$Att(Q,K,V) = \text{Softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$ ~~Attention~~ (Q,K,V) = Softm

formula,

① Self attention: allow each token in a sequence to took at other token to pay attention.

② we create ③ matrices

Query: Q - Represent what this token is looking for.

Key: Key - what this token offer

value: V - Actual Information.

→ It Computed by mul input emb with learned weighted matrix.

$$Q = XW_Q, \quad K = XW_K, \quad V = XW_V$$

③ The Attention Score:

$$\text{Score} = Q \times K^T$$

④ Scale the Score: $\sqrt{d_k}$

dimention of the key vector

$$\text{Scaled Sore} = \frac{QK^T}{\sqrt{d^k}}$$

⑤ Apply Softmax

$$\text{Att Weig} = \text{Softmax}\left(\frac{QK^T}{\sqrt{d^k}}\right).$$

⑥ weighted Sum: OutPut = AttWeigh x V

mple: I love AI

emb. size = 2 (make simple)

seq len = 3

(2)

(1)

$$X = \begin{array}{c c} & \begin{array}{cc} f_1 & f_2 \end{array} \\ \begin{array}{c} I \\ \text{love} \\ AI \end{array} & \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 1 \end{bmatrix} \end{array}$$

(2) Weight Materices:

Assume:

$$W_Q = \begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix}, \quad W_K = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}, \quad W_V = \begin{bmatrix} 1 & 2 \\ 0 & 1 \end{bmatrix}$$

(3) for "I" [1, 0]

$X_1 \times Q_I = [1,0] \times \begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix} = [1, 0]$

$X_1 \times K_I = [1,0] \times \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} = [0, 1]$

$X_1 \times V_I = [1,0] \times \begin{bmatrix} 1 & 2 \\ 0 & 1 \end{bmatrix} = [1, 2]$

for "love" (0,1)

$X_2 \times Q_L = [0,1] \times W_Q = [1,1]$

$X_2 \times K_L = [0,1] \times \cancel{K}W_K = [1, 0]$

$X_3 \times K_L = [0,] = W_K = [0, 1]$

for AI

$X_3. [1,1] \times Q_A = [1,1] \times W_Q = [2,1]$

$X_3 \times K_L = [1,1] \times W_K = = [1,]$

$X_3 \times V_L = [1,1] \times W_V = [1,3]$

$$Q = \begin{bmatrix} 1 & 0 \\ 1 & 1 \\ 2 & 1 \end{bmatrix}$$

$$K = \begin{bmatrix} 0 & 1 \\ 1 & 0 \\ 1 & 1 \end{bmatrix}$$

$$V = \begin{bmatrix} 1 & 2 \\ 0 & 1 \\ 1 & 3 \end{bmatrix}$$

$$QK^T = \begin{bmatrix} 0 & 1 & 1 \\ 1 & 1 & 2 \\ 1 & 2 & 3 \end{bmatrix}$$

3

Q: is matrix of queries (one row per word)
K: matrix of key (" ")
K^T = Transpose of K

$$Q = \begin{bmatrix} 1 & 0 \\ 1 & 1 \\ 2 & 1 \end{bmatrix}, \quad K = \begin{bmatrix} 0 & 1 \\ 1 & 0 \\ 1 & 1 \end{bmatrix}$$

Transpose of K

$$\begin{bmatrix} 0 & 1 & 1 \\ 1 & 0 & 1 \end{bmatrix} \Rightarrow \cancel{\text{Row2}}$$

$\text{Row1} = [1, 0] \times \begin{bmatrix} 0 & 1 & 1 \\ 1 & 0 & 1 \end{bmatrix} \Rightarrow (0, 1, 2)$

$\text{Row2} = [1, 1] \times [K^T] = (1, 1$

why T 3 - - - - - [1

inner dimention Should Match

3x2 3x2 (Not)
3x2 2x3 (Yes)

$Q \times K^T$ (dot Product)

$$QK^T = \begin{bmatrix} 0 & 1 & 1 \\ 1 & 1 & 2 \\ 1 & 2 & 3 \end{bmatrix}$$

3 Scaled by $\sqrt{d_k}$

$\sqrt{d_k} = \sqrt{2} \approx 1.41$

$$\frac{QK^T}{d\sqrt{k}}$$

divide each element by 1.41

$$\text{Scaled Score} = \begin{bmatrix} 0 & 0.71 & 0.71 \\ 0.71 & 0.71 & 0.71 \\ 0.71 & 1.41 & 2.12 \end{bmatrix}$$

1 -> softmax (row wise) $\cancel{\frac{1}{1+e^{-x}}} = \frac{e^{x_i}}{\sum_j e^{x_j}}$

Row1 $[0, 0.71, 0.71]$

$e^0 = 1, e^{0.71} \approx 2.03$

Sum $= 1 + 2.03 + 2.03 = 5.08$

R1 S.M $[0.1978, 0.4011, 0.4011]$

$$\text{softmax}(x_i) = \frac{e^{x_i}}{\sum_j e^{x_j}}$$

$e^0 = 1, \quad e^{0.71} = 2.03, \quad e^{0.71} = 2.03$

$$[1, 2.03, 2.03]$$

Sum $= 5.08$

divide $= \left[\frac{1}{5.08}, \frac{2.03}{5.08}, \frac{2.03}{5.08}\right]$

$[0.1978, 0.4011, 0.4011]$

word 1 = ("I") => 19.8% itself att

=> 40% love

=> 40% AI

⑤ : Multiply by V

Output Attweight × V

Row 1

$$[0.198, 0.401, 0.401] \times \begin{bmatrix} 1 & 2 \\ 0 & 1 \\ 1 & 3 \end{bmatrix} = [0.599, 2.000]$$

$$\text{Final Output} \begin{bmatrix} 0.599 & 2.000 \\ 0.752 & 2.255 \\ 0.716 & 2.292 \end{bmatrix}$$

Notes

1 → borrow for A1 due to big value in A1