Transformers

Greg Strabel

March 29, 2022

1 Preliminaries

Given two matrices $A \in \mathbb{R}^{n \times m}$ and $B \in \mathbb{R}^{m \times k}$

$$(XY)_{ij} = \sum_{l=1}^{m} X_{il} Y_{lj} = X_{i \cdot} Y_{\cdot j} \tag{1}$$

Therefore

$$(XY)_{\cdot j} = \sum_{l=1}^{m} Y_{lj} X_{\cdot l} \tag{2}$$

so that the columns of XY are linear combinations of the columns of X and

$$(XY)_{i \cdot} = \sum_{l=1}^{m} X_{il} Y_{l \cdot} \tag{3}$$

so that the rows of XY are linear combinations of the rows of Y.

2 Dot-Product Attention

Definition 2.1 (Dot-Product Attention)**.** Given $Q \in \mathbb{R}^{d_l \times d_k}$, $K \in \mathbb{R}^{d_s \times d_k}$ and $V \in \mathbb{R}^{d_s \times d_v}$

$$\text{Attention}\left(Q, K, V\right) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right) V \in \mathbb{R}^{d_l \times d_v} \tag{4}$$

3 Multi-head Attention

Given input matrices $X_Q \in \mathbb{R}^{d_l \times d_{eq}}$, $X_K \in \mathbb{R}^{d_s \times d_{ek}}$ and $X_V \in \mathbb{R}^{d_s \times d_{ev}}$ and weight matrices

$$\begin{aligned} &\{W_Q^i \in \mathbb{R}^{d_e \times d_k}\}_{i=1}^h \\ &\{W_K^i \in \mathbb{R}^{d_e \times d_k}\}_{i=1}^h \\ &\{W_V^i \in \mathbb{R}^{d_e \times d_v}\}_{i=1}^h \end{aligned} \tag{5}$$

$$W_O \in \mathbb{R}^{hd_v \times d_o} \tag{6}$$

we define

$$\begin{aligned} Q^i &= X_Q W_Q^i \in \mathbb{R}^{d_l \times d_k} \\ K^i &= X_K W_K^i \in \mathbb{R}^{d_s \times d_k} \\ V^i &= X_V W_V^i \in \mathbb{R}^{d_s \times d_v} \end{aligned} \tag{7}$$

$$A^i = \text{Attention}\left(Q^i, K^i, V^i\right) \in \mathbb{R}^{d_l \times d_v} \tag{8}$$

$$\text{Multihead}\left(X_Q, X_K, X_V\right) = \text{concat}\left(A_1, ..., A_h\right) W_O \in \mathbb{R}^{d_l \times d_o} \tag{9}$$

In Attention Is All You Need