|  |
| --- |
| **전이학습을 이용한 딥러닝 기반의 드노보 펩타이드 서열 분석** |
| 김태훈1, 서성빈O1, 이경준1, 이상정1, 김현우2, 박희진1  1한양대학교 소프트웨어대학 컴퓨터소프트웨어학부  2한국과학기술정보연구원 슈퍼컴퓨팅응용센터  dtc03012@gmail.com, zzp66u@gmail.com, lgj961001@gmail.com, othertics@hanyang.ac.kr, pardess@kisti.re.kr, hjpark@hanyang.ac.kr |
| Deep learning-based de novo peptide sequencing using transfer learning |
| Taehun Kim1, Seongbeen SeoO1, Gyeongjun Lee1, Sangjeong Lee1, Hyunwoo Kim2, Heejin Park1  1Department of Computer Science, Hanyang University  2Center for Supercomputing Applications, Korea Institute of Science and Technology Information |
| **요 약**  단백체학에서는 펩타이드 서열 동정을 위해 펩타이드의 질량 스펙트럼을 분석한다. 본 논문에서는 데이터베이스 없이 질량 스펙트럼을 분석하는 드노보 펩타이드 서열분석에 기계번역 딥러닝 모델인 트랜스포머를 적용했다. 이론 질량 스펙트럼을 이용한 선행 학습 모델을 바탕으로 실제 질량 스펙트럼의 특징 벡터를 추출하여 새로운 모델의 입력 값으로 사용하는 전이 학습을 진행했다. 선행 학습은 이론 질량 스펙트럼 중 m/z 값 만을 입력 값으로 사용했으며, 실제 질량 스펙트럼을 학습할 때는 (m/z, intensity) 쌍의 데이터를 사용해 선행 학습 모델을 거쳐서 나온 m/z의 특징 벡터에 intensity를 임베딩하여 더하는 방법을 사용했다. 아미노산 단계에서 75%의 정확도와 펩타이드 단계에서 26%의 정확도를 보여주어 높은 정확도는 아니지만 트랜스포머 모델을 이용한 펩타이드 서열 분석의 가능성을 볼 수 있었다. 정확도를 높이기 위해 데이터를 추가하고 실제 데이터에 intensity의 정보를 추가하는 여러 방법을 찾아 최적의 하이퍼 파라미터를 구하는 것이 앞으로 연구할 과제이다. | | |

**1. 서 론**

단백체학(Proteomics)에서 드노보 펩타이드 서열 분석(De novo peptide sequencing)[1]은 단백질 특성 분석(Protein characterization)을 위한 핵심 기술이다. 펩타이드 서열 분석에서는 질량 분석(MS, Mass Spectrometry)을 통해 얻어낸 펩타이드의 질량 스펙트럼(Mass spectrum)을 분석하여 펩타이드를 동정한다. 질량 분석 과정에서 단백질은 펩타이드로 분해되고, 분해되어 생성된 펩타이드를 이온화 시킨 후 질량 분석기를 거치면 질량 스펙트럼을 얻을 수 있다. 이온화된 펩타이드에 물리적 충돌을 가하여 펩타이드를 한번 더 쪼갠 후 질량 스펙트럼을 얻는 것을 탠덤 질량분석(Tandem mass spectrometry, MS/MS)이라 하고[2], 탠덤 질량 분석을 통해 얻어낸 질량 스펙트럼을 이용하여 펩타이드 서열을 분석한다.

최근 딥러닝 모델의 발달로 데이터베이스 없이 질량 스펙트럼을 분석하는 드노보 펩타이드 서열 분석에도 다양한 형태의 딥러닝 모델이 적용되고 있다. 주어진 질량 스펙트럼을 대표할 수 있는 특징 벡터(Feature vector)를 CNN(Convolution Neural Network)계열 모델로 추출하고, RNN(Recurrent Neural Network)계열 모델에서 예측 값을 출력하는 방법과 같이 CNN계열과 RNN계열이 결합된 모델들이 소개되었고, DeepNovo[3]과 DeepNovoV2[4]와 같은 모델들이 있다.

드노보 펩타이드 서열 분석은 질량 스펙트럼에서 피크(peek)간의 상대적 위치를 보고 펩타이드 서열로 바꾸는 것이기 때문에 기계번역(Machine translation)에 대입하여 생각해 볼 수 있다. 2017년 구글에서 발표한 기계 번역 모델 트랜스포머(Transformer)[5]는 이전의 기계 번역 모델들을 압도하는 성능을 보여주었고, 현재 자연어 처리 분야에서 가장 광범위하게 사용되는 모델이다.

본 논문에서는 딥러닝을 기반으로 한 드노보 펩타이드 서열 분석에 트랜스포머 모델을 적용했다. 실제 질량 스펙트럼(이하 실제 데이터)의 노이즈를 정제하기 위해 노이즈가 정제된 이론 질량 스펙트럼(이하 이론 데이터)을 이용하여 선행 학습(Pre-training)을 먼저 진행했다. 그리고 선행 학습 모델을 통해 실제 데이터의 특징 벡터를 추출하여 최종적인 출력 값을 예측하는 방식의 전이 학습(Transfer learning)을 진행하였다.

**2. 방 법**

**2.1 선행 학습**

실제 데이터의 경우 노이즈를 포함하고 있기 때문에 노이즈를 정제시킬 방법이 필요하다. 실제 데이터만으로는 노이즈를 특정할 수 없기 때문에 노이즈가 정제된 이론 데이터로 선행 학습을 진행하여 선행 학습 모델을 만들었다. 선행 학습 모델에는 노이즈가 아닌 피크들의 정보들을 담고 있기 때문에 이 모델을 통해 실제 데이터의 특징 벡터를 추출하면 노이즈를 정제할 수 있다.

선행 학습에서 이론 데이터를 학습할 때 그림 1과 같은 트랜스포머 모델을 사용했다. 트랜스포머에서 인코더(Encoder)의 입력 값은 문장이기 때문에 질량 스펙트럼 한 개에 해당한다고 볼 수 있으므로 m/z 값들을 그대로 입력 값으로 사용하면 된다.

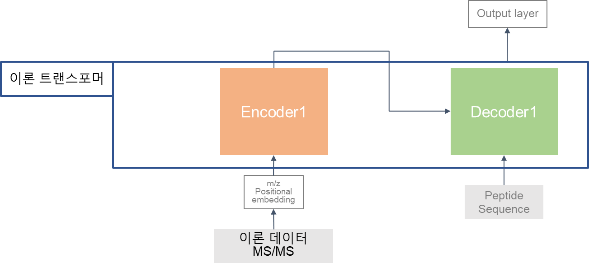
****

그림 1 이론 데이터 선행 학습 모델

**2.2 전이 학습**

전이 학습의 경우 그림 2와 같이 이론 데이터를 학습한 선행 학습 모델 위에 새로운 트랜스포머 모델을 쌓은 모델을 사용했다. 선행 학습 모델의 Encoder1과 Decoder1의 출력 값을 특징 벡터로 사용하여 새로운 모델의 입력 값으로 사용한다. 실제 데이터의 경우 질량 스펙트럼에서 intensity가 중요한 의미를 가지므로 intensity 정보를 추가해 줄 필요가 있다. 선행 학습 모델을 통과한 실제 데이터의 특징 벡터에 intensity를 임베딩한 벡터를 더해주는 방식을 사용했다.

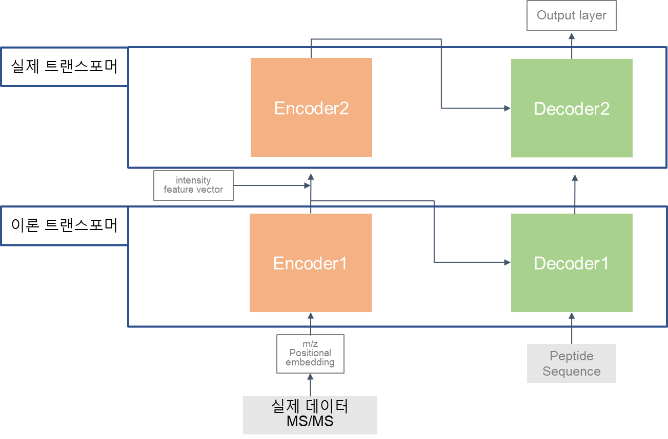


그림 2 실제 데이터 전이 학습 모델

**2.3 데이터 셋**

**2.3.1 이론 데이터**

이론 데이터의 경우 데이터베이스 탐색 방법에서 펩타이드 서열을 분석할 때 사용하는 방법[6]을 이용하여 데이터를 생성한다. 따라서 노이즈 없이 실제 펩타이드에 해당하는 피크의 정보만 존재한다.

**2.3.2 실제 데이터**

실제 데이터의 경우 proteomeTools에서 제공하는 합성 펩타이드 데이터셋[7]을 이용한다. 특정 펩타이드를 실제로 합성하여 탠덤 질량 분석을 통해 질량 스펙트럼을 얻어낸다. 실제 실험을 통해 얻어낸 데이터이기 때문에 펩타이드를 나타내는 피크 외에 노이즈와 허용 오차(tolerance)들이 존재한다.

**3. 결 과**

선행 학습 트랜스포머 모델의 인코더와 디코더 수는 2개로 진행하였고, 전이 학습 트랜스포머 모델의 인코더와 디코더 수는 4개로 진행하였다. 선행 학습에는 740만 개의 이론 데이터를 이용했고, 전체 전이 학습에는 150만 개의 실제 데이터를 사용하였다. 그림 3은 이론 데이터를 이용한 선행 학습 모델의 정확도이다. 15 에폭(epoch)부터 정확도가 99.9%에 수렴했기 때문에 더 이상 학습을 진행하지 않았다. 그림 4는 실제 데이터를 이용한 전이 학습 모델의 정확도이다. 각 펩타이드에서 아미노산 단계에서의 정확도는 약 75% 정도이고 펩타이드 단계에서의 정확도는 약 26%이다.

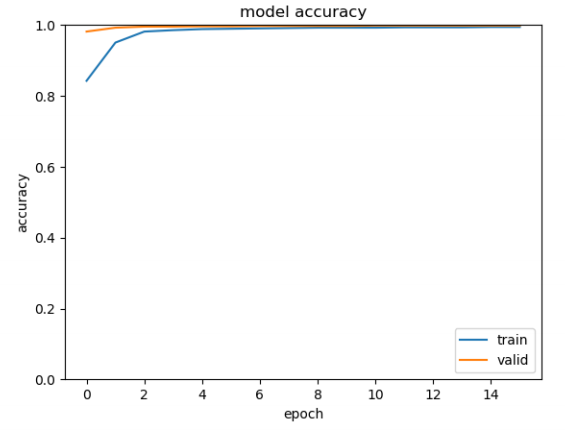


그림 3 선행 학습 모델의 정확도

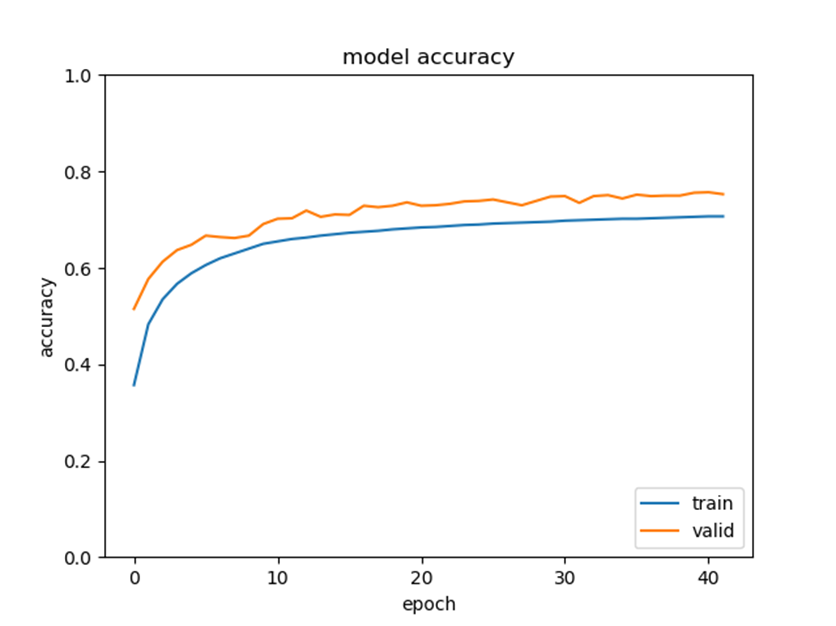


그림 4 전이 학습 모델의 정확도

**4. 결론 및 향후 연구**

드노보 펩타이드 서열 분석에 트랜스포머 모델을 적용하여 전이 학습을 진행하였다. 실험 결과에서는 40 에폭만에 아미노산 단계에서 75%의 정확도를 보이며 펩타이드 서열 분석에 가능성을 볼 수 있었다. 정확도를 높이기 위해 데이터를 추가하고 실제 데이터의 Intensity의 정보를 추가하는 여러 방법을 찾아 최적의 하이퍼 파라미터(Hyper parameter)를 구하는 것이 앞으로 연구할 과제이다.

**5. 참고 문헌**

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

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

[3] Tran, Ngoc Hieu, et al. "De novo peptide sequencing by deep learning." *Proceedings of the National Academy of Sciences* 114.31: 8247-8252, 2017

[4] Qiao, Rui, et al. "DeepNovoV2: Better de novo peptide sequencing with deep learning." *arXiv preprint arXiv:1904.08514*, 2019

[5] Vaswani, Ashish, et al. "Attention is all you need." *Advances in neural information processing systems*. 2017.

[6] Eng, Jimmy K., Tahmina A. Jahan, and Michael R. Hoopmann. "Comet: an open‐source MS/MS sequence database search tool." *Proteomics* 13.1: 22-24, 2013

[7] Zolg, Daniel P., et al. "Building ProteomeTools based on a complete synthetic human proteome." *Nature methods* 14.3:259-262, 2017