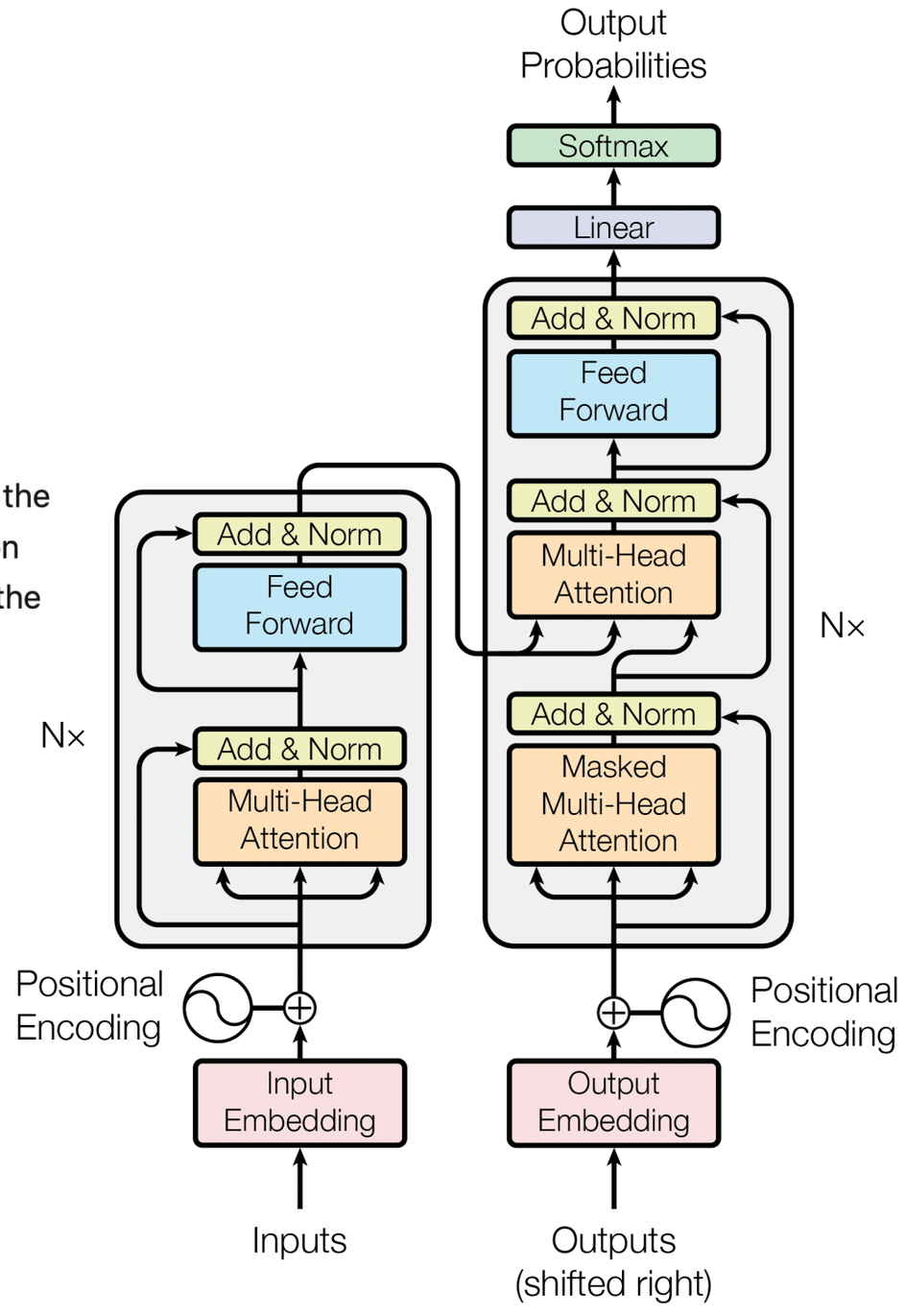
Attention

Introduction

$$
\begin{aligned}
h_0 &= UW_{\text{embed}} + W_{\text{position}} \\
h_l &= \text{transformer-block}(h_{l-1}) \forall l \in [1, n] \\
P(u) &= \text{softmax}(h_n W_{\text{embed}}^T)
\end{aligned}
$$

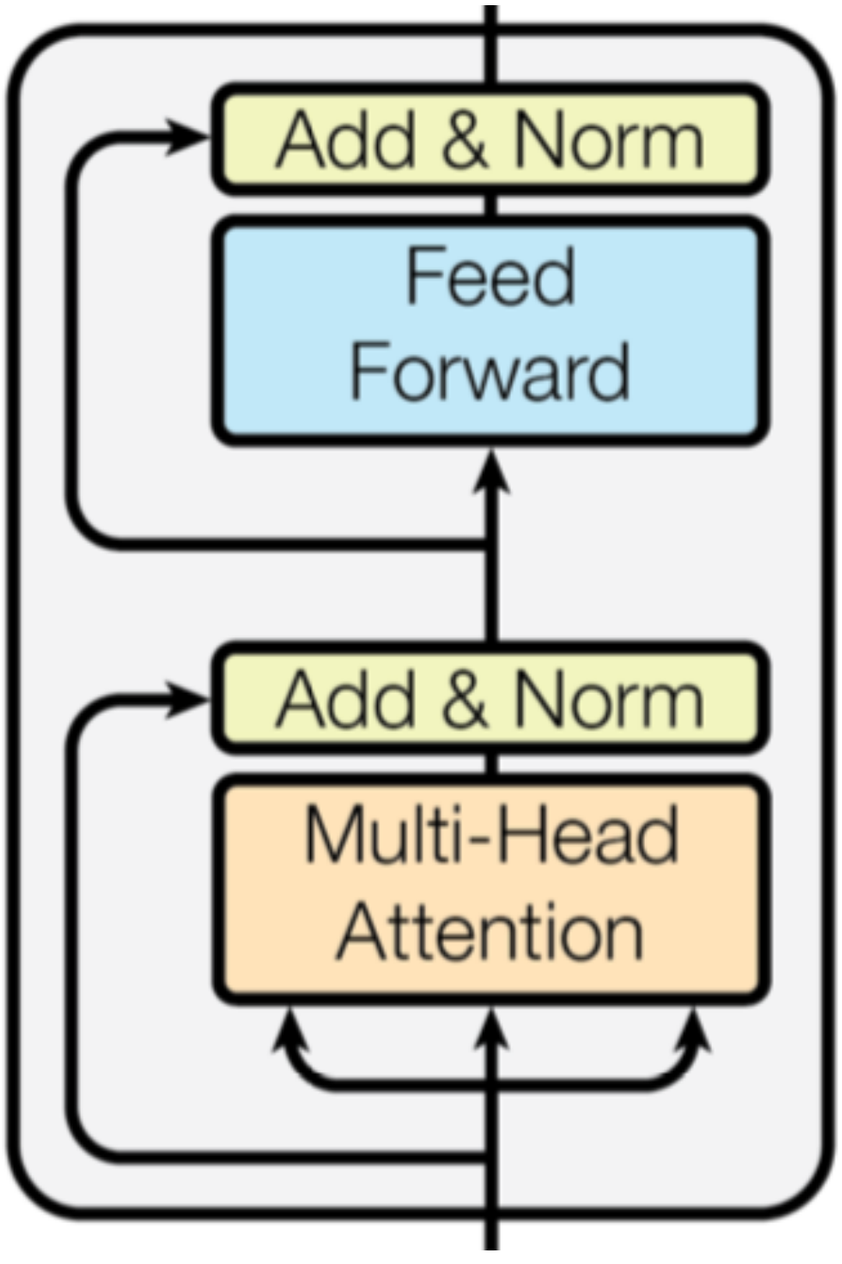
where $U = (u_{-k}, \ldots, u_{-1})$ is the context vector of tokens(one hot encodings), n is the number of layers, W_{embed} is the token embedding matrix, and W_{position} is the position embedding matrix. Let d_n be the length of the input and d_{model} be the dimension of the embedding. $W_{embed} \in \mathbb{R}^{d_{ntokens} \times d_{\text{model}}}$ and $U \in \mathbb{R}^{d_n \times d_{\text{ntokens}}}$. So $h_0 \in \mathbb{R}^{d_n \times d_{\text{model}}}$.



TransFormer Block

transformer-block:
input: h_{in}
output: h_{out}
$h_{mid} = \text{LayerNorm}(h_{in} + \text{MultiHead}(h_{in}))$
$h_{out} = \text{LayerNorm}(h_{mid} + \text{FFN}(h_{mid}))$

where $h_{in}, h_{out} \in \mathbb{R}^{d_n \times d_{\text{model}}}$. d_n is the length of the input and d_{model} is the dimension of the model (e.g., embedding dimension). LayerNorm is layer normalization.

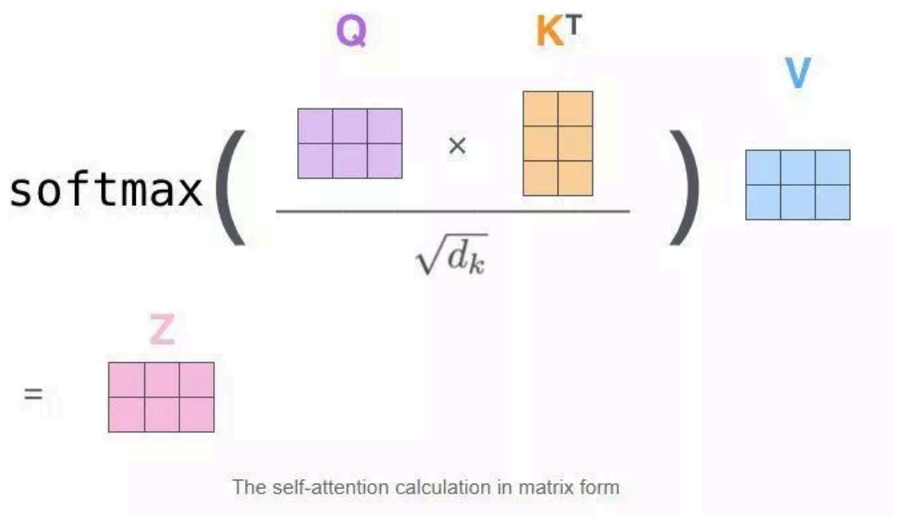


Attention

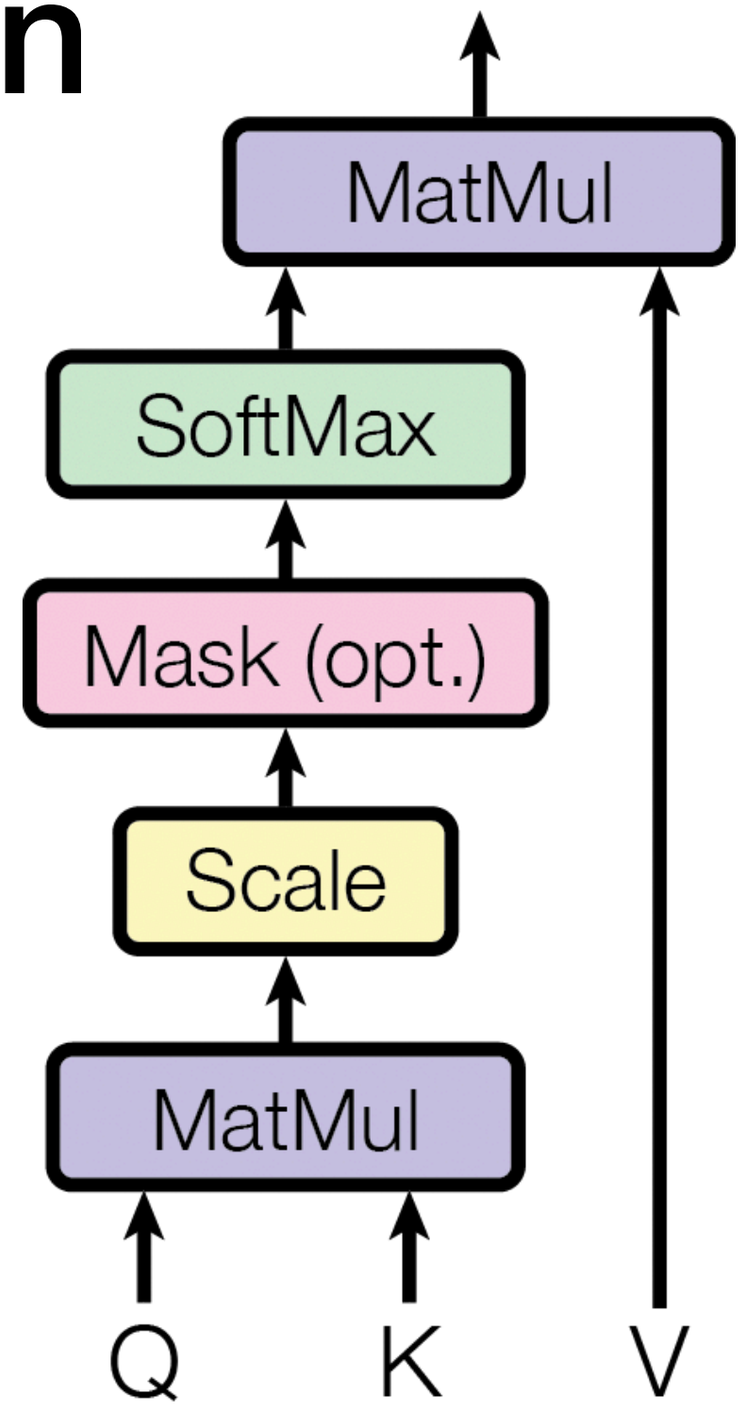
Scaled Dot-Product Attention

$$\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V$$

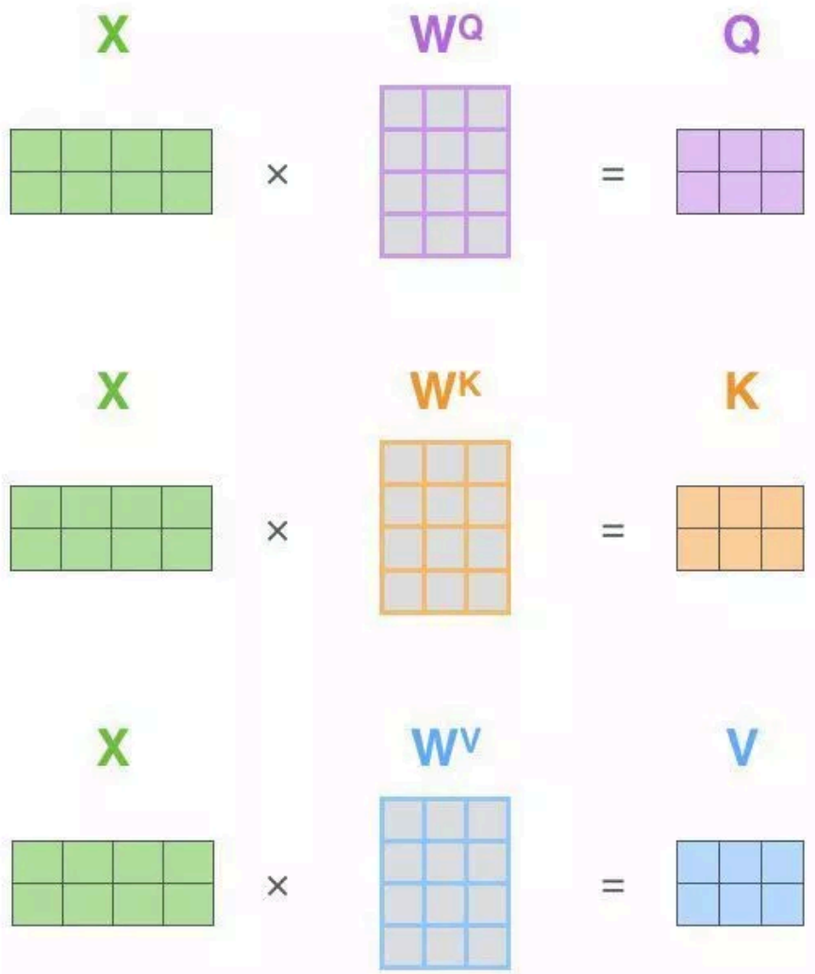
where the $Q, K \in \mathbb{R}^{d_n \times d_k}$ and $V \in \mathbb{R}^{d_n \times d_v}$



The self-attention calculation in matrix form



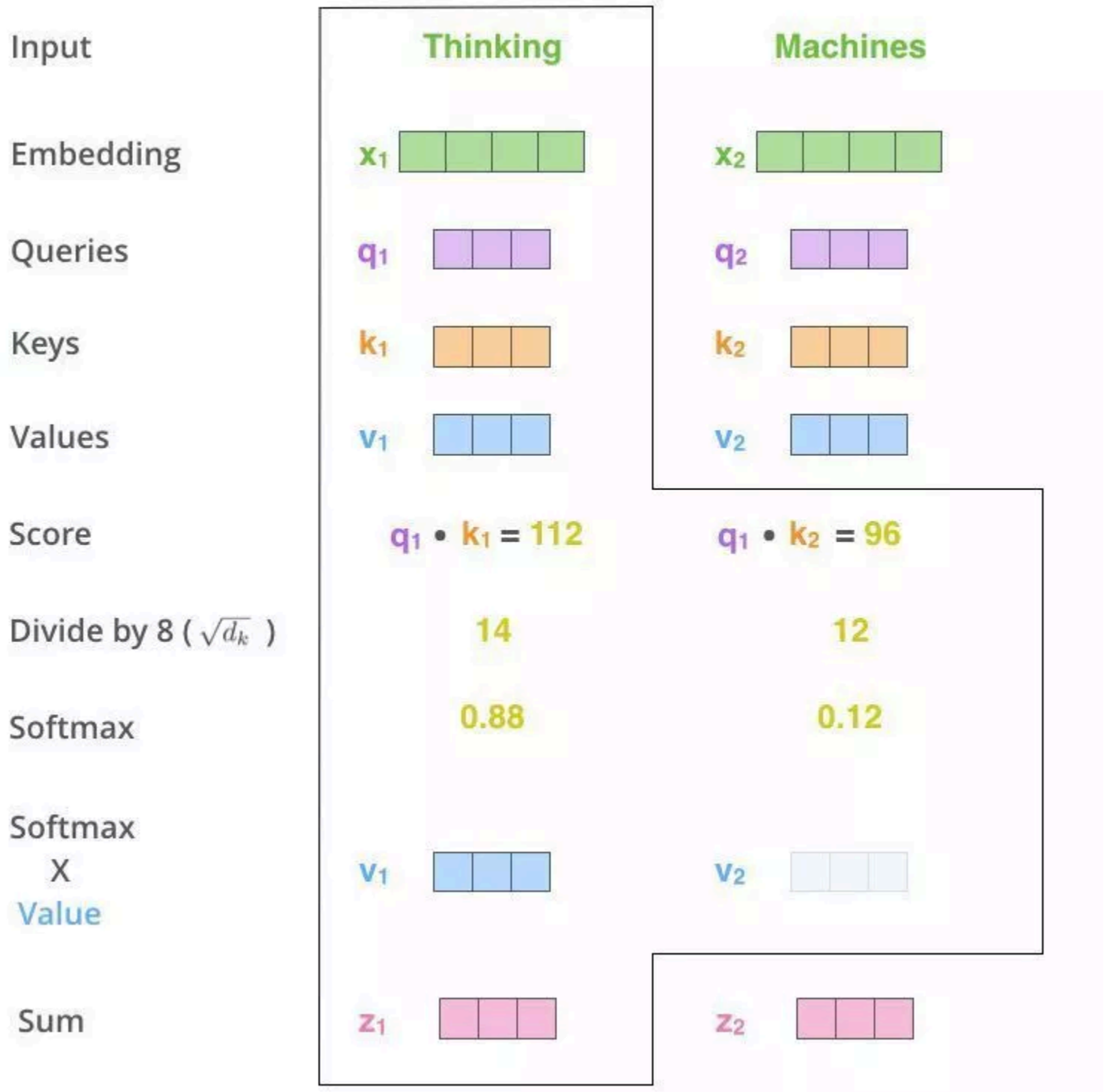
输入X经历3次线性变换得到Q,K,V



Every row in the X matrix corresponds to a word in the input sentence. We again see the difference in size of the embedding vector (512, or 4 boxes in the figure), and the q/k/v vectors (64, or 3 boxes in the figure)

具体运算细节

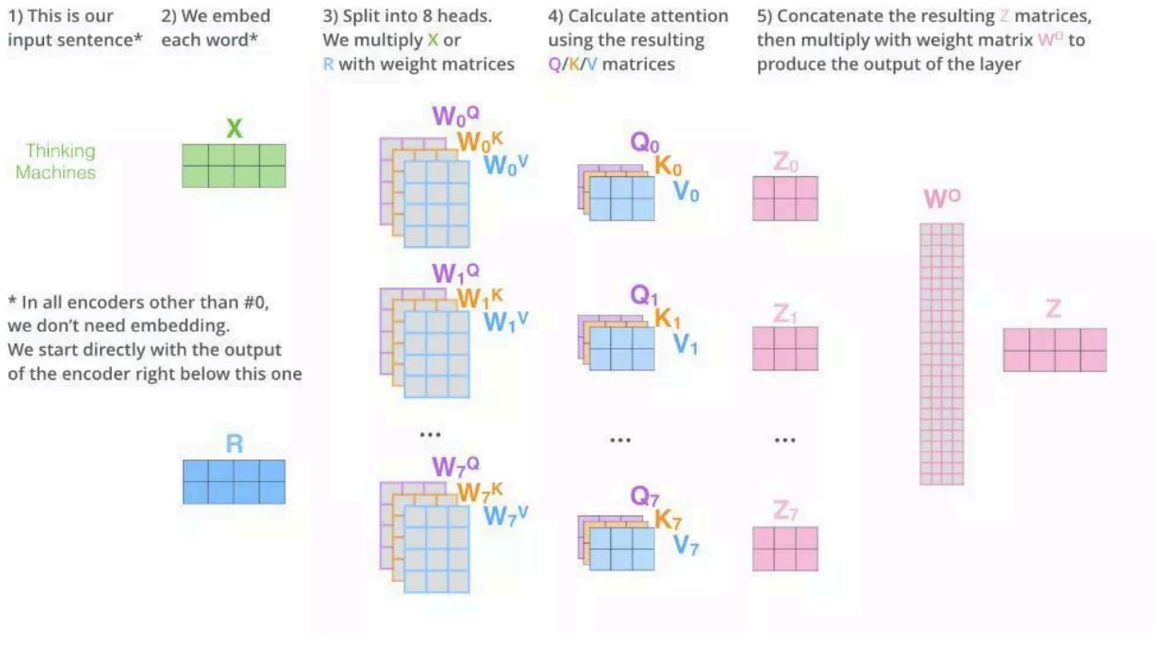
由X到Z

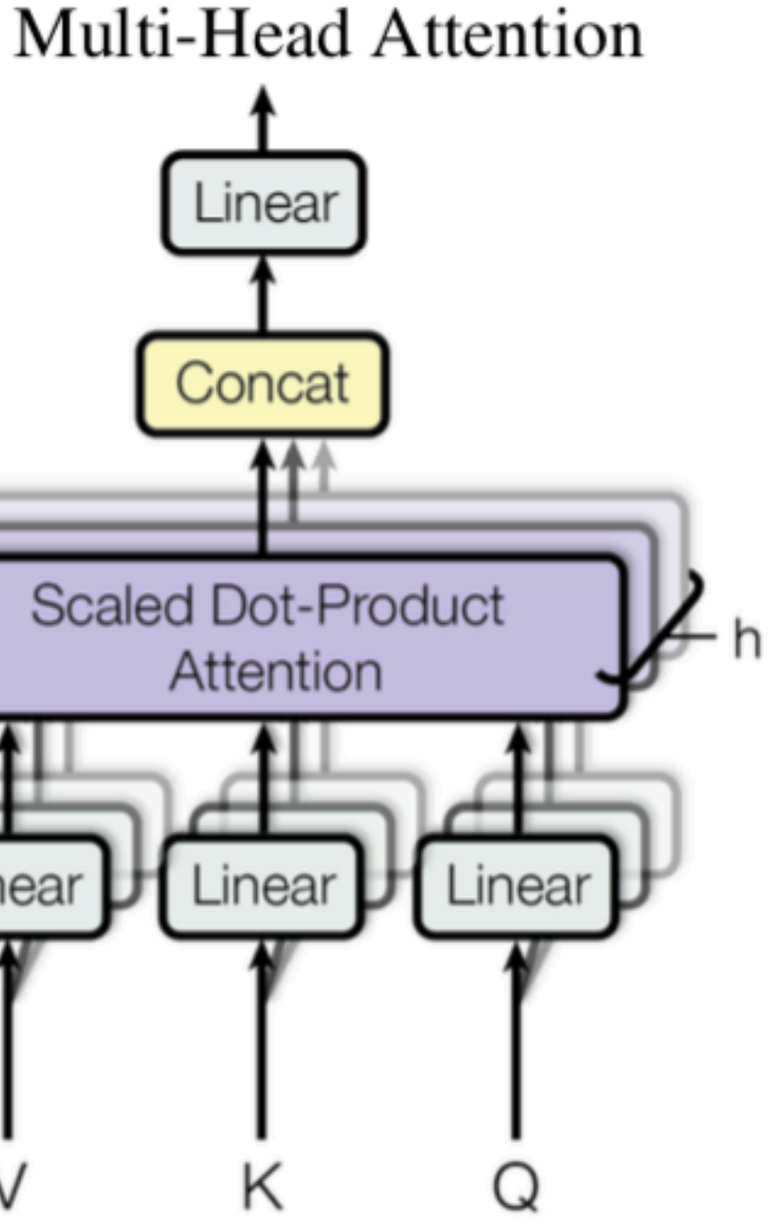

Input
Thinking
Machines
Embedding
x_1
x_2
Queries
q_1
q_2
Keys
k_1
k_2
Values
v_1
v_2
Score
$q_1 \cdot k_1 = 112$
$q_1 \cdot k_2 = 96$
Divide by 8 ($\sqrt{d_k}$)
14
12
Softmax
0.88
0.12
Softmax X Value
v_1
v_2
Sum
z_1
z_2

Multi-Head Attention

$$\text{MultiHead}(h) = \text{Concat}[head_1, \dots, head_m]W^O$$
$$\text{where } head_i = \text{Attention}(Q, K, V)$$
$$\text{where } Q, K, V = hW_i^Q, hW_i^K, hW_i^V$$

where m is the number of heads, $h \in \mathbb{R}^{d_n \times d_{\text{model}}}$ is the input, the $W_i^Q \in \mathbb{R}^{d_{\text{model}} \times d_k}, W_i^K \in \mathbb{R}^{d_{\text{model}} \times d_k}, W_i^V \in \mathbb{R}^{d_{\text{model}} \times d_v}$, and $ $W^O \in \mathbb{R}^{m*d_v \times d_{\text{model}}}$. The output of the Attention is $head_i \in \mathbb{R}^{d_n \times d_v}$ and the output of the MultiHead is $\in \mathbb{R}^{d_n \times d_{\text{model}}}$.





THX

And any questions?