Variants of Attention

1. Dot product version

$e_i = s^T h_i$ (assumes that $d_1 = d_2$ where $s \in \mathbb{R}^{d_1}$, $h_i \in \mathbb{R}^{d_2}$.)

2. Multiplicative attention

$e_i = s^T W h_i$

$W \in \mathbb{R}^{d_1 \times d_2}$

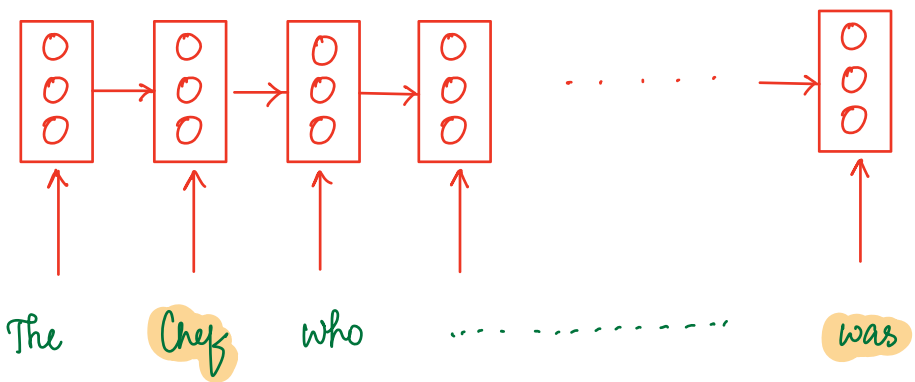
$d_1 = 2$, $d_2 = 3$

$$\underset{1\times 2}{[s_1\ s_2]} \underset{2\times 3}{\begin{bmatrix} w_{11} & w_{12} & w_{13} \\ w_{21} & w_{22} & w_{23} \end{bmatrix}} \underset{3\times 1}{\begin{bmatrix} h_1 \\ h_2 \\ h_3 \end{bmatrix}}$$

3. Additive attention :

$e_i = v^T \tanh(W_1 h_i + W_2 s)$

Self Attention



(i) Hard to learn long-distance dependencies.

(ii) Lack of parallelizability.

If not recurrence, then what? — Attention

Attention within a single sentence. — Self attention.

We have:

query $q_1, q_2, \cdots, q_T$

key $k_1, k_2, \cdots, k_T$

value $v_1, v_2, \cdots, v_T$

eg. $k_i = q_i = v_i = w_i$.

Self attention:

$$e_{ij} = q_i^T k_j$$

$$a_{ij} = \frac{\exp(e_{ij})}{\sum \exp(e_{ij})}$$

$$\alpha_i = \sum_j a_{ij} v_j$$

attention:

$$e_{tj} = s_t^T h_j$$

$$a_{tj} = \frac{\exp(e_{tj})}{\sum_j \exp(e_{tj})}$$

$$\alpha_t = \sum_j a_{tj} h_j$$

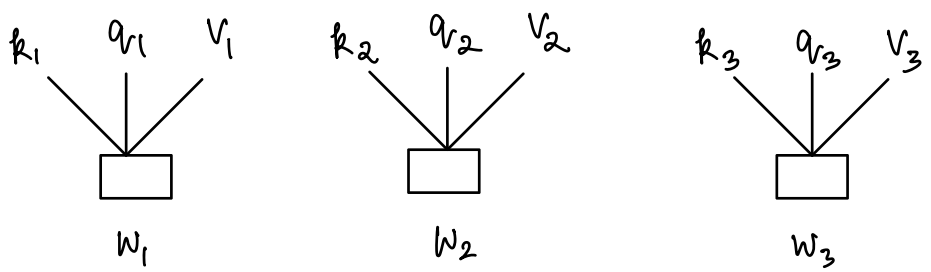
e_{11}	e_{12}	e_{13}
e_{21}	e_{22}	e_{23}
e_{31}	e_{32}	e_{33}

$e_{11} = q_1^T k_1 = w_1^T w_1$ $\quad a_{11} = \text{soft}(e_{11})$

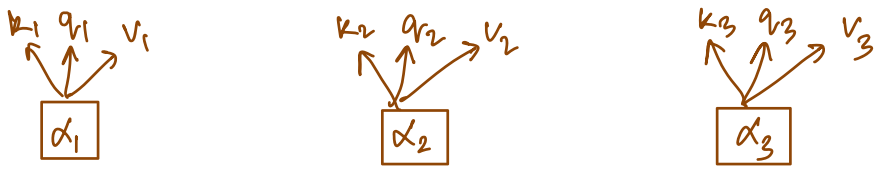
$e_{12} = q_1^T k_2 = w_1^T w_2$ $\quad a_{12} = \text{soft}(e_{12})$

$e_{13} = q_1^T k_3 = w_1^T w_3$ $\quad a_{13} = \text{soft}(e_{13})$

$$\alpha_1 = a_{11} v_1 + a_{12} v_2 + a_{13} v_3 = a_{11} w_1 + a_{12} w_2 + a_{13} w_3.$$



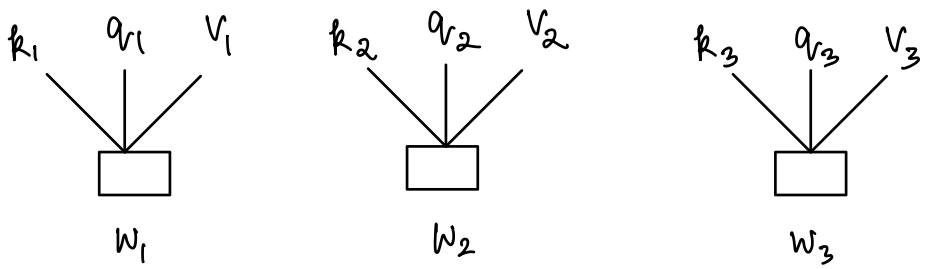
Self - Attention



a_{11} a_{12} a_{13} a_{21} a_{22} a_{23}

e_{12}

Self - Attention



$q^T k$

$a_{11} w_1 + a_{12} w_2 + a_{13} w_3$

$a_{11} w_1 + a_{13} w_3 + a_{12} w_3$

Problems

1) Doesn't have an inherent notion of order.

2) No non-linearity

3) Need to ensure that we do not look at the future words.

Solution

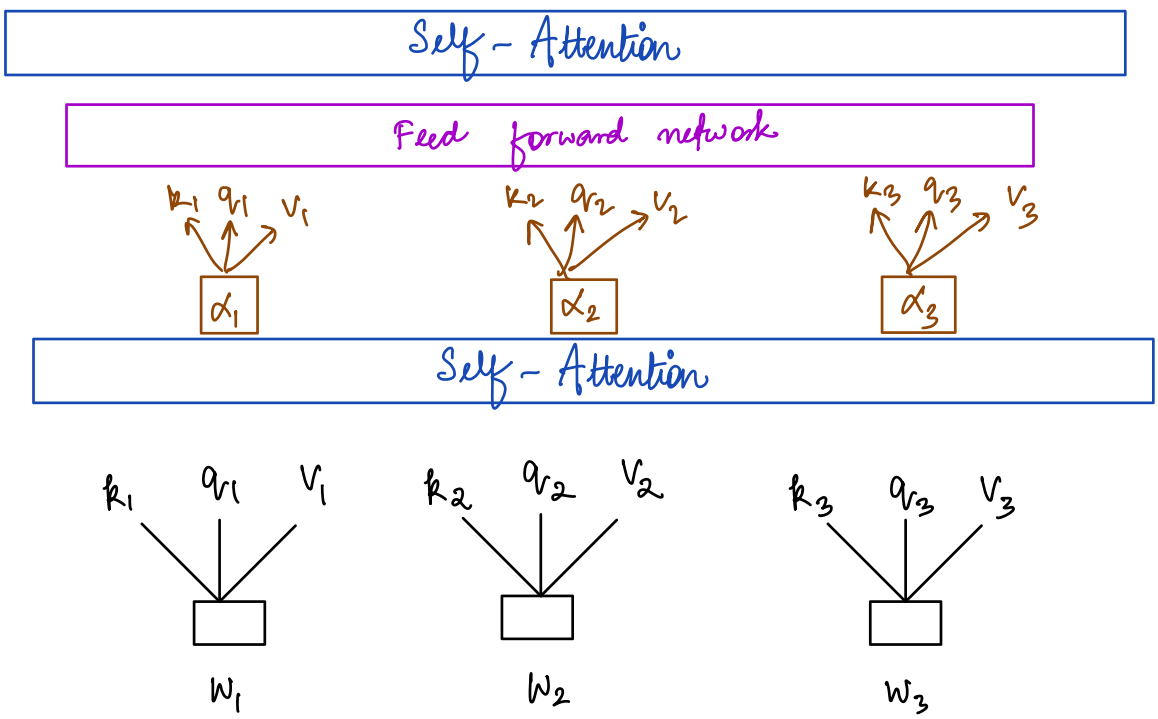
1) Positional encoding / vector $p_i \in \mathbb{R}^d$

$p_1, p_2, \ldots, p_T$

$$\tilde{q}_i = q_i + p_i$$
$$\tilde{k}_i = k_i + p_i$$
$$\tilde{v}_i = v_i + p_i$$

These days the vectors p_i's are learnt.

2) Add a FF network to post process each vector



(3)

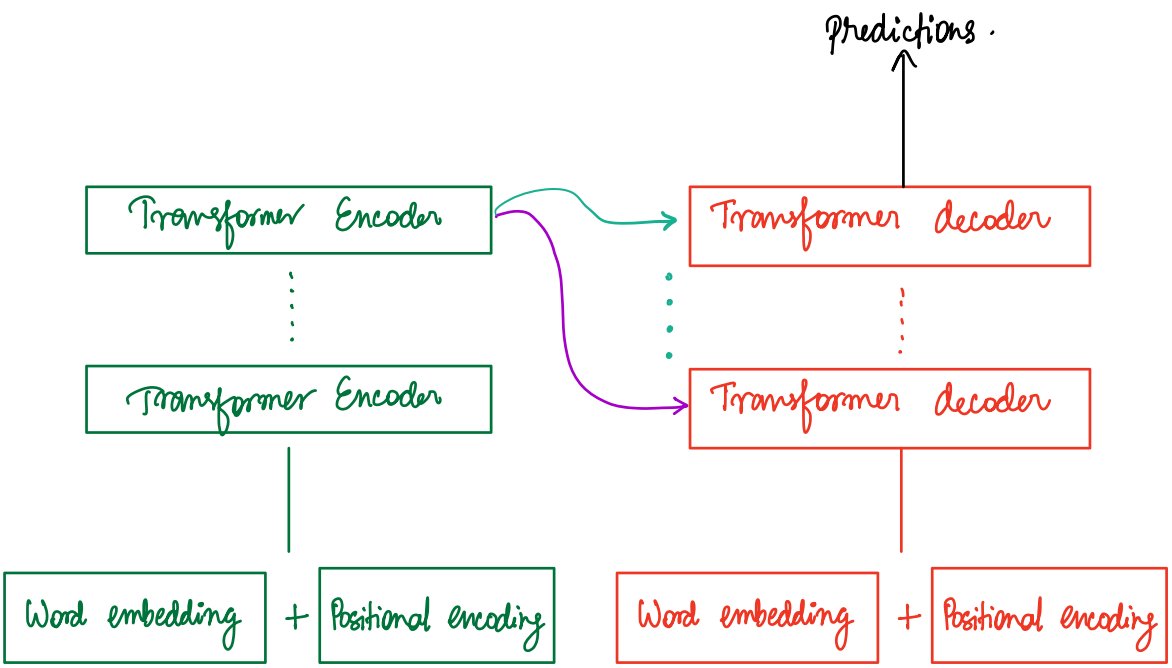
	j →	START	Tomorrow	9	will	fly.
			0	1	2	3
i ↓		START				
0	q_i	Tomorrow	✓	✗	✗	✗
1		9	✓	✓	✗	✗
2		will	✓	✓	✓	✗
3		fly	✓	✓	✓	✗

$j \geq i$

9:10 AM

Masking $e_{ij} = \begin{cases} q_i^T k_j & j < i \\ -\infty & j \geq i \end{cases}$

$$a_{ij} = \frac{\exp(e_{ij})}{\sum \exp(e_{ij})}$$



1. key - query - value attention

$$k_i = K w_i$$
$$q_i = Q w_i$$
$$v_i = V w_i$$

2. Multi-head attention. attend to multiple places in a single layer.