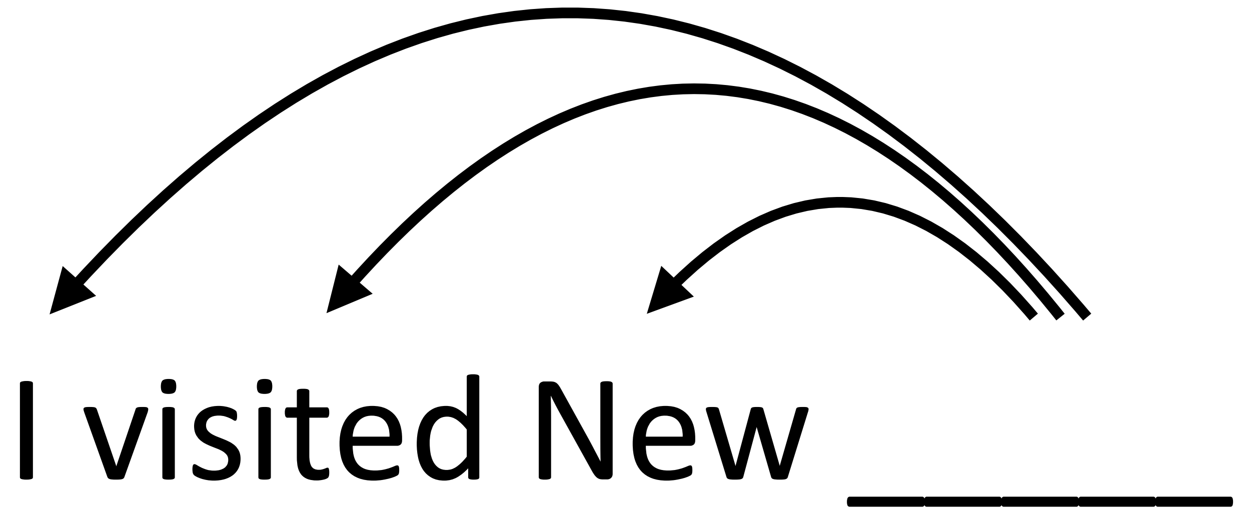
Peaky Attention

- Although attention can theoretically learn to attend to multiple tokens, in practice softmax distributions become peaked:

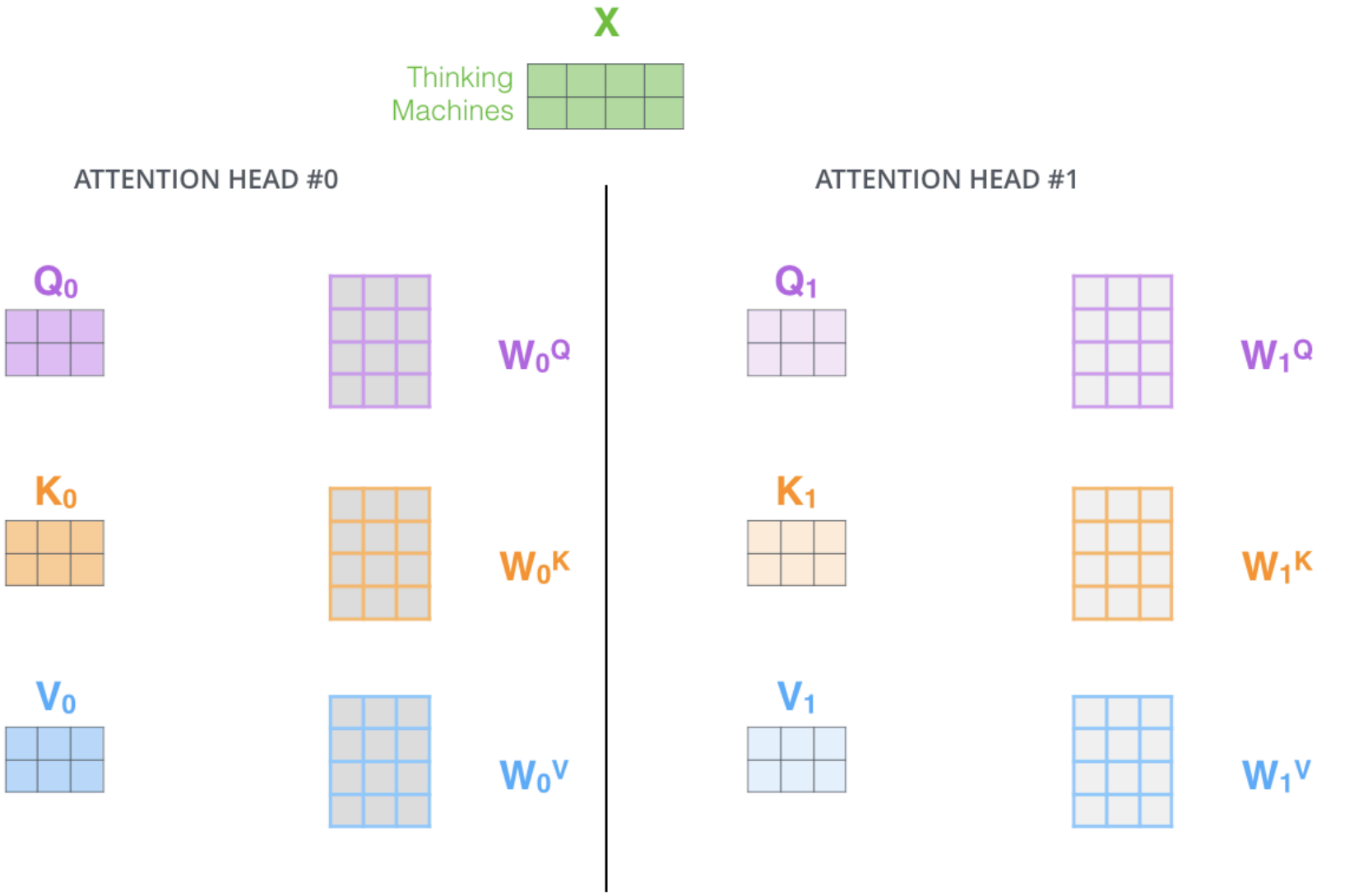


I visited New ______

- Having many layers of self-attention can help, but what about within a single layer?

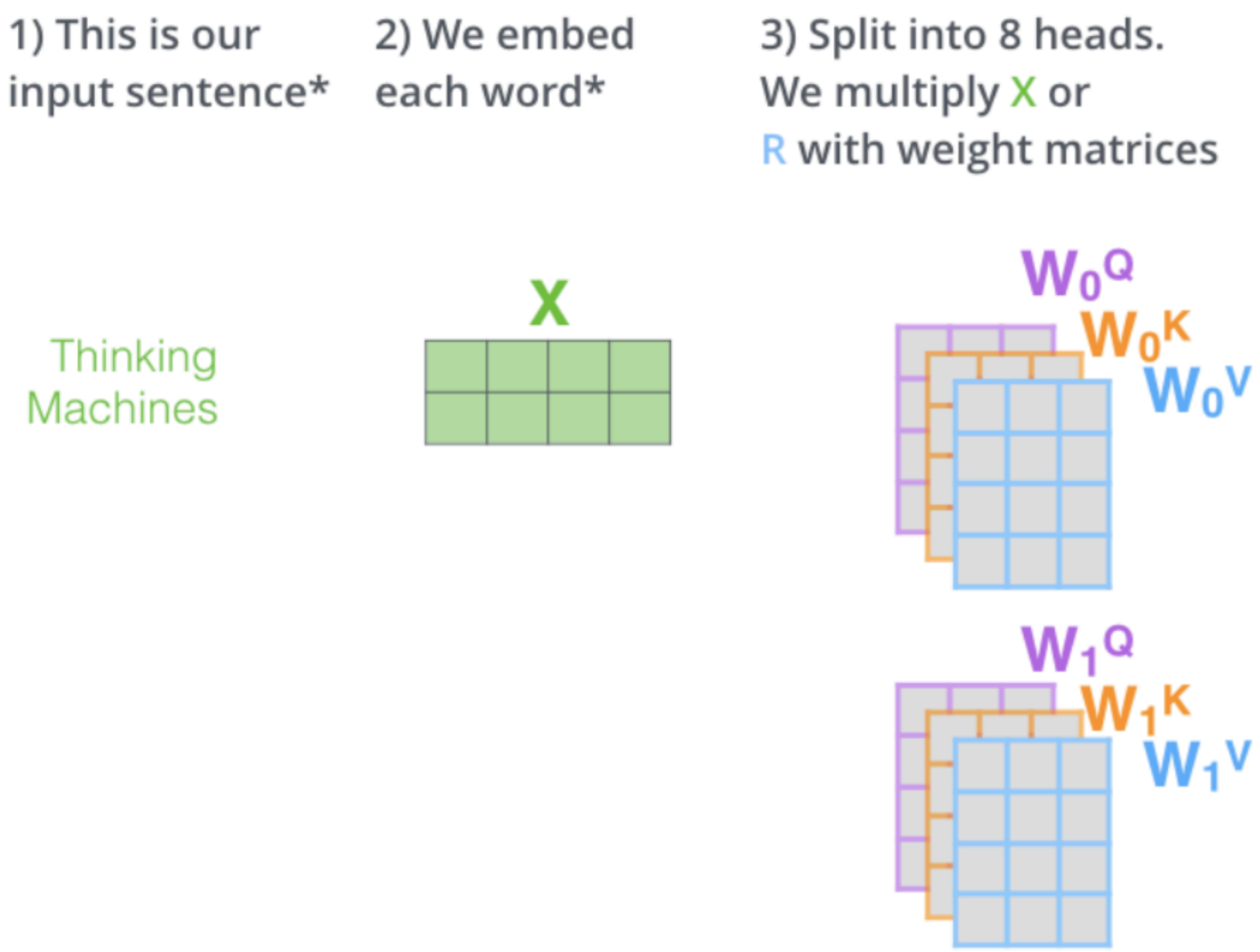
Multi-head Self-Attention

- Solution: multiple *heads* to do independent “copies” of attention



Alammar, *The Illustrated Transformer*

Multi-head Self-Attention



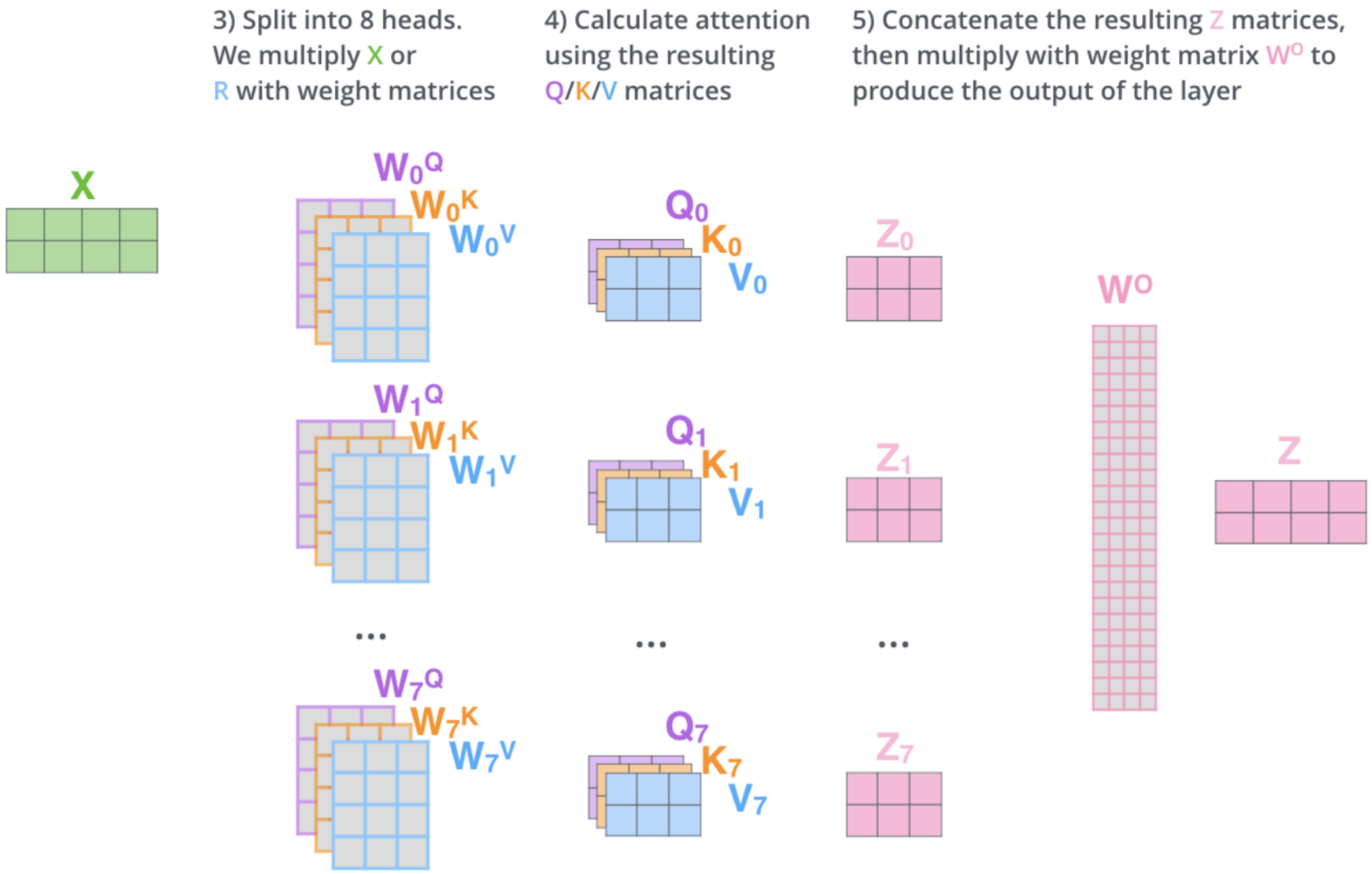
1) This is our input sentence*

2) We embed each word*

3) Split into 8 heads. We multiply X or R with weight matrices

Alammar, *The Illustrated Transformer*

Multi-head Self-Attention



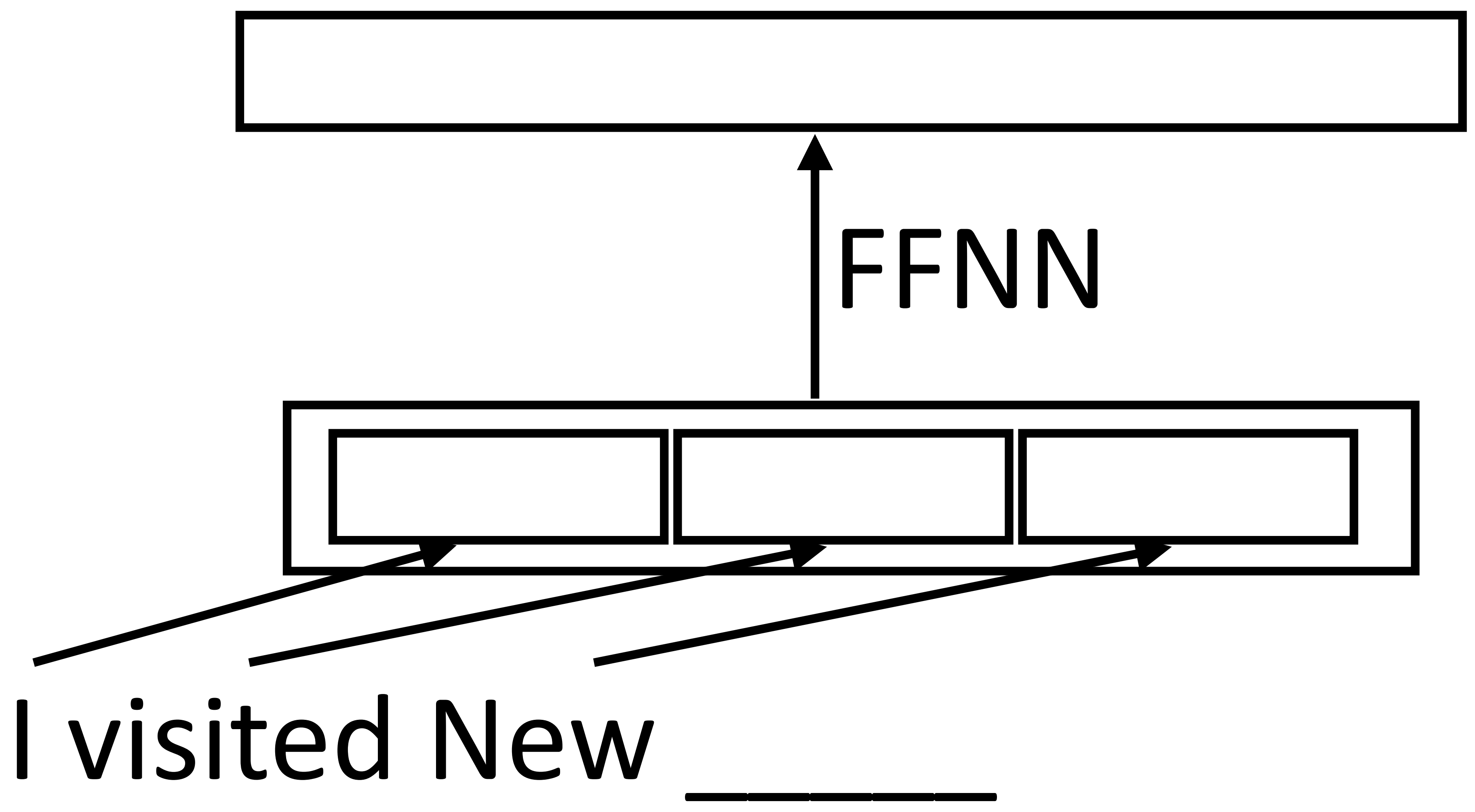
3) Split into 8 heads. We multiply X or R with weight matrices

4) Calculate attention using the resulting Q/K/V matrices

5) Concatenate the resulting Z matrices, then multiply with weight matrix W^O to produce the output of the layer

Alammar, *The Illustrated Transformer*

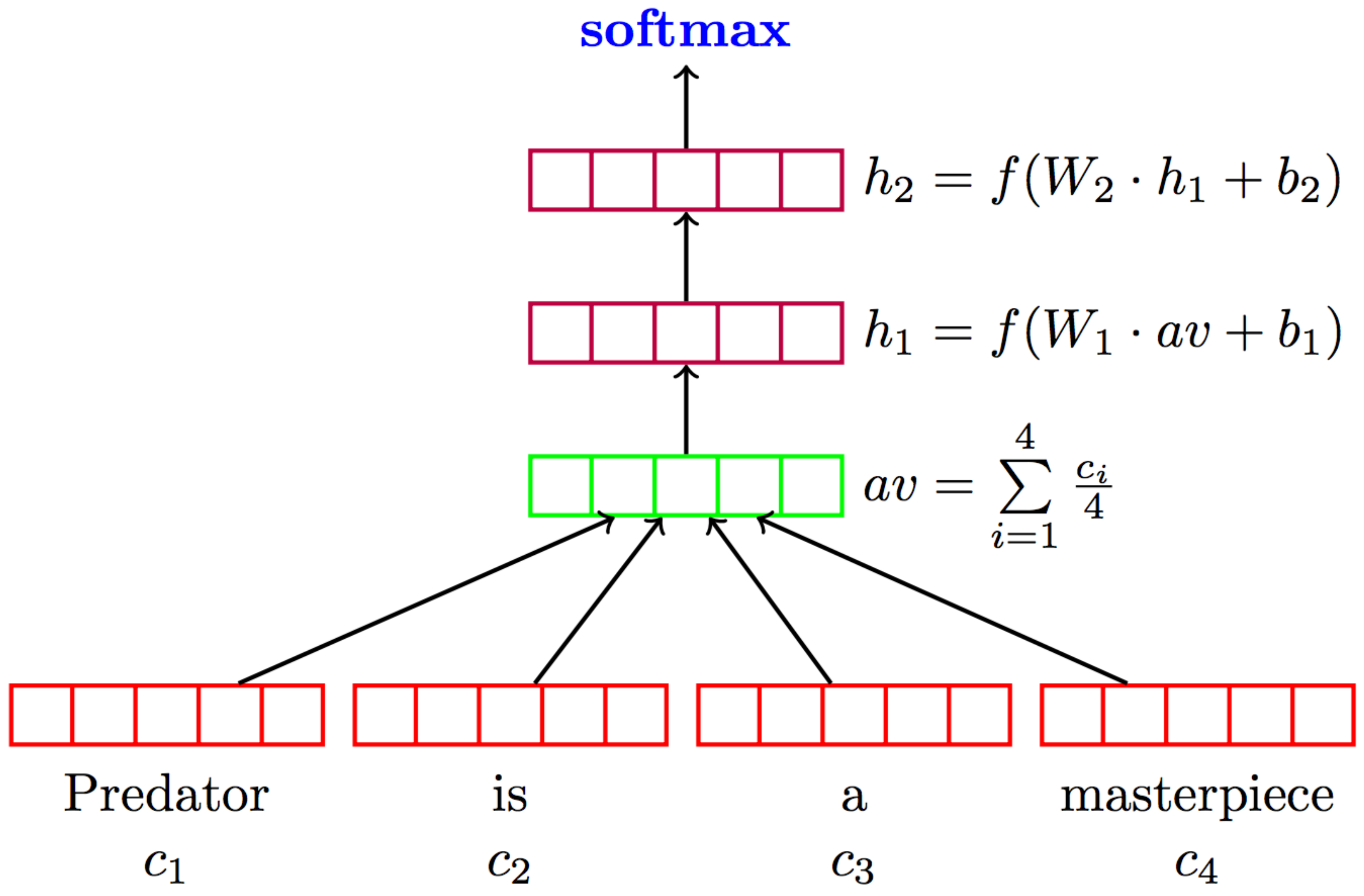
What does this give us?



FFNN

I visited New ______

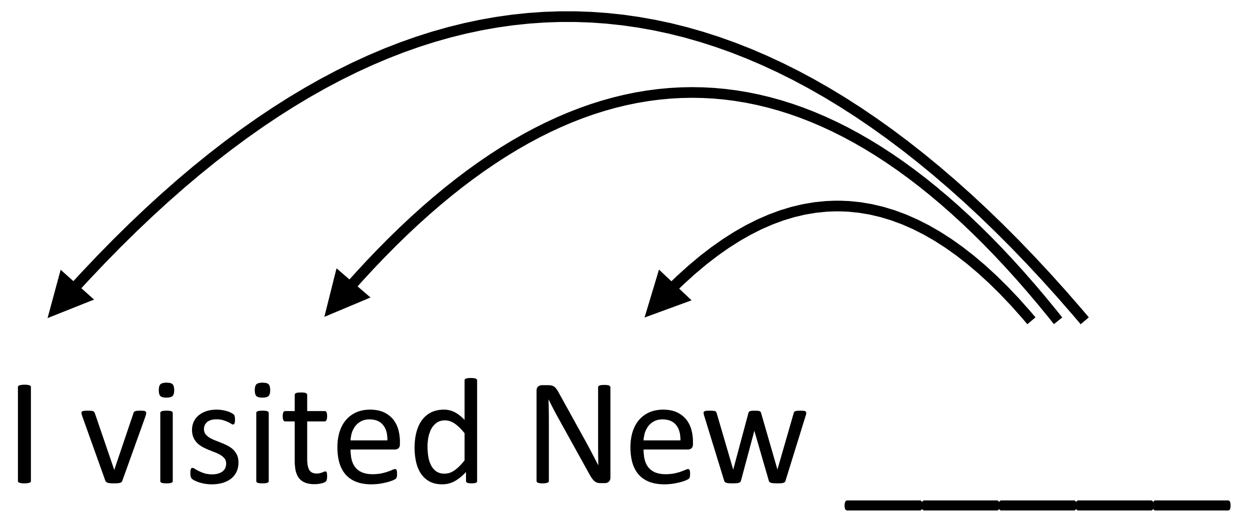
FFNN



softmax
$h_2 = f(W_2 \cdot h_1 + b_2)$
$h_1 = f(W_1 \cdot av + b_1)$
$av = \sum_{i=1}^{4} \frac{c_i}{4}$
Predator
c_1
is
c_2
a
c_3
masterpiece
c_4

DAN

Self-attention:



I visited New ______

- One head can attend to *New*, another can attend to *visited*; over the course of multiple layers, this information can combine