generate global dependencies between inputs & outputs (sequences). A self attention module takes n-inputs & returns n outputs. It allows the inputs to interact with each other & find out where they should pay most attention

Example:

Starts with some embedding vector (feature vector) of dimension 4

(1, 4)

, Say we have 3 of these inputs.

$$X = \begin{bmatrix} [\text{Input 1}] \\ [\text{Input 2}] \\ [\text{Input 3}] \end{bmatrix} = \begin{bmatrix} [X_1] \\ [X_2] \\ [X_3] \end{bmatrix} = \begin{bmatrix} x_{11} & x_{12} & x_{13} & x_{14} \\ x_{21} & x_{22} & x_{23} & x_{24} \\ x_{31} & x_{32} & x_{33} & x_{34} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 1 & 0 \\ 0 & 2 & 0 & 2 \\ 1 & 1 & 1 & 1 \end{bmatrix}$$

Every one of the 3 inputs need 3 representations, key, query & value. Let's say these representations have dimension 3 (1, 3), our weight matrices must have shape (1, 3) as $\underset{\text{input}}{(1,4)} \times \underset{\text{weights}}{(4,3)} \to \underset{\text{output}}{(1,3)}$

That is to say:

$$Q = XW^q \qquad K = XW^k \qquad V = XW^v$$

For our example shape of $W^q, W^k, W^v \Rightarrow (4,3)$, given $X \Rightarrow (1,4)$

& is wanting a $(1,3)$ representation,

& given: $W^q = \begin{bmatrix} 1&0&1 \\ 1&0&0 \\ 0&0&1 \\ 0&1&1 \end{bmatrix}$ $W^k = \begin{bmatrix} 0&0&1 \\ 1&1&0 \\ 0&1&0 \\ 1&1&0 \end{bmatrix}$ $W^v = \begin{bmatrix} 0&2&0 \\ 0&3&0 \\ 1&0&3 \\ 1&1&0 \end{bmatrix}$

$(4,3)$ $(4,3)$ $(4,3)$

$$\underset{(\underline{3},4)\ (4,\underline{3})}{X\ W^q} = \begin{bmatrix} 1&0&1&0 \\ 0&2&0&2 \\ 1&1&1&1 \end{bmatrix} \begin{bmatrix} 1&0&1 \\ 1&0&0 \\ 0&0&1 \\ 0&1&1 \end{bmatrix} = \begin{bmatrix} [1\ 0\ 2] \\ [2\ 2\ 2] \\ [2\ 1\ 3] \end{bmatrix} = \begin{bmatrix} [Q_1] \\ [Q_2] \\ [Q_3] \end{bmatrix} = Q$$

$(3,3)$

$$\underset{(\underline{3},4)\ (4,\underline{3})}{X\ W^k} = \begin{bmatrix} 1&0&1&0 \\ 0&2&0&2 \\ 1&1&1&1 \end{bmatrix} \begin{bmatrix} 0&0&1 \\ 1&1&0 \\ 0&1&0 \\ 1&1&0 \end{bmatrix} = \begin{bmatrix} [0\ 1\ 1] \\ [4\ 4\ 0] \\ [2\ 3\ 1] \end{bmatrix} = \begin{bmatrix} [K_1] \\ [K_2] \\ [K_3] \end{bmatrix} = K$$

$(3,3)$

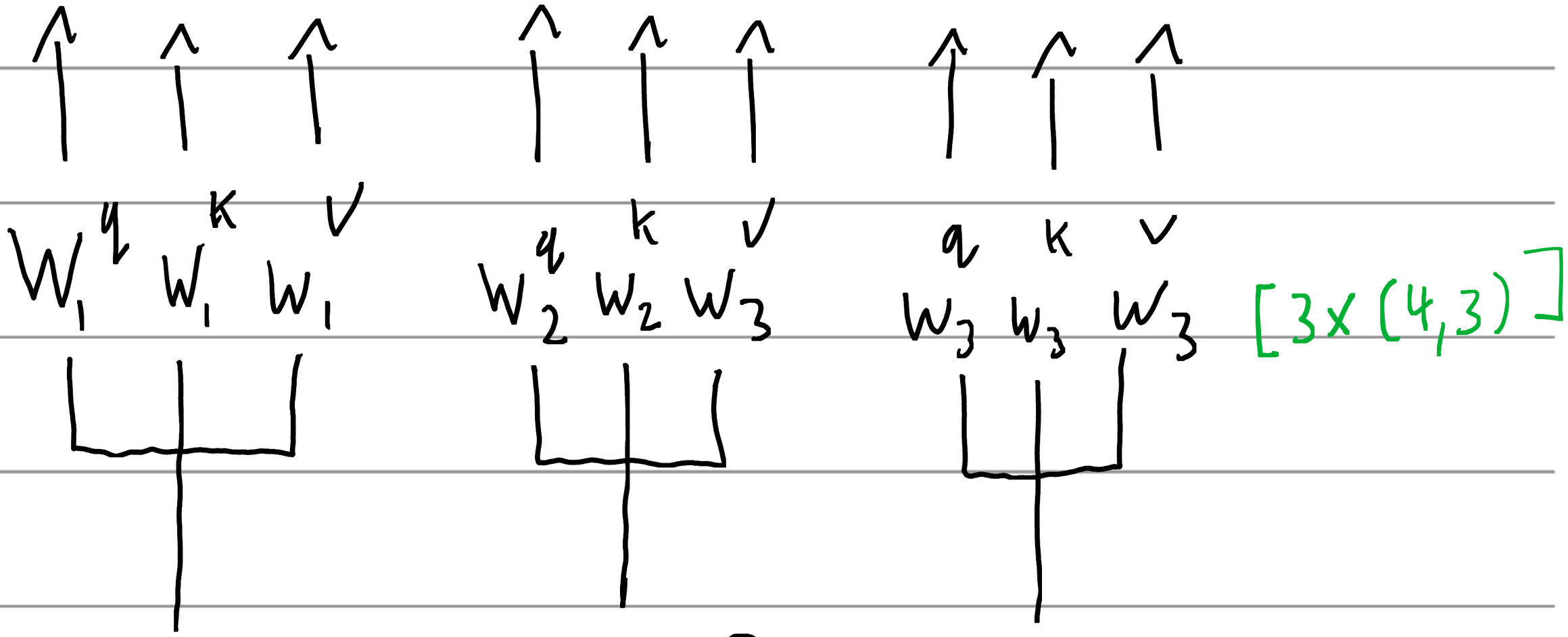
$$\underset{(3,4)\ (4,3)}{X\ W^v} = \begin{bmatrix} 1&0&1&0 \\ 0&2&0&2 \\ 1&1&1&1 \end{bmatrix} \begin{bmatrix} 0&2&0 \\ 0&3&0 \\ 1&0&3 \\ 1&1&0 \end{bmatrix} = \begin{bmatrix} [1\ 2\ 3] \\ [2\ 8\ 0] \\ [2\ 6\ 3] \end{bmatrix} = \begin{bmatrix} [V_1] \\ [V_2] \\ [V_3] \end{bmatrix} = V$$

$(\underline{3},\underline{3})$

Or diagrammatically

[102] [011] [123] [222] [440] [280] [213] [231] [263] [3x(1,3)]

Q_1 K_1 V_1 Q_2 K_2 V_2 Q_3 K_3 V_3 $Q = XW^Q$ $K = XW^K$ $V = XW^V$



[3x(4,3)]

[1010] [0202] [1111] [3x(1,4)]

Input 1 Input 2 Input 3

X_1 X_2 X_3

$(1,4) \times (4,3) \rightarrow (1,3)$

Input Weights Representations

$(3,4) \times (4,3) \rightarrow (3,3)$

Input Weight representation

x3 for Q, K, T

To obtain attention scores we take the dot product between each inputs query & all keys (including itself), we obtain 3 attention scores for each input (for this example)

$$Q = \begin{bmatrix} [\ Q_1\ ] \\ [\ Q_2\ ] \\ [\ Q_3\ ] \end{bmatrix} = \begin{bmatrix} q_{11} & q_{12} & q_{13} \\ q_{21} & q_{22} & q_{23} \\ q_{31} & q_{32} & q_{33} \end{bmatrix} \qquad K = \begin{bmatrix} [\ K_1\ ] \\ [\ K_2\ ] \\ [\ K_3\ ] \end{bmatrix} = \begin{bmatrix} K_{11} & K_{12} & K_{13} \\ K_{21} & K_{22} & K_{23} \\ K_{31} & K_{32} & K_{33} \end{bmatrix}$$

$$K^T = \begin{bmatrix} [K_1^T] & [K_2^T] & [K_3^T] \end{bmatrix} = \begin{bmatrix} K_{11} & K_{21} & K_{31} \\ K_{12} & K_{22} & K_{32} \\ K_{13} & K_{23} & K_{33} \end{bmatrix}$$

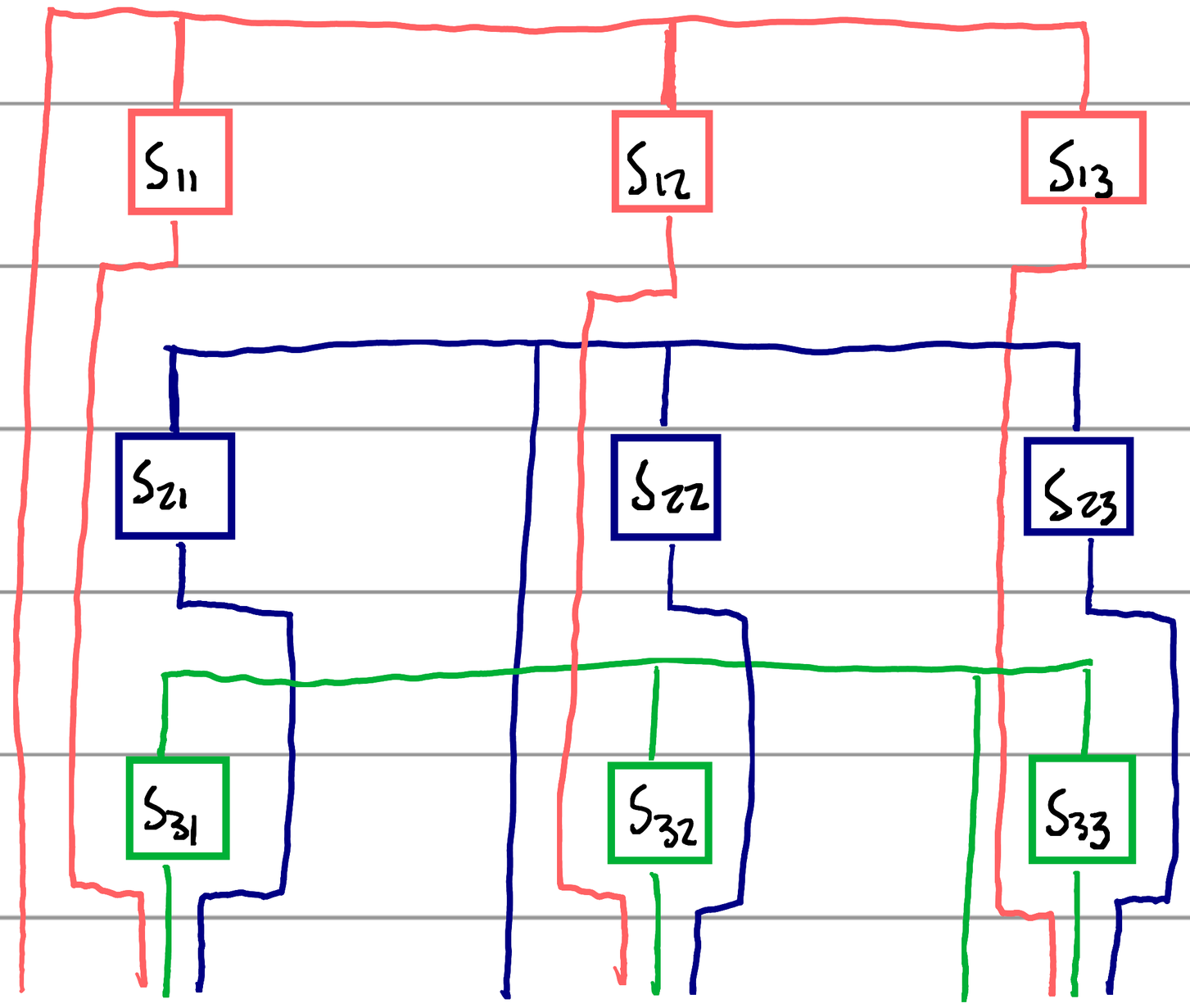
Similarity score $S = Q K^T$

$(3,3) \quad (3,3) \quad (3,3)$

$$S = \begin{bmatrix} [\ Q_1\ ] \\ [\ Q_2\ ] \\ [\ Q_3\ ] \end{bmatrix} \begin{bmatrix} [K_1^T] & [K_2^T] & [K_3^T] \end{bmatrix}$$

$$S = \begin{bmatrix} Q_1 K_1^T & Q_1 K_2^T & Q_1 K_3^T \\ Q_2 K_1^T & Q_2 K_2^T & Q_2 K_3^T \\ Q_3 K_1^T & Q_3 K_2^T & Q_3 K_3^T \end{bmatrix} = \begin{bmatrix} S_{11} & S_{12} & S_{13} \\ S_{21} & S_{22} & S_{23} \\ S_{31} & S_{32} & S_{33} \end{bmatrix}$$

Or diagramatically:



$S = QK^T$

[102] [011] [123] [222] [440] [280] [213] [231] [263] (1,3) [3x(1,3)]

Q_1 K_1 V_1 Q_2 K_2 V_2 Q_3 K_3 V_3 $Q = XW^q$ $K = XW^k$ $V = XW^v$

W_1^q W_1^k W_1^v W_2^q W_2^k W_3^v W_3^q W_3^k W_3^v (4,3) [3x(4,3)]

[1010] [0202] [1111] (1,4) [3x(1,4)]

Input 1 Input 2 Input 3

X_1 X_2 X_3

$Y = \text{attention}(Q, K, V)$ $\quad d_x$ is the embedding vector dimension

$= \text{softmax}\left(\frac{QK^T}{\sqrt{d_x}}\right)V$ $\quad$ for this example $d_x = 4$

$= \text{softmax}\left(\frac{S}{\sqrt{d_2}}\right)V = S''V$

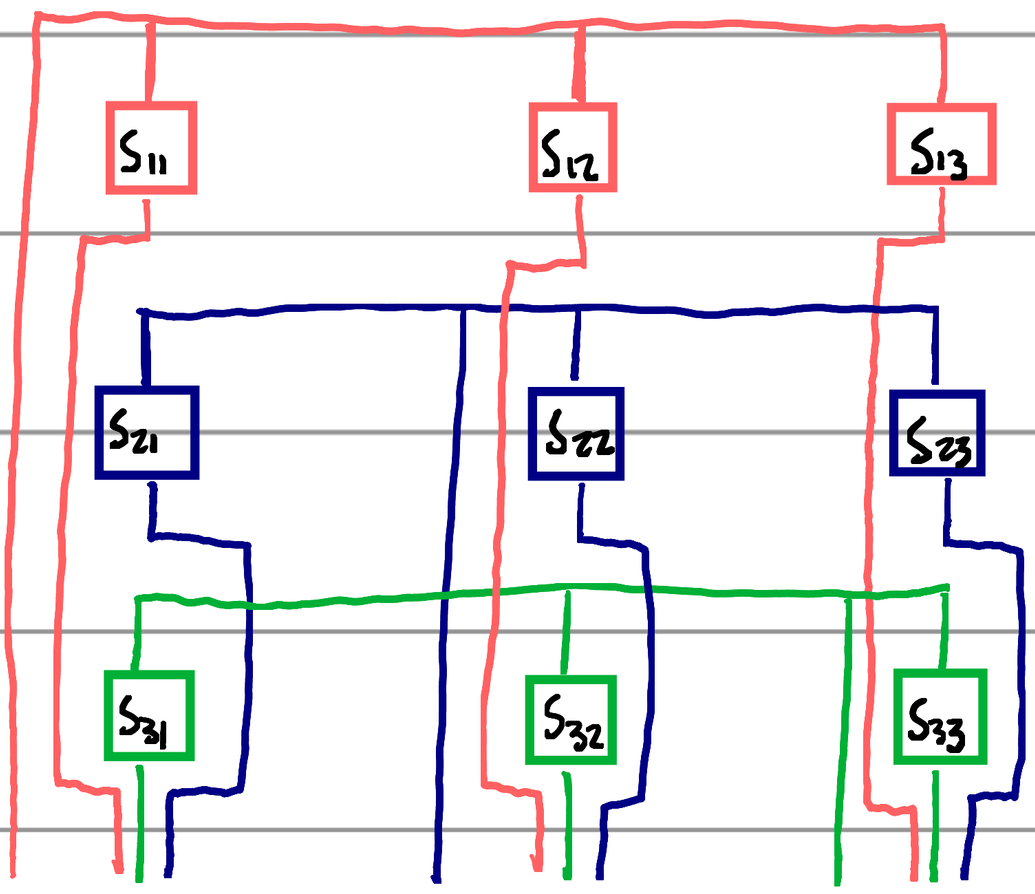
S^{11} $[3 \times 3]$ $\quad S'' = \text{softmax}\left(\frac{S}{\sqrt{d_2}}\right)$

Softmax

S' $[3 \times 3]$ $\quad S' = \frac{S}{\sqrt{d_2}}$

Normalise by $\sqrt{d_2}$

$[3 \times 3]$ $[$



$S = QK^T$

[102] [011] [123] [222] [110] [280] [213] [231] [263] (1,3) $\quad [3 \times [3 \times (1,3)]] = [3 \times (3 \times 3)]$

Q_1 K_1 V_1 $\quad Q_2$ K_2 V_2 $\quad Q_3$ K_3 V_3 $\quad Q = XW^Q$ $\quad K = XW^K$ $\quad V = XW^V$

W_1^Q W_1^K W_1^V $\quad W_2^Q$ W_2^K W_2^V $\quad W_3^Q$ W_3^K W_3^V $(4,3)$ $\quad [3 \times (4,3)]$

[1010] [0202] [1111] $(1,4)$ $\quad [3 \times (1,4)]$

Input 1 $\quad$ Input 2 $\quad$ Input 3

X_1 $\quad X_2$ $\quad X_3$