[Lec 14] Transformation and Self Attention for Generative Models

Transformer and Self-Attention

가변 길이의 data를 고정 크기의 vector 또는 matrix로 표현하는 것은 필수적임

RNN계열의 모델은 병렬화가 불가능, Long-term Dependency 반영 모함

Self Attention

병렬화가 가능함, 각 token이 최단거리로 연결되어 long-term dependency 해결

Self-Attention 과정

1. input을 각각 linear transform , query key value 생성
2. Query와 key pair의 dot product 계산
3. Scaling 적용
4. Softmax function 적용
5. Softmax output을 weight로 value vector들의 weighted sum 산출

→ 각 token을 sequence 내 모든 token과의 연관성을 기반으로 재표현 하는 과정으로 해석 가능

Multi-head Attention

한 문장 내에 다양한 정보 존재, 한번의 attention으로 모든 정보를 적절히 반영하기 어려움

scaled dot-product attention을 여러번 적용하여 concatenation

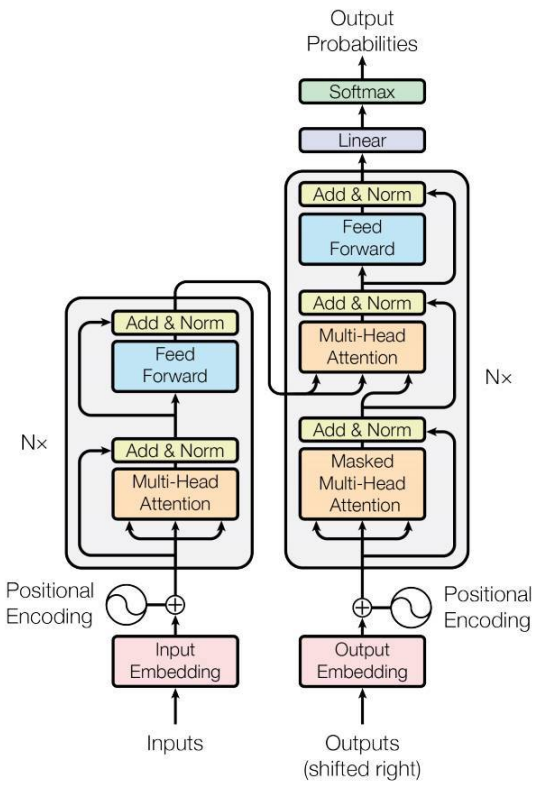

Image Transformer and Local Self-Attention

Task

- Unconditional Image Generation
- Conditional Image Generation
- Super Resolution

Problem

- Image 전처리 → RGB Embedding or Convolution
- Image의 Sequence 정의 → Raster Scan Order
- 고차원 Image에 대한 self-attention을 어떻게 효율적으로 정의? → Local Self-Attention

Music Transformer and Relative Positional Self-Attention

Relative Positional Self Attention : Query와 Key의 Sequence 내 거리를 attention weight에 반영

$$Relative\ Attention = Softmax\left(\frac{QK^T + \boxed{S^{rel}}}{\sqrt{D_h}}\right)V$$

- Token의 절대적 위치를 반영하는 대신 각 query마다 가지는 key와 상대적 위치를 attention score에 반영
- Absolute positional Encoding과 함께 사용 가능