|  |
| --- |
| **[[1]](#footnote-1)트랜스포머 모델을 활용한 드노보 펩타이드 서열 분석** |
| 황득연O1 한승도1 이상정1 김현우2 박희진1  1한양대학교 소프트웨어대학 컴퓨터소프트웨어학부  2한국과학기술정보연구원 슈퍼컴퓨팅응용센터  [dyh0624@naver.com](mailto:dyh0624@naver.com), [1212sd@naver.com](mailto:1212sd@naver.com), [othertics@hanyang.ac.kr](mailto:othertics@hanyang.ac.kr), [pardess@kisti.re.kr](mailto:pardess@kisti.re.kr), [hjpark@hanyang.ac.kr](mailto:hjpark@hanyang.ac.kr) |
| De Novo Peptide Sequencing by Transformer |
| Deukyeon HwangO1 Seungdo Han1 Sangjeong Lee1 Hyunwoo Kim2 Heejin Park1  1Department of Computer Science, Hanyang University  2Center for Supercomputing Appliance, Korea Institute of Science and Technology Information |
| **요 약**  드노보를 활용한 펩타이드 시퀀싱은 펩타이드 서열 분석을 위해 자주 사용되는 방법 중 하나로, 스펙트럼의 m/z 및 intensity 등 고유한 특징을 이용하여 펩타이드를 분석하는 방법이다. 현재는 드노보 펩타이드 서열 분석을 CNN, RNN, LSTM등을 통해 분석하는 방법들이 여러 논문에서 소개되고 있다. 본 논문에서는 구글의 신경망 모델인 트랜스포머 모델을 펩타이드 서열 분석에 활용하였다. 트랜스포머 모델에 적용하기 위해서 총 약 400만개의 데이터들 중 스코어 순으로 선별하여 tfrecord 파일을 통해 학습 데이터셋과 검증 데이터셋을 나누어 저장하였고, 약 3주간 학습을 진행하였다. 실험 결과 아미노산 단계의 정확도는 약 71%이며 펩타이드 단계의 정확도는 약 15%를 보였다. 펩타이드 단계의 정확도가 높은 정확도는 아니었지만 학습을 진행하면서 많은 학습효과를 보였기 때문에 앞으로 더 다양하게 하이퍼 파라미터와 더 많은 데이터를 통해서 추가적인 학습을 진행한다면 더 높은 정확도를 보일 것이라고 판단한다. | | |

**1. 서 론**

단백질 그 자체의 서열 정보를 얻기 위해선 많은 어려움이 따르기 때문에 펩타이드 서열 정보를 얻는 것으로 대체 한다. 펩타이드 서열 정보를 얻기 위해 단백질을 펩타이드 단위로 소화(digestion)한 후 이온화를 진행한다. 그리고 질량 분석기를 통해 탠덤 질량 스펙트럼(Tandem mass spectrum, MS/MS)을 이용해 펩타이드 서열을 분석한다[1].

펩타이드 분석에는 크게 두가지 방법이 있는데 데이터베이스 탐색[2] 방법과 드노보(De novo) 방법[3]이다. 데이터베이스 탐색 방법은 이미 존재하는 펩타이드의 정보를 보다 빠르고 정확하게 분석할 수 있지만 모든 펩타이드의 정보를 가지고 있지 않다는 한계점이 존재한다. 그러므로 본 실험에서는 질량 스펙트럼을 분석하여 펩타이드를 분석하는 드노보 방법을 통해 실험을 진행하였다.

최근에는 딥러닝에 대한 많은 학술 연구가 진행이 되고, 꾸준히 발전하면서 다양한 분야에 딥러닝이 적용이 되고 있다. 드노보 펩타이드 시퀀싱(peptide sequencing) 에서도 마찬가지로 딥러닝을 활용한 실험이 꾸준히 등장하고 있고, 가장 대표적인 사례로는 CNN과 RNN을 결합하여 분석하는 방법인 딥노보(DeepNovo)를 활용한 펩타이드 서열 분석 방법이 있다[4].

본 논문에서는 딥러닝 모델 중 하나로 구글에서 발표한 신경망 모델인 트랜스포머를 활용하여 펩타이드 서열을 분석하였다[5].

**2. 본 론**

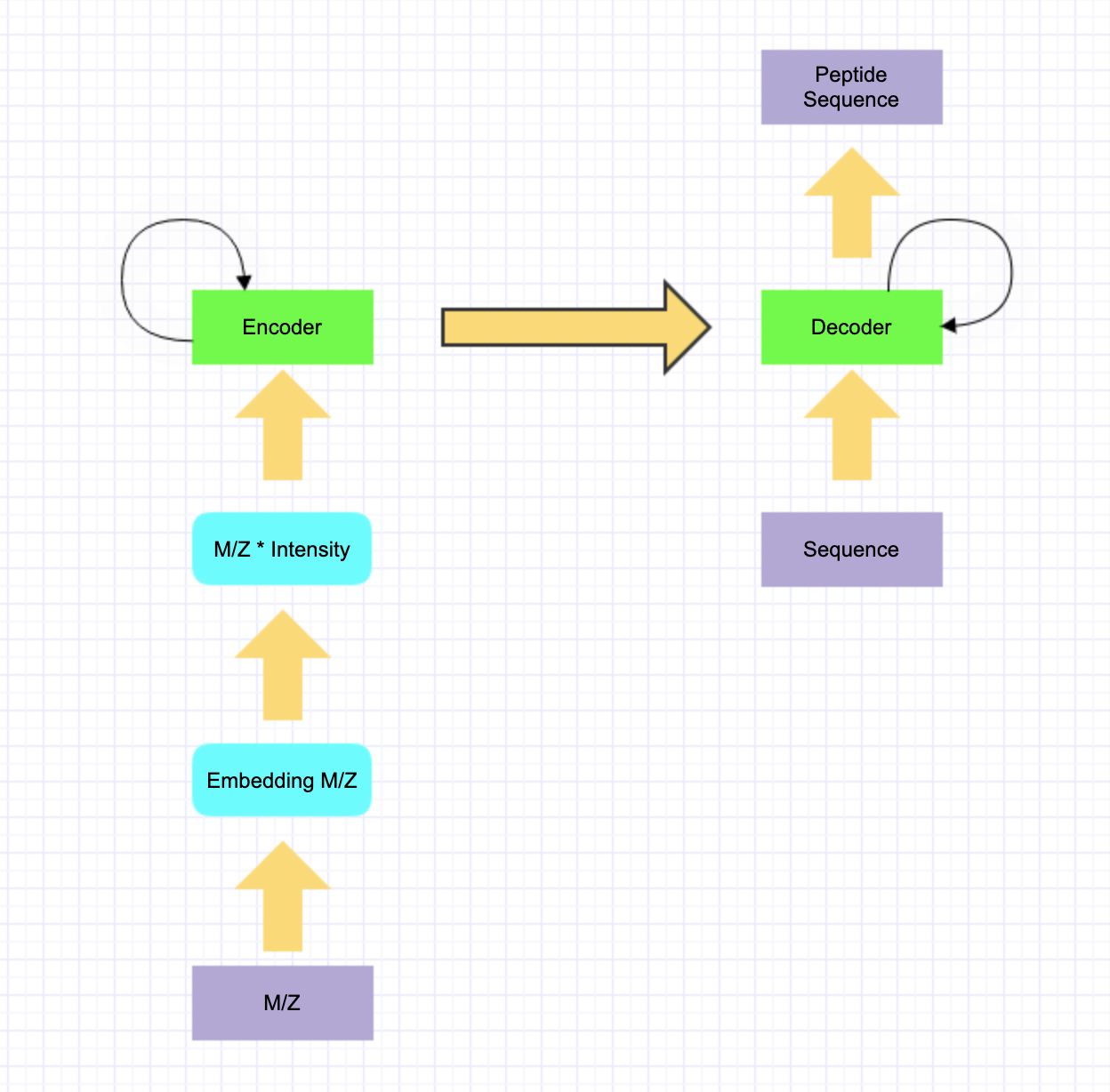
**2.1 트랜스포머**

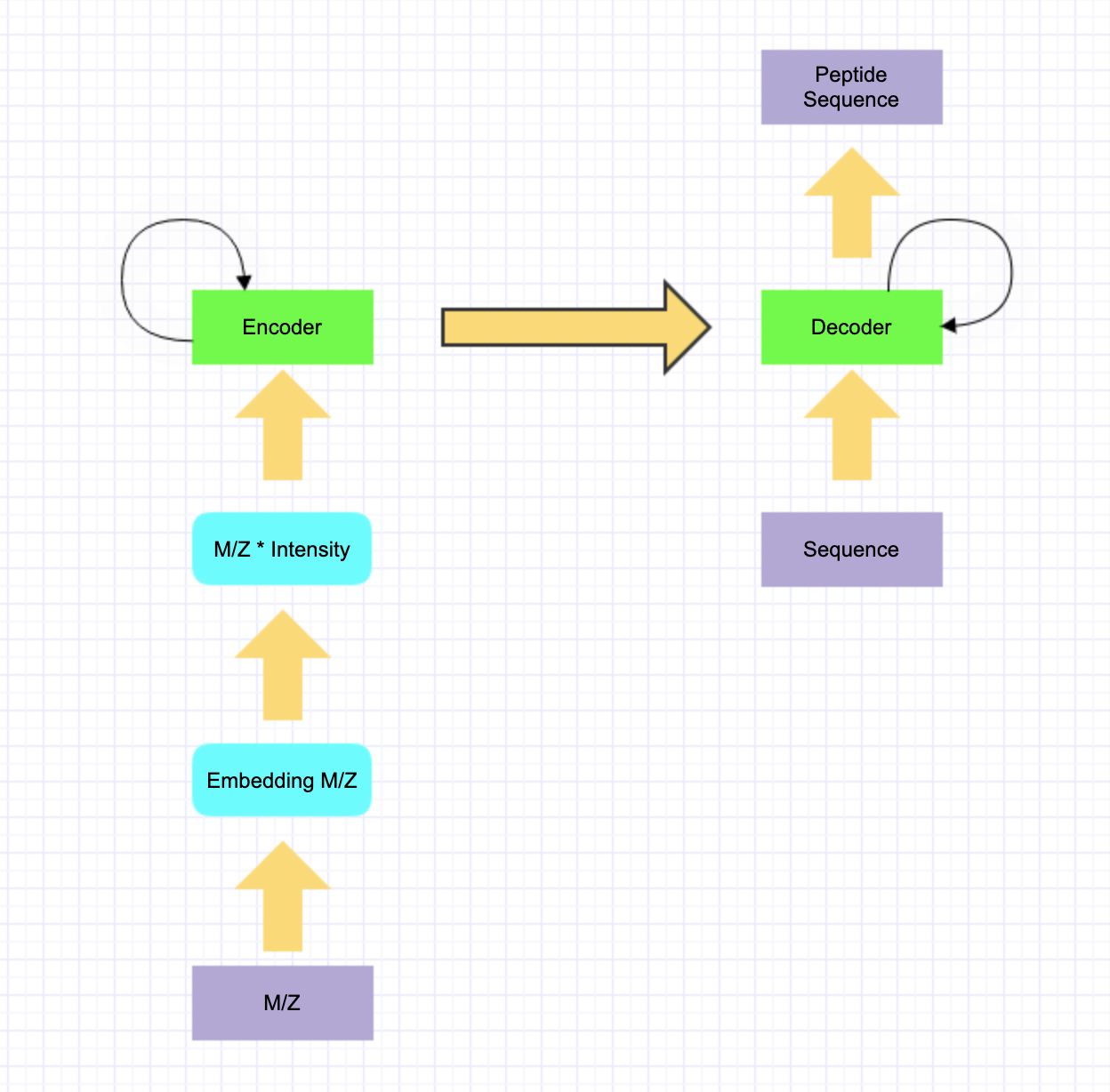
최근 가장 활발한 연구가 보이고 있는 딥러닝 분야는 기계 학습의 한 방법으로, 학습과정동안 인공 신경망으로서 예시 데이터에서 얻은 일반적인 규칙을 독립적으로 학습한다. 딥러닝은 현재 이미지분류 , 객체 탐지, 세그먼테이션(Segmentation), 자연어 처리, 알파고와 같은 강화학습, 딥페이크 등 수많은 분야에서 활용이 되고 있으며, 트랜스포머는 자연어 처리 분야 뿐만 아니라 여러 딥러닝 연구 분야에서 가장 적극적으로 활용이 되고 있는 모델 중 하나이다.

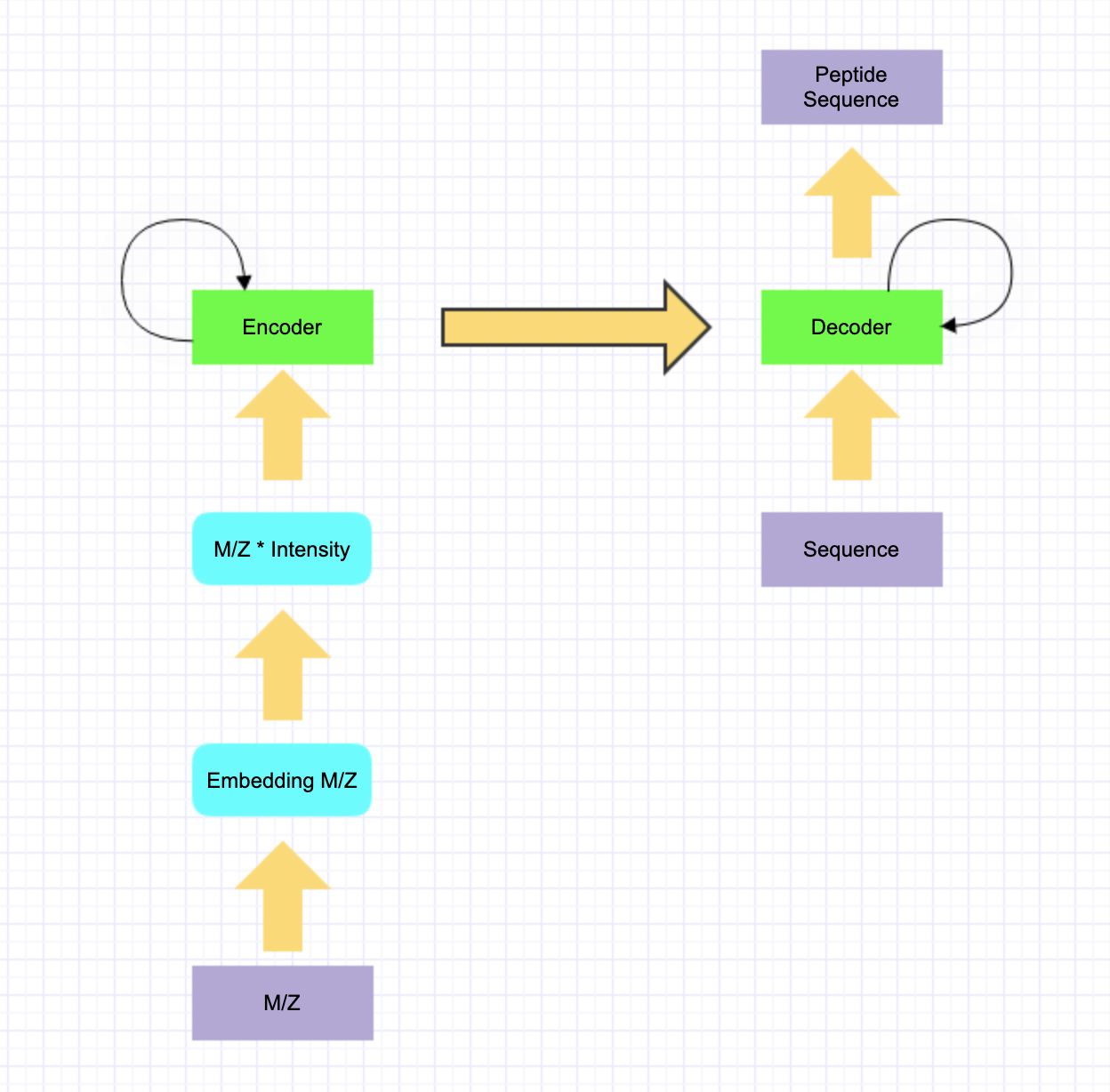
트랜스포머는 2017년 구글에서 ‘Attention is all you need’라는 논문을 통해 발표한 신경망 모델로 인코더-디코더 구조를 가지고 논문 제목처럼 어텐션으로 구현된 모델이다. 트랜스포머는 일반적으로 문장을 입력 받고 문장을 출력하기 때문에 펩타이드 서열 분석 실험을 진행하기 위해서 데이터를 트랜스포머 모델 내부로 입력 가능하도록 가공해 주어야 한다. 전체적인 모델의 구성은 그림 1과 같다.

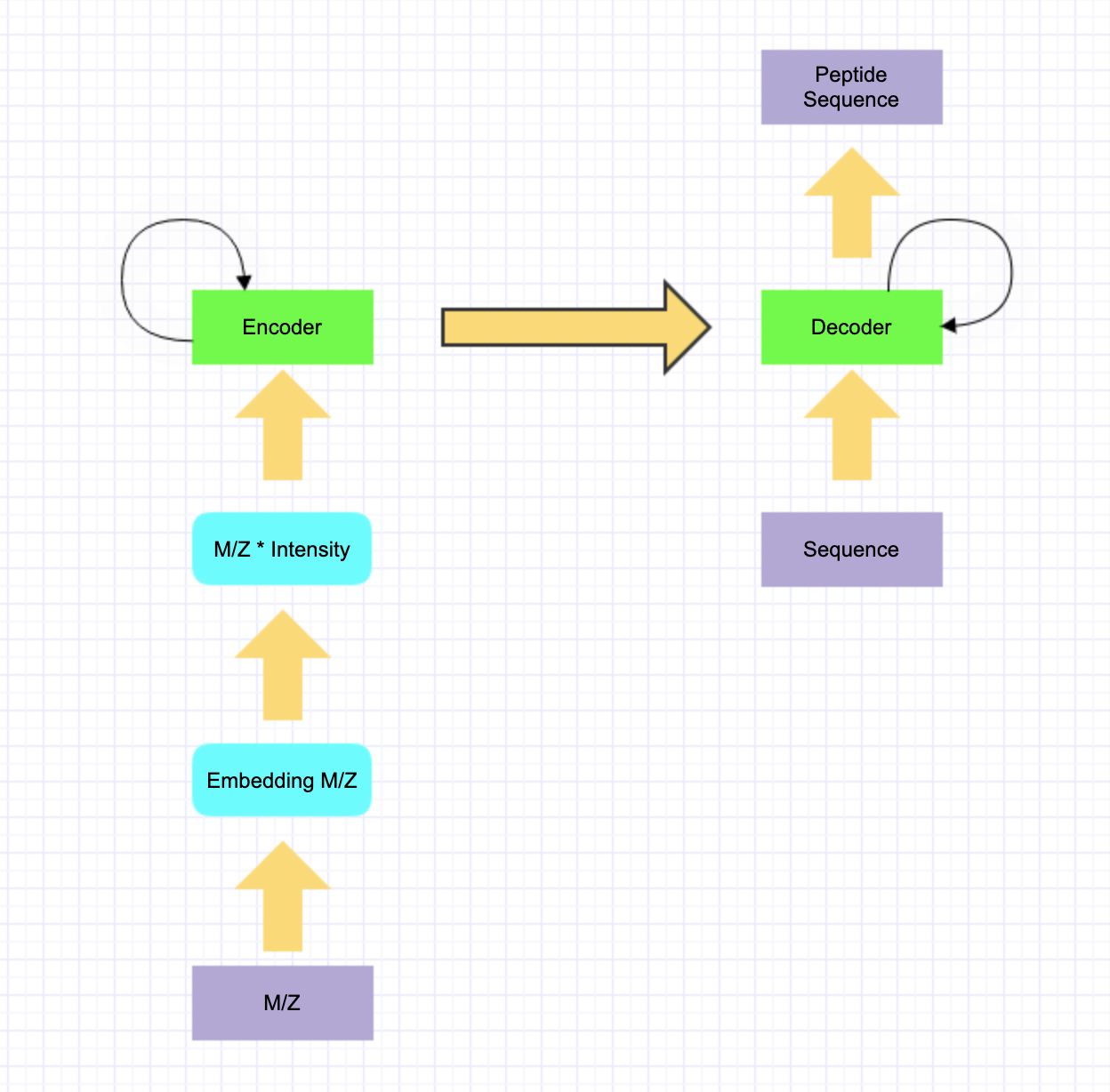
Transformer

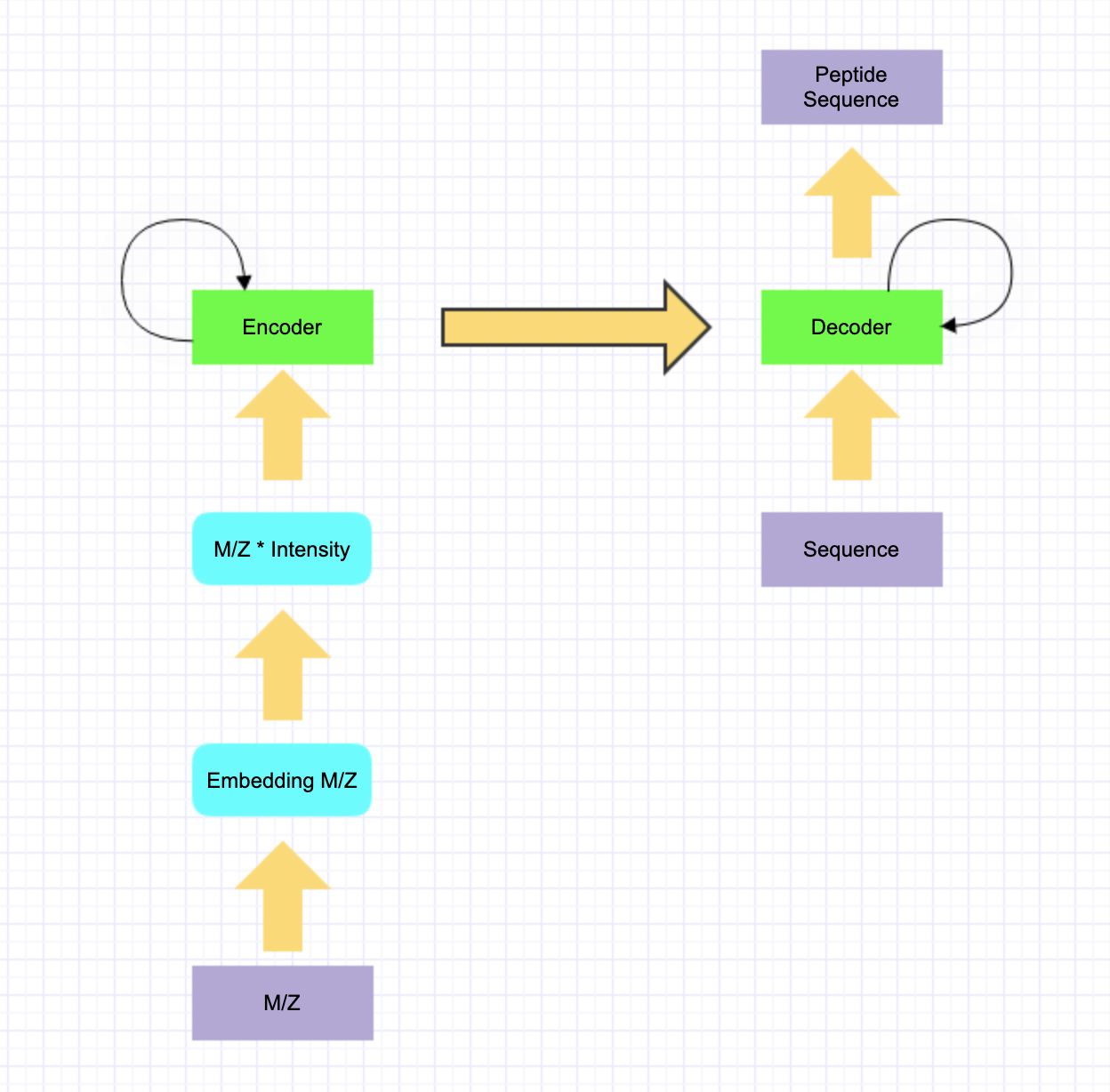
Peptide Sequence

Decoder ****

Encoder ****

Sequence****

M/Z \* Intensity ****

Embedding M/Z ****

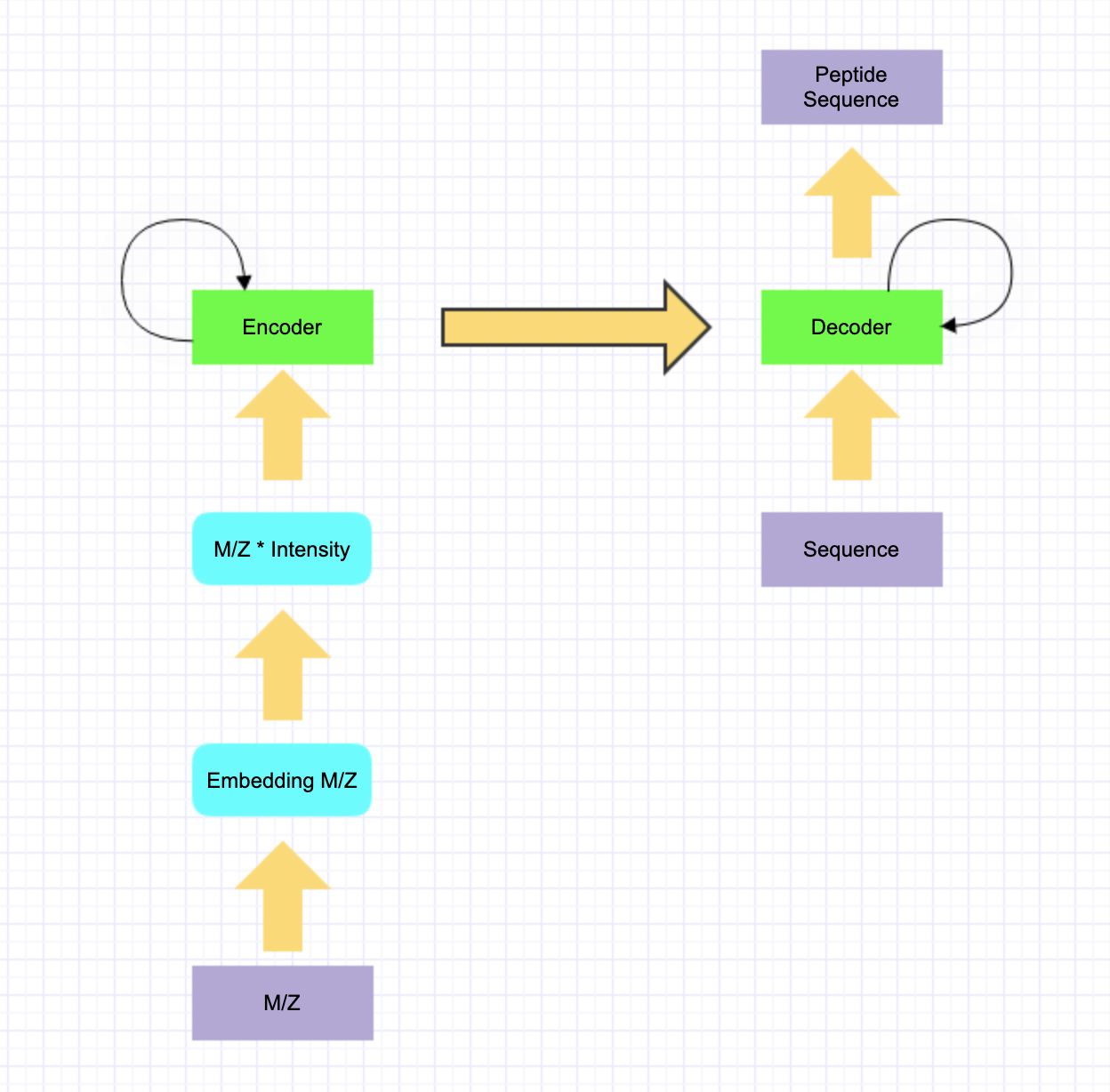
M/Z ****

그림1. 펩타이드 서열 분석을 위한 트랜스포머 모델.

위 실험에서는 먼저 데이터로 주어진 m/z와 Intensity를 가지고 트랜스포머 모델의 인풋이 될 탠덤 질량 스펙트럼을 구해주기 위해 m/z와 Intensity를 모두 트랜스포머 모델 내부로 가져와 스펙트럼 데이터를 가공하고 인코더의 인풋으로 입력해 주었다. 스펙트럼 데이터 가공 절차는 모델 내부로 가져온 m/z 데이터에 대해서 포지셔널 임베딩을 진행해주고, 임베딩이 이루어진 m/z 데이터와 Intensity 데이터를 곱해줌으로써 스펙트럼 데이터를 생성하였으며, 이를 인코더의 입력 값으로 사용을 하였다.

**2.2 데이터셋**

본 논문에서는 ProteomeTools에서 제공하는 합성 펩타이드(Synthetic Peptide) 데이터 셋을 이용하였다[6]. 약 400만개의 데이터 중 같은 펩타이드 서열 정보를 가진 데이터에서 스코어가 높은 3개를 뽑아 중복 없이 랜덤하게 학습 데이터셋(90%)과 검증 데이터셋 (10%) 으로 나눠 tfrecord 파일에 저장하였고, 검증 데이터셋 중에 약 7000개의 데이터를 뽑아 펩타이드 서열 분석을 위한 테스트 데이터셋으로 사용하였다. 데이터 중 변형(modification)이 이루어져 메티오닌(M)이 산화된 데이터는 같은 서열 정보라도 다른 서열 정보로 취급 하였고 새로운 아미노산 인덱스를 지정해주었다.

**3. 결 과**

모델 학습은 비교를 위해 하이퍼 파라미터 값만 약간 수정하여 2번을 진행하였다.

|  |  |  |
| --- | --- | --- |
|  | 실험1 | 실험2 |
| D\_model | 128 | 128 |
| dff | 1024 | 256 |
| Num of layers | 2 | 2 |
| Num of heads | 8 | 4 |
| Epoch | 160 | 160 |
| Train accuracy | 66.5% | 64% |
| Valid accuracy | 61% | 59% |
| Peptide accuracy | 16% | 12% |

표1. 트랜스포머 하이퍼 파라미터 및 정확도

학습은 두 실험 모두 구글 코랩 프로에서 제공하는 GPU와 코랩에 기본적으로 세팅이 되어 있는 파이썬과 Tensorflow, Keras를 사용하였고, 학습 데이터셋은 약 38만개의 데이터, 검증 데이터셋은 약 4만개의 데이터로 각각 구성이 되어 있으며 공통적으로 160EPOCH 까지 학습을 진행하였다.

실험 결과 표1에서 보이는 바와 같이 학습과 유효성 테스트에서 정확도가 실험1이 실험2에 비해 높은 학습 효과와 정확도를 보였기 때문에 실험1의 학습을 추가적으로 550 EPOCH까지 진행하였다. 그 결과, 그림2와 같이 학습의 정확도는 71%까지 상승하였고, 유효성 테스트에서는 64%의 정확도를 보였다.

이후 실험1에 의해 학습이 이루어진 모델을 이용하여 펩타이드 서열 분석에 적용시켜보았을 때, 펩타이드 서열의 정확도는 약 16%의 정확도를 보였다.

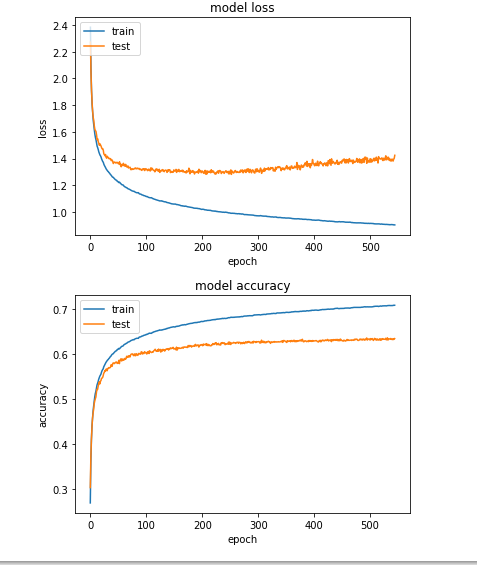


그림 2. 트랜스포머 모델의 정확도

**4. 결 론**

펩타이드 서열을 분석하기 위해 신경망 딥러닝 모델 중 하나인 트랜스포머 모델을 사용하였다. 인풋으로 입력할 MS/MS를 생성하기 위해서 m/z와 intensity를 모두 트랜스포머 내부로 가져와 임베딩된 m/z와 intensity를 곱해주었다. 실험결과 펩타이드 단계의 서열 분석에서는 높은 정확도는 아니었지만 더 많은 데이터, 여러 하이퍼 파라미터 적용, MS/MS 생성 방법 변경(m/z와 intensity간의 다양한 연산)등의 다양한 시도를 통해 추가적인 실험을 진행할 필요가 있다고 판단된다.

**5. 참고문헌**

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

[2] 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 : 5593-5599, 2002.

[3] 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.

[4] Tran, Ngoc Hieu, et al. “De novo peptide sequencing by deep learning.” Proceedings of the National Academy of Sciences:8247-8252, 2017.

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

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

1. "본 연구는 과학기술정보통신부 및 정보통신기획평가원의 SW중심대학지원사업의 연구결과로 수행되었음"(2016-0-00023) [↑](#footnote-ref-1)