1. 논문 정리

De novo 를 활용한 펩타이드 시퀀싱 (펩타이드 서열 분석)
(Database 탐색과 차이 : 새로운 데이터에 대응할 수 있는가 ?)
스펙트럼의 m/z , intensity 등 고유한 특징을 이용하여 펩타이드를 분석하는 방법
Transformer 모델을 활용하였음 .

서론에서 ,
단백질 서열 정보를 얻고자 함 => 펩타이드 단위로 나누어서 서열 분석
단백질 => 펩타이드 대체하는 과정
Digestion => ionization => 질량분석기, 텐덤 질량 스펙트럼 이용하여 펩타이드 서열 분석

본론에서 ,
Transformer => Attention 기법으로 구현된 모델
Transformer 의 input 은 문장이기 때문에, 데이터를 가공하여야 한다.
데이터로 주어진 m/z , intensity 를 위치 임베딩하여 intensity 와 곱해서 스펙트럼 데이터 생성 (Transformer 의 input)

---

2. 간략한 조사

1) De novo Sequencing : 정렬에 사용할 수 있는 참조 서열이 없는 새로운 시퀀싱
m/z : 펩타이드 이온의 질량을 전하량으로 나눈 값
Intensity : 펩타이드가 얼마나 많이 나타났는지

2) Transformer

Attention Mechanism 만을 사용 ,
멀티헤드 self-attention 을 이용해 Sequential computation 을 줄이고,
더 많은 단어들 간의 Dependency 를 모델링

- 인코더 (그림의 좌측)
Input : 단어 ID의 시퀀스로 표현된 문장의 배치 (배치 사이즈 , 최대 길이)
(Attention is all you need 논문에 따르면) 각 단어를 512차원으로 인코딩하여 출력한다 (배치 사이즈 , 최대 길이 , d=512)
그림의 구조를 N 개 반복한다

- 디코더 (그림의 우측)
Input : 단어 ID의 시퀀스로 표현된 Target 문장
인코더의 출력이 N개의 디코더에 모두 주입되고,
타임 스텝(단어의 위치) 마다 가능한 다음 단어에 대한 확률을 출력

출력 크기는 (배치 사이즈 , 최대 길이 , d = 타깃 언어 어휘 개수)

- 임베딩 층 2개, 스킵 연결 5 * N 개 , 피드포워드 모듈(정규화 층, 밀집 층 2개) 2 * N 개로 이루어져 있고, 마지막 출력 층은 밀집 층이다 (Softmax Activation Function 사용)

- 모든 층은 타임 스텝에 대하여 독립적이다 (time-distributed)
  각 단어는 다른 모든 단어에 대하여 독립적으로 처리된다

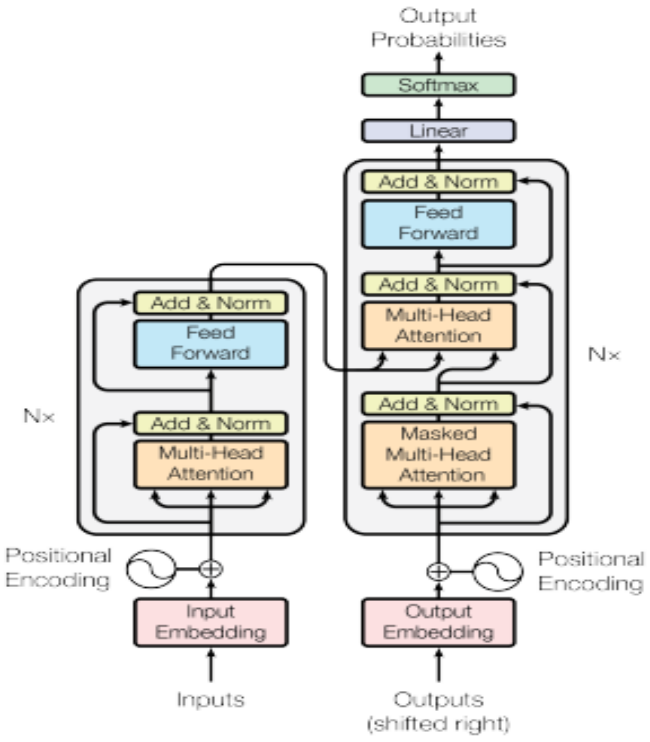
  인코더의 멀티 헤드 어텐션 층 => 각 단어와 동일한 문장에 있는 다른 단어의 관계를 인코딩한다 (self-attention)

  디코더의 Masked 멀티 헤드 어텐션 층도 동일한 작업을 수행하지만, 각 단어는 이전에 등장한 단어에만 attention 할 수 있다

  디코더의 멀티 헤드 어텐션 층 => 디코더가 입력 문장에 있는 단어에 대해 attention 한다

- 위치 인코딩 : 문장에 있는 단어의 위치를 나타내는 밀집 벡터
  n번째 위치 인코딩이 각 문장에 있는 n번째 단어의 단어 임베딩에 더해진다
  결과 : 모델이 각 단어의 위치를 알 수 있다

  +) 멀티 헤드 어텐션 층은 단어의 순서나 위치를 고려하지 않고, 다른 층들은 time-distributed 하기 때문에 각 단어의 위치를 알 수 없다.



3) 위치 인코딩 (positional encoding)

어텐션 층으로 진입하기 전에, input 으로 주어질 단어 vector 안에 단어의 위치 정보를 포함시킨다.

위치 정보를 벡터로 표현하고, 각 단어에 위치 임베딩을 추가하여 문장 내 위치 정보를 표시한다.

학습자료

- Transformer 관련

  Attention Is All You Need (2017) , https://arxiv.org/abs/1706.03762

- Attention 관련

  Neural Machine Translation (2014) , https://arxiv.org/abs/1409.0473

  Effective Approaches to Attention-based Neural Machine Translation (2015) , https://arxiv.org/abs/1508.04025

참고문헌

· 오젤리앙 제롱 지음, 박해선 옮김, 핸즈온 머신러닝, 한빛미디어, 2020

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