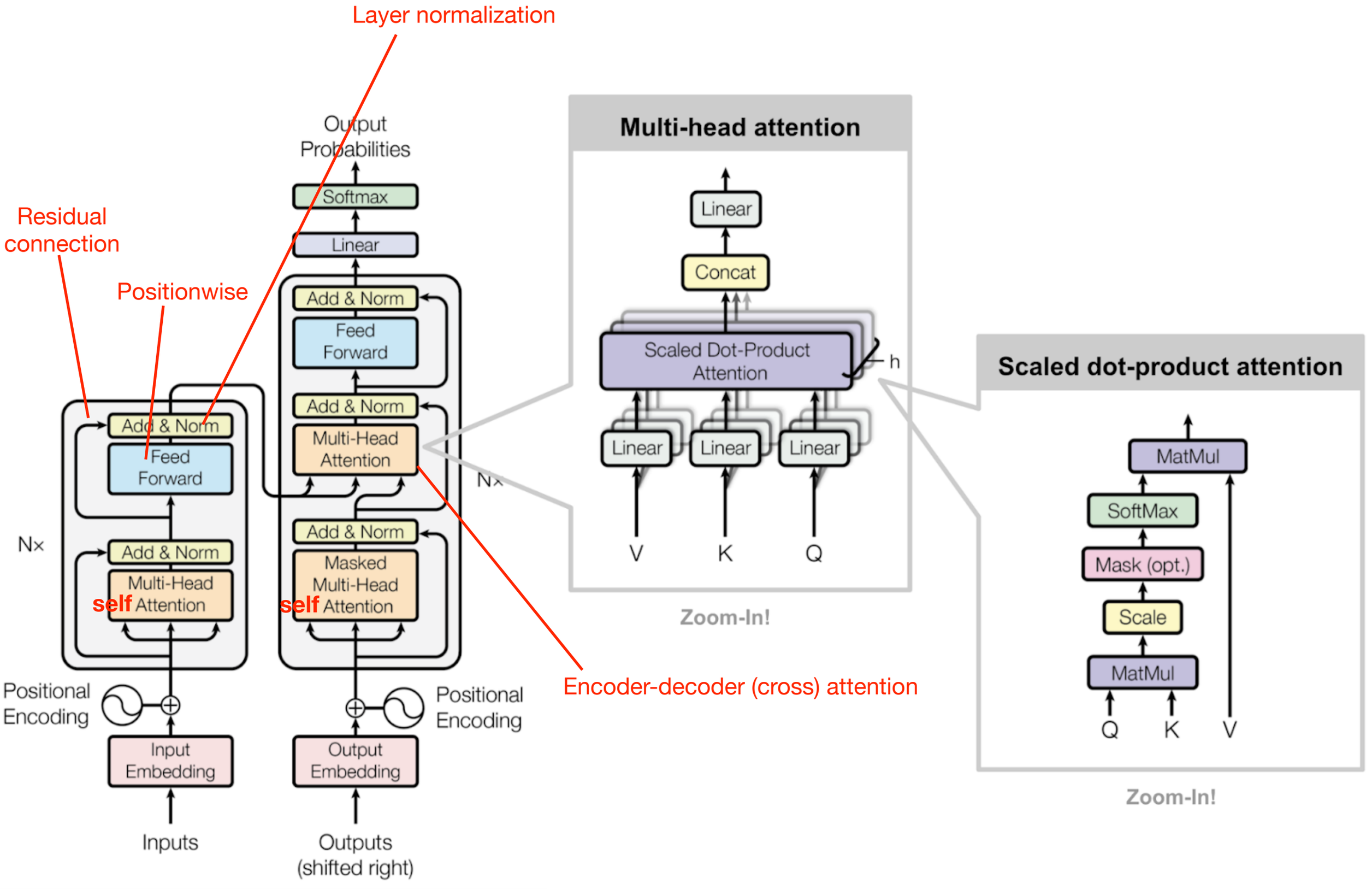
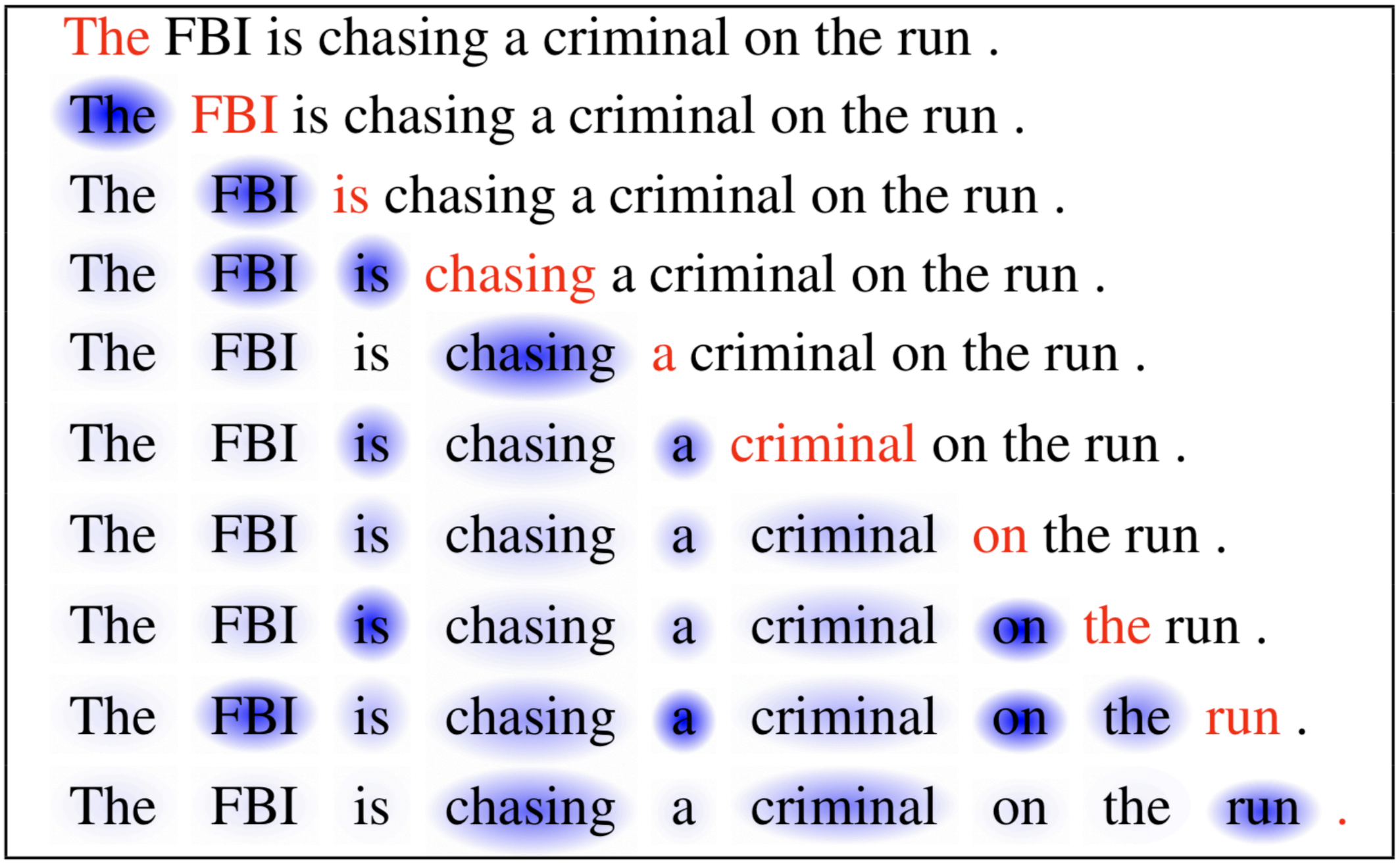
Attention and Transformer (Big Picture)



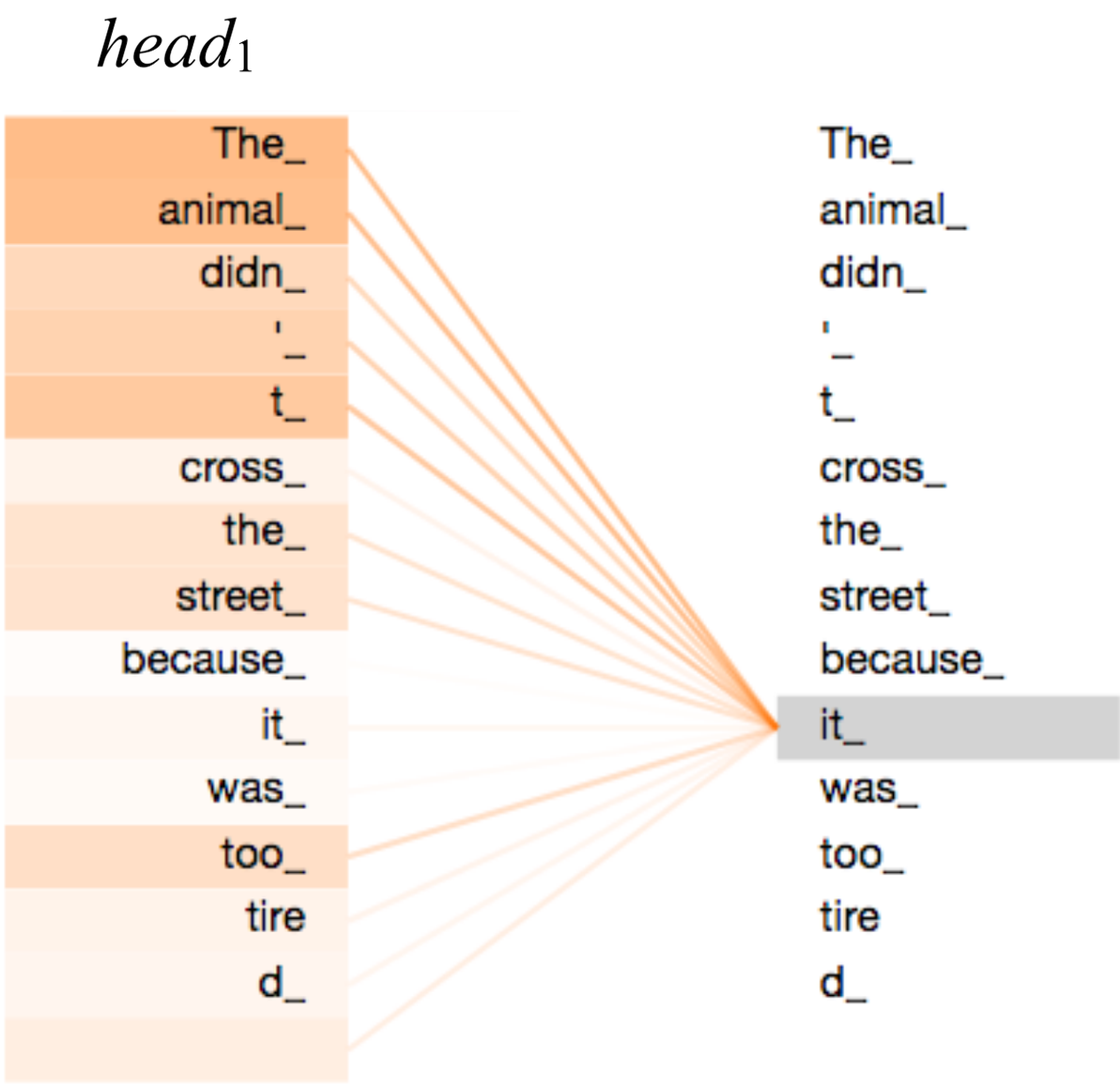
Vaswani et al. “Attention is all you need” 2017

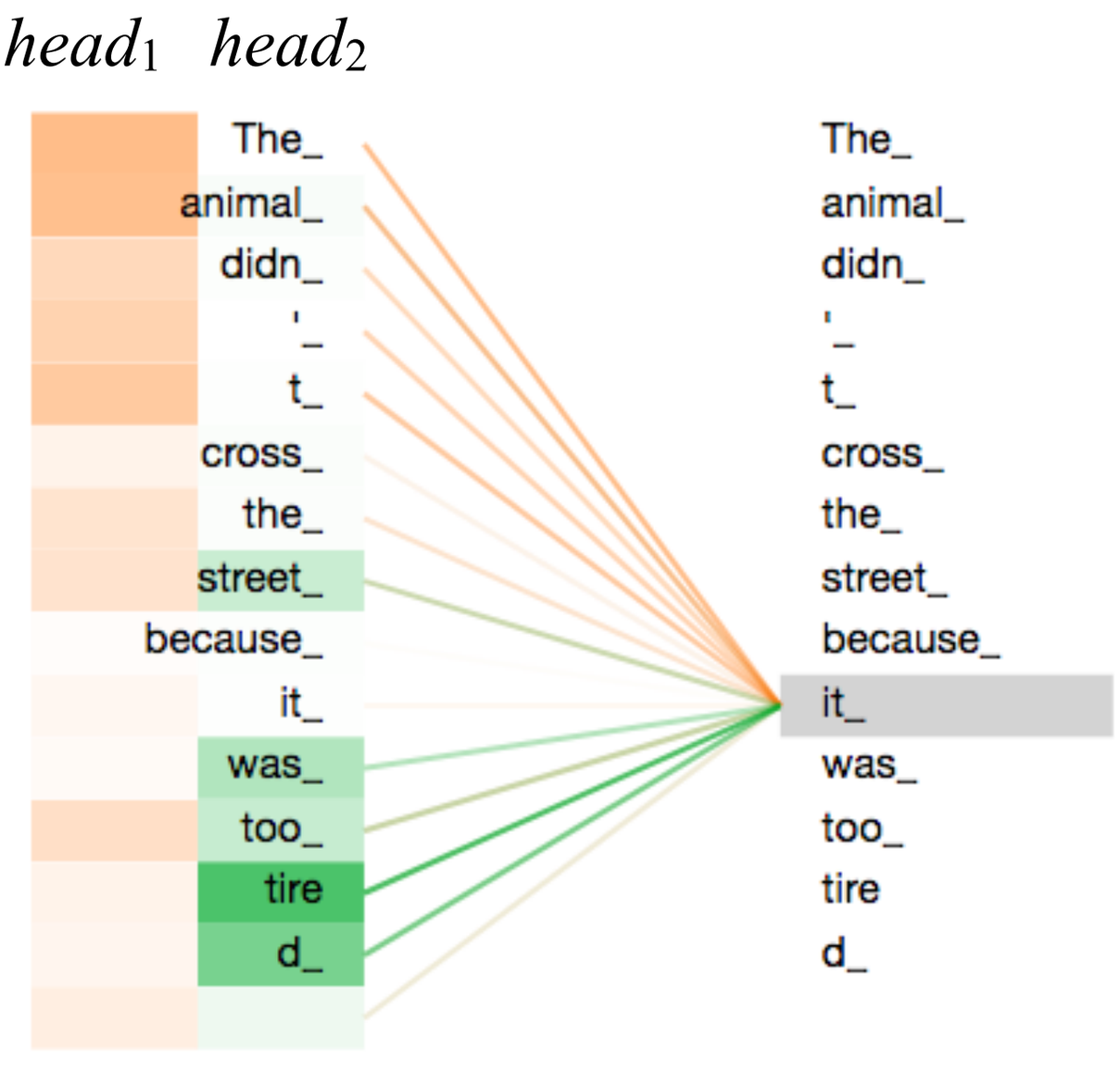
An Illustration of Self-attention

The current word (Query) is in red, the size of blue shade indicates the attention weight.



Attention and Multi-head Attention (Intuition)





Positional Encoding

X & P : Seq x Emb matrices

P :

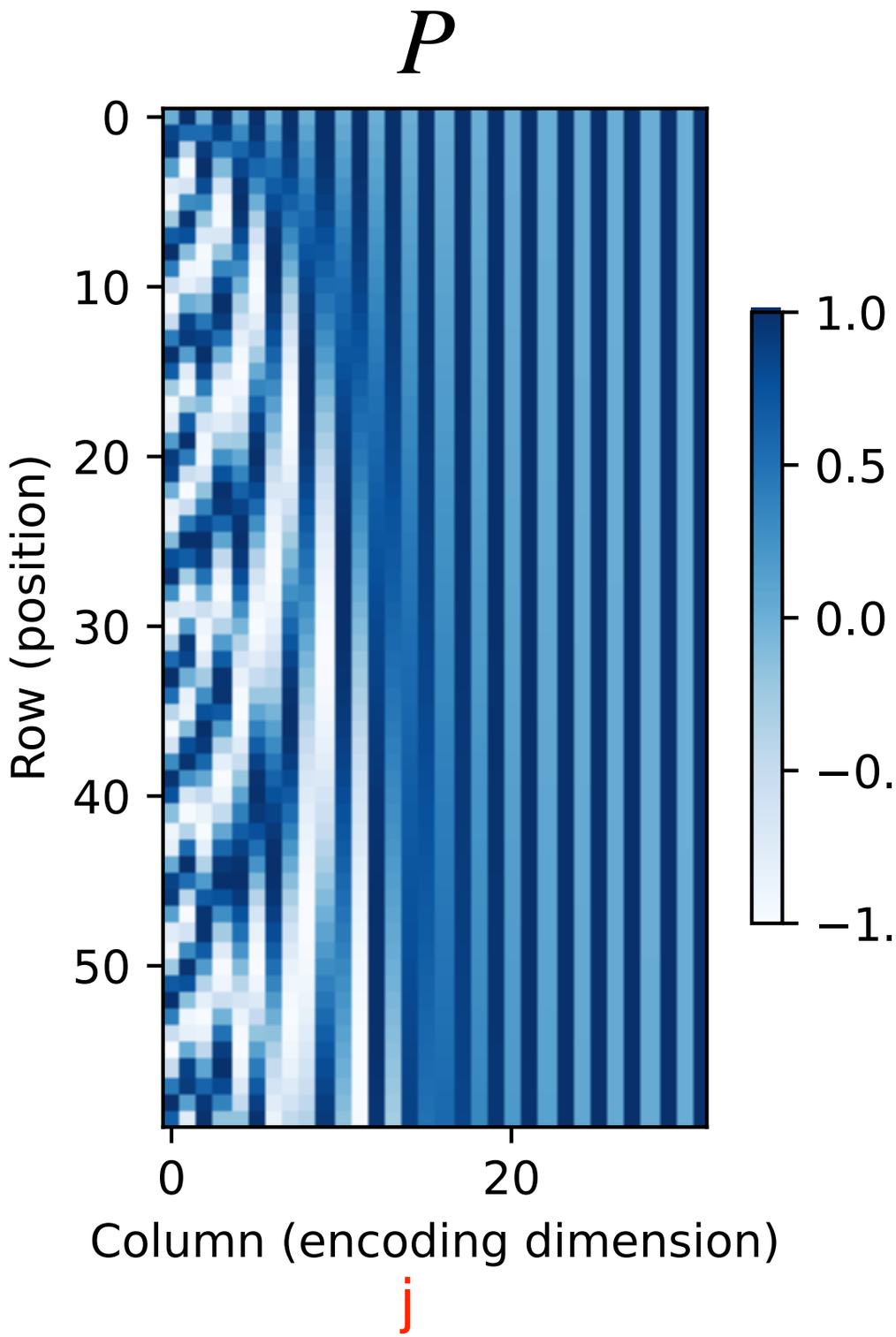
Positional encoding:

$$p_{i,2j} = \sin\left(\frac{i}{10000^{2j/d}}\right),$$

$$p_{i,2j+1} = \cos\left(\frac{i}{10000^{2j/d}}\right).$$

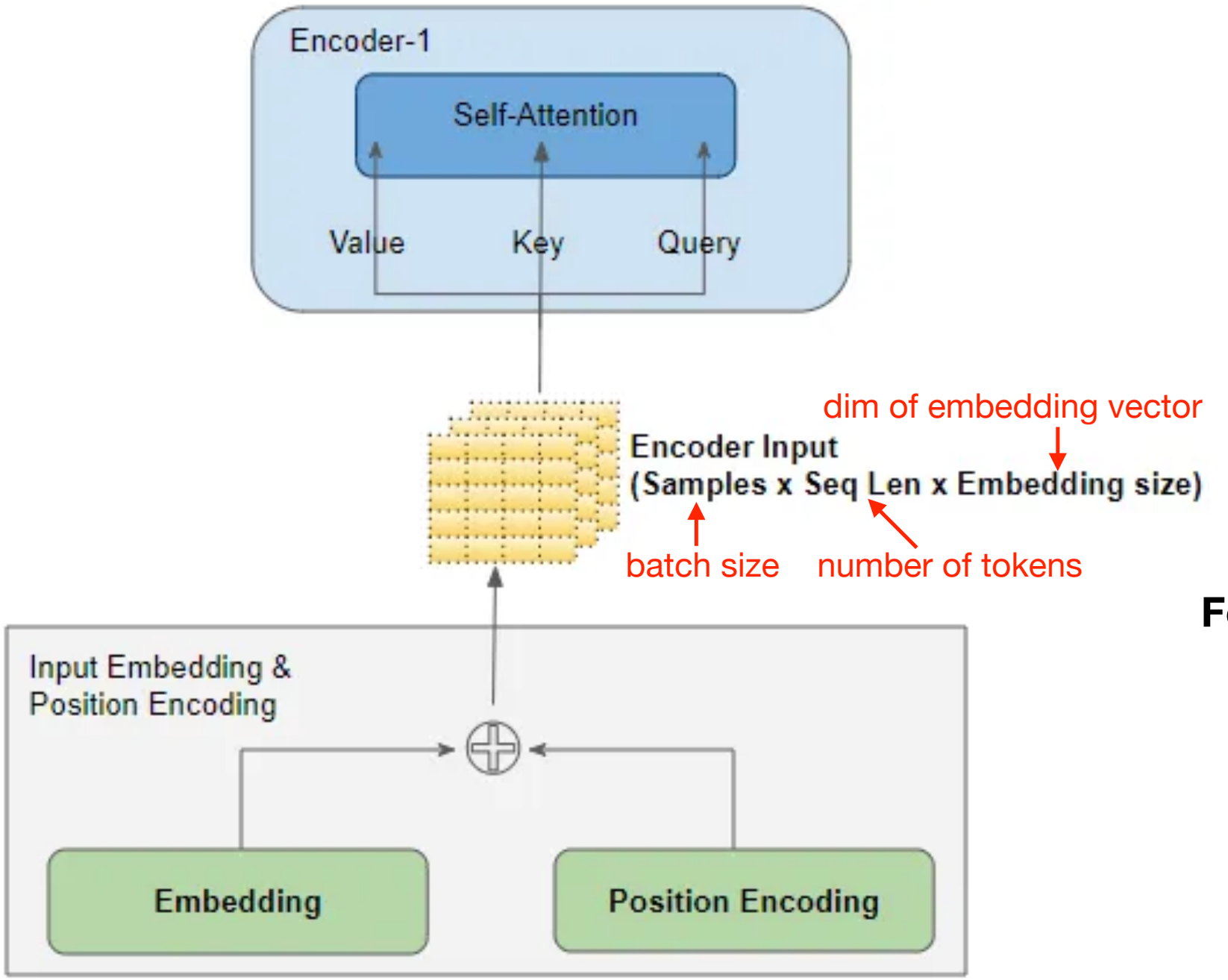
Binary encoding:

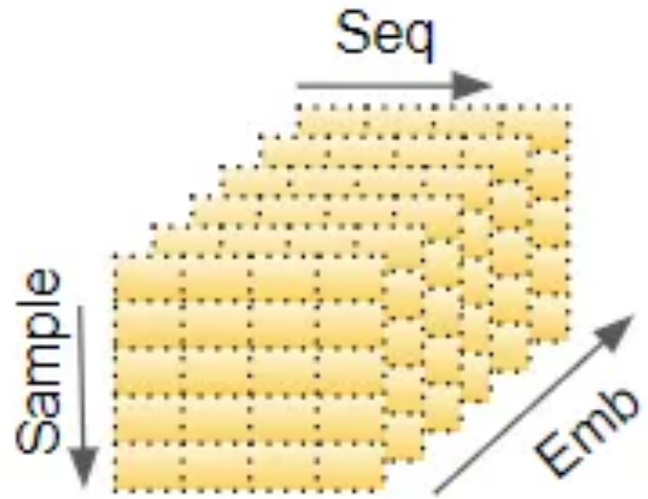
```
0 in binary is 000
1 in binary is 001
2 in binary is 010
3 in binary is 011
4 in binary is 100
5 in binary is 101
6 in binary is 110
7 in binary is 111
```



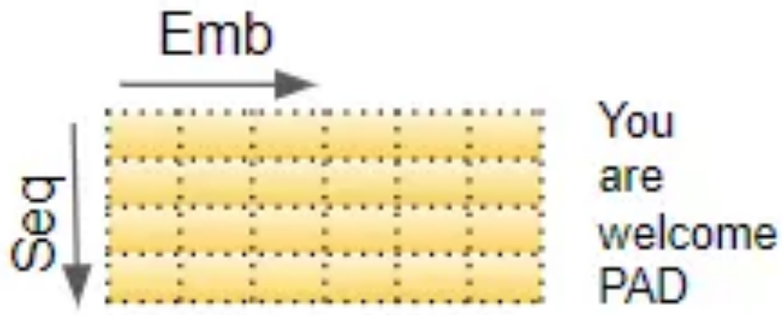
“Dive into DL” Sec 11.6.3

Multi-head Self Attention (Input Layer)





For clarity, consider batch size = 1

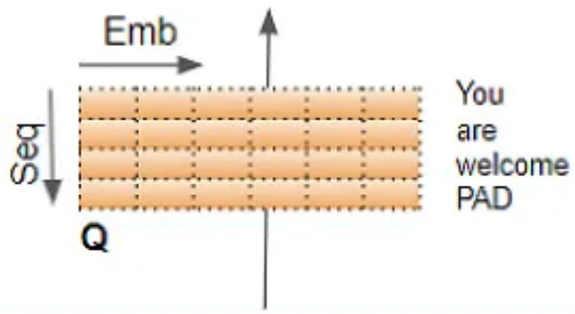


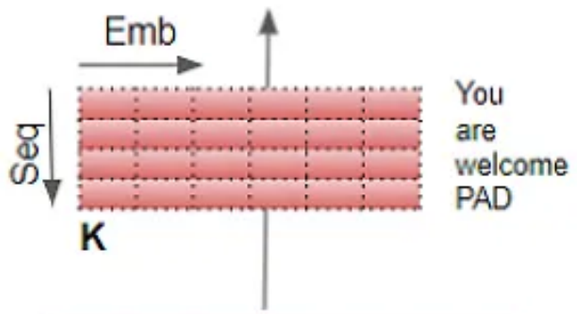
Doshi, “Transformers explained visually” Toward Data Science

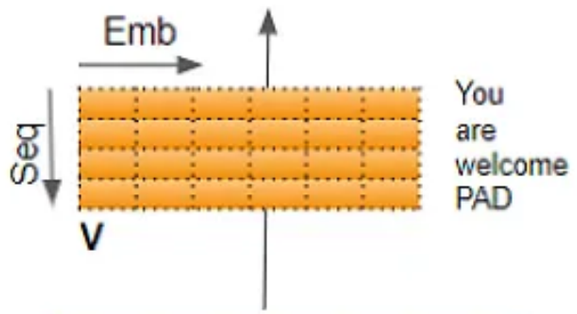
Multi-head Self Attention (Linear Layer)

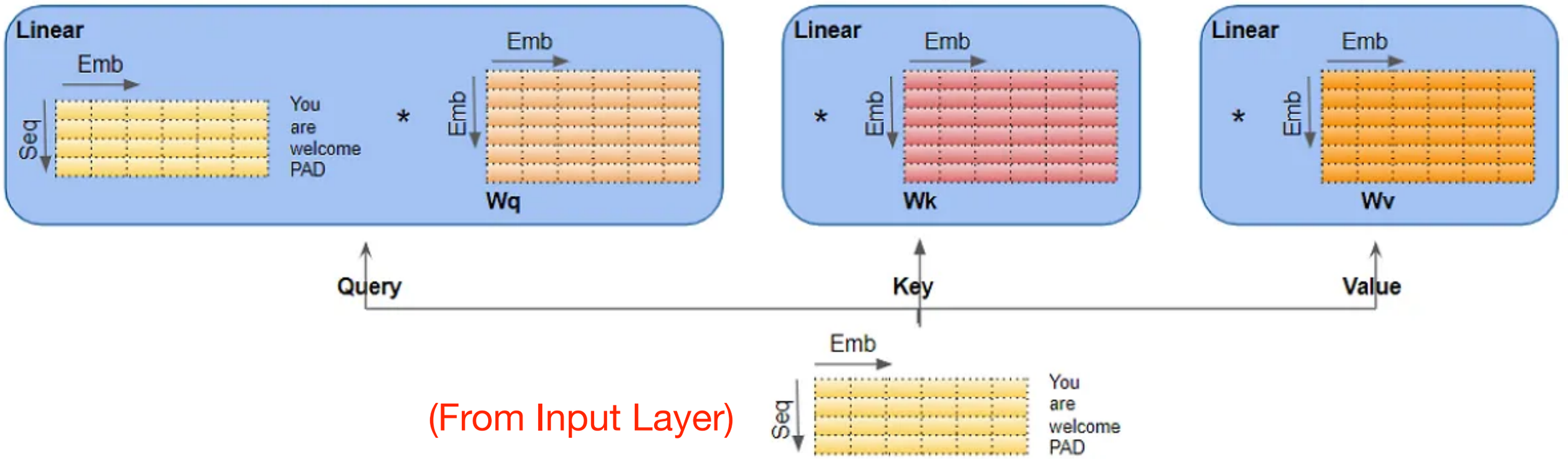
Remarks:

1) At this stage, Wq, Wk & Wv haven't mixed tokens yet
2) Wq, Wk & Wv mix up the embedding dimensions
3) No multi-head yet

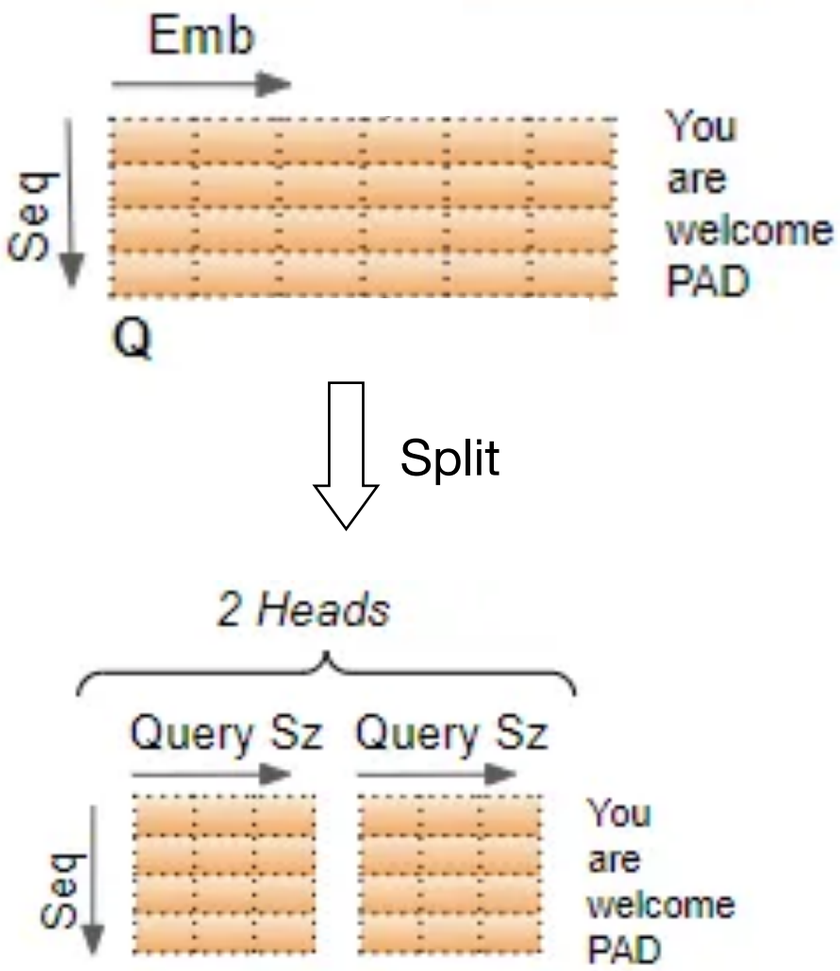








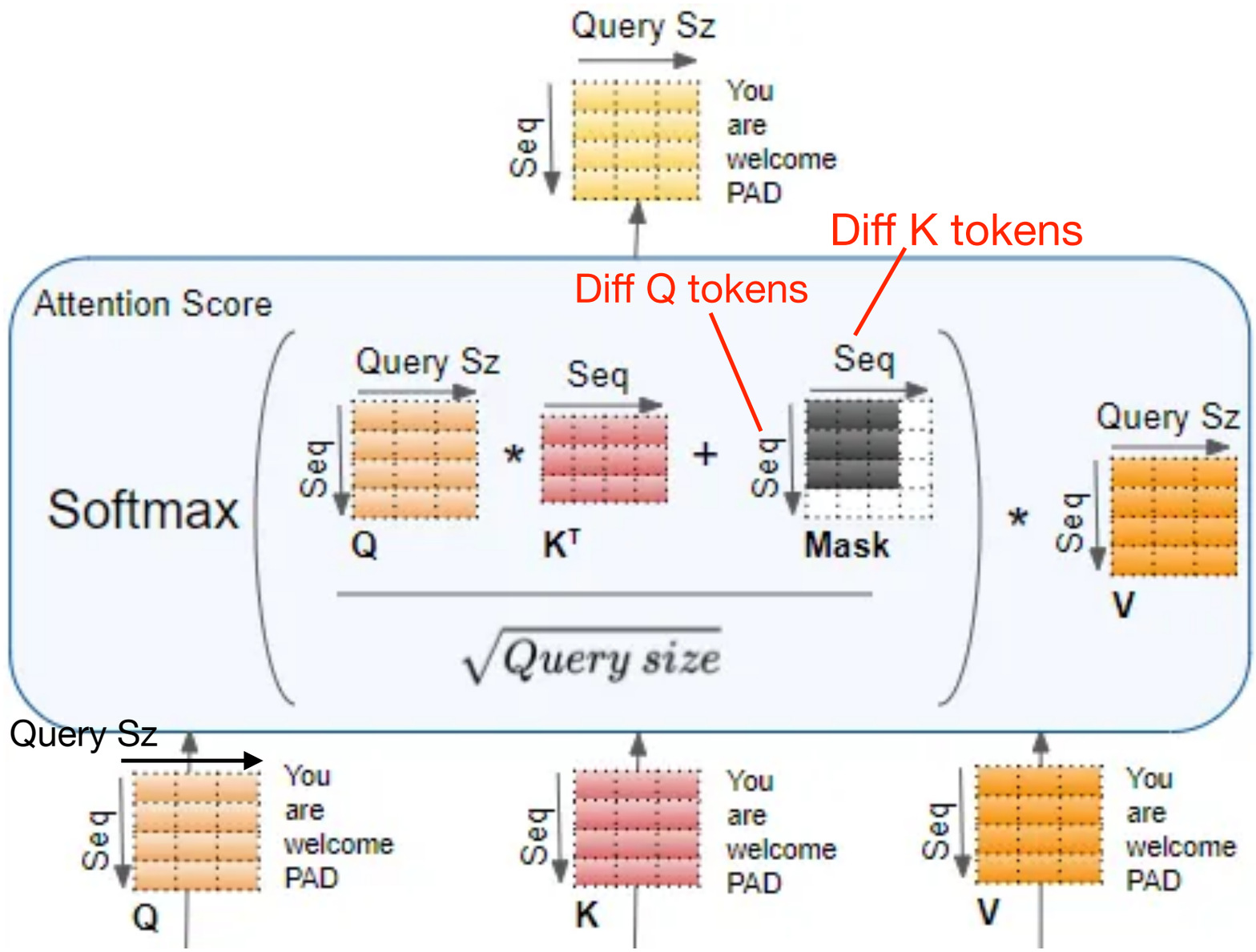
Multi-head Self Attention



Similar for K & V

Single head

A Single Head Attention Score



PAD: Padding tokens

Multi-head Self Attention (Summary)