# ATTENTION

- Self-attention (intra-attention): attention mechanism relating different positions of a single sequence in order to compute a representation of the sequence.
  - + Usages: reading comprehension, abstractive summarization, textual entailment, and learning task-independents sequence representations.
- Attention function: mapping a query and a set of key-value ~~to amp~~ pairs to an output
  - + Query; keys, values, and output: vectors
  - + Output: weighted sum of the values
    - ↳ compatibility function of query and keys

⊕ Scaled ~~dot~~ Dot Product Attention:

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

+ $Q$: matrix of queries
+ $K$: matrix of keys
+ $V$: matrix of values
+ $d_k$: keys of dimension
+ softmax: $\sigma(\vec{z})_i = \dfrac{e^{z_i}}{\sum_{j=1}^{k} e^{z_j}}$ (✱)

+ $q$ and $k$ are independent random variable with mean 0 and variance 1 ⇒ Dot-product: $q \cdot k = \sum_{i=1}^{d_k} q_i k_i$ has mean 0 and variance $d_k$

⇒ Extremely small gradients ⇒ Scale factor: $\dfrac{1}{\sqrt{d_k}}$

⊕ Multi-Head Attention:

$$\boxed{\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, \ldots, \text{head}_h)W^O}$$

+ $\text{head}_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V)$
+ $W_i^Q \in \mathbb{R}^{d_{model} \times d_k}$
+ $W_i^K \in \mathbb{R}^{d_{model} \times d_k}$
+ $W_i^V \in \mathbb{R}^{d_{model} \times d_k}$
+ $W^O \in \mathbb{R}^{hd_v \times d_{model}}$

+ $d_{model}$: dimension keys, values, and queries
+ $d_k, d_v$: dimension
+ $d_v$: dimension output
+ $h$: parallel attention layers

- Perform the attention layer in parallel

⊕ Application:

- Decoder → queries
  Encoder → keys, values
  } → Position in the decoder to attend over all positions in the input sequence