Attention

Input

where x_{ij} represents an element
***e.g*, a word in a sentence**

$$\mathbf{x}_1 = [x_1^1, 0, \cdots, 0]$$
$$\vdots$$
$$\mathbf{x}_i = [x_i^1, x_i^2, \cdots, x_i^i, 0, \cdots, 0]$$
$$\vdots$$
$$\mathbf{x}_N = [x_N^1, x_N^2, \cdots, x_N^N]$$

$$\mathbf{X} = \begin{bmatrix} x_{11} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ x_{N1} & \cdots & x_{NN} \end{bmatrix}$$

Query, key & value

where $\mathbf{W}_{\{Q,K,V\}}$ represents parameters for

e.g., the mapping $\mathbf{x}_i \mapsto \mathbf{q}_i$

$$\mathbf{q}_i = \mathbf{x}_i \mathbf{W}_Q$$
$$\mathbf{k}_i = \mathbf{x}_i \mathbf{W}_K$$
$$\mathbf{v}_i = \mathbf{x}_i \mathbf{W}_V$$

$$\mathbf{Q} = \mathbf{X} \mathbf{W}_Q$$
$$\mathbf{K} = \mathbf{X} \mathbf{W}_K$$
$$\mathbf{V} = \mathbf{X} \mathbf{W}_V$$

Score

where score_{ij} represents the similarity between the queries and keys

$$\text{score}_{ij} = \mathbf{q}_i \cdot \mathbf{k}_j \quad \text{for } j = 1, \cdots, d_k$$

$$\mathbf{QK}^T$$

note that any similarity metric may be used (here; dot product)

Scale and softmax

$$\text{weighting}_{ij} = \text{softmax}\left(\frac{\mathbf{q}_i \cdot \mathbf{k}_j}{\sqrt{d_k}}\right)$$

$$\text{for } j = 1, \cdots, d_k$$

$$\text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}}\right)$$

note that any scaling may be used (here; root of dim) to alter weights

Attention

$$\text{attention}_i = \sum_{j=1}^{N} \text{weighting}_{ij} \mathbf{v}_j$$

$$= \sum_{j=1}^{N} \text{softmax}\left(\frac{\mathbf{q}_i \cdot \mathbf{k}_j}{\sqrt{d_k}}\right) \mathbf{v}_j$$

$$\text{attention} = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}}\right)\mathbf{V}$$