Fast Attention

Junhao Jiang
DMU

Yi Zuo
DMU & BDAI

Abstract

Fast Attention based on Manhattan Distance.

1 Attention

$$\text{Attention}(Q, K, V) = \text{softmax}(\frac{Q \cdot K^T}{\sqrt{d_k}}) \times V \tag{1}$$

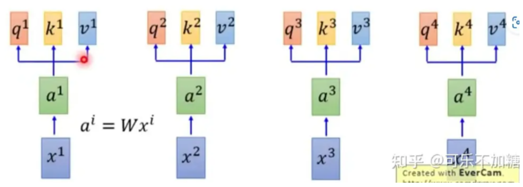

Figure 1: From input to hidden state of q, k, v

Input hidden state (Figure 1):

$$h_q^i = w_{xq} \cdot x \tag{2}$$

$$h_k^i = w_{xk} \cdot x \tag{3}$$

$$h_v^i = w_{xv} \cdot x \tag{4}$$

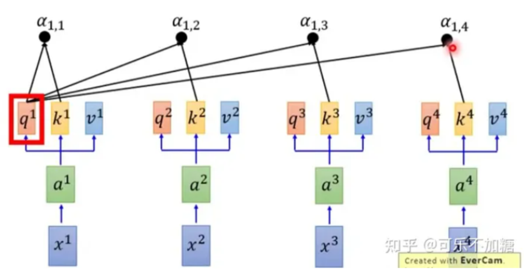

Figure 2: Hidden state of (q·k)

Query hidden state (Figure 2):

$$h_q^a = w_q^a \cdot h_q^i \tag{5}$$

$$h_k^a = w_k^a \cdot h_k^i \tag{6}$$

$$h_{qk}^a = h_q^a \cdot h_k^a \tag{7}$$

31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.

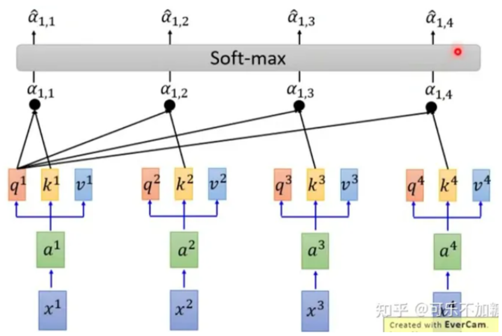

Figure 3: Hidden state of softmax

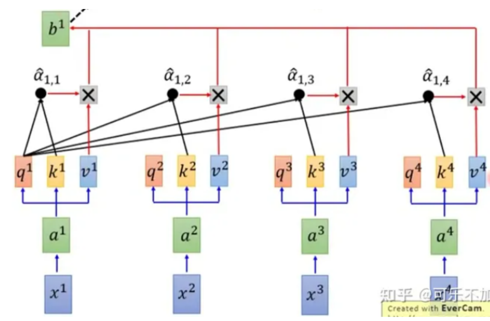

Figure 4: Hidden state of value muliplication

Query hidden state of softmax (Figure 3):

$$h_{qk}^{b} = \frac{h_{qk}^{a}}{\sum_{j=1}^{K} h_{qj}^{b}} \tag{8}$$

Query hidden state of value muliplication (Figure 4):

$$h_{qkv}^{c} = \sum_{j=1}^{V} h_{qk}^{b} \times h_{vj}^{i} \tag{9}$$

Attention is an adaptive full-connected hidden layer as shown in Figure 5 (left). $q \cdot v$ **and** $softmax$ **provide adaptive measures of relevance of current new input with other inputs as shown in Figure 5 (right).**

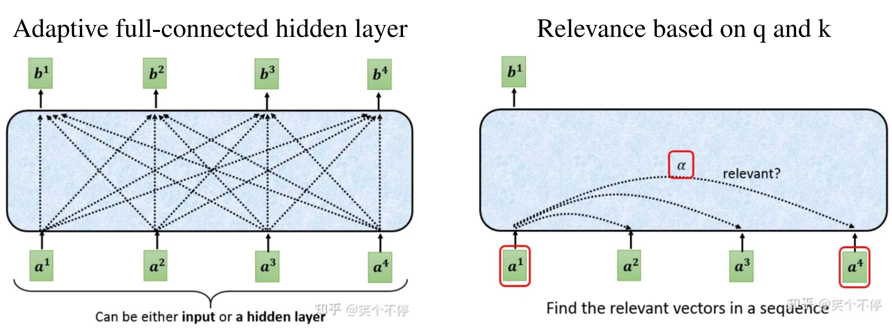

Figure 5: The main idea of attention

2 Fast Attention

2.1 Demerits of attention

1. The relevance of $q_1 \cdot k_1$ and $q_1 \cdot k_2$ is measure by q_1. Therefore, it is necessary to keep h_q^a and h_k^a with same dimension. However, as a query, h_q^a is no need to contain high dimension.
2. $softmax$ provides new assignment of hidden state weights with new inputs, which need to be calculated in high computing cost.

2.2 Idea of fast attention

2.2.1 Replacement of q·k by log function

According to Eq. (10), we use log function to transform $h_q^a \cdot h_k^a$ in Eq. (7). However, the computing cost has been reduced. It is still need to keep h_q^a and h_k^a with same dimension for the addition calculating of $\log h_q^a + \log h_k^a$.

$$h_q^a \cdot h_k^a \simeq \log(h_q^a \cdot h_k^a) = \log h_q^a + \log h_k^a \tag{10}$$

2.2.2 Replacement of (q,k) relevance by Manhattan distance

For hidden states of (q_1, k_1) and (q_1, k_2), we use Manhattan distance to measure similarity between them. The relevance can be describe as distance in Eq. (11).

$$\Delta d_{21}^a = (h_q^{a1}, h_k^{a2}) - (h_q^{a1}, h_k^{a1}) = (\log h_{q1}^a - \log h_{q1}^a) + (\log h_{k2}^a - \log h_{k1}^a) = \log h_{k2}^a - \log h_{k1}^a \tag{11}$$

Therefore, each adaptive hidden weight is unnecessary to be relative with query q, and h_{qk}^b in Eq. (8) is also no need to be calculated. The final output hidden state can be obtained by Eq. (12).

$$h_{qkv}^c = h_{k1}^a \times h_{v1}^i + \sum_{j=2}^{V} \Delta d_{j1}^a \times h_{vj}^i \tag{12}$$