템을 사용하였으며, 코어 레벨 스펙트럼을 얻기 위해 Al Kα X선 광원이 ($h\nu$ = 1486.6 eV) 사용되었다.[3,4]

2. 모델 훈련을 위한 인공 데이터 생성

인공 신경망 학습에 요구되는 막대한 양의 데이터에 비해서, 실험에서 얻을 수 있는 데이터의 양에는 한계가 있다. 따라서, 본 연구에 있어서 가장 먼저 요구되는 것은 신경망의 학습을 위한 대규모 훈련 데이터의 생성이었다. 이를 위해서, X선 회절 또는 PES 에서의 개개의 여기 피크를 피팅하는데 널리 사용되는 유사 Voigt 함수를 사용하였으며 [16], 실제 에너지 피크를 생성하는데 사용한 함수는 다음과 같다.

$$f(\omega;\omega_0,I_0,w)=0.7e^{-\frac{w_G^2(\omega-\omega_0)^2}{w^2}}+0.3\frac{1}{1+\frac{w_L^2(\omega-\omega_0)^2}{w^2}} \quad (1)$$

위 식에서 ω는 진동수이고, w_G = 1.634, w_L = 2.195로 고정된다. 따라서 위 식에 의해 만들어지는 피크는 최대 세기 I_0, 피크의 폭을 나타내는 단일 변수 w, 그리고 피크의 위치인 ω_0 3개의 변수로 나타낼 수 있으며, 일정 구간 내에서 이 3개의 변수를 무작위로 생성하여 피크를 생성한다. 이 때, 실제 실험에서 발생할 수 있는 피크 모양의 좌우 비대칭성은 고려하지 않았다. 개개의 훈련 데이터는 진동수 구간 내에서 1 3개의 피크를 가지며, 피크의 갯수 또한 무작위로 결정된다. 마지막으로, I_0 의 최대 5% 정도의 노이즈를 더하여 실제 실험서 발생하는 노이즈 효과를 모방하였다. (Fig. 1 참조) 다만, 실제 측정에서 발생할 수 있는 배경값은 무시하였으나, 배경값을 포함한 1차원 진동수 영역에서의 신호 처리 또한 CNN 방법론을 통해 효과적으로 수행될 수 있음이 밝혀져 있다 [17]. 학습을 위하여 진동수 ω는 ω_i ($0 \le i \le 400$) 의 값들로 이산화되어, 기계학습을 위해 입력되는 데이터는 $\{f_n(\omega_i)\}$ 의 벡터로서 표현될 수 있다 ($1 \le n \le N$, N은 학습 데이터셋 내 데이터 총 수).

3. 학습을 위한 심층 신경망 모델 구조 설정 및 훈련

다음 단계로서, 훈련의 대상이 되는 신경망의 구조를 설정하여야 한다. 심층 신경망 모델 중 CNN이 사진이나 동영상 등 2차원 화상의 특성 추출에 매우 뛰어난 성능을 보인다는 점에 착안하여, 본 연구에서는 PES 스펙트럼을 1차원 데이터로 간주하고 신경망 모델의 데이터 입력 부분에 여러 겹의 CNN 배열을 적용하였다. (Fig. 2 참조) 신경망의 앞 부분에는 Conv1D (1차원 합성곱) 및 MaxPooling1D (데이터 크기를 1/2로 줄이는 다운샘플링) 층이 3중으로 배치되어 입력 데이터의 차원을 효과적으로 축소하는 한 편 적절한 깊이의 특성 추출이 효과적으로 이루어질 수 있도록 하였다. 다음 단계로서 GlobalMaxPooling1D(전역적 맥스풀링) 층을 사용하여 추출된 특성값들을 1차원 벡터로서 변환 후, 해당 특성값 벡터를 입력받아 원하는 목표 변수 4종 — 스펙트럼 내 피크 수, 각 피크의 세기 I_0와 중심 위치 ω_0, 그리고 에너지 폭 w — 을 추출 및 출력하는 FCL 층 4개를 연결하였다. 여기에서 스펙트럼 내 피크 수 및 각 피크의 중심 위치들은 나머지 변수들을 예측하는데 필요한 정보이기 때문에, 이 두 값들은 먼저 추출되어 각 피크 세기와 에너지 폭을 추출해 내는 FCL 층의 입력값으로 batch-normalization 층을 거쳐 제공되도록 하였다. 모든 층에서의 활성화 함수는 ReLU (Rectified Linear Unit) 함수를 사용하였다. 참고로, 본 연구에서는 피크의 수가 최대 3개인 데이터들을 생성하여 모델을 훈련하였으며, 피크 수가 이보다 작은 데이터의 경우 목표 변수들의 값들이 피크 수 만큼만 예측되고 나머지는 0으로 출력되도록 모델을 학습하였다.

모델 훈련을 위한 총 입력 데이터 수는 2천만개, 그리고 학습을 위한 에포크 수는 200개로 조정하였으며, 모델 가중치 최적화를 위해서는 Adam [18] 알고리즘을 사용하였다. 또한 학습과정에서의 과대적합(overfitting)을 방지하기 위해, 학습 과정에서 과대적합이 일정한 에포크 이상으로 지속적으로 발생할 경우 학습률(learning rate)를 감소시키는

[3] 보다 자세한 실험 방법을 위해서는 Ref. 13 를 참고하시오.
[4] 본 연구에 사용된 코드는 다음 링크에서 다운로드 받을 수 있다: [link]

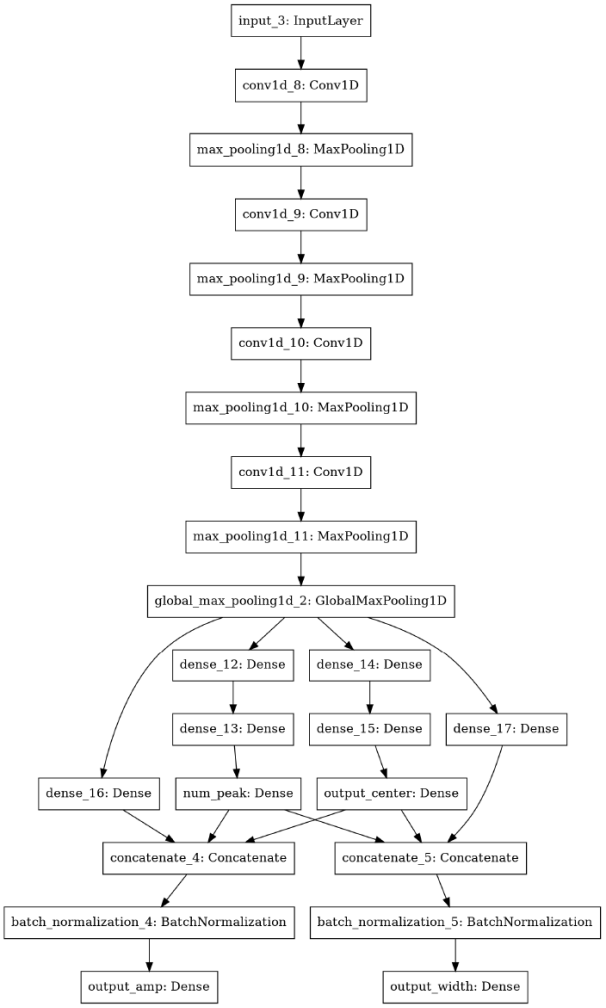

Fig. 2. (Color online) Diagram illustrating our model structure, where InputLayer at the top receives training dataset $\{f_n(\omega_i)\}$. Initially data are processed via three consecutive one-dimensional convolution and max-pooling layers for feature selection and extraction, which are then passed to four FCL layers to finally obtain four target variables; number of peaks, intensities, peak centers, and widths of each peak.

장치 및 항상 최적의 손실함수 감소를 보인 상태에 대해서만 신경망 모델을 저장하는 장치를 콜백 함수를 통해 삽입하였다. 모델의 과대적합 관련하여 한 가지 언급해야 할 것은, 위에서 설명한 모델보다 더욱 복잡한 구조를 가진 모델이 더 나은 성능을 보여줄 것으로 예상할 수 있으나, 실제로는 보다 복잡한 구조 및 더 많은 수의 가중치 변수들을 가진 모델들이 더 과대적합에 빠지기 쉬운 경향을 보였기 때문에 그렇지 않았다는 점이다. 이를 해결하기 위해 유사한 성능을 보이는 여러 가지 모델 중 가장 간단한 구조를 갖는 신경망 모델을 선정하였으며, 이러한 고려들을 통해 모델의 과대적합을 억누르고 올바른 방식으로 손실함수를 감소케 하였다. (Fig. 3 참조) 모델 내 총 학습 가능한 가중치 변수

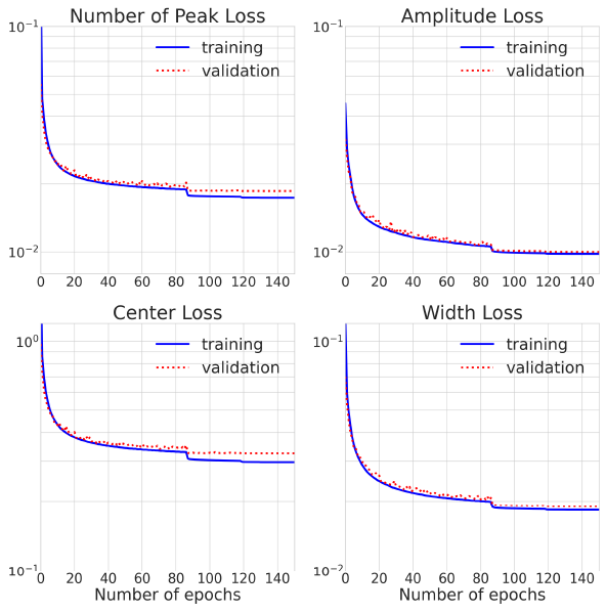

Fig. 3. (Color online) Values of loss functions as a function of training epoch. Four loss functions are shown; (top left) number of peaks, (top right) amplitudes I_0, (bottom left) centers ω_0, and (bottom right) widths w of each peak. Total loss is given as a sum of the four losses. All loss functions almost saturate after 120 epochs, where the learning rate of Adam optimizer is decresed below 10^{-4}. Overfitting, which can be signaled by an increase of validation loss, was well-suppressed as can be seen from the data.

갯수는 약 17만 3천개이다.

4. 훈련된 모델의 시험 데이터 적용 결과

Figure 4는 상기 과정들을 통해 훈련을 마친 모델에 테스트 데이터셋을 적용한 결과들이다. 결과 중 위쪽 2개 및 왼쪽 아래 패널은 2개의 피크로 이루어진 테스트 데이터를 사용한 결과이며, 오른쪽 아래 패널은 피크 3개로 이루어진 테스트 데이터에서 나온 결과이다. 위 4개의 결과에서 확인할 수 있듯이, 무작위로 생성된 테스트 데이터셋에 대해서 상기한 훈련된 모델이 잘 작동하여 비교적 좋은 피팅 결과를 이끌어 내는 것을 볼 수 있다. 물론, 이는 피크의 위치 및 대략적인 세기를 입력 후 회귀 분석을 통해 이끌어 낸 결과들에 비해서는 정확도가 떨어진다 할 수 있으나, 인간의 개입 및 배경지식 없이 기계학습된 모델만을 사용한 결과로서는 인상적이라 할 수 있다. 특히, Fig. 4의 오른쪽 위 패널의 경우, 두 개의 피크 중심이 상당히 가깝게 위치해 있는데다 하나의 피크의 세기가 약한 어려운 데이터임에도 불구하고 두 개의 피크의 위치 및 세기를 잘 잡아내는 것을 확인할 수 있다.

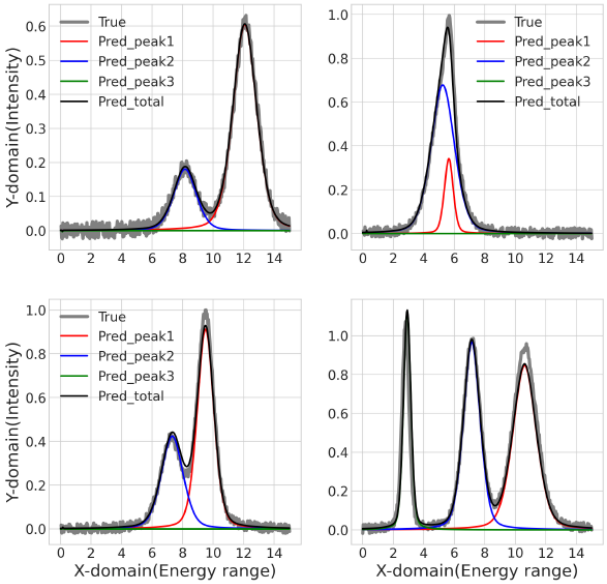

Fig. 4. (Color online) Values of loss functions as a function of training epoch. Four loss functions are shown; (top left) number of peaks, (top right) amplitudes I_0, (bottom left) centers ω_0, and (bottom right) widths w of each peak. Total loss is given as a sum of the four losses. All loss functions almost saturate after 120 epochs, where the learning rate of Adam optimizer is decresed below 10^{-4}. Overfitting, which can be signaled by an increase of validation loss, was well-suppressed as can be seen from the data.

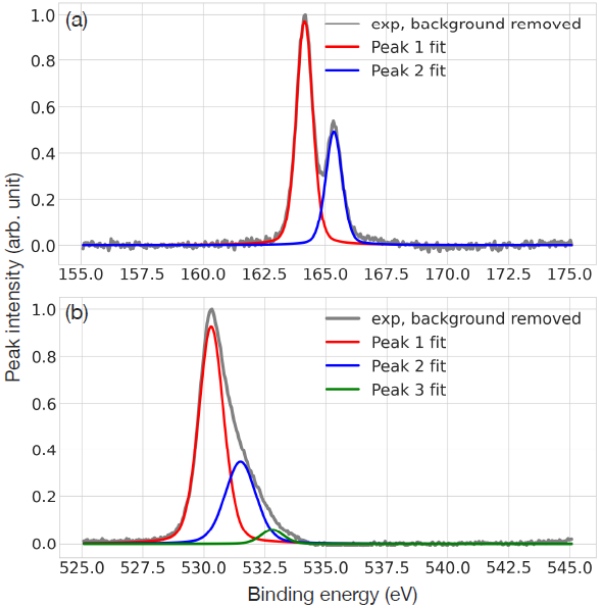

Fig. 5. (Color online) Core level X-ray PES spectra of (a) S $2p$ states of P3HT molecules on ITO substrate and (b) O $1s$ states of ITO. In both panels gray curves are measured spectra with backgrounds subtracted, and red, blue, and green curves are fitted peaks with pseudo-Voigt functions using IGOR PRO package (Eq. 1). In (a), red and blue curves corresponds to excitations from S $2p_{3/2}$ and $2p_{1/2}$ levels, respectively. In (b), energy differences between each peak are induced by the shift of oxygen $1s$ level by different chemical environments such as oxygen deficiency and hydrogen bondings at surfaces [19].

참고로, 피크의 갯수가 하나로 고정된 경우에는, CNN 없이 FCL 층만을 통해 신경망 모델을 구성하더라도 매우 정확한 결과를 얻을 수 있었다. 하지만 피크의 갯수가 2개 이상으로 주어지며 또한 노이즈가 섞여 있는 결과들에 대해서는, 유사한 최적화 가능 가중치 변수 숫자를 갖춘 모델들을 비교했을 때 CNN+FCL 층의 조합이 FCL만으로 구성된 모델에 비해 보다 나은 성능을 보여주는 것을 확인할 수 있었다. 이는 노이즈 처리 및 신호의 국소적 특성 추출에 유리한 CNN 층의 장점이 나타난 결과라 해석할 수 있으리라 생각된다.

5. 훈련된 모델의 실제 광전자분광학 데이터 적용 결과

마지막으로, 학습된 모델을 ITO 기판 위에 증착된 P3HT 분자들 및 ITO 자체에 대한 코어 레벨 X선 PES (XPS) 스펙트럼 측정 결과에 적용해 보았다. Figure 5(a), (b)는 각각 P3HT 분자들의 S $2p$ 및 ITO 기판 자체의 O $1s$ 코어 레벨 XPS 측정 결과이며, 각각의 그림에서 빨간색, 파란색 및 초록색 커브들은 IGOR PRO 패키지를 이용하여 피팅된 피크들을 나타낸다. 분석을 용이하게 하기 위하여 측정된 스펙트럼에서 배경 신호를 제거하였으며, 또한 스펙트럼의 최대 값을 1이 되도록 데이터를 재규격화 하였다.

Figure 6은 P3HT에서 측정된 XPS 스펙트럼 및 IGOR PRO 패키지 피팅을 통해 얻은 각 피크 커브들을 위 과정을 통해 얻은 신경망 모델로부터 얻은 결과들과 비교하고 있다. Figure 6(a)는 배경 신호를 제거한 실험 결과(회색 선)를 신경망 모델을 통해 도출한 피크들의 합(검은색)과 비교하고 있으며, 이 두 결과가 상대적으로 잘 일치하고 있음을 보여주고 있다. Figure 6(b)는 실험 결과로부터 IGOR PRO 패키지 피팅을 통해 얻어낸 황의 $2p_{1/2}$ (어두운 파란색 선) 및 $2p_{3/2}$ (어두운 빨간색) 피크들과 신경망 모델이 얻어낸 피크들을 (밝은 파란색 및 빨간색) 비교하고 있으며, 이를 통해 P3HT에서는 각 피크의 위치 및 세기, 그리고 피크들의 폭을 신경망 모델이 비교적 잘 얻어내고 있음을 알 수 있다. 이는 P3HT에서는 S $2p_{1/2}$ 와 $2p_{3/2}$ 상태 사이의 에너지 차이가 (~ 1.2 eV) 피크의 폭보다 충분히 크기 때문에 [13], 두 피크가 비교적 잘 구분되고 있기 때문으로 추정된다.

마지막으로 Fig. 7은 ITO에서 측정된 XPS 스펙트럼 및 IGOR PRO 패키지 피팅을 통해 얻은 각 피크 커브들을 신

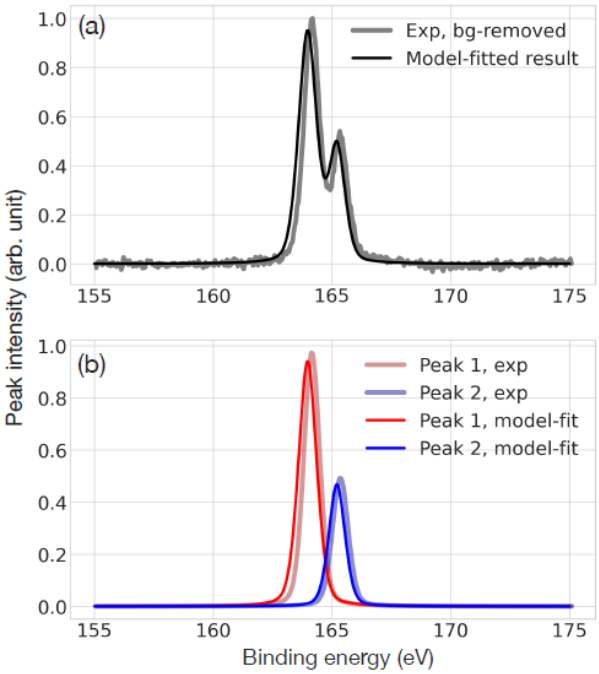

Fig. 6. (Color online) (a) Comparison between the measured (gray) and model-generated (black) P3HT spectra. (b) Comparison between the human-fitted (dark colors) and model-predicted (bright colors) results for each peak.

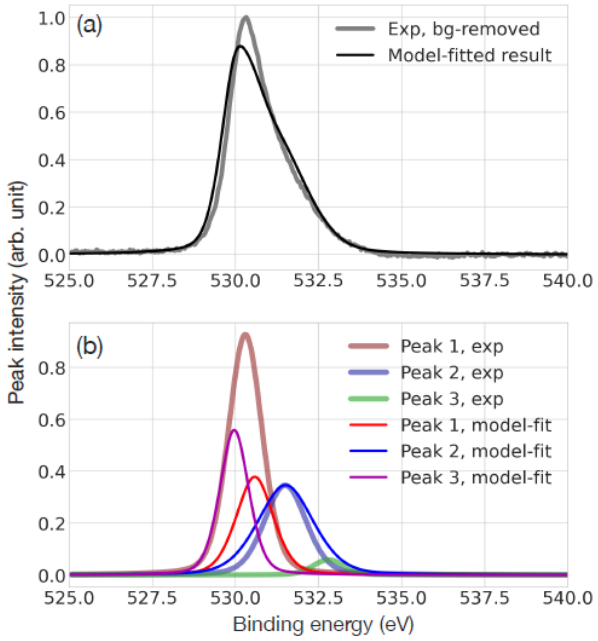

Fig. 7. (Color online) (a) Comparison between the measured (gray) and model-generated (black) ITO spectra. (b) Comparison between the human-fitted (dark colors) and model-predicted (bright colors) results for each peak.

경망 모델 예측 결과들과 비교하여 보여주고 있다. Figure 5(b)에서 보여진 것처럼, ITO 표면에 존재하는 산소 결핍 및 수소 결합의 존재는 서로 다른 3개의 산소 $1s$ 여기 상태 피크를 만들어낸다 [19]. 단, Fig. 7(a) 및 5(b)에서 확인할 수 있듯이, P3HT의 경우와는 달리 각 피크들의 폭에 비해 피크들 사이의 에너지 차이가 크지 않아 서로 다른 피크들이 충분히 잘 구분되지 않으며, 이 때문에 신경망 모델이 전체 스펙트럼으로부터 서로 다른 피크들을 분리해 해는 것에 어려움이 있을 수 있음을 예상할 수 있다. Figure 7(b)는 IGOR PRO 패키지 피팅 결과 및 신경망 모델이 도출한 결과들을 비교하고 있으며, 위에서 예상한 바와 같이 신경망 모델이 각 피크의 위치 및 세기를 잘 도출해 내고 있지 못함을 확인할 수 있다. 특히, 530.29 eV에 위치한 첫번째 피크를 (어두운 빨간색), 신경망 모델은 서로 다른 2개의 근접한 피크들의 (밝은 보라색 및 빨간색) 합으로 잘못 해석하고 있으며, 대신 532.8 eV에 위치한 세번째 피크의 존재를 (어두운 초록색) 신경망 모델이 놓치고 있는 것을 확인할 수 있다.

P3HT	Fitted	Model-predicted	**ITO**	Fitted	Model-predicted
Peak 1			Peak 1		
ω_0	164.15	163.97	ω_0	530.29	530.58
w	0.789	0.683	w	1.199	0.999
Peak 2			Peak 2		
ω_0	165.35	165.21	ω_0	531.5	531.51
w	0.786	0.609	w	1.483	1.496
			Peak 3		
			ω_0	532.8	529.97
			w	1.095	0.779

Table 1. Comparison between fitted and model-predicted parameters from P3HT and ITO spectra.

Table 1은, 신경망 모델에서 얻어낸 파라미터들을 연구자가 패키지 피팅을 통해 얻은 결과들과 비교하고 있으며, 현재 사용한 초보적인 수준의 신경망 모델에 개선의 여지가 아직 많이 남아 있음을 보여주고 있다.

III. 결 론

본 연구를 통해 기계학습 및 심층신경망 방법론의 물리학 문제 및 실험 결과의 분석에 대한 적용 가능성을 확인할 수 있었으나, 실제 연구를 위한 사용 가능한 도구로서 활용되기 위해서는 보다 심도 있는 모델 구성 및 파라미터 튜닝, 그리고 더 많은 양의 훈련 데이터가 필요함 또한 확인할 수 있었다. 특별히, 비선형 회귀분석 방법론을 통한 1차원 분광학 데이터의 피크 피팅은 이미 그 기술적 성숙도가 상당

수준에 이른 상태이며, 이에 대해 정량적인 정확도의 우위를 점하기 위해서는 보다 정교한 모델 및 파라미터 선정, 그리고 더 많은 학습용 데이터가 필요하리라 생각된다. 실제로, 본 연구와 유사하지만 보다 복잡한 피크 피팅 및 물질 내 원소 구성비 예측을 목표로 한 이전 선행 연구의 경우 [20], 기계 학습 모델을 통해 예측된 물질 조성비가 실제 측정 결과와 비교하여 10% 이상의 큰 오차를 보이는 것을 확인할 수 있다. 물론 이러한 오차의 주된 원인으로서는 모델 훈련을 위한 가상의 분광학 데이터 생성의 어려움을 들 수 있을 것이나, 기계학습을 통해 어느 정도의 오차가 허용이 되는 분류 문제가 아닌 정량적 수치를 예측하는 문제에 접근하는 것의 어려움을 엿볼 수 있는 한 가지 예라 생각된다. 다만, 최근에 발표되어 주목을 받고 있는 crystal graph CNN 모델의 예와 같이 [9], 학습 데이터를 기계학습에 적합한 형태로 변환할 수 있을 때 물질의 전체 에너지와 같은 정량적인 성질 또한 비교적 높은 정밀도로 예측 가능한 사례도 존재하며, 이를 통해 신경망 모델의 학습 뿐 아니라 적절한 데이터의 표현형을 찾는 것 또한 중요한 문제인 것을 알 수 있다.

마지막으로, 본 연구와 같은 기계학습 방법론은, 1차원 피팅 문제처럼 사람이 접근하기 비교적 쉬운 문제에는 그 이득이 크지 않을 수도 있으나, 분광학 데이터 도메인의 차원이 2차원 이상으로 확장되는 (진동수 + 운동량) 각분해 광전자 분광학 (angle-resolved PES, ARPES) 데이터를 분석하는데는 유용하게 사용될 수 있음을 강조하고자 한다. 본 연구의 후속 연구로서, 현재의 모델의 정확도 및 성능을 개선하는 한 편 2차원 이상의 물질에서의 ARPES 데이터에서 전자 스펙트럼 및 전자간 상호작용에 의한 자기에너지(self-energy)를 추출하는 연구를 진행할 계획이며, 장기적으로는 이를 분광학 데이터베이스와 접목하여 [21][5] 전자간 상호작용이 흥미로운 물성을 유발하는 신물질 탐색 및 디자인에 유용하게 사용할 수 있으리라 기대하고 있다.

[5] 예를 들어 NOMAD (https://nomad-lab.eu/) 데이터베이스 등.

감사의 글

본 논문은 강원대학교 신임교수연구비 지원 프로그램 및 한국연구재단 신진연구자지원사업 (NRF-2020R1C1C1005900)의 지원을 통해 작성되었음을 밝히며, 또한 여러 가지 유용한 조언을 제공해 주신 이주용, 이훈표 교수님, 윤홍기 박사님께 감사를 드립니다.

REFERENCES

[1] A. Karpatne *et al.*, IEEE Trans. Knowl. Data Eng. **29**, 2318 (2017).
[2] A. Krizhevsky, I. Sutskever and G. E. Hinton, NIPS'12: Proceedings of the 25th International Conference on Neural Information Processing Systems **1**, 1097 (2012).
[3] Y. LeCun, Y. Bengio and G. Hinton, Nature **521**, 436 (2015).
[4] D. Silver *et al.*, Nature **529**, 484 (2016).
[5] A. Esteva *et al.*, Nature **542**, 115 (2017).
[6] G. Carleo *et al.*, Rev. Mod. Phys. **91**, 045002 (2019).
[7] J. Carrasquilla, R. G. Melko, Nat. Phys. **13**, 431 (2017).
[8] G. Carleo, M. Troyer, Science **355**, 602 (2017).
[9] T. Xie, J. C. Grossman, Phys. Rev. Lett. **120**, 145301 (2018).
[10] K. T. Butler *et al.*, Nature **559**, 547 (2018).
[11] A. Zunger, Nat. Rev. Chem. **2**, 0121 (2018).
[12] B. Sanchez-Lengeling, A. Aspuru-Guzik, Science **361**, 360 (2018).
[13] H. Kim, H. Lee and H. Lee, Jpn. J. Appl. Phys. **57**, 071601 (2018).
[14] https://keras.io *Keras* (2015).
[15] TensorFlow: Large-scale machine learning on heterogeneous systems, https://www.tensorow.org/
[16] T. Ida, M. Ando and H. Toraya, J. Appl. Cryst. **33**, 1311 (2000).
[17] M. N. Schmidt, T. S. Alstrm, M. Svendstorp and J. Larsen, Proceedings of 2019 IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2019 **2757**, (2019).
[18] D. P. Kingma, J. Ba, arXiv preprint arXiv:1412.6980, (2014).

[19] C. Donley *et al.*, Langmuir **18**, 450 (2002).
[20] G. Drera, C. M. Kropf and L. Sangaletti, Mach. Learn.: Sci. Technol. **1**, 015008 (2020).
[21] R. P. Xian *et al.*, rXiv:1909.07714, (2019).