

Keys

$$k_m = \beta_k + W_k x_m$$

$$k_1, \dots, k_N$$

Information retrieval terminology:
“key” is something associated with the object to retrieve
(e.g., video title)

Dot-product self-attention

$$a(x_m, x_n) = \text{softmax}_m(k_1^T q_n, \dots, k_N^T q_n) = \frac{\exp(k_m^T q_n)}{\sum_{l=1}^N \exp(k_l^T q_n)}$$

Attention $a(x_m, x_n)$ measures the attention paid by x_n to x_m and has probability-like properties:

- 1) $a(\cdot, x_n) \geq 0$
- 2) $\sum_m a(x_m, x_n) = 1$

Dot product acts as a similarity measure between a key and a query

Queries

$$q_n = \beta_q + W_q x_n$$

$$q_1, \dots, q_N$$

Information retrieval terminology:
“query” is something used to search for the object
(e.g., text in the search bar)

The n -th output of SA

$$sa_n(x_1, \dots, x_N) = \sum_{m=1}^N a(x_m, x_n) v_m$$

Softly selecting the value with the highest attention paid by x_n

$$v_1, \dots, v_N$$

Values

$$v_m = \beta_v + W_v x_m$$

Information retrieval terminology:
“value” is the object we want to retrieve (e.g., video)