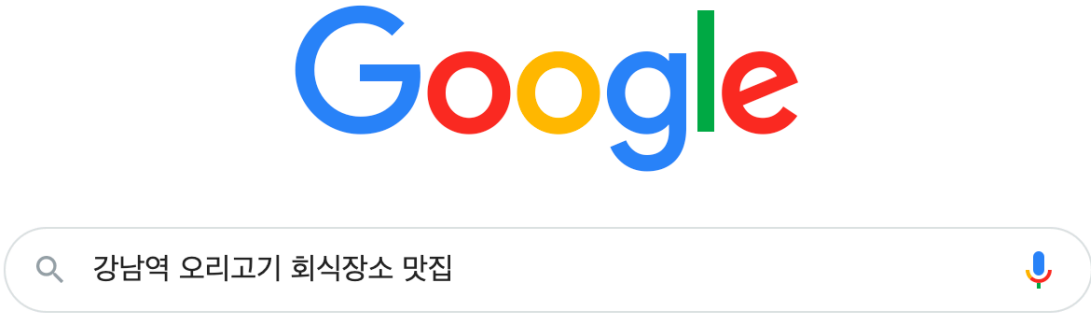
Transformer: Multi-head Attention

Ki Hyun Kim

nlp.with.deep.learning@gmail.com

Attention: Query Generation

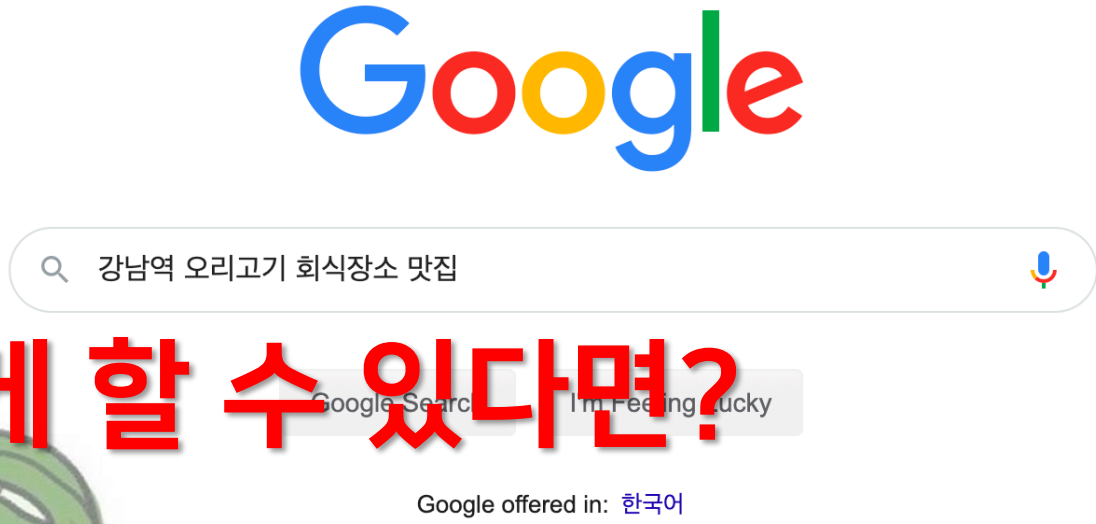
- Example:



마음의 상태(state)를 잘 반영하면서
좋은 검색 결과를 이끌어내는 쿼리를 얻기 위함

Attention: Query Generation

- Example:

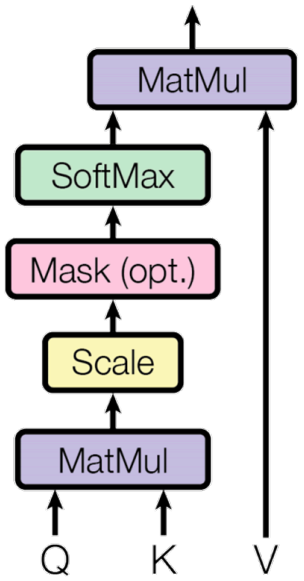


만약 검색을 다양하게 할 수 있다면?

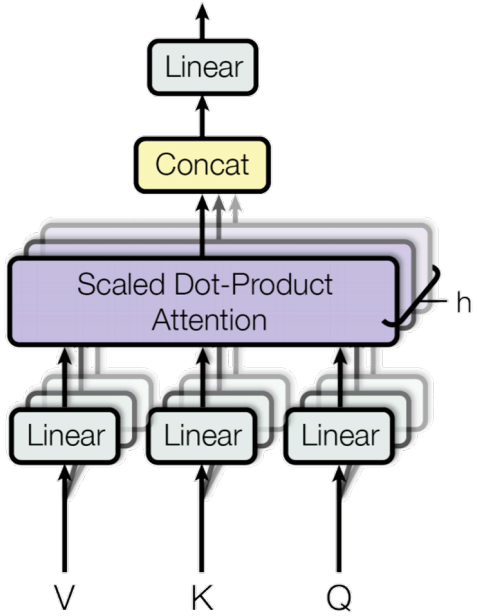
마음의 상태(state)를 잘 반영하면서
좋은 검색 결과를 이끌어내는 쿼리를 얻기 위함

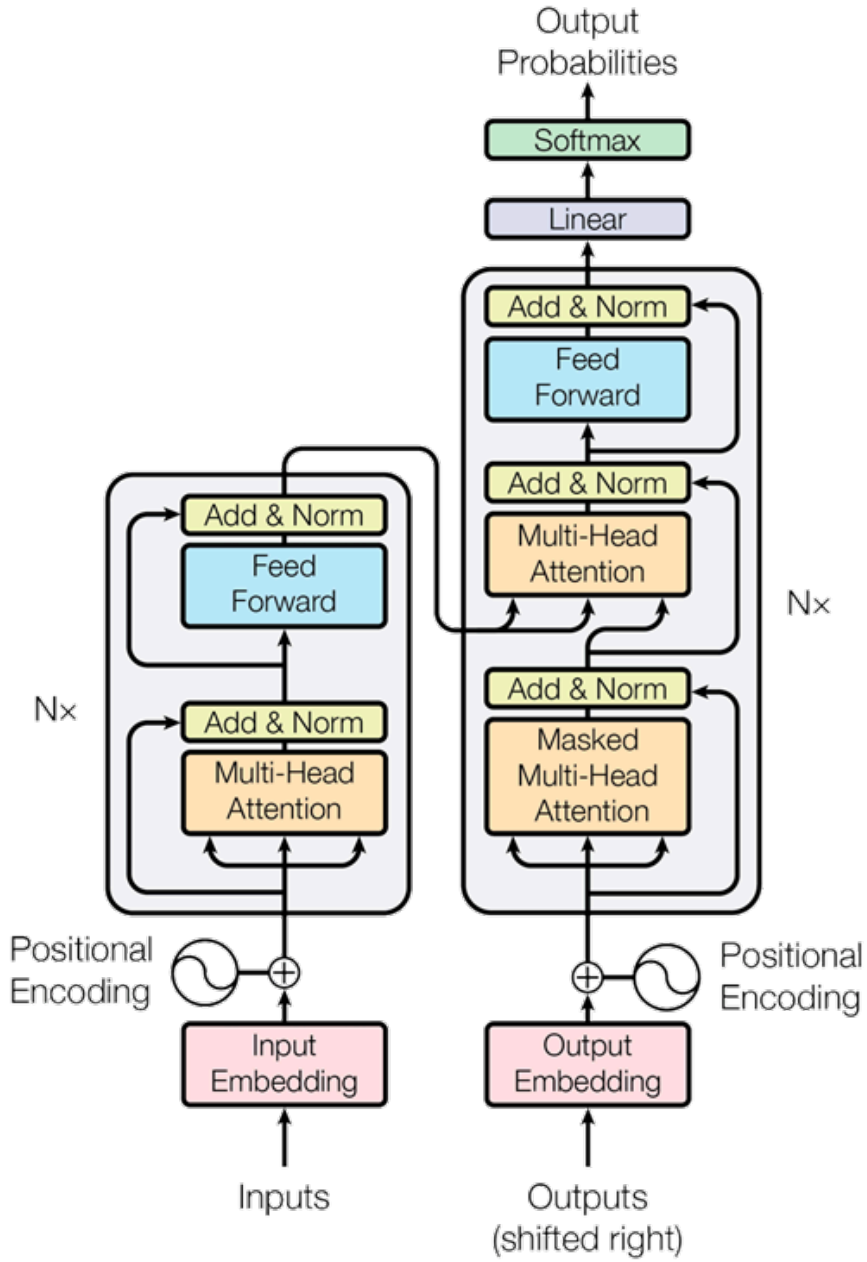
Transformer & Attention

Scaled Dot-Product Attention



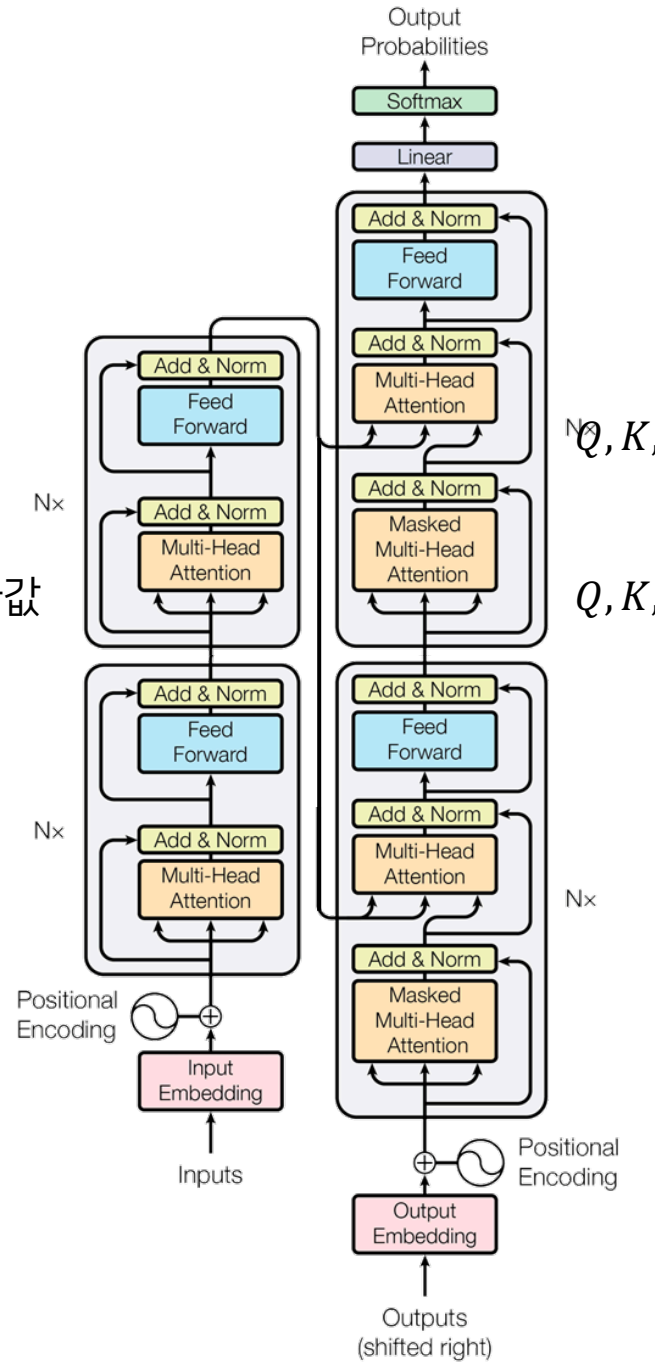
Multi-Head Attention





굳이 그려보자면…

Q, K, V: 이전 블럭의 결과값



Q, K, V: 인코더의 최종 결과값

Q, K, V: 이전 블럭의 결과값

Equations

In case of Q, K and V come from same origin,
$|Q| = |K| = |V| = (\text{batch\_size}, n \text{ or } m, \text{hidden\_size})$.

In case of Q and K, V come from different origin,
$|Q| = (\text{batch\_size}, n, \text{hidden\_size})$
$|K| = |V| = (\text{batch\_size}, m, \text{hidden\_size})$.

$$\text{Attention}(Q, K, V) = \text{softmax}(\frac{Q \cdot K^\intercal}{\sqrt{d_{\text{head}}}}) \cdot V$$

$$\text{MultiHead}(Q, K, V) = [head_1; \cdots; head_h] \cdot W^O$$

$$\text{where } head_i = \text{Attention}(Q \cdot W_i^Q, K \cdot W_i^K, V \cdot W_i^V),$$

$$\text{and } W_i^Q \in \mathbb{R}^{d_{\text{model}} \times d_{\text{head}}}, W_i^K \in \mathbb{R}^{d_{\text{model}} \times d_{\text{head}}}, W_i^V \in \mathbb{R}^{d_{\text{model}} \times d_{\text{head}}}, W^O \in \mathbb{R}^{(h \times d_{\text{head}}) \times d_{\text{model}}}.$$

$$d_{\text{head}} = d_{\text{model}} / h = 64$$

$$h = 8, d_{\text{model}} = 512$$

Summary

- Previous method: attention in sequence to sequence
  - Query를 잘 만들어 key-value를 잘 matching 시키자

- Multi-head Attention
  - 여러 개의 query를 만들어 다양한 정보를 잘 얻어오자

- Attention 자체로도 정보의 encoding과 decoding이 가능함을 보여줌