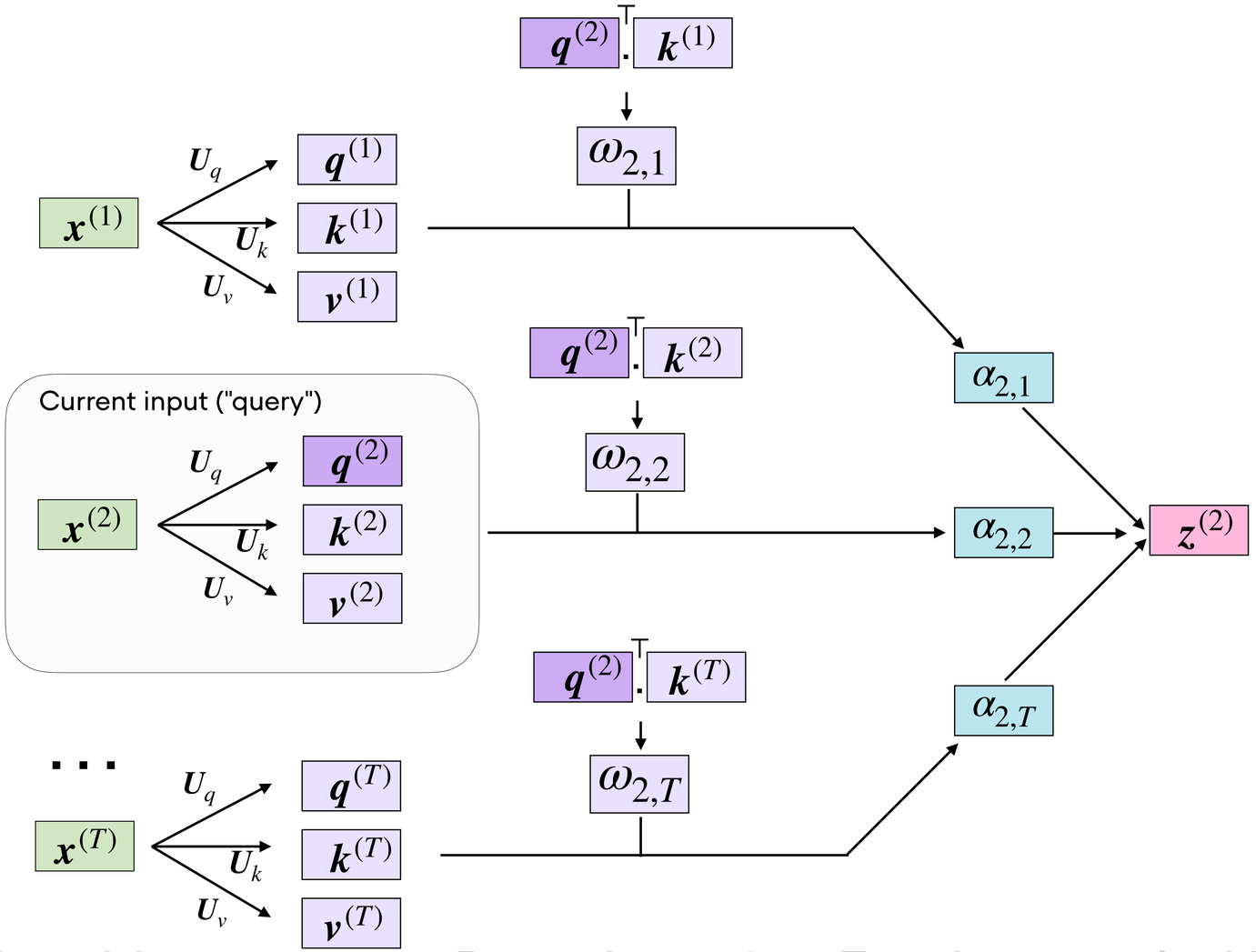
8.5

Understanding Self-Attention

Part 3: From Self-Attention to Multi-Head Attention

Sebastian Raschka and the Lightning AI Team



where $z^{(2)} = \sum_{j=1}^{T} \alpha_{2,j} v^{(j)}$

This scaled dot-product attention mechanism is used in the multi-head attention blocks

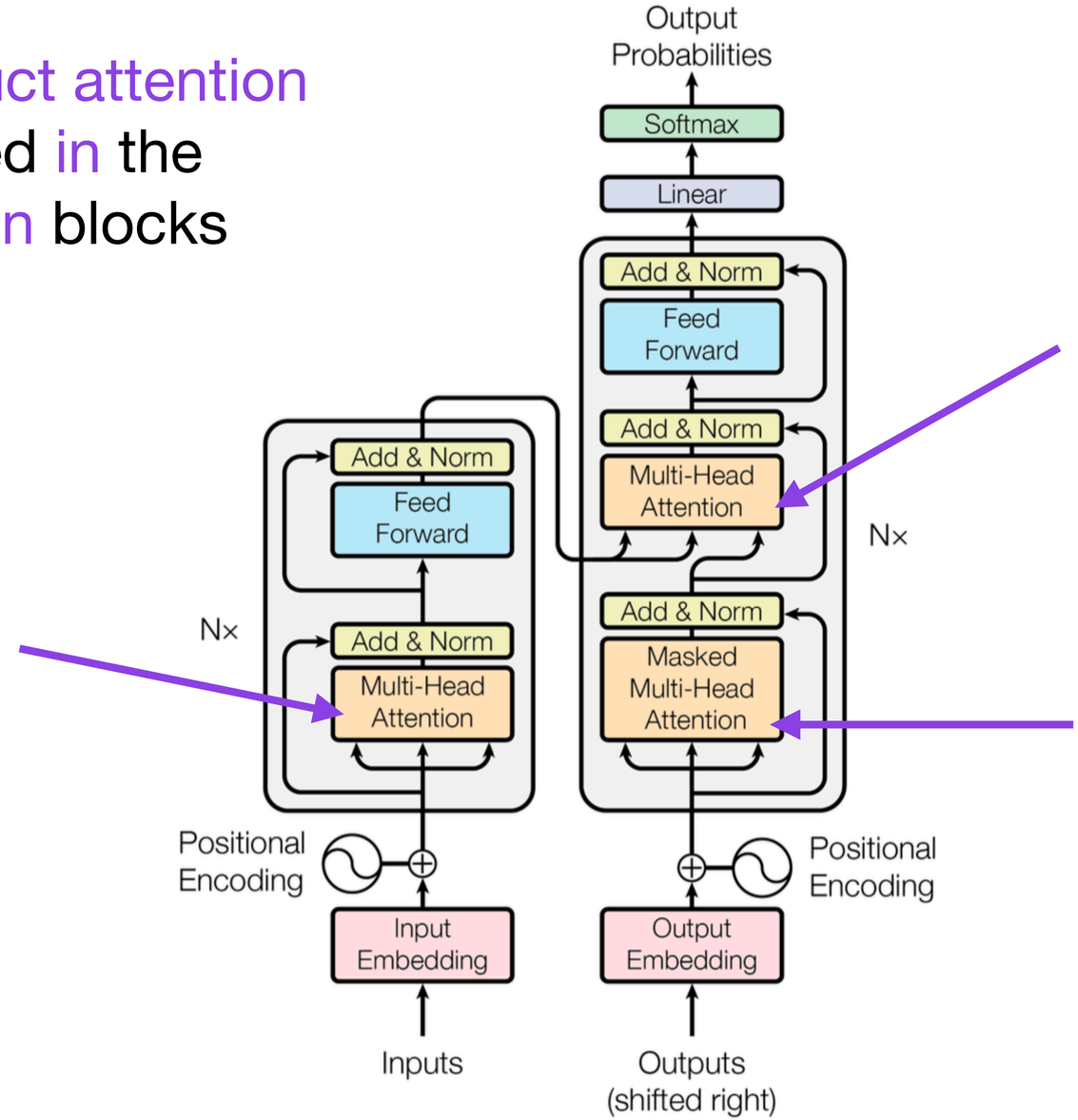
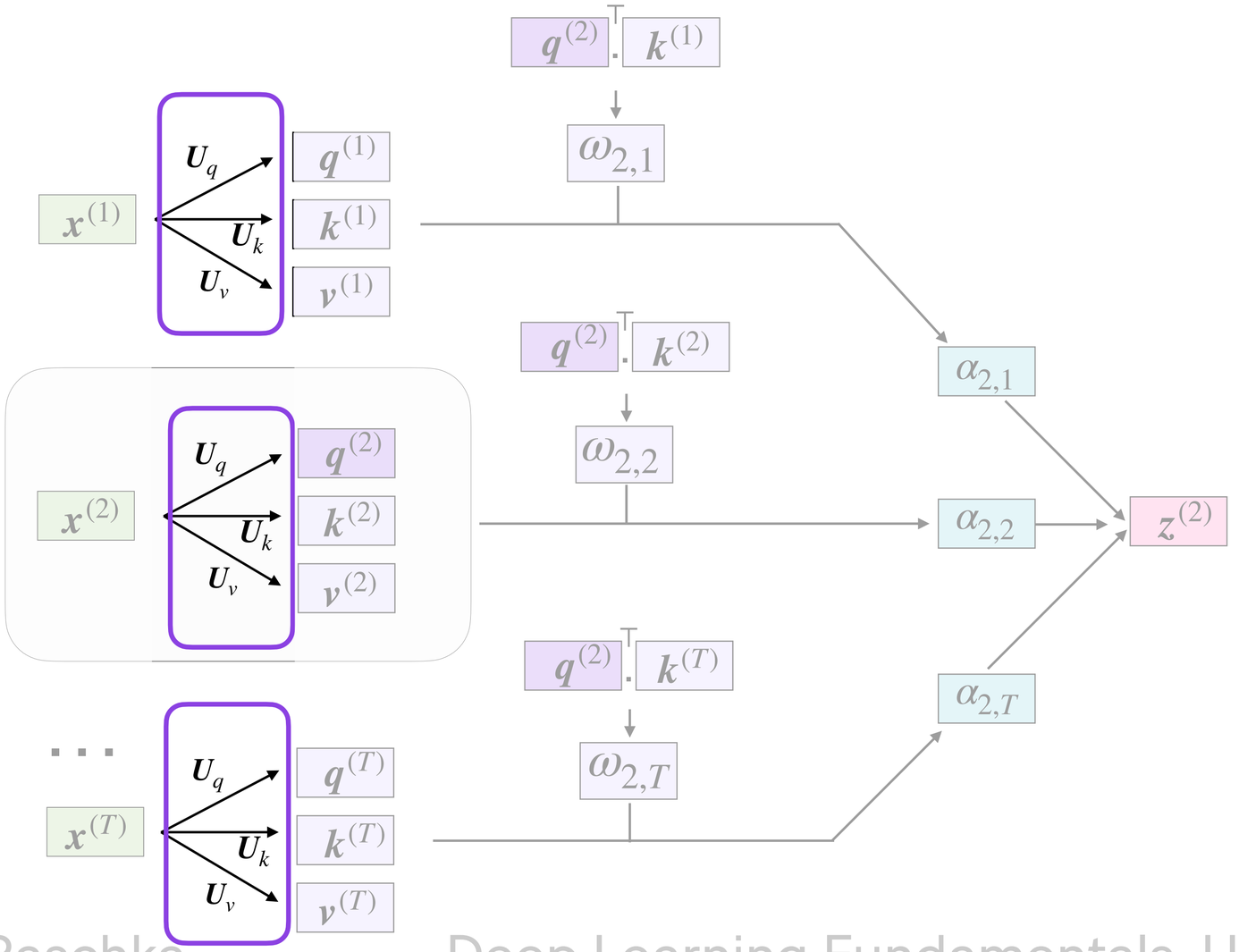


Figure 1: The Transformer - model architecture.

Previously, we used a set of 3 matrices

$$U_q \quad U_k \quad U_v$$

We use the same matrices U_q, U_k, U_v here



where $z^{(2)} = \sum_{j=1}^{T} \alpha_{2,j} \, v^{(j)}$

Previously, we used a set of 3 matrices

Let's add the index "1" U_{q_1} U_{k_1} U_{v_1}

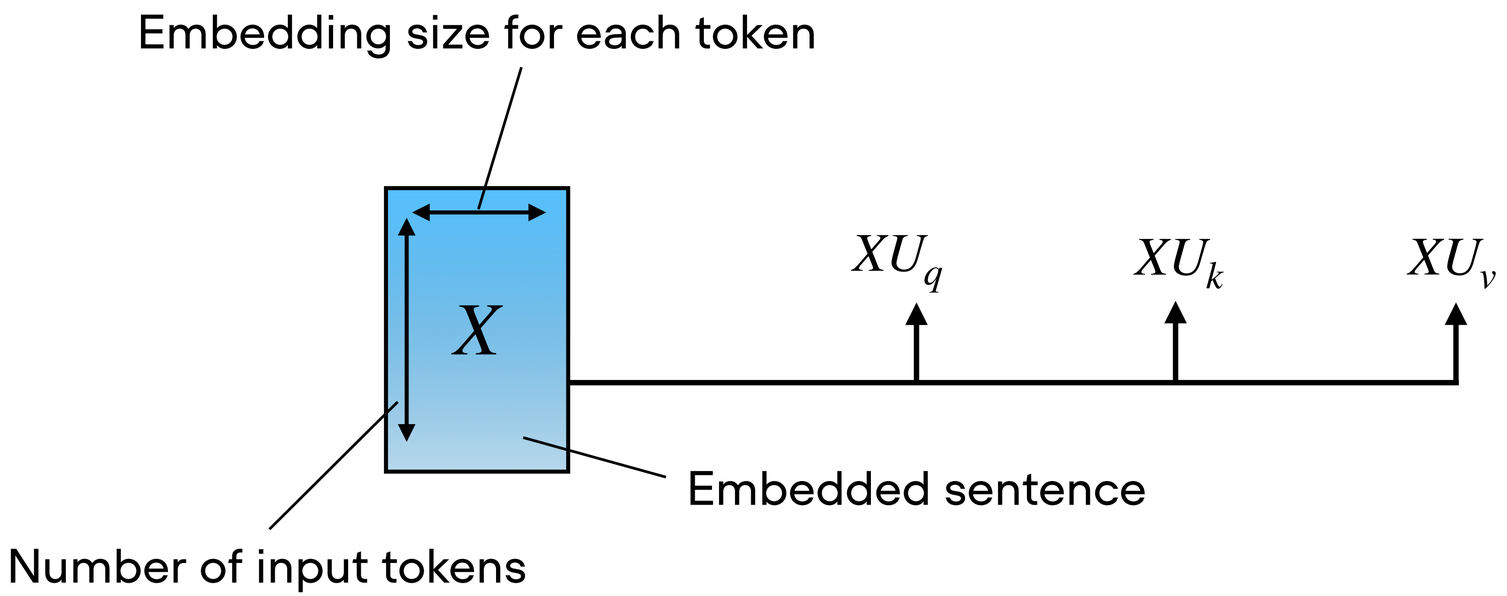
Previously, we used a set of 3 matrices

Let's add the index "1" $U_{q_1} \quad U_{k_1} \quad U_{v_1}$

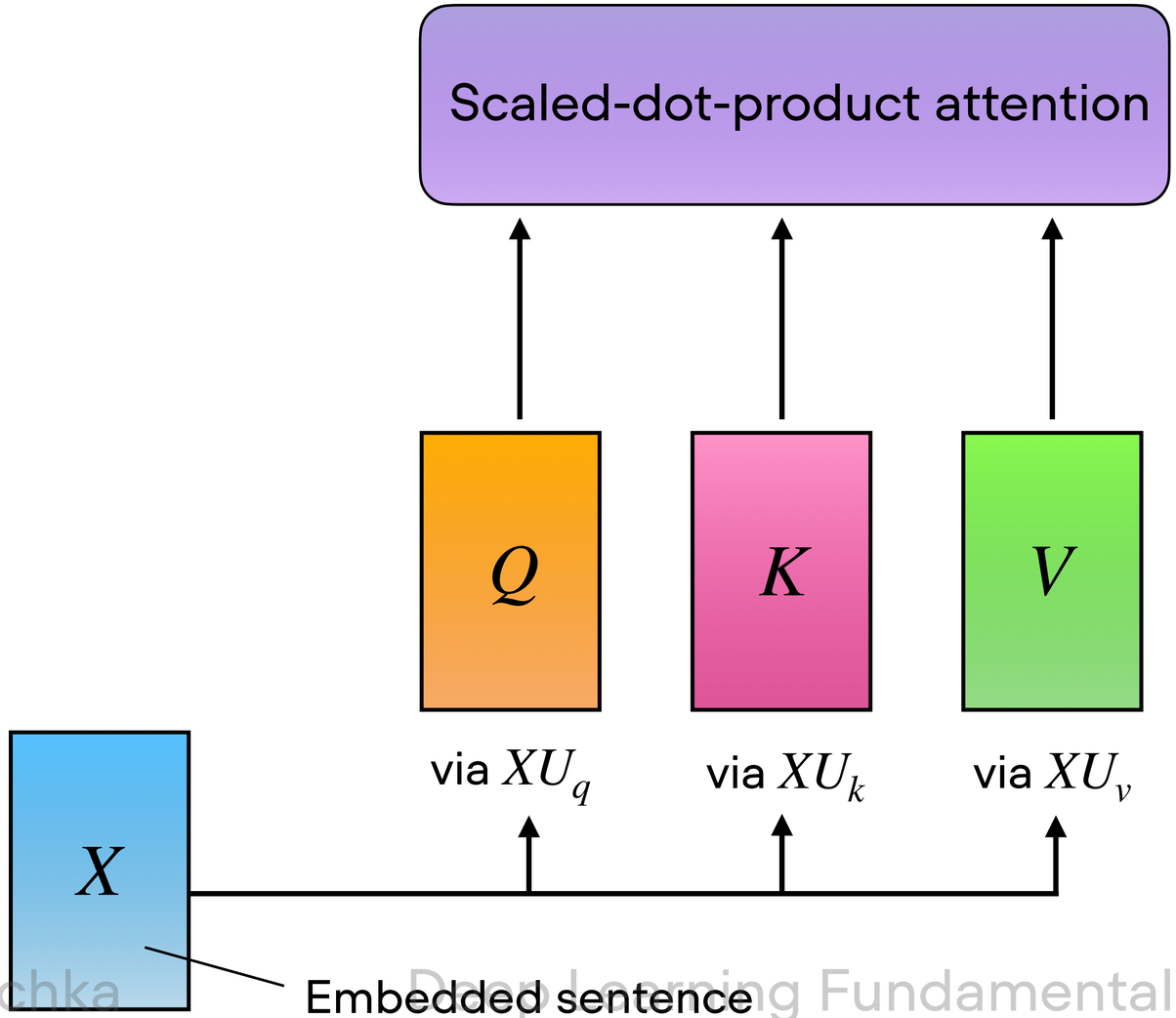
In multi-head attention, we have stack of h matrices

$$\begin{matrix} U_{q_2} & U_{k_2} & U_{v_2} \\ U_{q_3} & U_{k_3} & U_{v_3} \\ & \dots & \\ U_{q_h} & U_{k_h} & U_{v_h} \end{matrix}$$

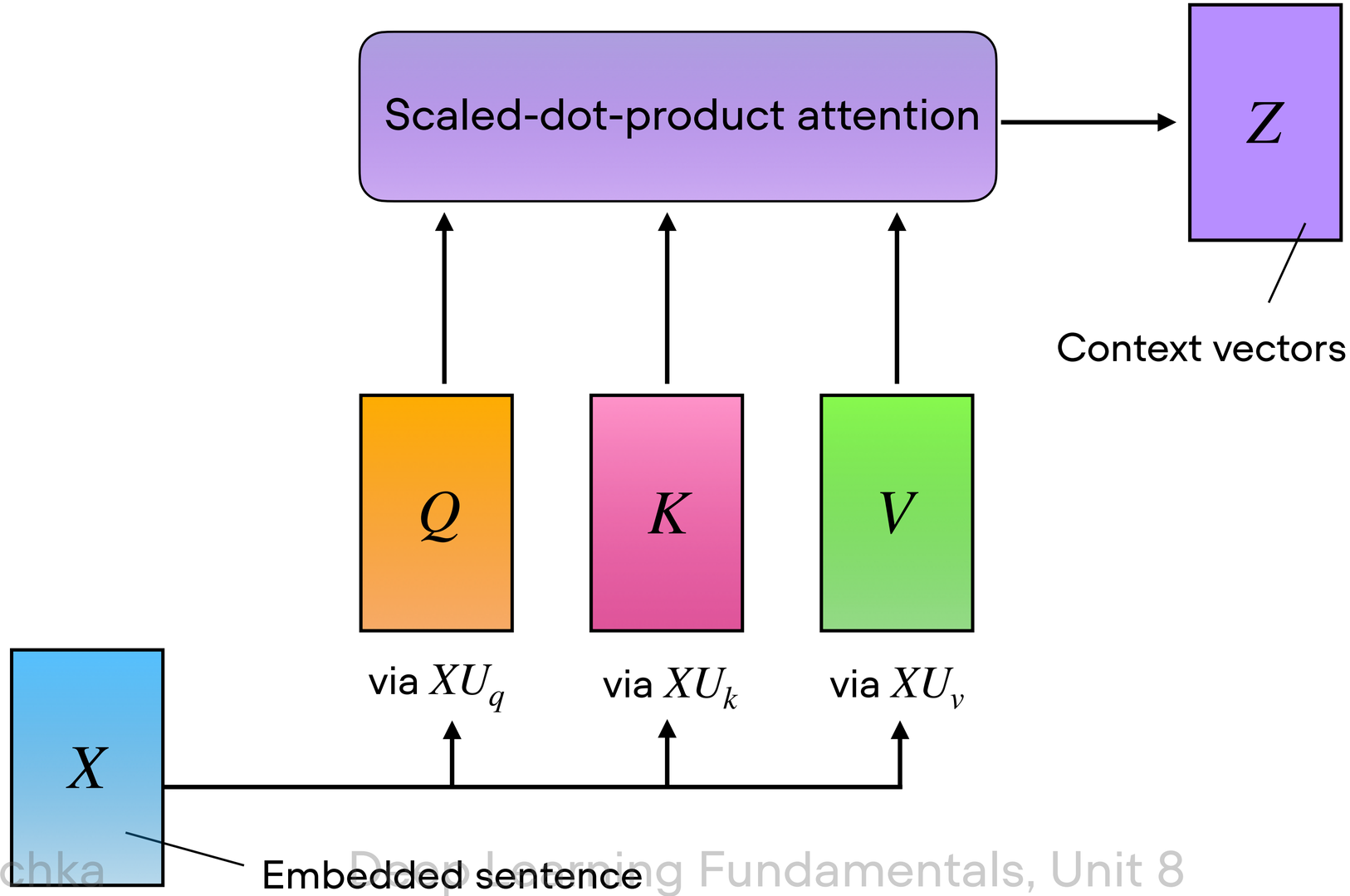
Recap: single-head attention



Recap: single-head attention

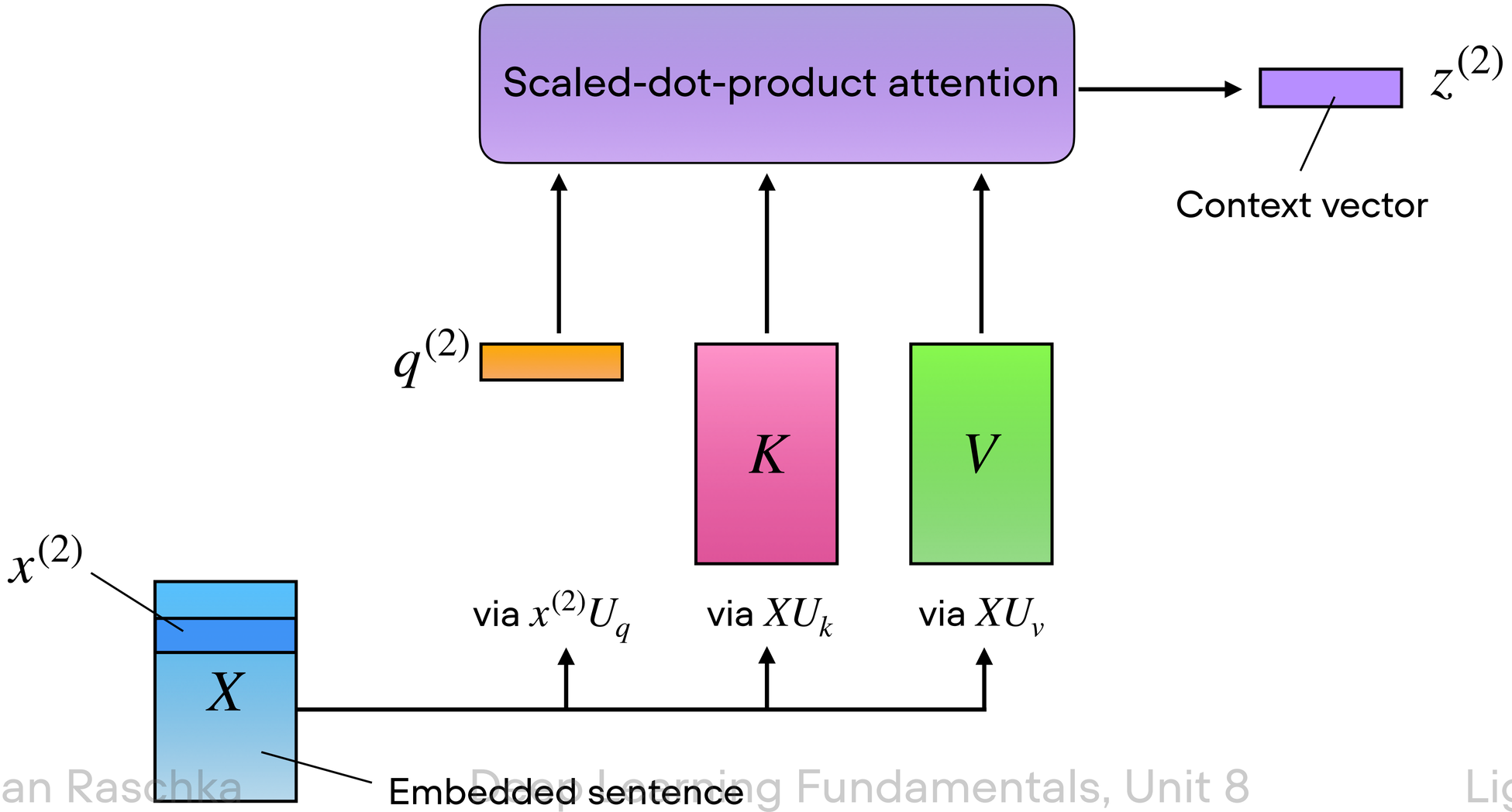


Recap: single-head attention



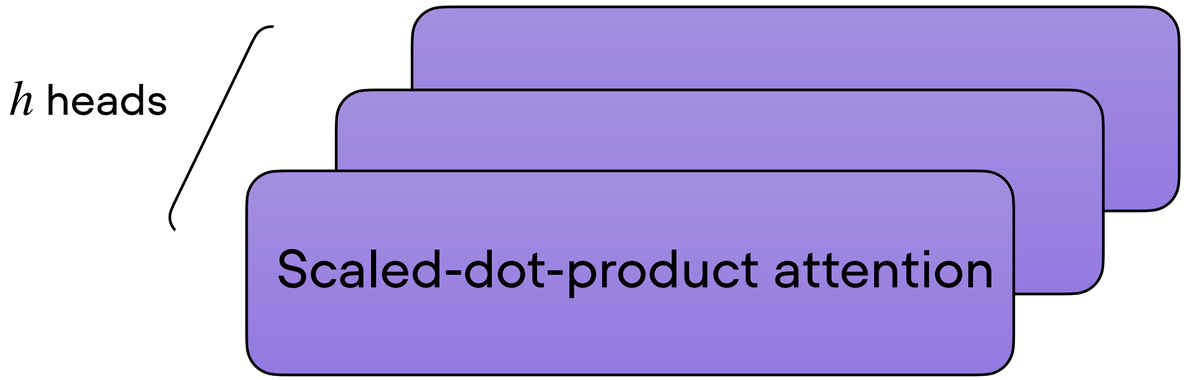
While attention can be carried out to generate all context vectors all at once, let us focus on the 2nd element for simplicity

Recap: single-head attention

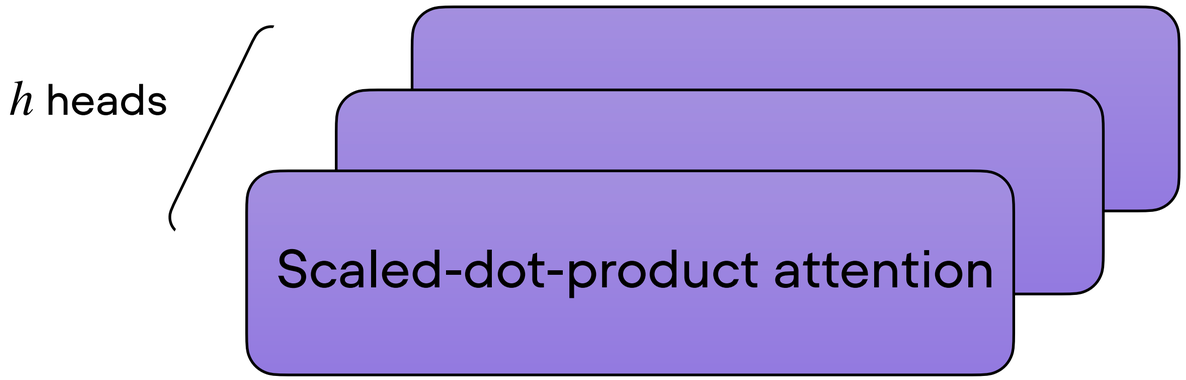


Let’s now go to multi-head attention

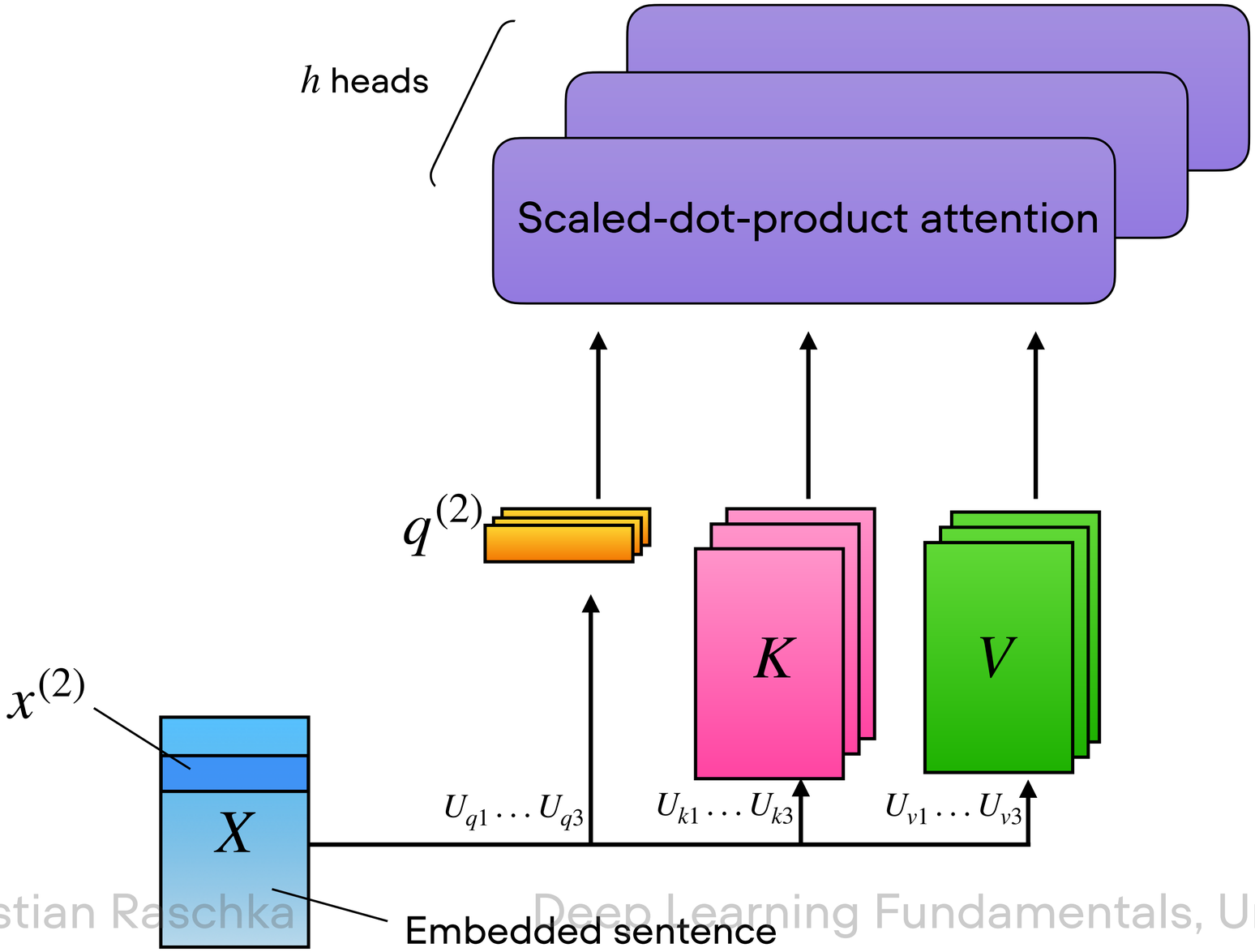
Multi-head attention

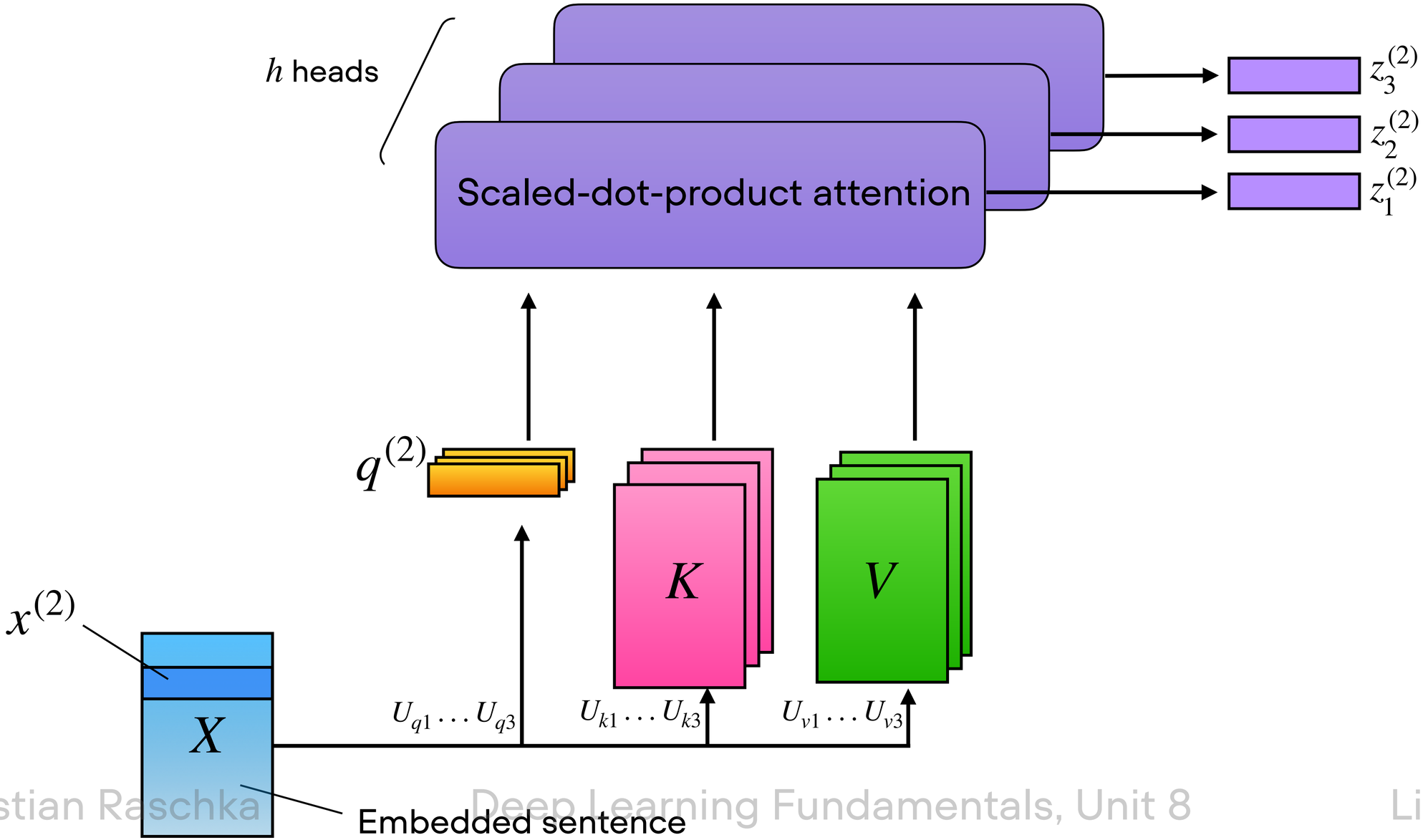


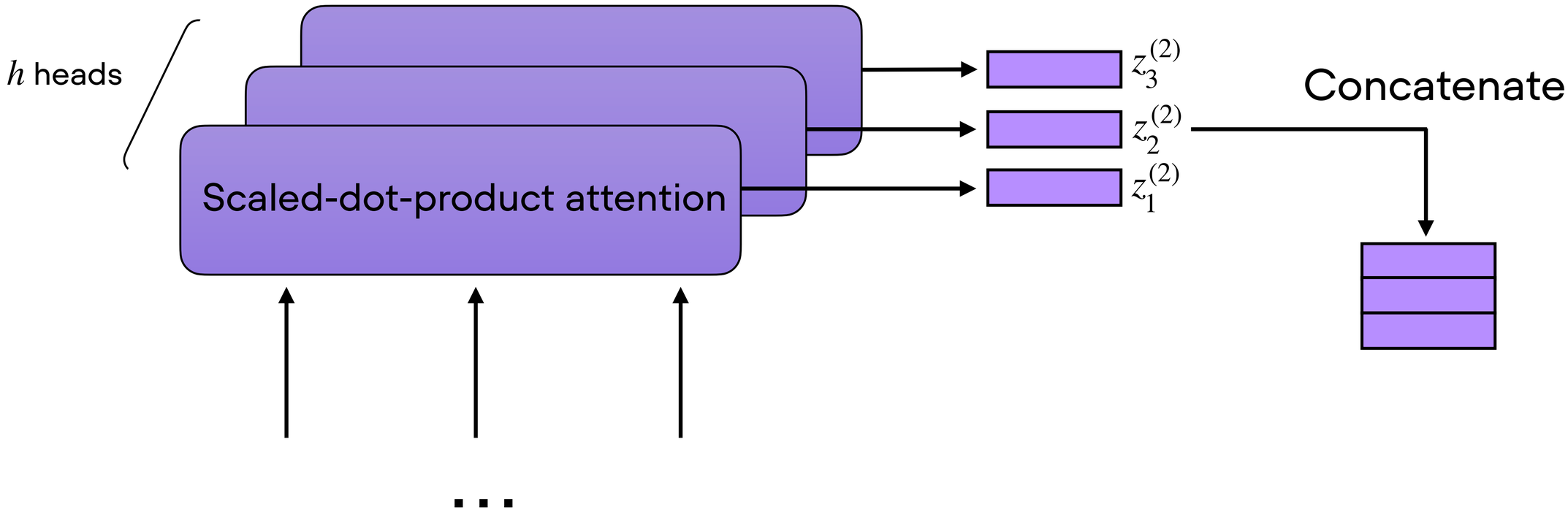
Multi-head attention

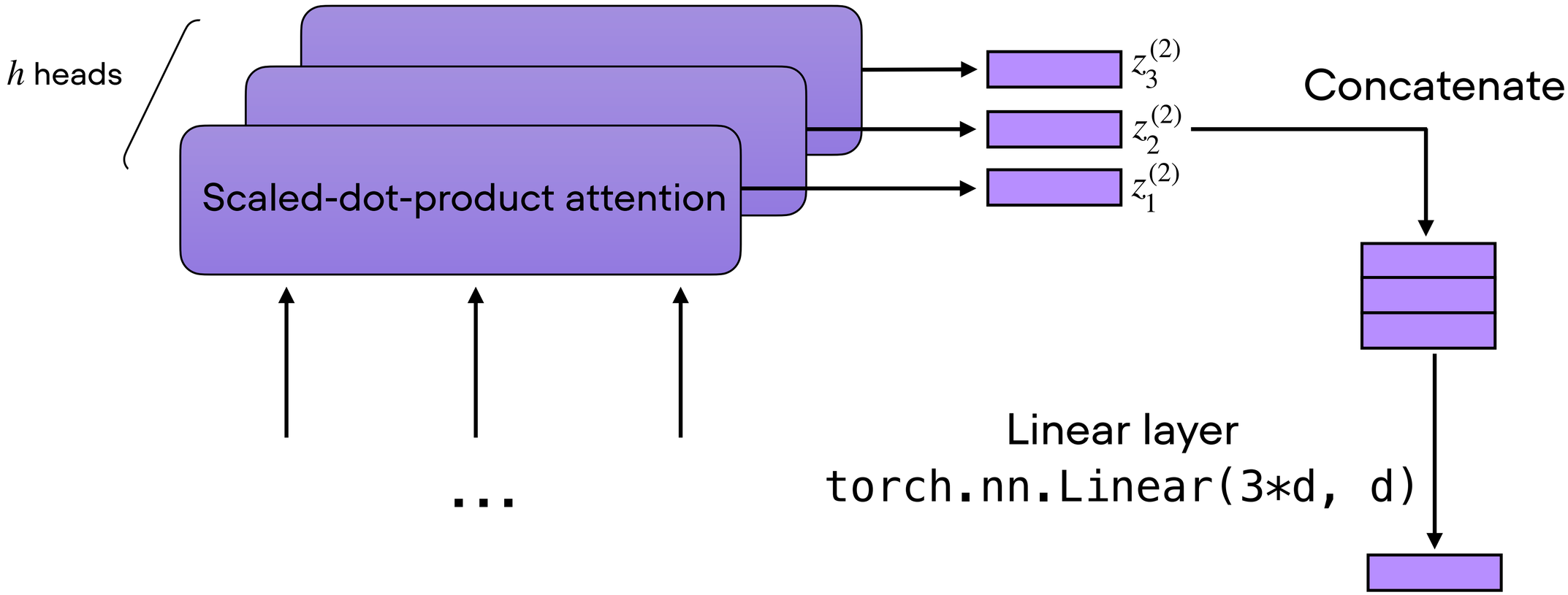


(think of it as having multiple convolutional filter to generate multiple output channels)


h heads
Scaled-dot-product attention
$q^{(2)}$
K
V
$x^{(2)}$
X
$U_{q1} \dots U_{q3}$
$U_{k1} \dots U_{k3}$
$U_{v1} \dots U_{v3}$
Embedded sentence


h heads
Scaled-dot-product attention
$z_3^{(2)}$
$z_2^{(2)}$
$z_1^{(2)}$
$q^{(2)}$
K
V
$x^{(2)}$
X
$U_{q1} \dots U_{q3}$
$U_{k1} \dots U_{k3}$
$U_{v1} \dots U_{v3}$
Embedded sentence


h heads
Scaled-dot-product attention
$z_3^{(2)}$
$z_2^{(2)}$
$z_1^{(2)}$
Concatenate
...


h heads
Scaled-dot-product attention
...
$z_3^{(2)}$
$z_2^{(2)}$
$z_1^{(2)}$
Concatenate
Linear layer
torch.nn.Linear(3*d, d)

We now covered multi-head attention!

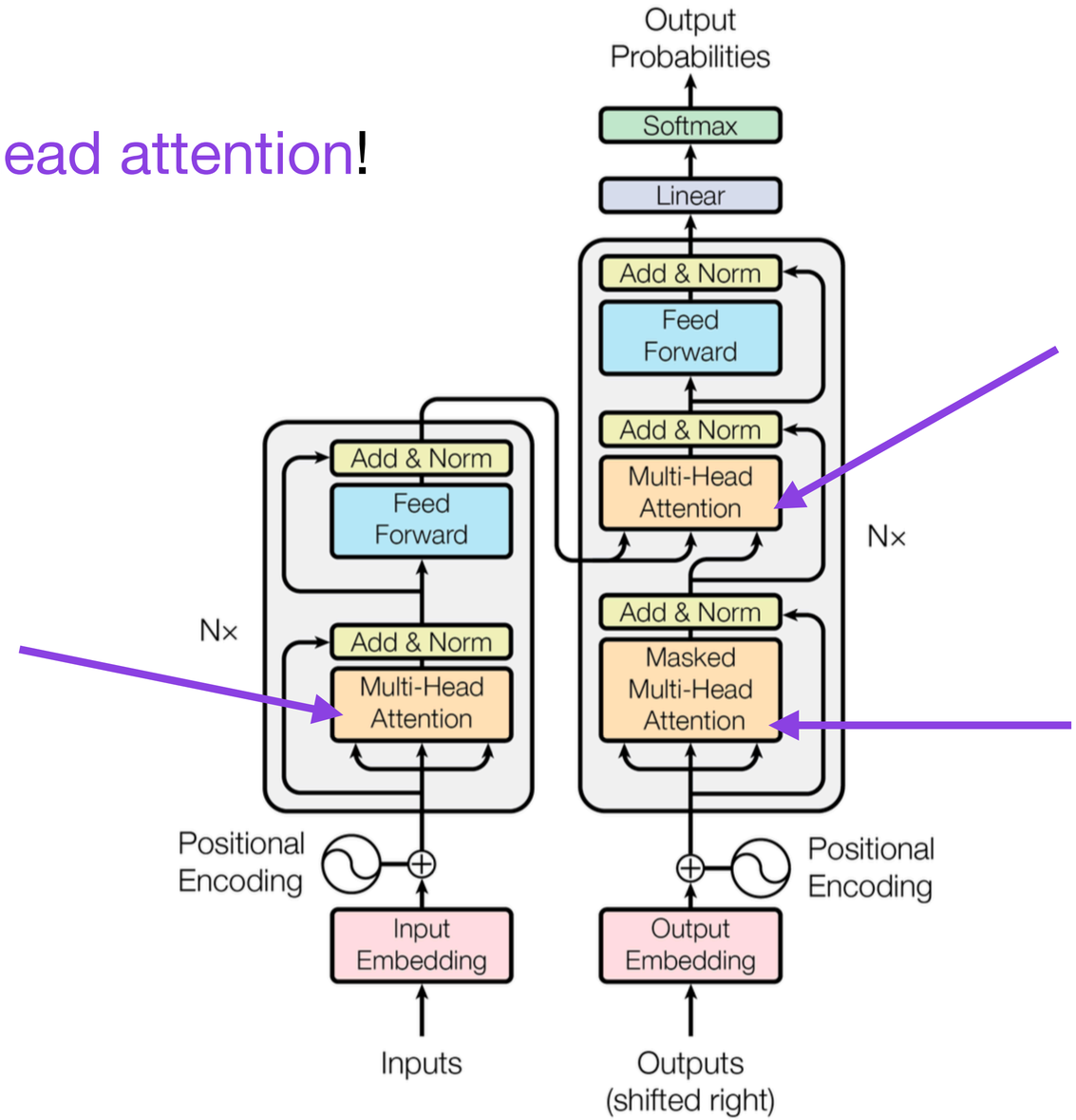


Figure 1: The Transformer - model architecture.

Next: Let's understand the other parts of the Transformer architecture