AIRI

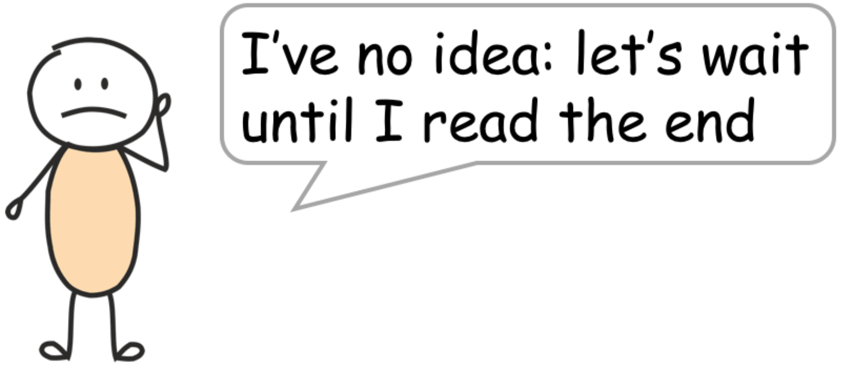
Lecture 5: Attention, transformer architecture

Leonid Sanochkin

RNN drawbacks

I arrived at the bank after crossing the … …street? …river?

What does **bank** mean in this sentence?



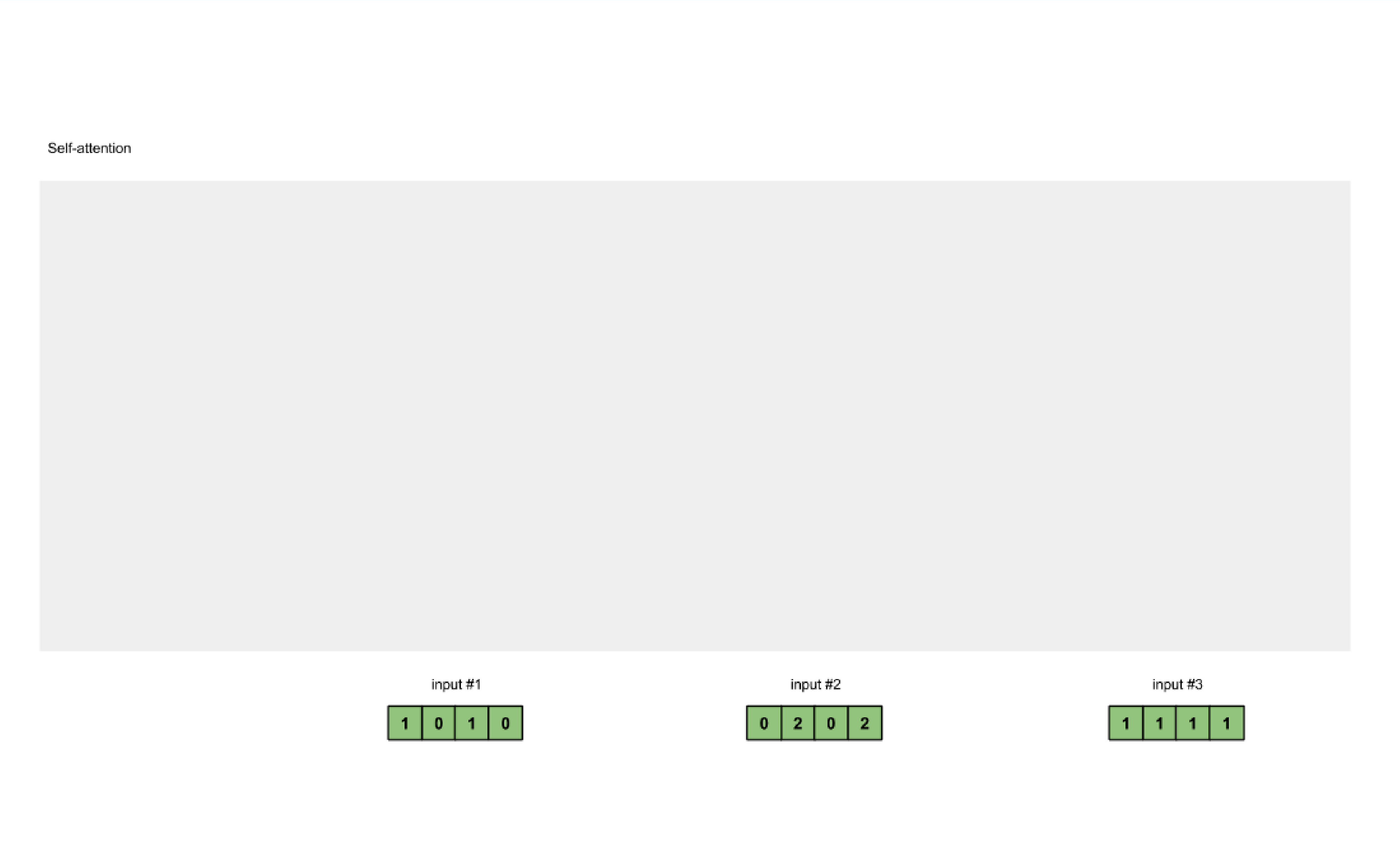
RNNs

O(N) steps to process a sentence with length N

Transformer

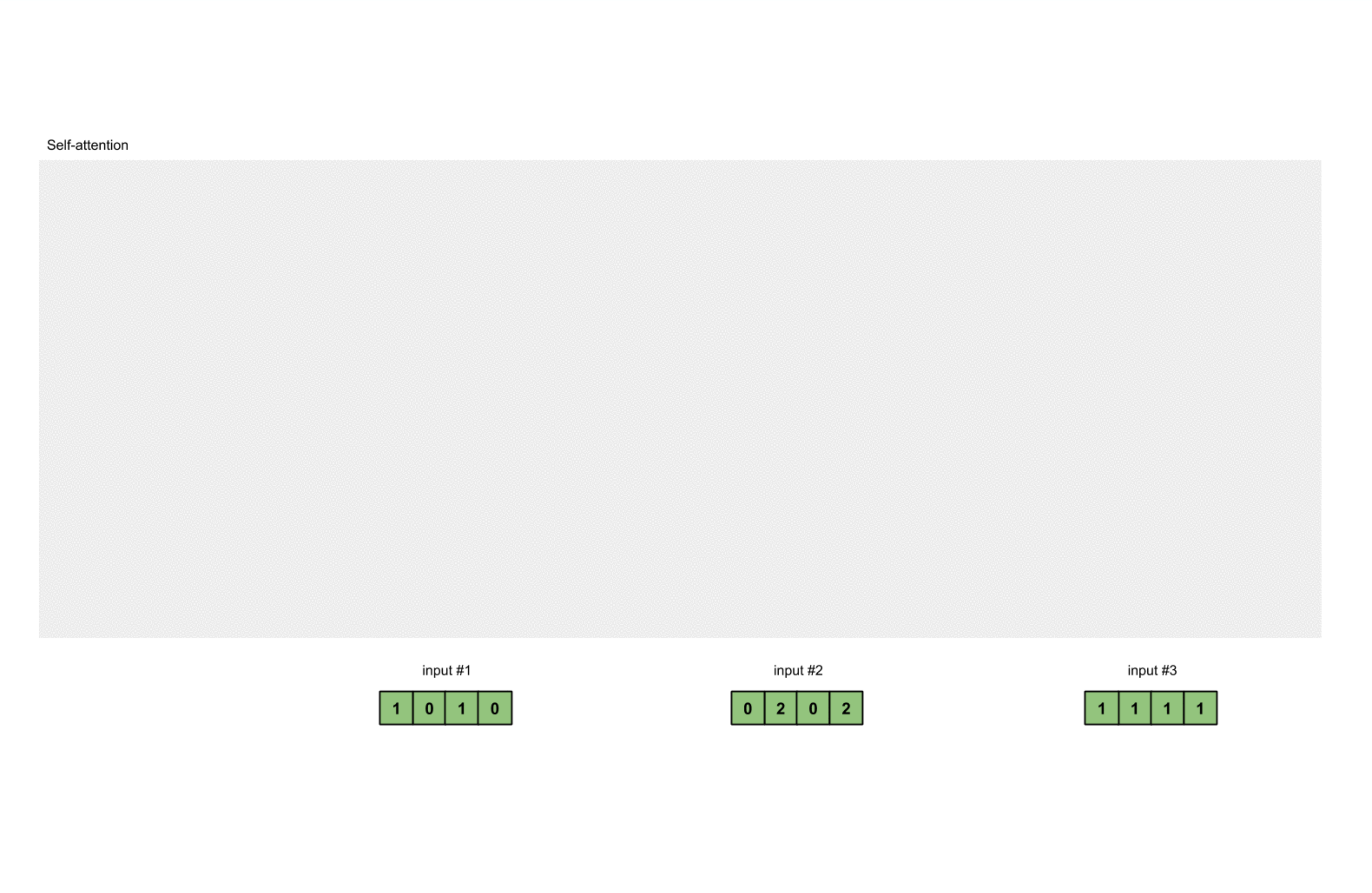
Constant number of steps to process any sentence

Self-attention



input #1	input #2	input #3
1 0 1 0	0 2 0 2	1 1 1 1

Self-attention



Self-attention

input #1

1	0	1	0

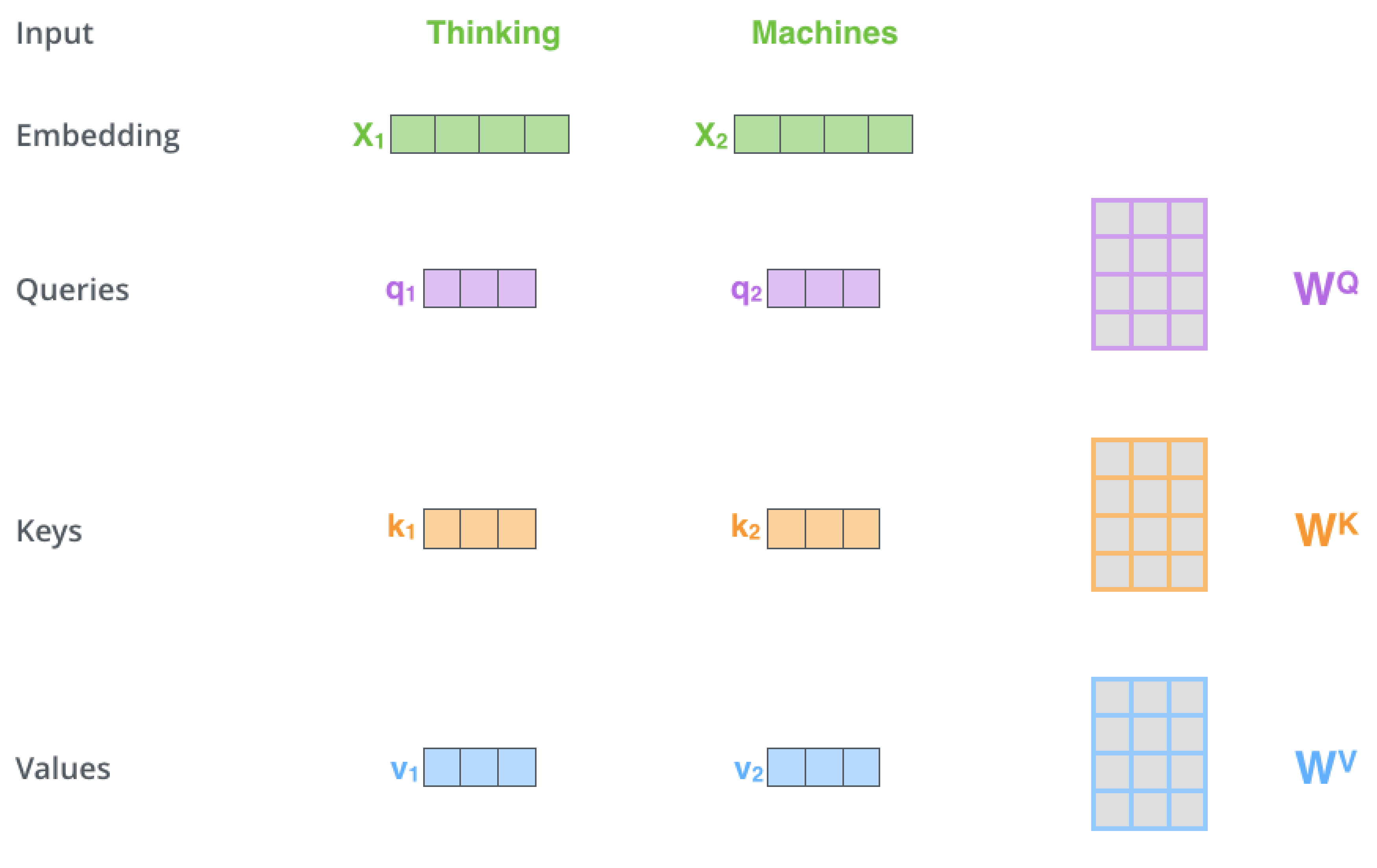
input #2

0	2	0	2

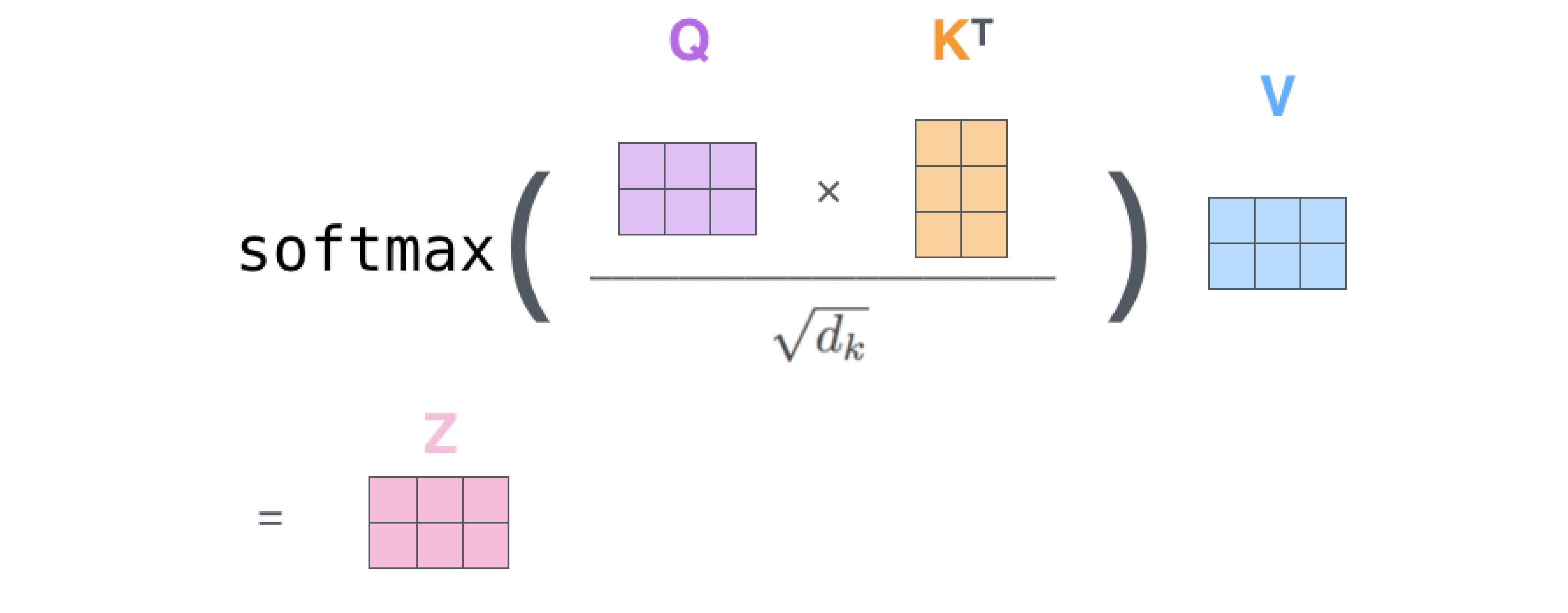
input #3

1	1	1	1

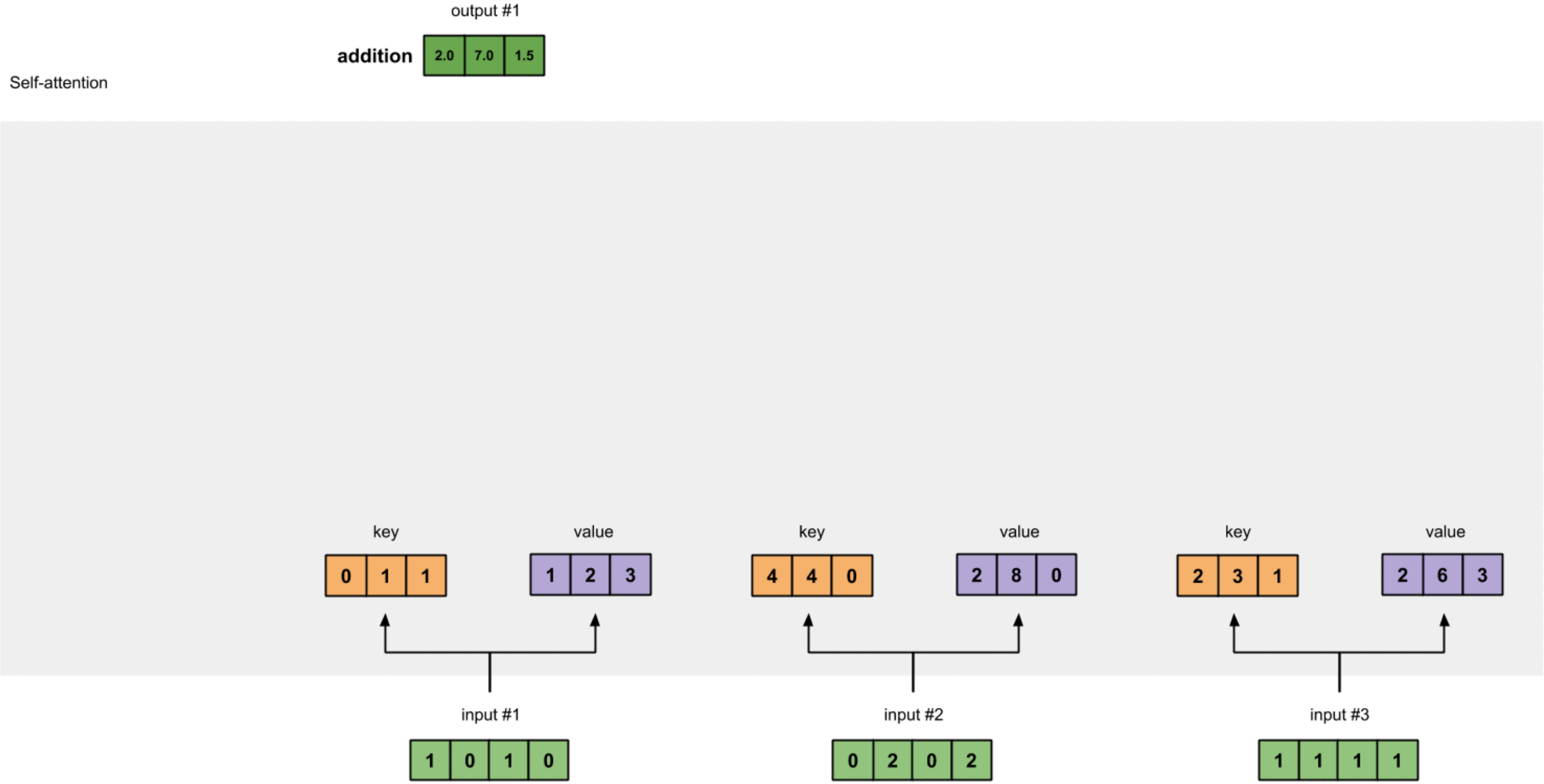
Self-attention



Self-attention

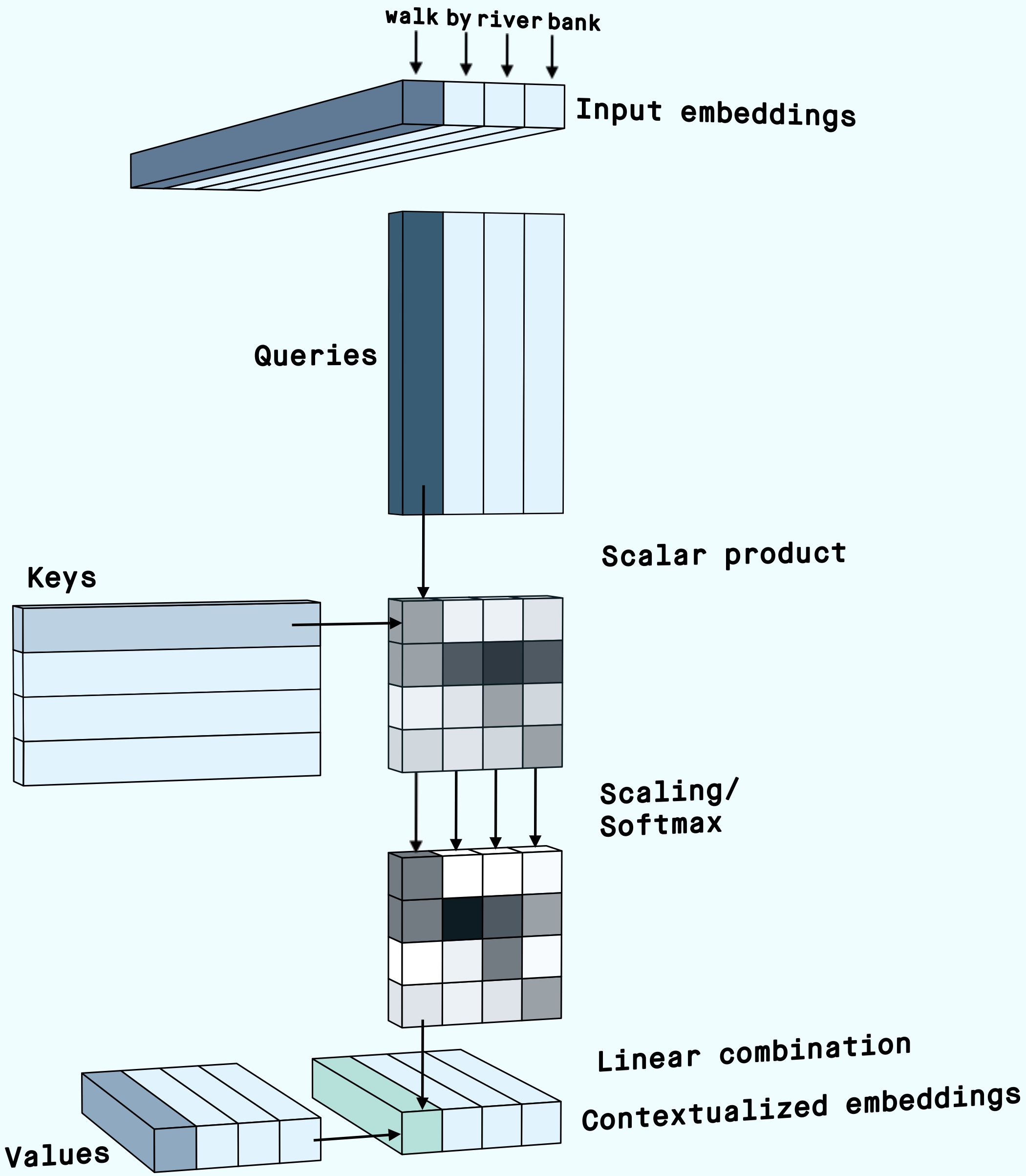


Self-attention

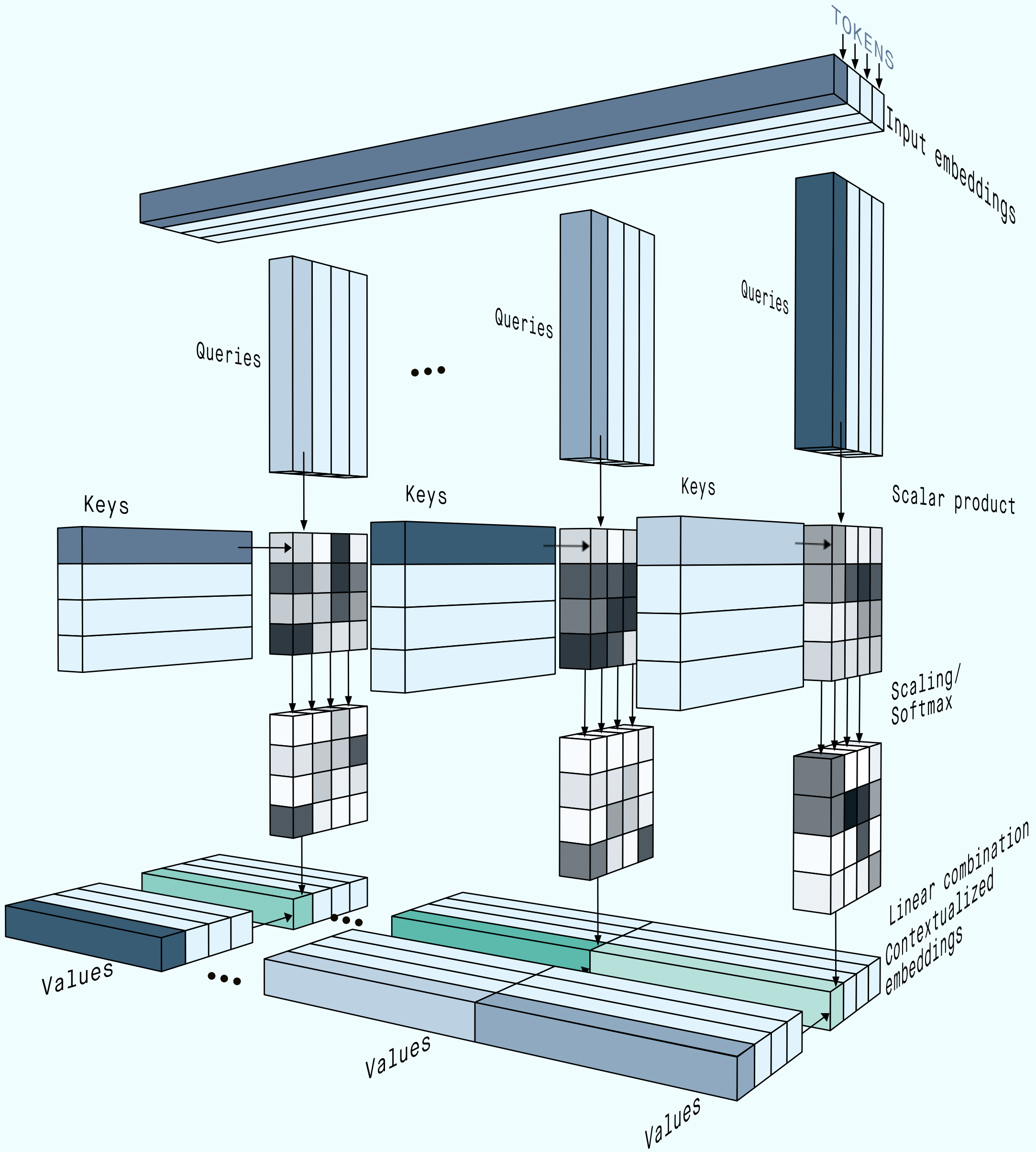


NATIONAL RESEARCH UNIVERSITY

Self-attention

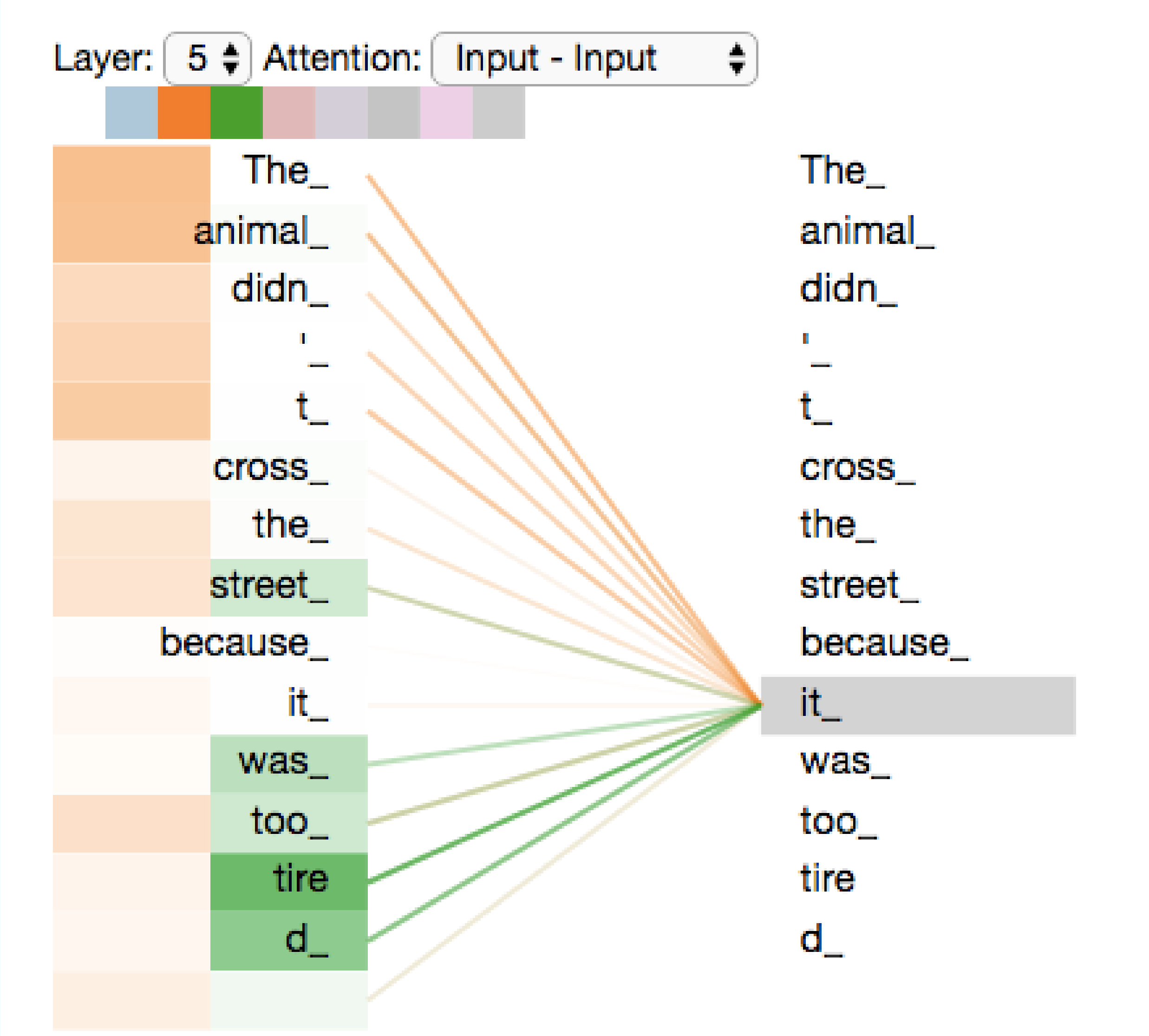


Multihead self-attention



NATIONAL RESEARCH UNIVERSITY

Multihead self-attention



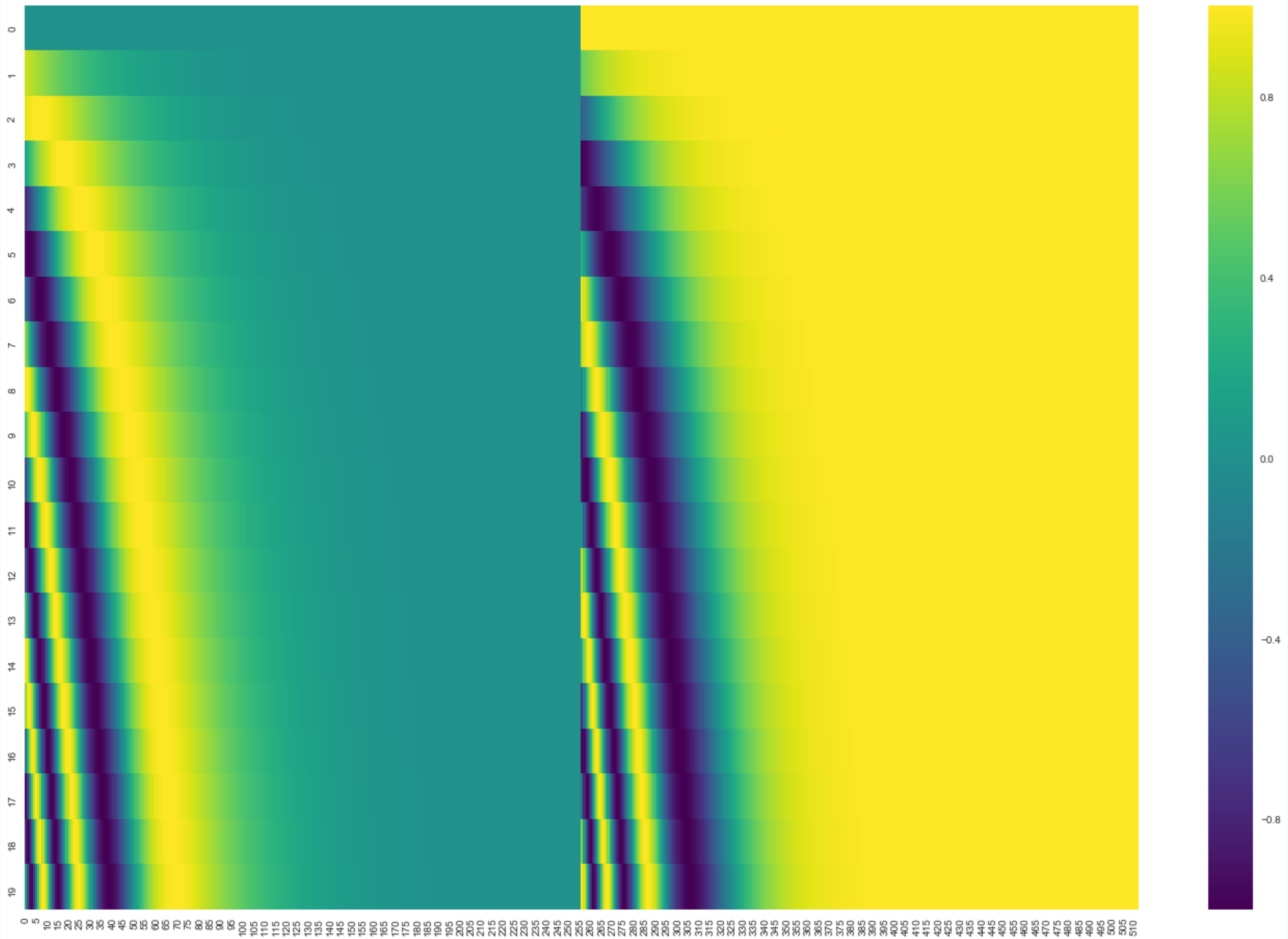
Layer: 5 Attention: Input - Input

Positional encodings

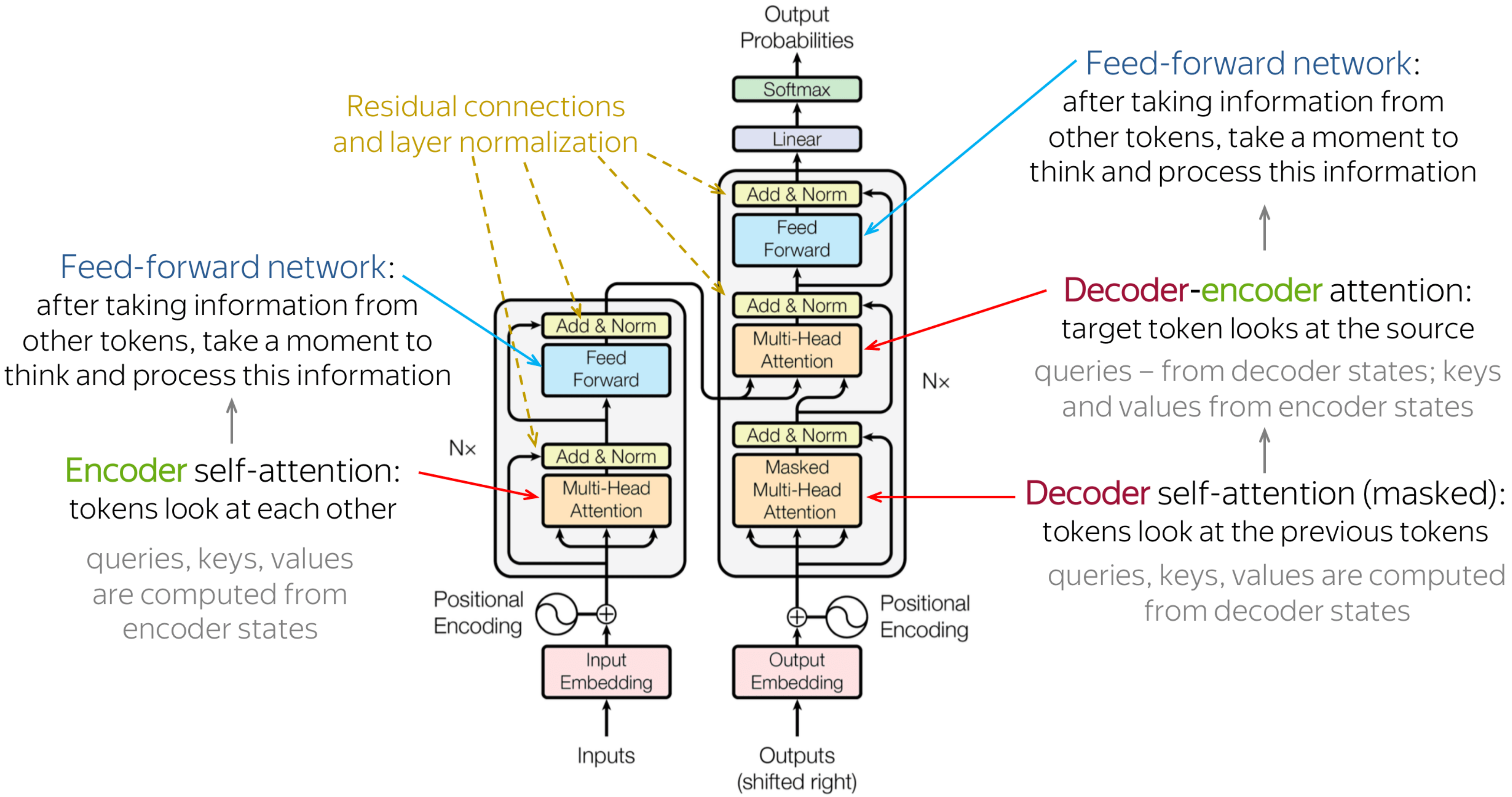
$$PE_{pos,2i} = \sin(pos/10000^{2i/d_{model}}),$$

$$PE_{pos,2i+1} = \cos(pos/10000^{2i/d_{model}}),$$

Positional encodings



Transformer architecture



Residual connections and layer normalization

Feed-forward network: after taking information from other tokens, take a moment to think and process this information

Encoder self-attention: tokens look at each other

queries, keys, values are computed from encoder states

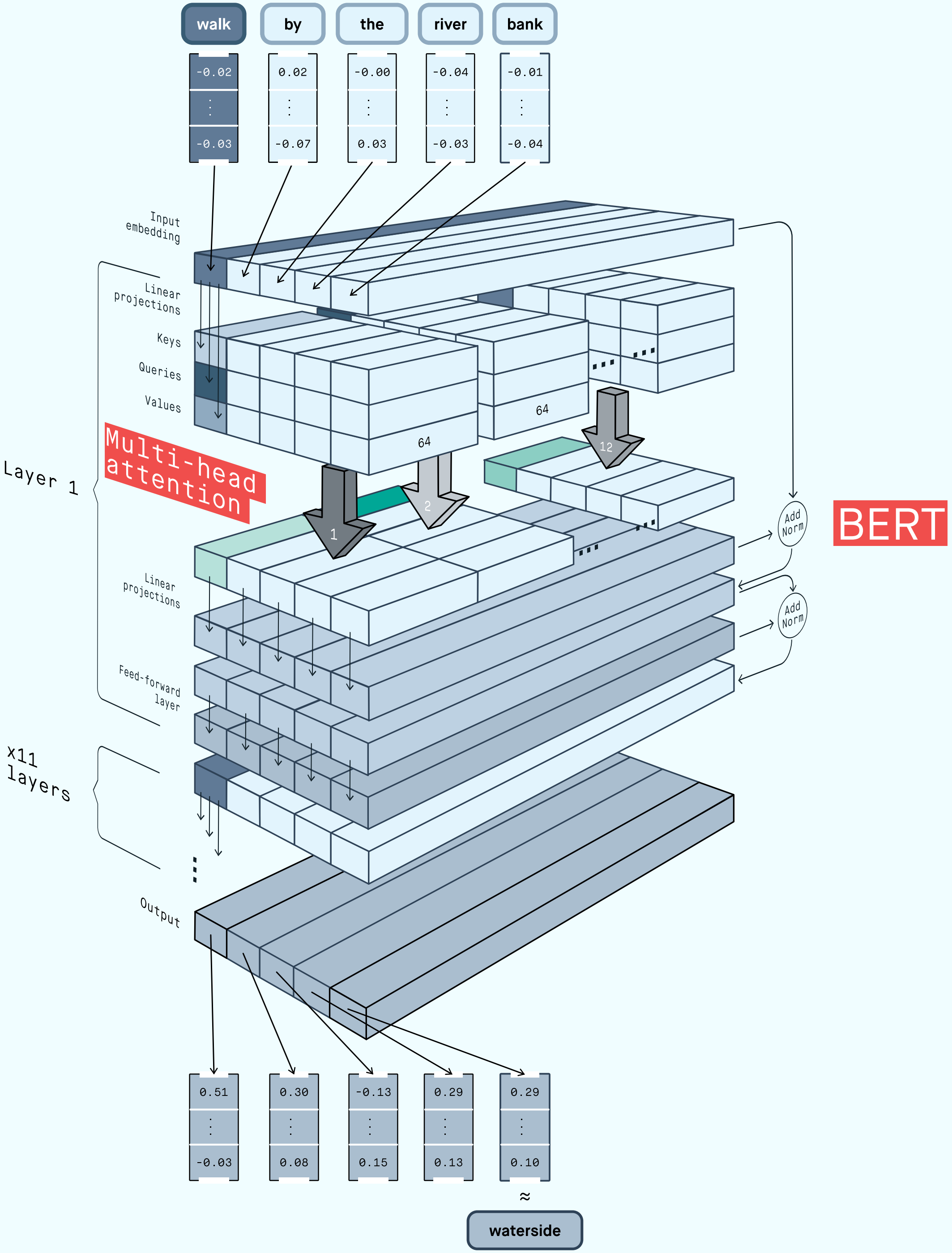
Feed-forward network: after taking information from other tokens, take a moment to think and process this information

Decoder-encoder attention: target token looks at the source

queries – from decoder states; keys and values from encoder states

Decoder self-attention (masked): tokens look at the previous tokens

queries, keys, values are computed from decoder states


walk
by
the
river
bank
-0.02
-0.03
0.02
-0.07
-0.00
0.03
-0.04
-0.03
-0.01
-0.04
Input embedding
Linear projections
Keys
Queries
Values
64
64
12
1
2
Multi-head attention
BERT
Add Norm
Add Norm
Layer 1
Linear projections
Feed-forward layer
x11 layers
Output
0.51
-0.03
0.30
0.08
-0.13
0.15
0.29
0.13
0.29
0.10
≈
waterside

AIRI

airi.net