트랜스포머 인코더를 활용한 탠덤 질량 스펙트럼의 노이즈 피크 판별

심승현 [1], 정희원 [1], 이상정 [1], 박희진 [1]
[1]한양대학교 소프트웨어대학 컴퓨터소프트웨어학부
lmazine97@hanyang.ac.kr, haley980406@naver.com,
othertics@hanyang.ac.kr, hjpark@hanyang.ac.kr

Transformer encoder-based noise detection of Tandem MS spectrum

Seunghyun Shim, Heewon Jung, Sangjeong Lee, Heejin Park
[1]Department of Computer Science, Hanyang University

요 약

단백체학에서 펩타이드를 분석하기 위해 탠덤 질량 분석을 활용한다. 탠덤 질량 분석의 결과로 얻어지는 스펙트럼에는 기계의 오차 등으로 발생하는 노이즈 피크가 존재한다. 노이즈 피크들은 스펙트럼을 펩타이드로 식별하는 데에 어려움을 야기하므로, 본 논문에서는 딥러닝 기법을 활용하여 스펙트럼 내의 피크들이 각각 노이즈 피크인지를 판별하고자 한다. 데이터 셋은 NIST (National Institute of Standards and Technology)에서 제공하는 Consensus Human HCD Libraries를 활용하였다. 학습 모델은 트랜스포머 모델의 인코더를 쌓아 올린 형태로 구성하였다. 임베딩된 데이터 셋의 질량 대 전하 비와 밀집도 값을 서로 더하여 모델의 입력 값으로 사용하였으며, 모델의 출력을 확률적으로 해석하여 노이즈 피크를 판별한다. 피크 단위에서 최대 97%의 정확도를 보였고, 스펙트럼 단위에서는 약 20%의 정확도를 보였다. 모델의 크기가 증가할수록 스펙트럼 단위에서의 정확도가 크게 증가하는 경향성을 보였다. 추후 다양한 실험을 통해 실제 분석에 범용적으로 활용될 수 있는 보다 정밀한 노이즈 피크 판별 모델을 구축하기를 기대한다.

1. 서 론

탠덤 질량 분석(Tandem mass spectrometry)은 단백체학(Proteomics)에서 펩타이드를 분석하기 위해 일반적으로 활용되는 기법이다 [1, 2]. 탠덤 질량 분석의 결과로 얻어진 스펙트럼은 데이터베이스 검색, 드 노보 시퀀싱 등의 방법을 통해 펩타이드로 식별된다 [3, 4]

스펙트럼은 펩타이드 조각(fragment)들을 의미하는 피크(peak)들로 구성되어 있으며, 각 피크들은 질량 대 전하 비(m/z)와 밀집도(intensity)값을 가진다.

스펙트럼에 존재하는 피크들 중, 기계의 오차 혹은 고립화 등에서 발생한 오류들로부터 얻어진 피크들이 존재한다 [5]. 이러한 피크들을 노이즈 피크(Noise peak)라고 한다. 노이즈 피크들은 스펙트럼을 통한 펩타이드 식별을 어렵게 한다 [6].

따라서 본 논문에서는 이러한 노이즈 피크를 탐지하고, 이를 제거함으로써 보다 정밀한 스펙트럼 분석을 가능케 하고자 한다. 지금까지 주로 알고리즘적인 접근을 통한 노이즈 피크 탐지 기법이 제안되어 왔으며 [7-12], 본 논문에서는 딥러닝 기법 중 하나인 트랜스포머 모델 [13]을 활용한 노이즈 피크의 탐지 모델을 제안한다.

2. 방 법

2.1 데이터 셋

본 논문에서는 NIST (National Institute of Standards and Technology)에서 제공하는 Consensus Human HCD Libraries를 활용하여 학습을 진행하였다 [14]. 약 91만 개의 스펙트럼으로 구성된 데이터 셋은 피크의 질량 대 전하 비와 밀집도 값, 그리고 해당 피크가 펩타이드의 어떤 이온에 해당하는지에 관한 정보를 제공한다. 본 논문에서 진행한 실험에서는 이온 정보가 존재하는 피크들과 그렇지 않은 피크들을 각각 라벨링하여 학습을 진행하였다.

2.2 데이터 전처리

딥 러닝에서, 학습 데이터에 긴 꼬리 현상(Long tail effect) 이 존재할 경우 모델의 학습 성능을 크게 위협할 수 있다 [15]. 데이터 셋에서 긴 꼬리 현상을 어느 정도 완화하기 위해, 학습 데이터 셋에서 질량 대 전하 비 값과 스펙트럼이 가지는 피크의 개수를 기준으로 양 극단에서 일부 데이터(전체 데이터의 약 4%)를 학습 대상에서 제외하였다.

스펙트럼에서 일반적으로 적용되는 허용 오차를 고려하

여, 질량 대 전하 비의 값을 소수점 둘째 자리까지 사용하였다. 밀집도 값은 스펙트럼 내에서 상대적인 값으로 변환하여 사용하였다. 임베딩된 데이터 셋의 질량 대 전하 비와 밀집도 값을 서로 더하여 모델의 입력값으로 사용하였다.

해당 데이터 셋을 8:1:1의 비율로 나누어 각각 학습(Train), 검증(Validation), 테스트(Test) 데이터의 역할을 수행하게 하였다.

2.3 학 습

본 논문에서는 트랜스포머 모델을 기반으로 모델을 구축하였다. 트랜스포머는 어텐션(Attention) 기법을 활용하여 입력 데이터 간의 연관도를 측정하며, 위치 인코딩(Positio-nal encoding)을 도입하여 입력 데이터 간의 상대적인 위치를 고려하였다. 본 논문에서는 질량 스펙트럼에서 각 피크들의 연관도를 측정하여 노이즈 피크를 판별한다.

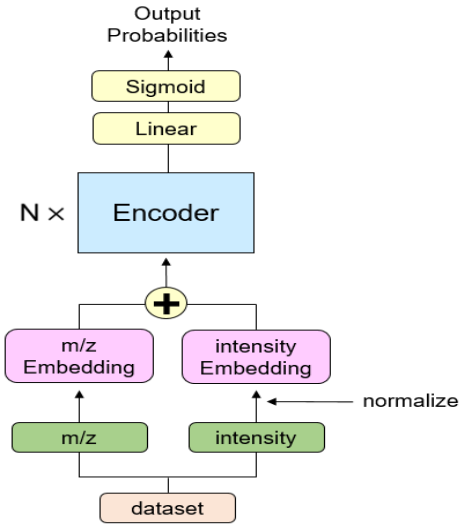

그림 1. 학습 모델 구조도

학습 모델은 그림 1 과 같이, 트랜스포머 인코더를 쌓아 올린 형태로 구성하였다. 스펙트럼 데이터에서, 각 피크들은 b이온, y이온 등 이온의 종류에 따라 연관성을 가진다. 단편화된 이온들의 질량 대 전하 비의 관계는 양방향적(Bidirectional)이다 [16]. 트랜스포머 인코더는 양방향적인 문맥을 반영하여 연관도를 측정하는 특징이 있으나, 트랜스포머 디코더의 경우 좌측 문맥만을 활용하여 연관도를 측정하므로 스펙트럼 데이터의 양방향적인 관계를 반영하지 못한다. 이에 따라 트랜스포머 디코더를 생략한 학습 모델을 구성하였다.

입력 값은 여러 개의 인코더 레이어와 밀집 층(Dense layer)을 통과한 후 시그모이드(Sigmoid) 함수를 통해 확률적으로 해석된다. 이를 토대로 해당 피크가 노이즈 피크인지 아닌지를 판별한다. 이는 이진 분류의 과정과 동일하며, 모델의 출력으로 얻어진 확률에 문턱값(Threshold)을 정하여 판별의 기준으로 삼았다.

3. 결 과

학습은 하이퍼 파라미터 값을 서로 다르게 하여 세 가지의 실험을 진행하였다. 하이퍼 파라미터(hyperparameter) 값은 표 1 과 같다.

Hyperparameters	실험 1	실험 2	실험 3
d_model	128		256
dff	128		256
Number of layers	4	6	4
Number of heads	4		
Batch size	128		
Dropout rate	0.1		
Loss	BCE Loss		

표 1. 하이퍼 파라미터

학습은 세 실험 모두 100에폭(epoch)까지 진행하였으며, 각 학습의 진행 양상은 그림 2, 3과 같다.

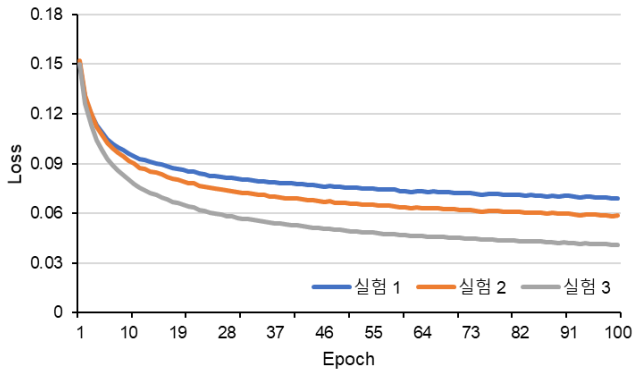

그림 2. 손실 그래프

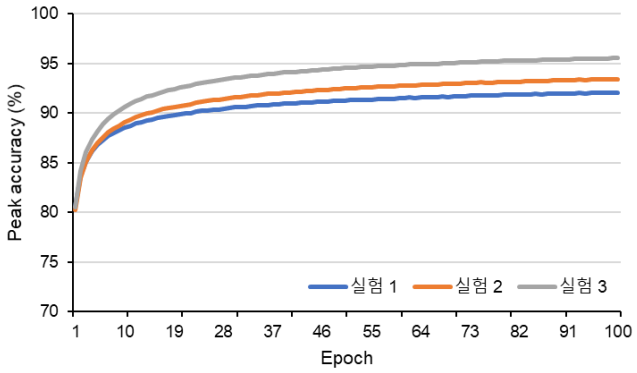

그림 3. 정확도 그래프

각 학습의 테스트 데이터 셋에서의 정밀도(Precision), 재현율(Recall), 정확도(Accuracy)는 표 2 와 같다. 표 2 에서, 스펙트럼 정확도(Spectrum Accuracy) 란 전체 스펙트럼 중 스펙트럼 내에 존재하는 모든 피크들의 노이즈 피크 여부를 정확하게 판단한 스펙트럼의 비율을 의미한다.

	실험 1	실험 2	실험 3
Precision	92.49%	94.17%	96.91%
Recall	94.99%	95.91%	97.26%
Accuracy	93.46%	94.85%	97.00%
Spectrum Accuracy	5.78%	9.92%	20.66%

표 2. 테스트 데이터 셋에서의 정밀도, 재현율 및 정확도

하이퍼 파라미터를 서로 다르게 하여 진행된 실험 1, 2, 그리고 3에서 모델은 피크 단위에서 각각 93.46%, 94.85%, 97.00%의 정확도를 보였다.

표 2와 같이, 전반적으로 모델의 크기에 비례한 성능 향상을 보였다. 특히, 스펙트럼 정확도는 모델의 크기가 증가할수록 가파르게 상승하였다.

4. 결론 및 향후 연구

본 논문에서는 탠덤 질량 스펙트럼의 노이즈 피크 판별에 트랜스포머 인코더를 활용하여 학습을 진행하였다. 가장 높은 정확도를 보인 실험 3에서, 피크 단위에서는 약 97%의 정확도를 보였고, 스펙트럼 단위에서는 약 20%의 정확도를 보였다. 전반적으로 모델의 크기가 증가할수록 정확도가 증가하는 모습을 보였고, 특히 스펙트럼 단위에서는 매우 가파른 상승폭을 보였다. 추후 하이퍼 파라미터 값의 변경, 모델 크기의 최적화, 그리고 다양한 데이터 셋을 학습에 이용함으로써 스펙트럼 단위에서의 높은 정확도 향상과 실제 분석에 범용적으로 사용할 수 있는 노이즈 피크 판별 모델을 구축할 수 있을 것이라 기대한다.

5. 참고 문헌

[1] Aebersold, Ruedi, and Matthias Mann. "Mass spectrometry-based proteomics." Nature 422.6928 (2003): 198-207.

[2] Steen, Hanno, and Matthias Mann. "The ABC's (and XYZ's) of peptide sequencing." Nature reviews Molecular cell biology 5.9 (2004): 699-711.

[3] MacCoss, Michael J., Christine C. Wu, and John R. Yates. "Probability-based validation of protein identifications using a modified SEQUEST algorithm." Analytical chemistry 74.21 (2002): 5593-5599.

[4] Kim, Sangtae, et al. "Spectral dictionaries: Integrating de novo peptide sequencing with database search of tandem mass spectra." Molecular & Cellular Proteomics 8.1 (2009): 53-69.

[5] Reiz, Beáta, et al. "Precursor mass dependent filtering of mass spectra for proteomics analysis." Protein and Peptide Letters 21.8 (2014): 858-863.

[6] McDonnell, Kevin, Enda Howley, and Florence Abram. "The impact of noise and missing fragmentation cleavages on de novo peptide identification algorithms." Computational and Structural Biotechnology Journal 20 (2022): 1402-1412.

[7] Zhang, Shu-Qin, et al. "Peak detection with chemical noise removal using Short-Time FFT for a kind of MALDI Data." Optimization and Systems Biology (2007).

[8] Xu, Hua, and Michael A. Freitas. "A dynamic noise level algorithm for spectral screening of peptide MS/MS spectra." BMC bioinformatics 11.1 (2010): 1-8.

[9] Zhang, Shuqin, et al. "A novel peak detection approach with chemical noise removal using short-time FFT for prOTOF MS data." Proteomics 9.15 (2009): 3833-3842.

[10] Zhurov, Konstantin O., et al. "Distinguishing analyte from noise components in mass spectra of complex samples: where to cut the noise?" Analytical chemistry 86.7 (2014): 3308-3316.

[11] Kast, Jürgen, et al. "Noise filtering techniques for electrospray quadrupole time of flight mass spectra." Journal of the American Society for Mass Spectrometry 14.7 (2003): 766-776.

[12] Gallia, Jason, et al. "Filtering of MS/MS data for peptide identification." BMC genomics 14.7 (2013): 1-9.

[13] Vaswani, Ashish, et al. "Attention is all you need." Advances in neural information processing systems 30 (2017).

[14] Sergey L. Sheetlin, et al. “Filtering and optimization of peptide tandem mass spectral libraries”, American Society for Mass Spectrometry (2020).

[15] Samy Bengio. “The battle against the long tail”. In Talk on Workshop on Big Data and Statistical Machine Learning. (2015)

[16] Devlin, Jacob, et al. "Bert: Pre-training of deep bidirectional transformers for language understanding." arXiv preprint arXiv:1810.04805 (2018).