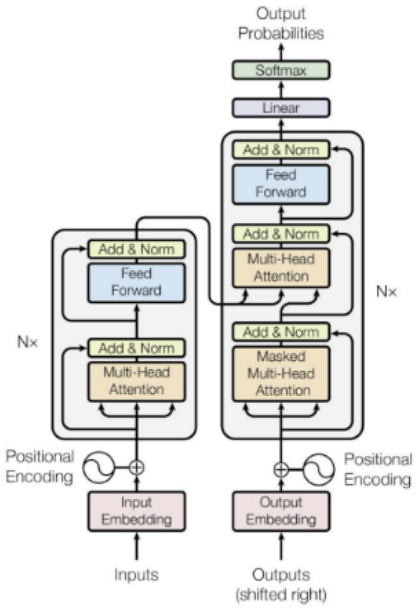
Attention Is All You Need



encoder

-n=6

-각 layer에 두 개의 하위 layer 존재

decoder

-n=6

-각 encoder 계층의 두 하위 계층 외에도 encoder 출력에 대해 다중 head attention을 수행하는 하위 layer로 구성

attention

-query vector, key vector, value vector 쌍을 output으로 mapping하는 과정

-value vector들을 각 weight를 통해 가중합을 하여 output 도출

-query를 기준으로 key vector과 query의 similarity를 계산

Scaled Dot-Product Attention

-compute the dot products of the query with all keys, divide each by 루트dk, and apply a softmax function to obtain the weights on the value

-the key and values are also packed together into matrices K and V

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

Multi-Head Attention

-instead of performing a single attention function → linearly project the queries, keys and values h times with different, learned linear projections to dk, dk and dv dimesions

-multi-head attention allows the model to jointly attend to information from different representation subspaces at different positions

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