lecture 14 preview

1. RNN, CNN, and Self-Attention

   Sequence Modeling 에서 가변길이의 데이터를 고정 크기의 데이터로 표현하는 것이 필수적임.

   시퀀스의 길이가 일정해야 계산속도를 향상 시킬 수 있기 때문

   RNN 모델의 단점

   - 병렬화 불가능
   - 긴 구간에 걸친 의존도를 반영하지 못함

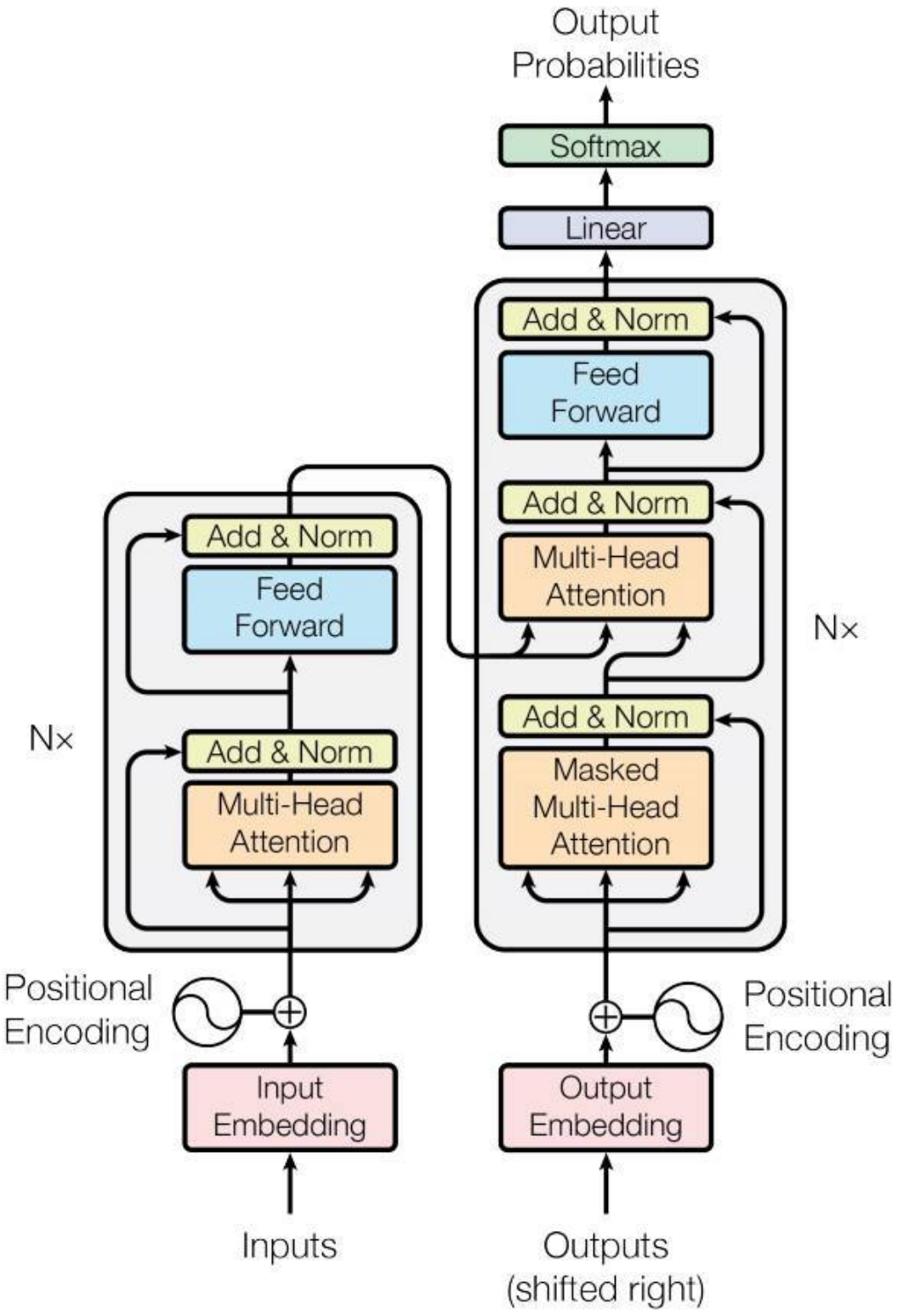
   CNN 모델의 단점

   - 병렬화 가능
   - 긴 구간의 의존도를 반영하기 위해 레이어가 많이 필요하다

   → Self-Attention은 병렬화가 가능하고, long-term dependency 문제를 해결할 수 있다

2. Transformer : self-attention 메커니즘을 극대화한 모델

- 연속적인 토큰을 연속적으로 처리하지 않아 병렬처리를 가능하게 만들었다
- 인코더, 디코더 블록으로 구성
- 각 블록에 6개씩 인코더와 디코더가 쌓여있다
- 인코딩 블록은 eos와 패딩을 제외한 토큰들을 self-attention 과 feed forward neural network에 전달하는 2단 구조를 가진다
- 디코딩 블록은 인코딩 블록의 아웃풋과 masked self-attention의 벡터를 입력으로 받아 self-attention을 진행하여 3단 구조를 가진다
- 디코딩 블록의 아웃풋은 sequential하게 생성되어 masked self-attention이라는 추가적인 작업이 필요하다



- input embedding : 단어를 벡터로 변환, 이를 첫 번째 인코더의 인풋으로만 사용하고 이후부터는 이전 단계의 아웃풋을 인풋으로 사용한다

- positional encoding : 단어의 위치정보를 저장한다. input embedding과 positonal encoding 벡터를 더한 값이 인코딩 블록의 인풋이 된다
- multihead attention, residual connection&normalization : 한 문장 내에서 존재하는 다양한 정보를 한 번의 attention만으로 반영하기는 어렵다. 하지만 하나의 어텐션이 한 가지 정보에만 집중하면 8 attention heads를 두어 정보를 attention score에 반영할 수 있다.
- position-wise feed-forward networks
- masked multi-head attention : 토큰들 사이의 인과관계를 부여
- final linear and softmax layer : FFNN의 일종, attention score를 probability로 전환

3. Image Transformer

   tasks

   - Unconditional Image Generation : 대규모의 데이터로 특정한 이미지를 제작
   - Class Conditional Image Generation : 클래스 각각의 임베딩 벡터를 입력으로 받아 이미지를 제작
   - Super Resolution : 저화질의 이미지를 입력으로 받아 고화질의 이미지를 출력

   - 이미지의 경우 dimension보다 sequence length가 커져 self-attention이 비효율적

   → Local Self-attention

4. Music Transformer

   tasks

   - 인풋 없이 다양한 데이터를 통해 음악을 생성
   - 음악의 앞 부분을 입력으로 받아 그 뒷부분을 생성

   - Relative Positional Self-Attention
     을 사용 - Relative Positional Vector를 더해 query와 key의 sequence 내 거리를 attention weight에 반영
   - query에 따라 positional attention score가 달라지도록 설정하여 positional encoding보다 풍부한 표현을 할 수 있다 → 마스킹한 보형이 삼각형이 되도록

reshape

- attention score에 relative positional attention score를 더하여 output