Date

Attention in Seq2Seq

Example

$h_1 = [0.1, 0.2, 0.3, 0.4]$
$h_2 = [0.2, 0.3, 0.4, 0.5]$
$h_3 = [0.3, 0.4, 0.5, 0.6]$

$S_t = [0.4, 0.5, 0.6, 0.7]$

1 → Attention Scores

$e^t = S_t \cdot h_{ti}$

$eh_1 = [0.4, 0.5, 0.6, 0.7][0.1, 0.2, 0.3, 0.4]$
$= 0.6$

$eh_2 = [0.4, 0.5, 0.6, 0.7][0.2, 0.3, 0.4, 0.5]$
$= 0.82$

$eh_3 = [0.4, 0.5, 0.6, 0.7][0.3, 0.4, 0.5, 0.6]$
$= 1.04$

$e^t = [0.6, 0.82, 1.04]$

2 → Attention Distribution

$\alpha^t = \text{softmax}(e^t)$

Date

$$ah_1 = e^{0.6} / e^{0.6} + e^{0.82} + e^{1.04} = 0.26$$

$$ah_2 = e^{0.82} / e^{0.6} + e^{0.82} + e^{1.04} = 0.32$$

$$ah_3 = e^{1.04} / e^{0.6} + e^{0.82} + e^{1.04} = 0.41$$

$$a^t = [0.26, 0.32, 0.41]$$

(3)

→ Attention output

$$a^t = \sum a_i^t h_i$$

$$a^t = 0.26 \begin{bmatrix} 0.1 \\ 0.2 \\ 0.3 \\ 0.4 \end{bmatrix} + 0.32 \begin{bmatrix} 0.2 \\ 0.3 \\ 0.4 \\ 0.5 \end{bmatrix} + 0.41 \begin{bmatrix} 0.3 \\ 0.4 \\ 0.5 \\ 0.6 \end{bmatrix}$$

$$= \begin{bmatrix} 0.213 \\ 0.312 \\ 0.411 \\ 0.51 \end{bmatrix}$$

$$a^t = [0.213, 0.312, 0.411, 0.51]$$

(4) New Encoder State $[[a_t], [S_t]]$

$$[[0.213, 0.312, 0.411, 0.51], [0.4, 0.5, 0.6, 0.7]]$$

Date

→ using multiplicative attention

$$W = \begin{bmatrix} 0.1 & 0.2 & 0.3 & 0.4 \\ 0.2 & 0.3 & 0.4 & 0.5 \\ 0.3 & 0.4 & 0.5 & 0.6 \\ 0.4 & 0.5 & 0.6 & 0.7 \end{bmatrix}$$

→ previous hidden vectors used and St also

So, $e_i = S^T W h_i$

$$Wh_1 = W \times \begin{bmatrix} 0.1 \\ 0.2 \\ 0.3 \\ 0.4 \end{bmatrix} = \begin{bmatrix} 0.3 \\ 0.4 \\ 0.5 \\ 0.6 \end{bmatrix}$$

$$Wh_2 = W \times \begin{bmatrix} 0.2 \\ 0.3 \\ 0.4 \\ 0.5 \end{bmatrix} = \begin{bmatrix} 0.4 \\ 0.5 \\ 0.6 \\ 0.7 \end{bmatrix}$$

$$Wh_3 = W \times \begin{bmatrix} 0.3 \\ 0.4 \\ 0.5 \\ 0.6 \end{bmatrix} = \begin{bmatrix} 0.5 \\ 0.6 \\ 0.7 \\ 0.8 \end{bmatrix}$$

$$\begin{aligned} e_1 &= S_t W h_1 \\ &= [0.4, 0.5, 0.6, 0.7][0.3, 0.4, 0.5, 0.6] \\ &= 1.04 \end{aligned}$$

$e_2 = S_t W h_2$

$$= [0.4, 0.5, 0.6, 0.7][0.4, 0.5, 0.6, 0.7]$$
$$= 1.26$$

$$e_3 = s^t W h_3$$
$$= [0.4, 0.5, 0.6, 0.7][0.5, 0.6, 0.7, 0.8]$$
$$= 1.48$$

$$e^t = [1.04,\ 1.26,\ 1.48]$$

$$\rightarrow \alpha h_1 = e^{1.04} / e^{1.04} + e^{1.26} + e^{1.48} = 0.26$$

$$\alpha h_2 = e^{1.26} / e^{1.26} + e^{1.04} + e^{1.48} = 0.33$$

$$\alpha h_3 = e^{1.48} / e^{1.26} + e^{1.04} + e^{1.48} = 0.41$$

$$\alpha^t = [0.26,\ 0.33,\ 0.41]$$

$$\rightarrow a^t = \sum \alpha_i^t h_i$$

$$= 0.26 \begin{bmatrix} 0.1 \\ 0.2 \\ 0.3 \\ 0.4 \end{bmatrix} + 0.33 \begin{bmatrix} 0.2 \\ 0.3 \\ 0.4 \\ 0.5 \end{bmatrix} + 0.41 \begin{bmatrix} 0.3 \\ 0.4 \\ 0.5 \\ 0.6 \end{bmatrix}$$

$$= \begin{bmatrix} 0.026 \\ 0.052 \\ 0.078 \\ \end{bmatrix} + \begin{bmatrix} 0.066 \\ 0.099 \\ 0.132 \\ \end{bmatrix} + \begin{bmatrix} 0.123 \\ 0.164 \\ 0.205 \\ 0.246 \end{bmatrix} = \begin{bmatrix} 0.215 \\ 0.315 \\ 0.415 \\ 0.515 \end{bmatrix}$$

Date

$$a^t = [0.215,\ 0.315,\ 0.415,\ 0.515]$$

New, encoder state

$$[[a_t];[S_t]]$$

$$[[0.215, 0.315, 0.415, 0.515] ; [0.4, 0.5, 0.6, 0.7]]$$

Q1) Given the following weight matrices (Wq, Wk, Wv) and embedding vectors (x1, x2, x3), calculate output of self-attention layer (z1 ,z2, and z3). [5 Marks]

$$W_q = \begin{matrix} 0 & 0.7 & 1 \\ 1 & 2 & 0.3 \\ 0.5 & 1 & 1 \end{matrix} , \quad W_k = \begin{matrix} 1.8 & 1 & 1 \\ 1 & 0.5 & 0.7 \\ 0.2 & 1.5 & 0.9 \end{matrix} , \quad W_v = \begin{matrix} 1 & 1.3 & 0.4 \\ 2 & 1 & 2 \\ 1 & 1.5 & 0.2 \end{matrix}$$

$$x1 = [1 \;\; 0.3 \;\; 0.4] \quad , \; x2 = [1.5 \;\; 0.5 \;\; 1] , \quad x3 = [0.3 \;\; 1 \;\; 0.8]$$

Self-attention Layer

$$W_q = \begin{bmatrix} 0 & 0.7 & 1 \\ 1 & 2 & 0.3 \\ 0.5 & 1 & 1 \end{bmatrix}, \quad W_v = \begin{bmatrix} 1 & 1.3 & 0.4 \\ 2 & 1 & 2 \\ 1 & 1.5 & 0.2 \end{bmatrix}$$

$$W_k = \begin{bmatrix} 1.8 & 1 & 1 \\ 1 & 0.5 & 0.7 \\ 0.2 & 1.5 & 0.9 \end{bmatrix}, \quad \begin{aligned} x_1 &= [1 \quad 0.3 \quad 0.4] \\ x_2 &= [1.5 \quad 0.5 \quad 1] \\ x_3 &= [0.3 \quad 1 \quad 0.8] \end{aligned}$$

So, $x = \begin{bmatrix} 1 & 0.3 & 0.4 \\ 1.5 & 0.5 & 1 \\ 0.3 & 1 & 0.8 \end{bmatrix}$

$$Q = x \times W_q = \begin{bmatrix} 1 & 0.3 & 0.4 \\ 1.5 & 0.5 & 1 \\ 0.3 & 1 & 0.8 \end{bmatrix} \begin{bmatrix} 0 & 0.7 & 1 \\ 1 & 2 & 0.3 \\ 0.5 & 1 & 1 \end{bmatrix}$$

Date

$$Q = \begin{bmatrix} 0.5 & 1.7 & 1.49 \\ 1 & 3.05 & 2.65 \\ 1.4 & 3.01 & 1.4 \end{bmatrix}$$

$$K = x \times W_K = \begin{bmatrix} 1 & 0.3 & 0.4 \\ 1.5 & 0.5 & 1 \\ 0.3 & 1 & 0.8 \end{bmatrix} \begin{bmatrix} 1.8 & 1 & 1 \\ 1 & 0.5 & 0.7 \\ 0.2 & 1.5 & 0.9 \end{bmatrix}$$

$$K = \begin{bmatrix} 2.18 & 1.75 & 1.57 \\ 3.4 & 3.25 & 2.75 \\ 1.7 & 2 & 1.72 \end{bmatrix}$$

$$V = x \times W_V = \begin{bmatrix} 1 & 0.3 & 0.4 \\ 1.5 & 0.5 & 1 \\ 0.3 & 1 & 0.8 \end{bmatrix} \begin{bmatrix} 1 & 1.3 & 0.4 \\ 2 & 1 & 2 \\ 1 & 1.5 & 0.2 \end{bmatrix}$$

$$V = \begin{bmatrix} 2 & 2.2 & 1.08 \\ 3.5 & 3.95 & 1.8 \\ 3.1 & 2.59 & 2.28 \end{bmatrix}$$

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

$$\text{softmax}\left(\begin{bmatrix} 0.5 & 1.7 & 1.49 \\ 1 & 3.05 & 2.65 \\ 1.4 & 3.01 & 1.4 \end{bmatrix} \begin{bmatrix} 2.18 & 3.4 & 1.7 \\ 1.75 & 3.25 & 2 \\ 1.57 & 2.75 & 1.72 \end{bmatrix}\right)^{1/\sqrt{3}}$$

Date

$$= \text{softmax}\begin{bmatrix} 3.69 & 6.54 & 3.93 \\ 6.74 & 11.89 & 7.14 \\ 6.07 & 10.62 & 6.24 \end{bmatrix}$$

$$= \begin{bmatrix} 0.05 & 0.88 & 0.07 \\ 0 & 0.99 & 0.01 \\ 0.01 & 0.98 & 0.01 \end{bmatrix}$$

$$Z = \begin{bmatrix} 0.05 & 0.88 & 0.07 \\ 0 & 0.99 & 0.01 \\ 0.01 & 0.98 & 0.01 \end{bmatrix} \begin{bmatrix} 2 & 2.2 & 1.08 \\ 3.5 & 3.95 & 1.8 \\ 3.1 & 2.59 & 2.28 \end{bmatrix}$$

$$\underset{V}{}$$

$$Z = \begin{bmatrix} 3.4 & 3.8 & 1.8 \\ 3.5 & 4.0 & 1.8 \\ 3.5 & 3.9 & 1.8 \end{bmatrix} \begin{matrix} \rightarrow Z_1 \\ \rightarrow Z_2 \\ \rightarrow Z_3 \end{matrix}$$