Scaled-dot Product Attention

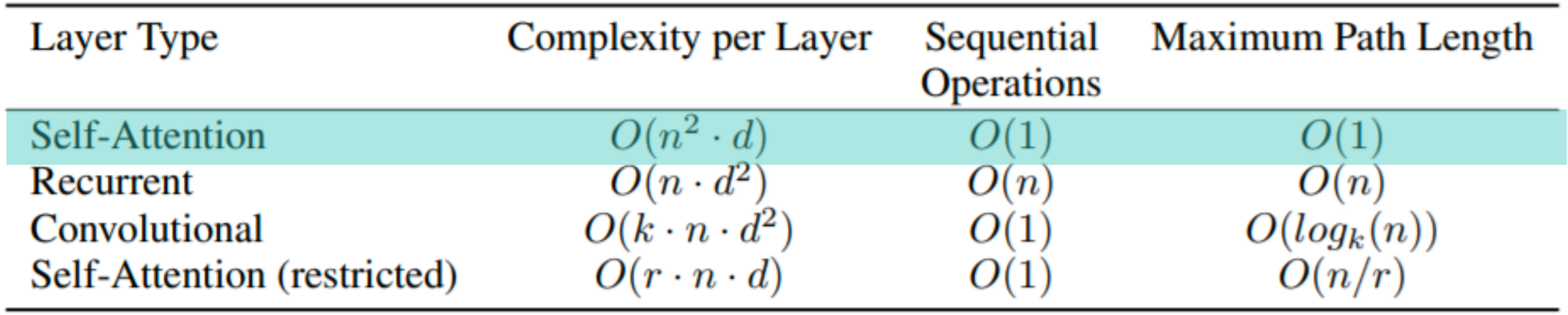
Attention in Transformer

Hyungjin Kim

Self Attention

Key concept used in Transformer

Self Attention



Layer Type	Complexity per Layer	Sequential Operations	Maximum Path Length
Self-Attention	$O(n^2 \cdot d)$	$O(1)$	$O(1)$
Recurrent	$O(n \cdot d^2)$	$O(n)$	$O(n)$
Convolutional	$O(k \cdot n \cdot d^2)$	$O(1)$	$O(log_k(n))$
Self-Attention (restricted)	$O(r \cdot n \cdot d)$	$O(1)$	$O(n/r)$

Scaled-dot Product Attention

Model Architecture

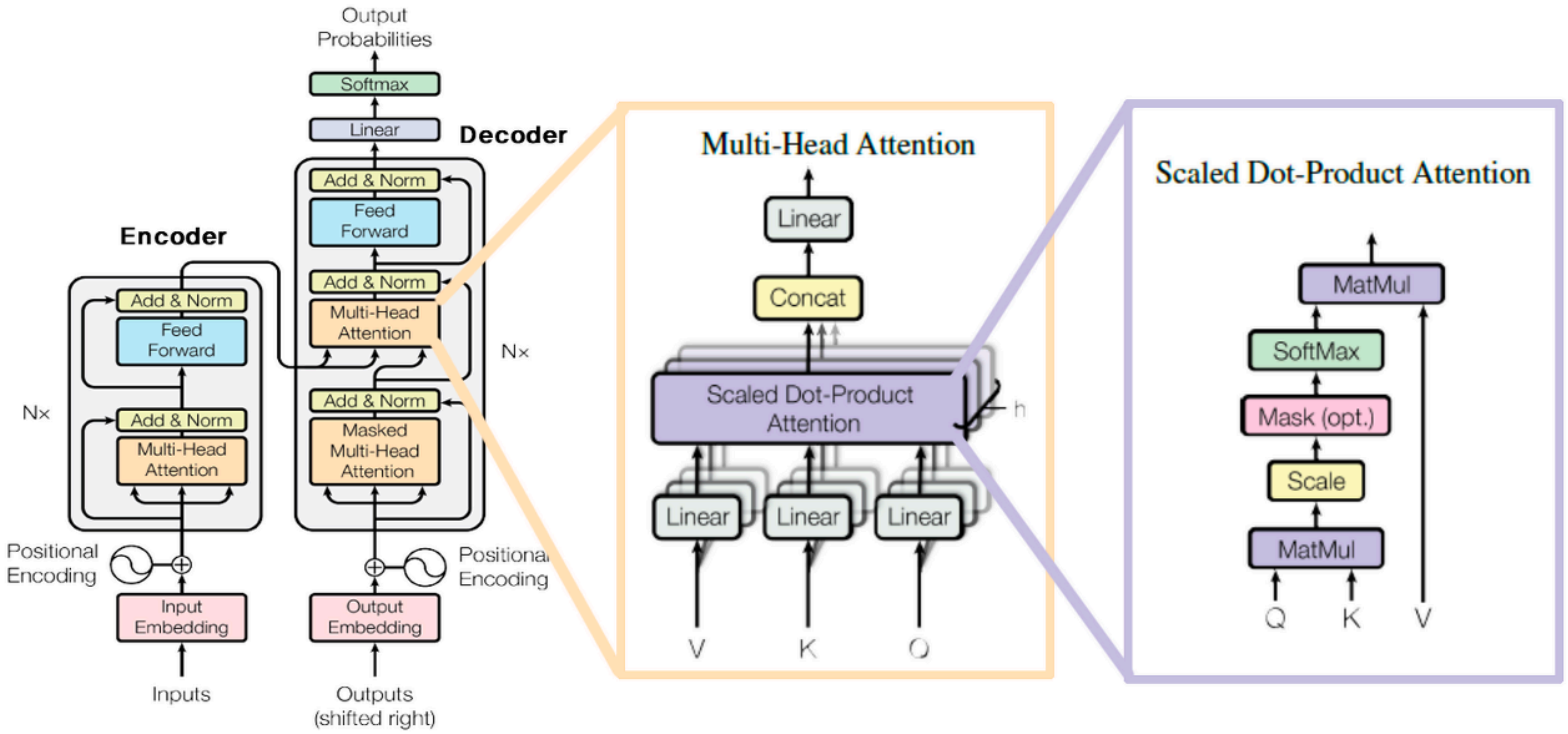
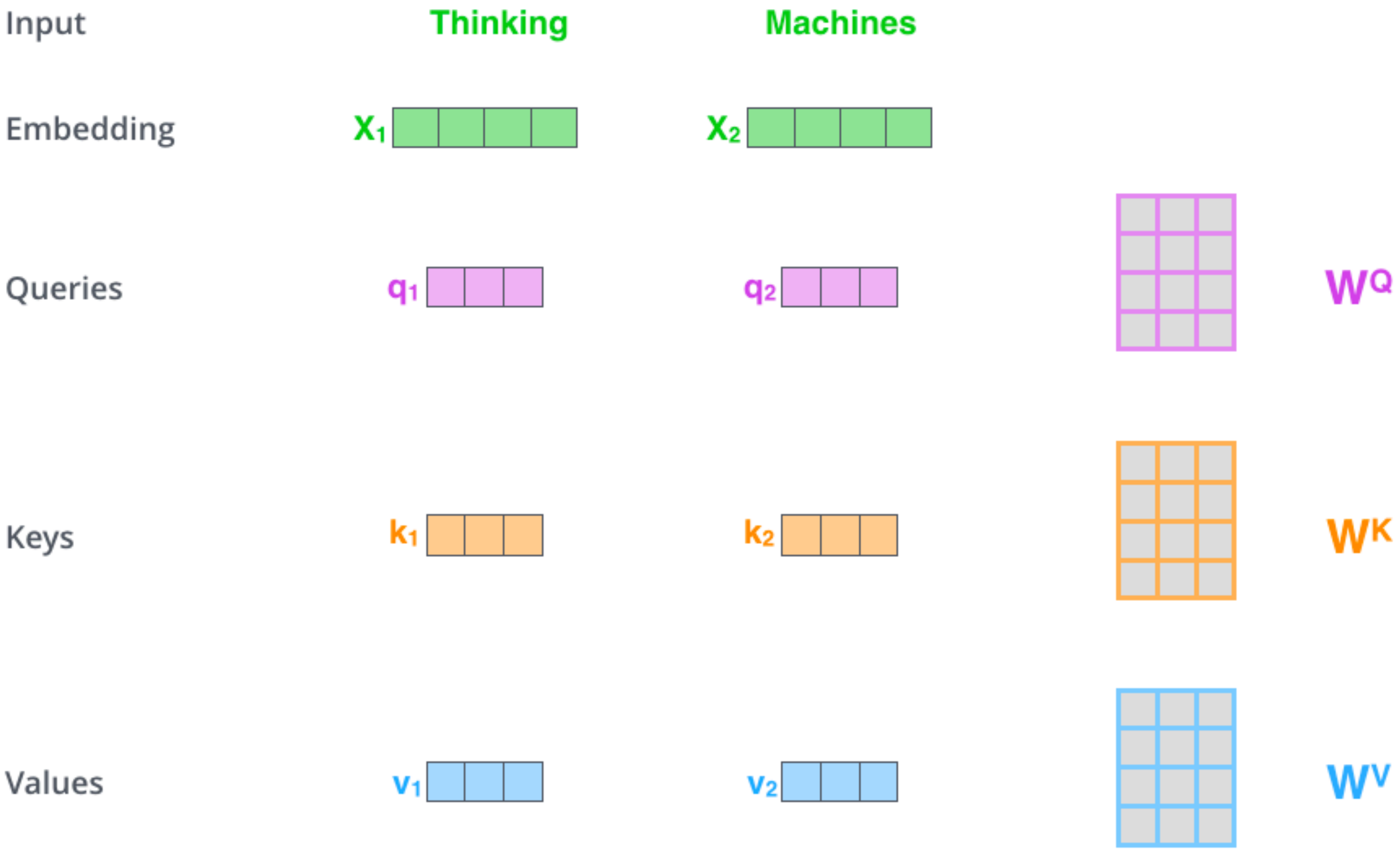


Figure 1: The Transformer - model architecture.

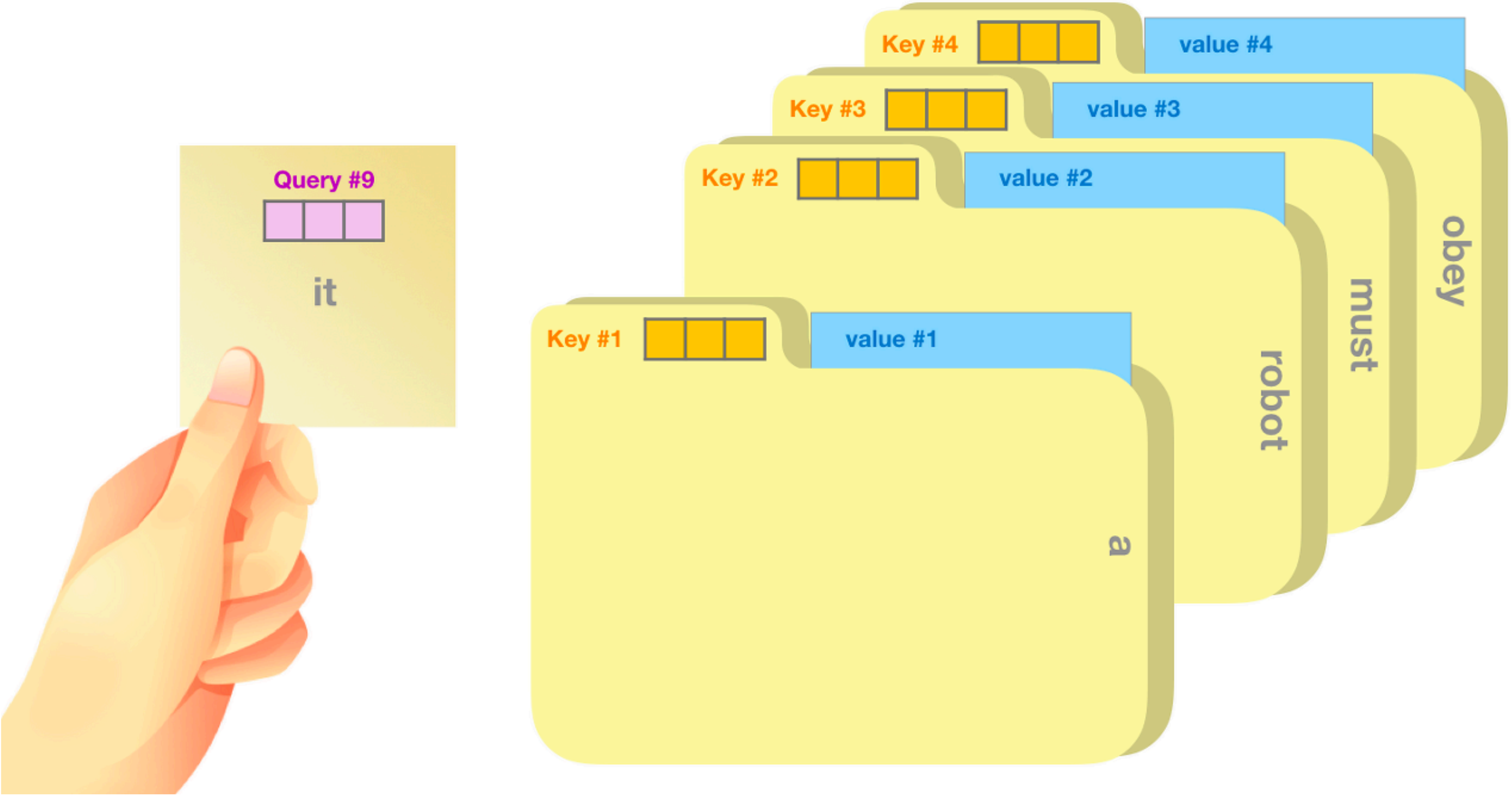
Scaled-dot Product Attention

Step by Step



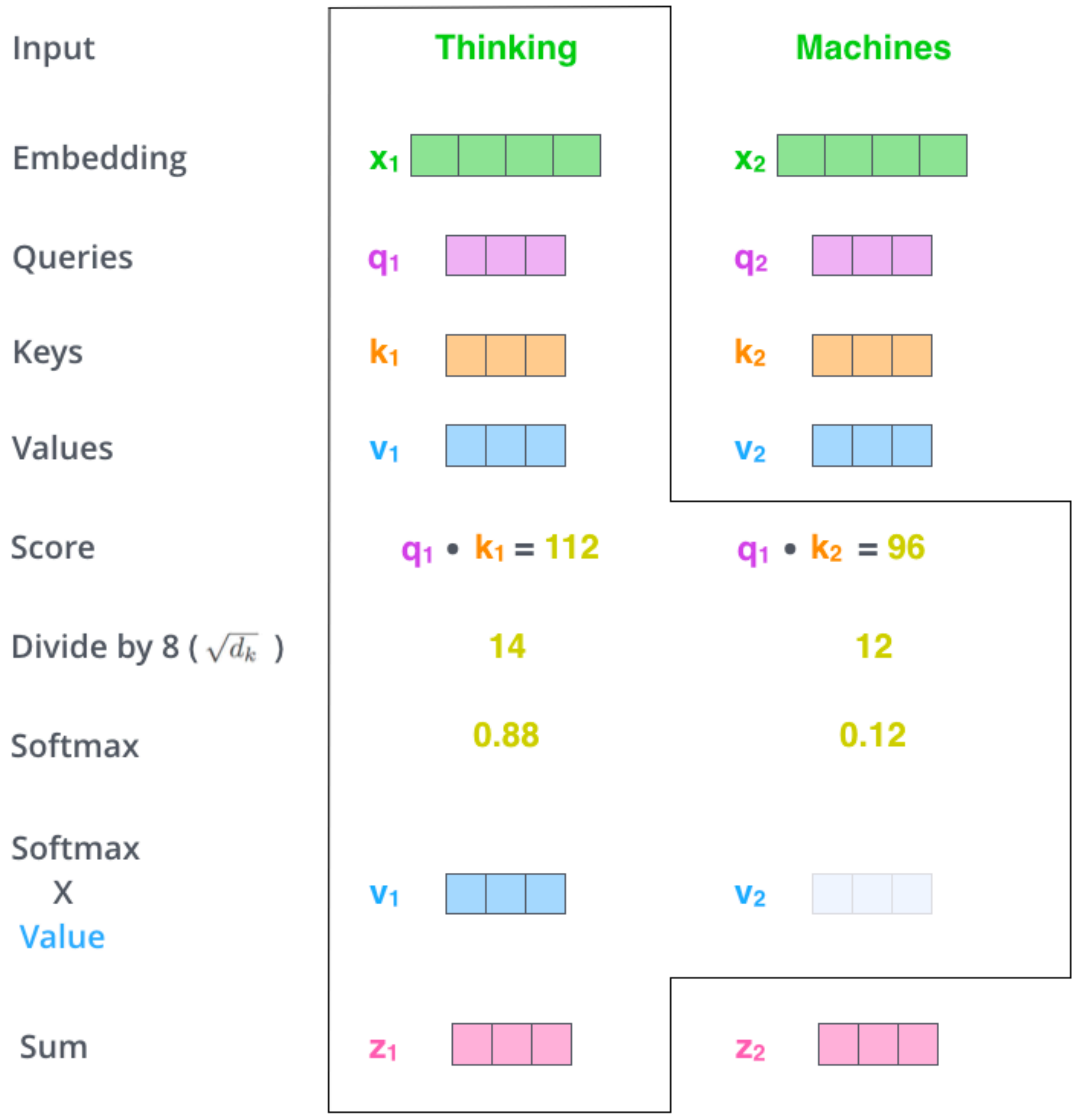
Scaled-dot Product Attention

Step by Step



Scaled-dot Product Attention

Step by Step



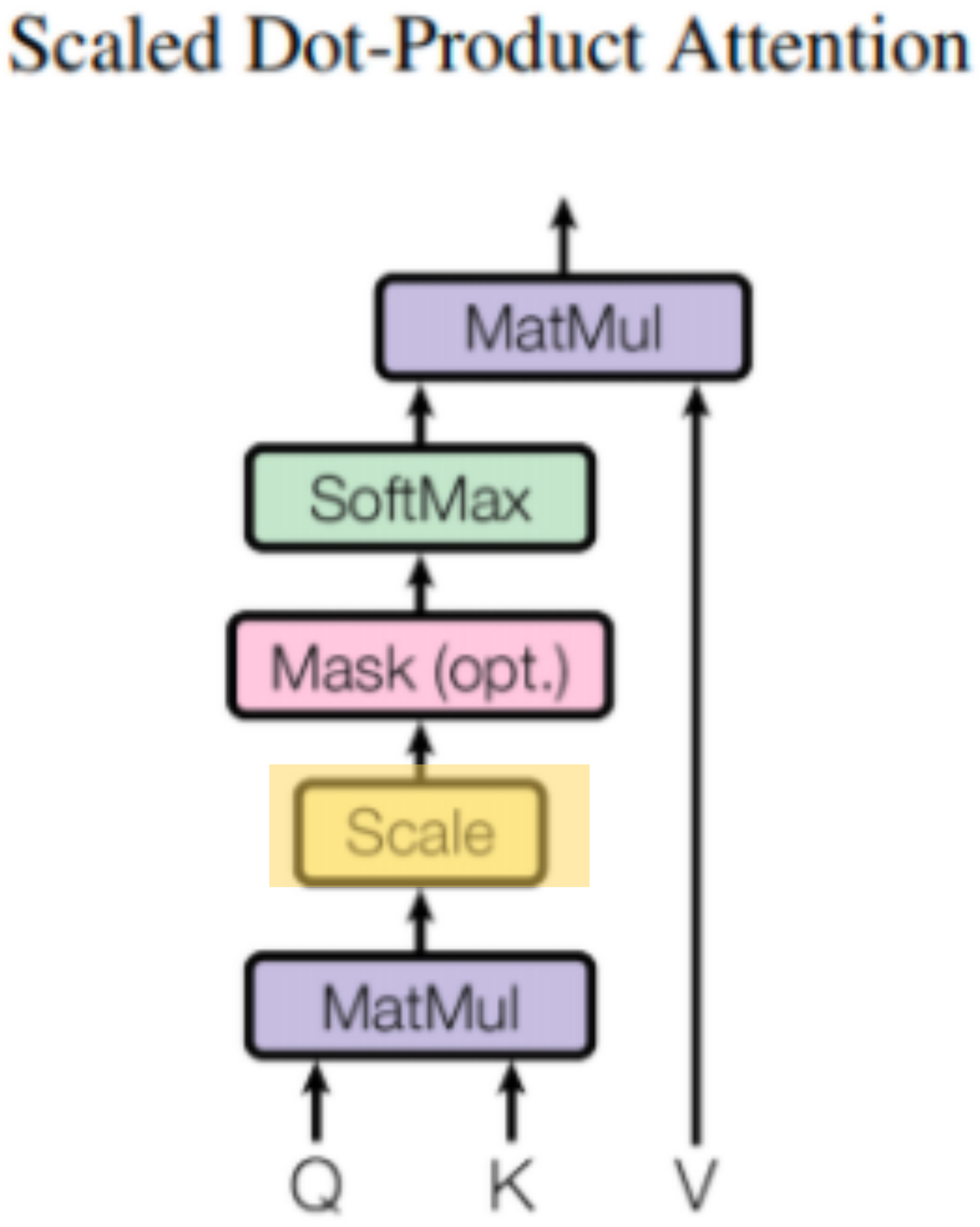
Input	Thinking	Machines
Embedding	x_1	x_2
Queries	q_1	q_2
Keys	k_1	k_2
Values	v_1	v_2
Score	$q_1 \bullet k_1 = 112$	$q_1 \bullet k_2 = 96$
Divide by 8 ($\sqrt{d_k}$)	14	12
Softmax	0.88	0.12
Softmax X Value	v_1	v_2
Sum	z_1	z_2

Scaled-dot Product Attention

How it works?

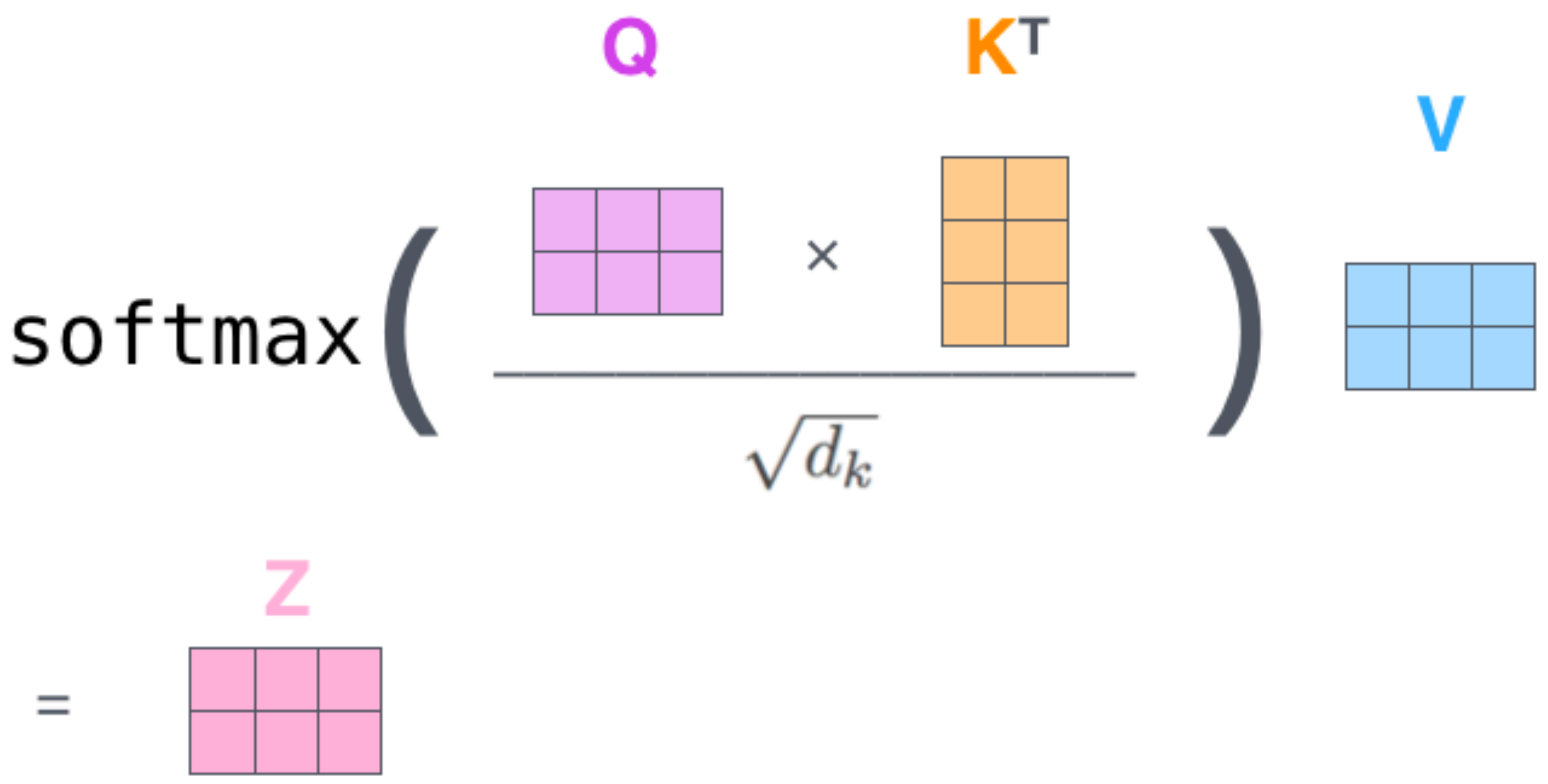
$$Attention(Q,K,V) = softmax(\frac{QK}{\sqrt{d_k}})V$$

$$Q = XW_q, V = XW_k, V = XW_v \qquad W_q, W_k, W_v \in \mathbb{R}^{d \times d_k}$$



Scaled-dot Product Attention

Step by Step



References

- https://welcome-to-dewy-world.tistory.com/108

- https://towardsdatascience.com/illustrated-self-attention-2d627e33b20a

- https://jalammar.github.io/illustrated-transformer/

- “Attention is all you need”
  Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł. & Polosukhin, I.
- (2017). Attention is all you need. *Advances in Neural Information Processing Systems* (p./pp. 5998--6008), .

E.O.D