# CVadvanced

Attention Is All You Need(2017) 논문리뷰

## Introduction

- RNN 및 CNN 대신 Attention mechanism에 전적으로 의존하여 입력과 출력 사이의 전역 의존성을 이끄는 모델 아키텍처인 Transformer을 제안
- 기존 모델의 장거리에 있는 의존성을 알기 취약하다는 단점을 해결 (순환없이 Attention mechanism만을 이용하여 의존성 찾음)

## Model Architecture

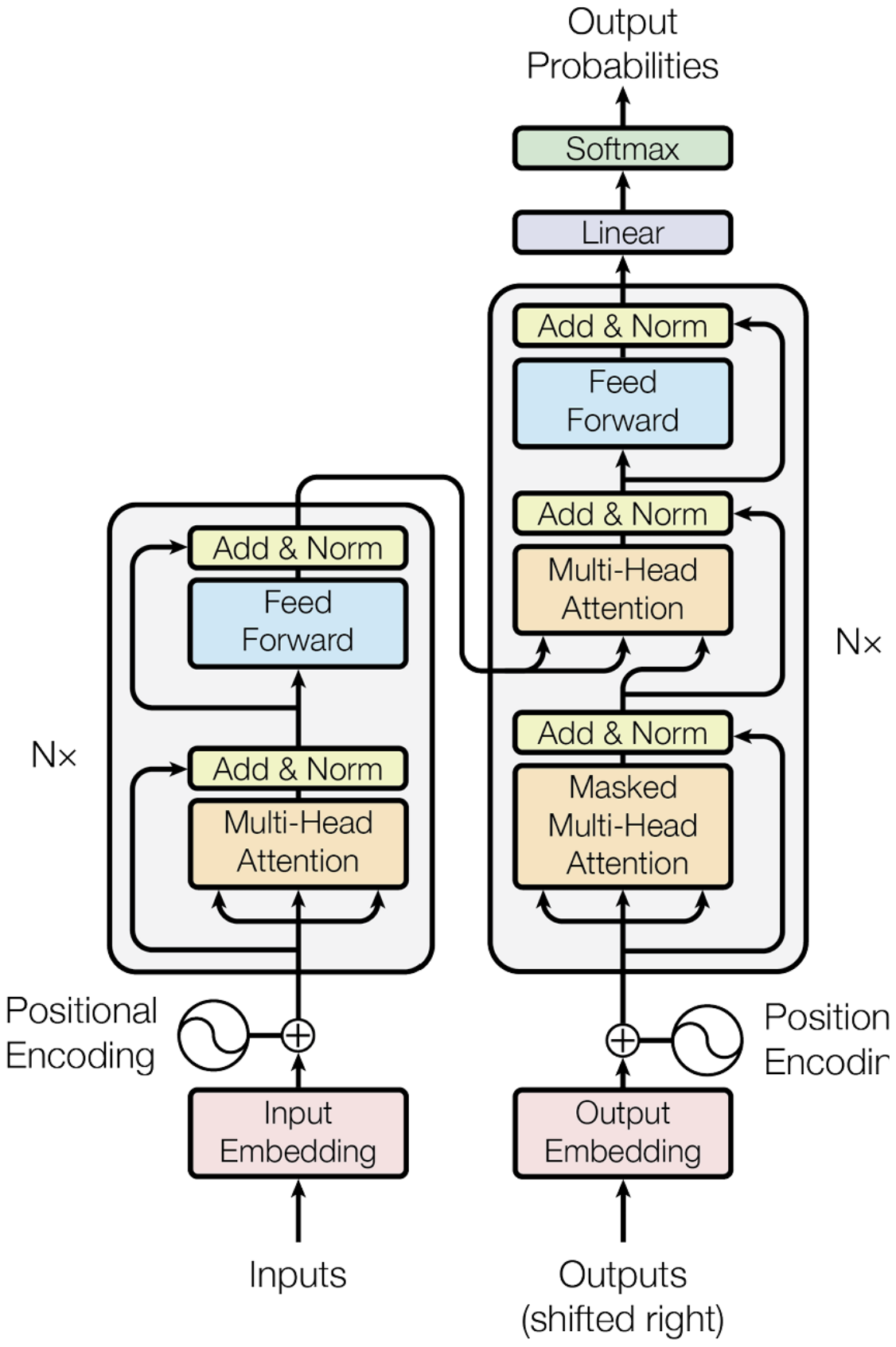



Figure 1: The Transformer - model architecture.

이 모델은 stacked sefl-attention 과 point-wise 완전 연결 레이어를 사용하여 인코더와 디코더 양쪽에 대해 위의 그림과 같이 전반적인 구조를 따르고 있음.

## Scaled Dot-Product Attention vs Multi-Head Attention

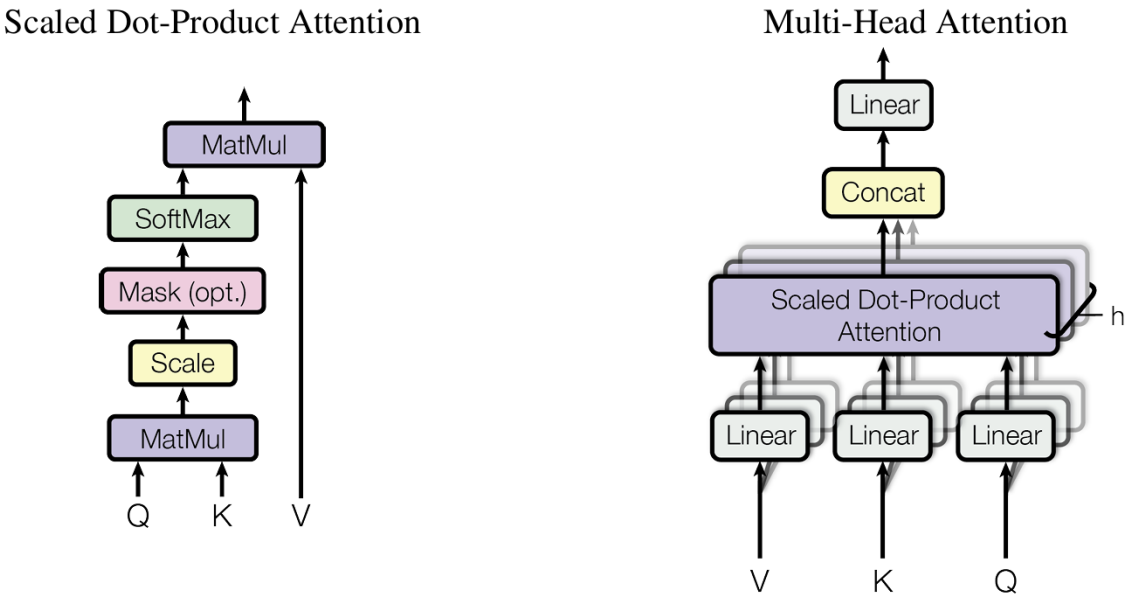
1. **Scaled Dot-Product Attention:**
    - Self-Attention 메커니즘에서 가장 중요한 연산 중 하나는 각 위치에서 모든 다른 위치에 대한 유사도를 계산하는 단계인데, 이때 사용되는 것이 Scaled Dot-Product Attention

$$\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V$$

    - **계산 과정** :
        1. Query(Q), Key(K), Value(V) 세 가지 선형 변환을 수행한다.
        2. Query와 Key 간의 내적을 계산하고, 이를 특정 스케일링 요소로 나누어주어 softmax 함수를 통과시킨다.
        3. 계산된 가중치를 사용하여 Value에 가중합을 수행한다.
    - 유사도를 스케일링하여 안정적인 학습을 도모하고, 각 위치 간의 관계를 학습하는 데에 사용
2. **Multi-Head Attention:**
    - Scaled Dot-Product Attention을 여러 개의 서로 다른 헤드로 나누어 병렬로 계산하고, 이를 다시 합치는 방식

$$\begin{aligned}\text{MultiHead}(Q, K, V) &= \text{Concat}(\text{head}_1, ..., \text{head}_\text{h})W^O \\ \text{where head}_\text{i} &= \text{Attention}(QW_i^Q, KW_i^K, VW_i^V)\end{aligned}$$

- **계산 과정:**
  1. 여러 개의 Attention 헤드를 생성하고, 각 헤드에 대해 독립적으로 Q, K, V를 선형 변환한다.
  2. 각 헤드에 대해 Scaled Dot-Product Attention을 계산한다.
  3. 계산된 결과를 다시 합쳐서 최종 결과를 얻는다.
- 여러 헤드를 사용하여 모델이 서로 다른 종류의 정보 및 관점을 학습할 수 있게 하며, 병렬 계산을 통해 효율적인 학습을 가능하게 함




# Positional Encoding

Self-Attention은 입력 sequence 내의 상대적 관계를 학습하는데, 이 매커니즘 자체는 순서 정보를 고려하지 않으므로 sequence 내 토큰의 상대적 또는 절대적 위치에 대한 정보를 주입해야함.

$$PE_{(pos,2i)} = sin(pos/10000^{2i/d_{\text{model}}})$$
$$PE_{(pos,2i+1)} = cos(pos/10000^{2i/d_{\text{model}}})$$

## Why Self-Attention?

- **계산양 감소 -** 입력 시퀀스의 모든 요소를 직접 연결하여 계산하는 방식으로, 다른 방법에 비해 상대적으로 효율적인 학습 가능(시퀀스 길이가 길어지면서 발**생하는 연산량 증가 문제 완화 가능)**
- **병렬처리 가능 -** 각 위치의 토큰 간의 관계를 독립적으로 계산할 수 있음. GPU 및 TPU와 같은 병렬 처리 장치를 활용하여 모델 학습 가속화 가능
- **장거리 의존성 증가 -** 각 위치에서 모든 다른 위치에 대한 관계를 고려하기에 먼 거리의 단어 간 의존성을 높은 효율성으로 학습할 수 있음
- **해석 가능한 모델 -** 각 위치 간의 관계를 직접적으로 계산하기에 모델의 예측에 대한 해석이 상대적으로 용이. 이는 모델이 어떤 입력에 어떻게 반응하는지 이해하고 해석하는데 도움