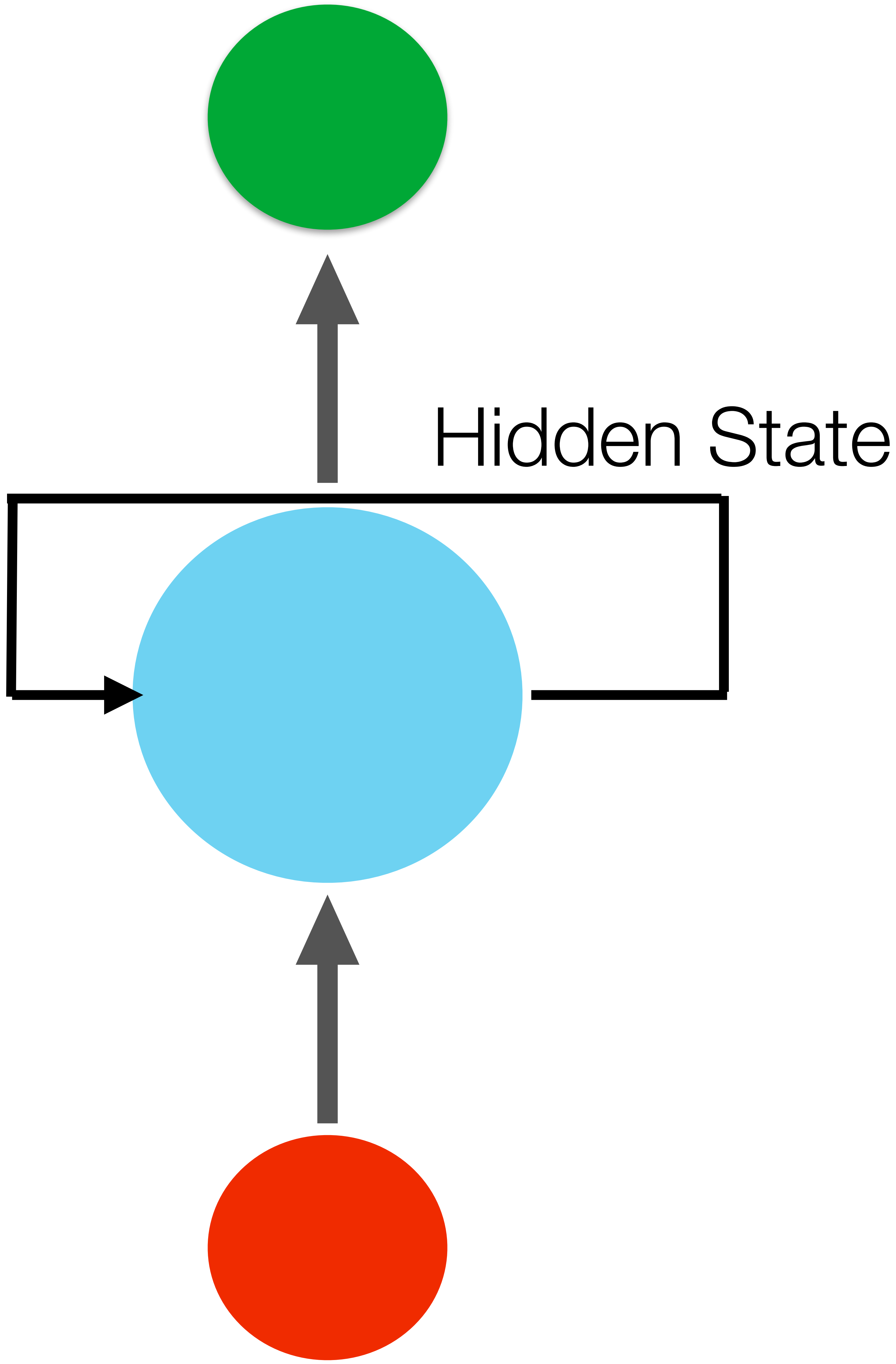
Attention Is All You Need

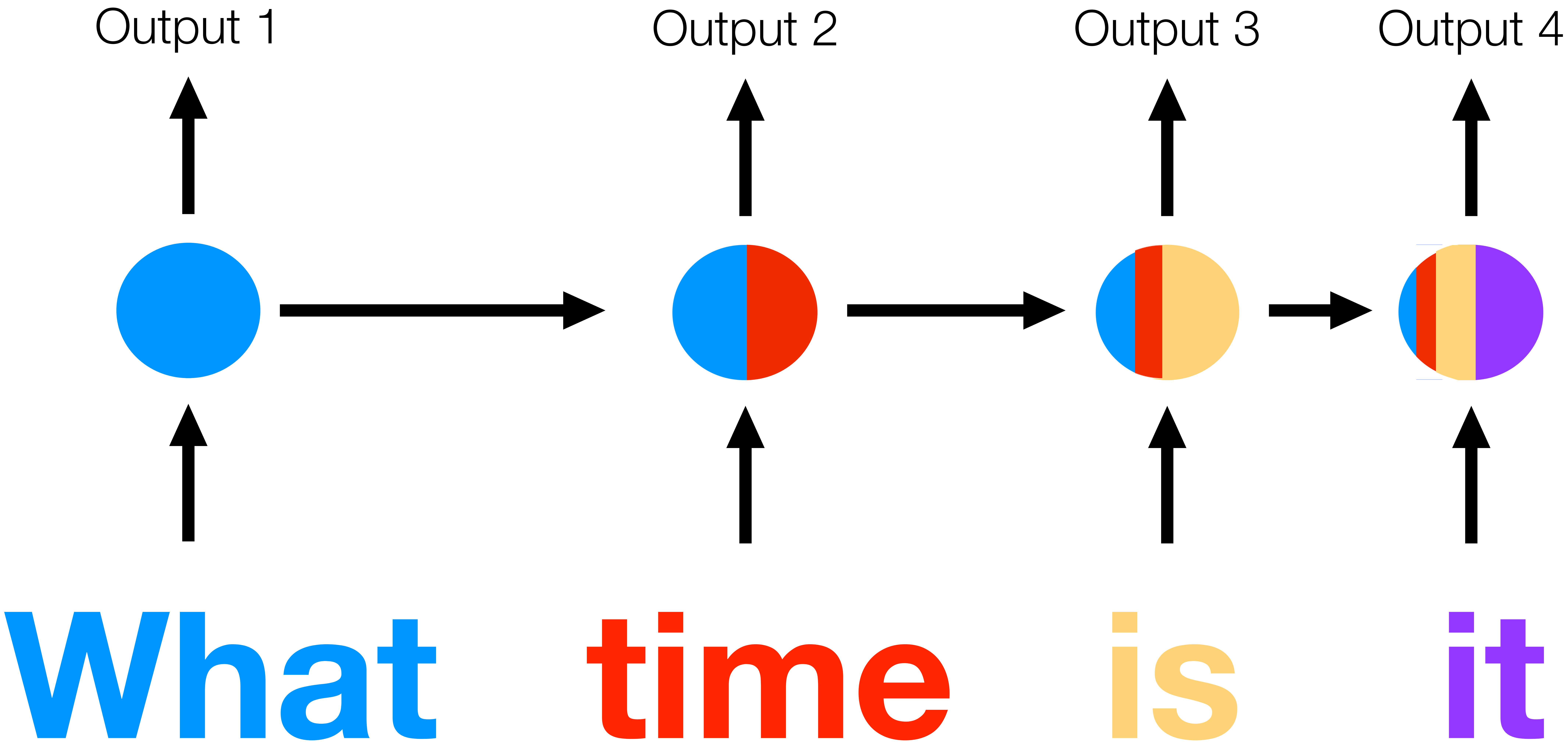
Matteo Omenetti

Google

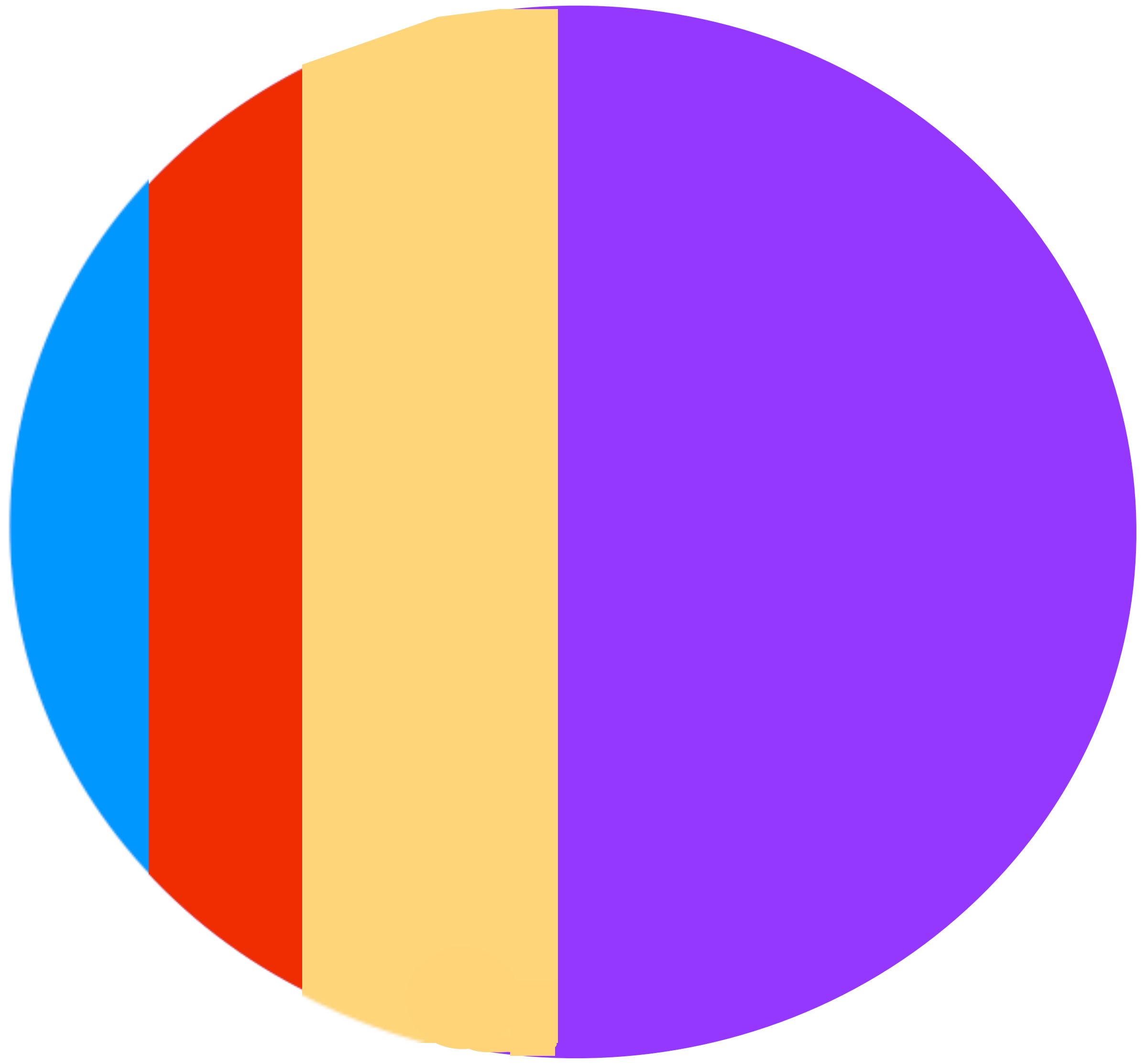
Recurrent Neural Networks (RNNs) (1)



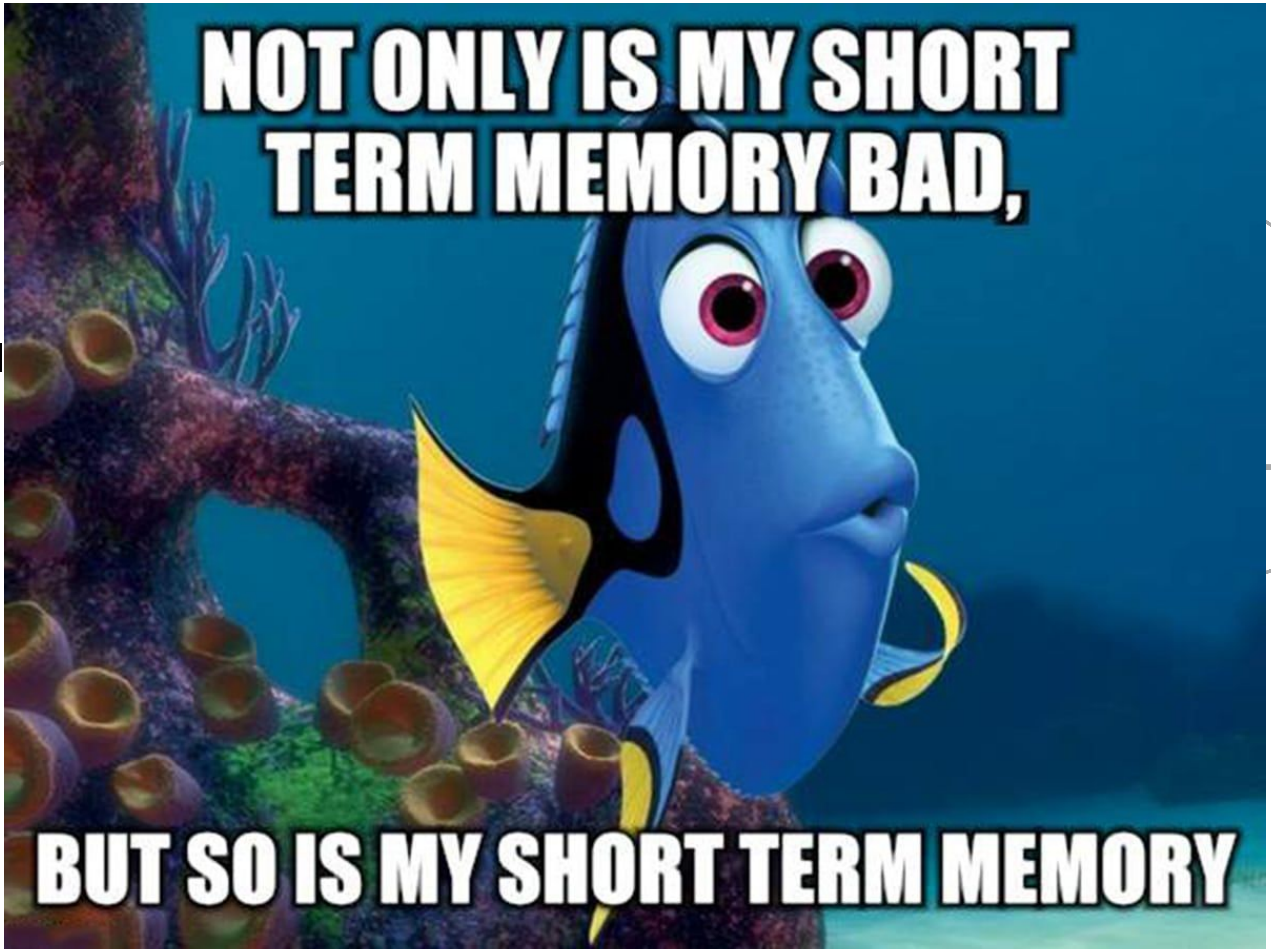
Recurrent Neural Networks (RNNs) (2)



Recurrent Neural Networks (RNNs) (3)



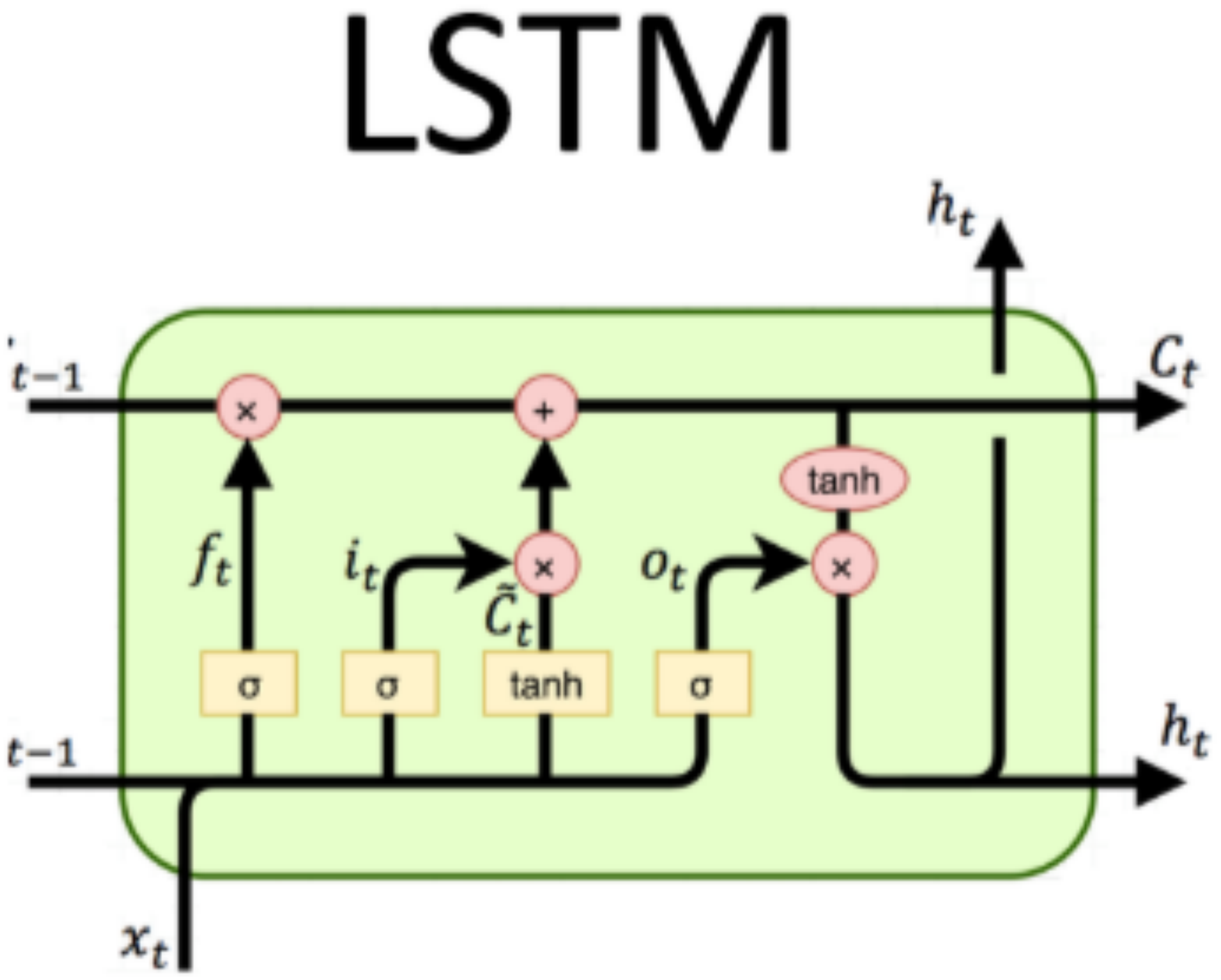
Recurrent Neural Networks (RNNs) (3)

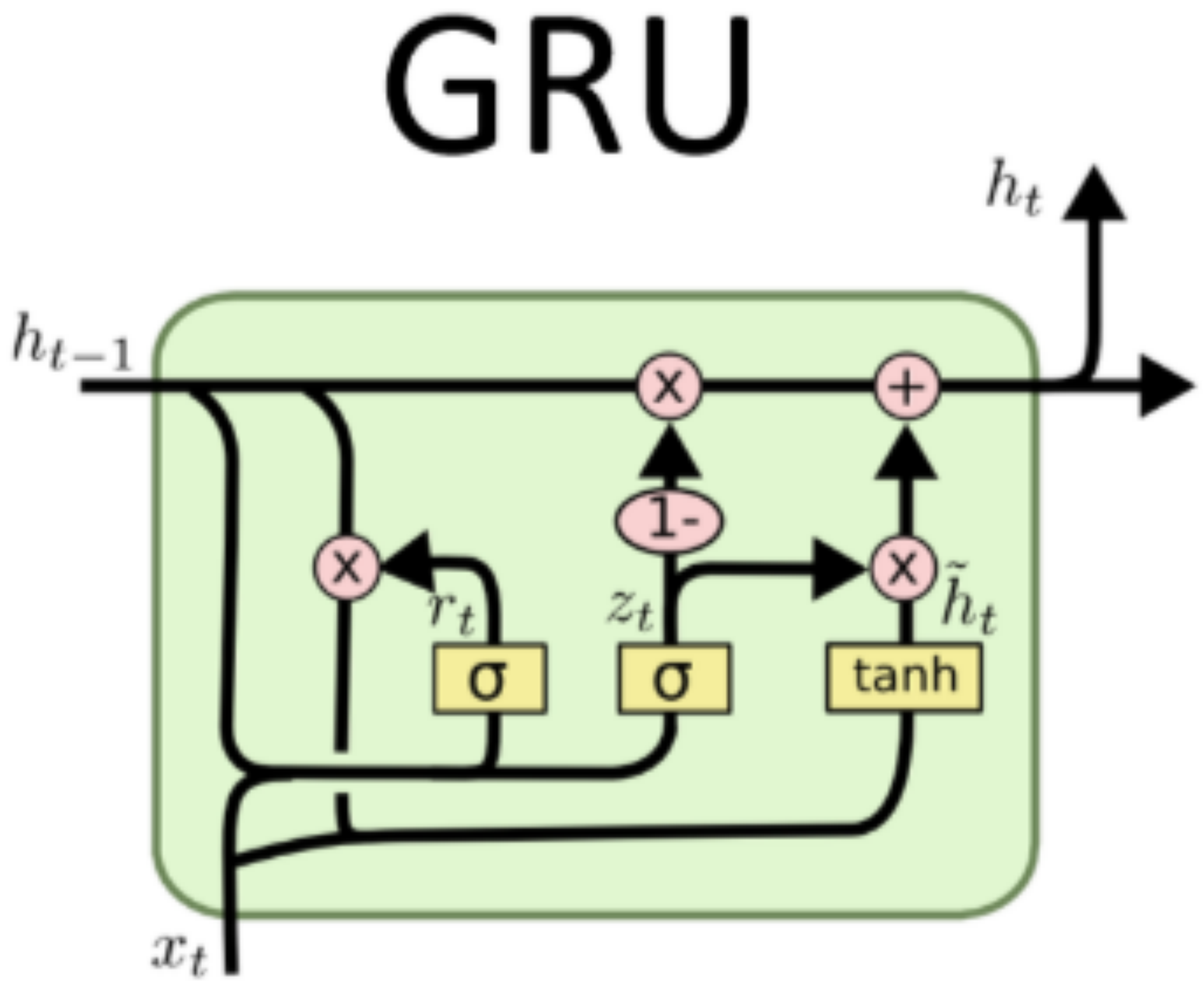

NOT ONLY IS MY SHORT TERM MEMORY BAD,
BUT SO IS MY SHORT TERM MEMORY

As aliens entered ou

of extraterrestrials…

Recurrent Neural Networks (RNNs) (4)





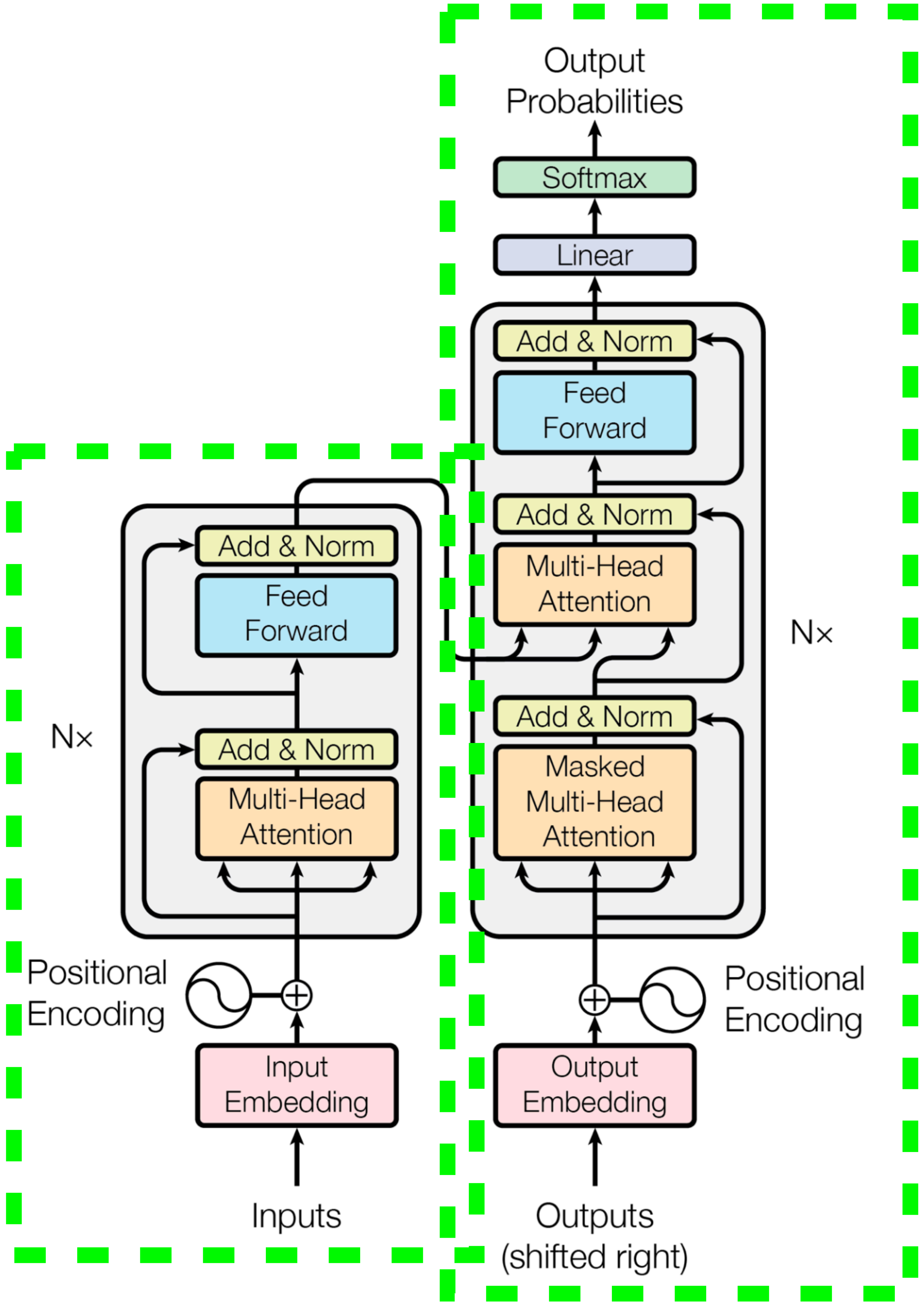
Recurrent Neural Networks (RNNs) (5)

As aliens entered our planet and began to colonize earth a certain group of extraterrestrials…

Transformers

As aliens entered our planet and began to colonize earth a certain group of extraterrestrials

The Architecture



The Example

🇺🇸 The cat is under the table

=

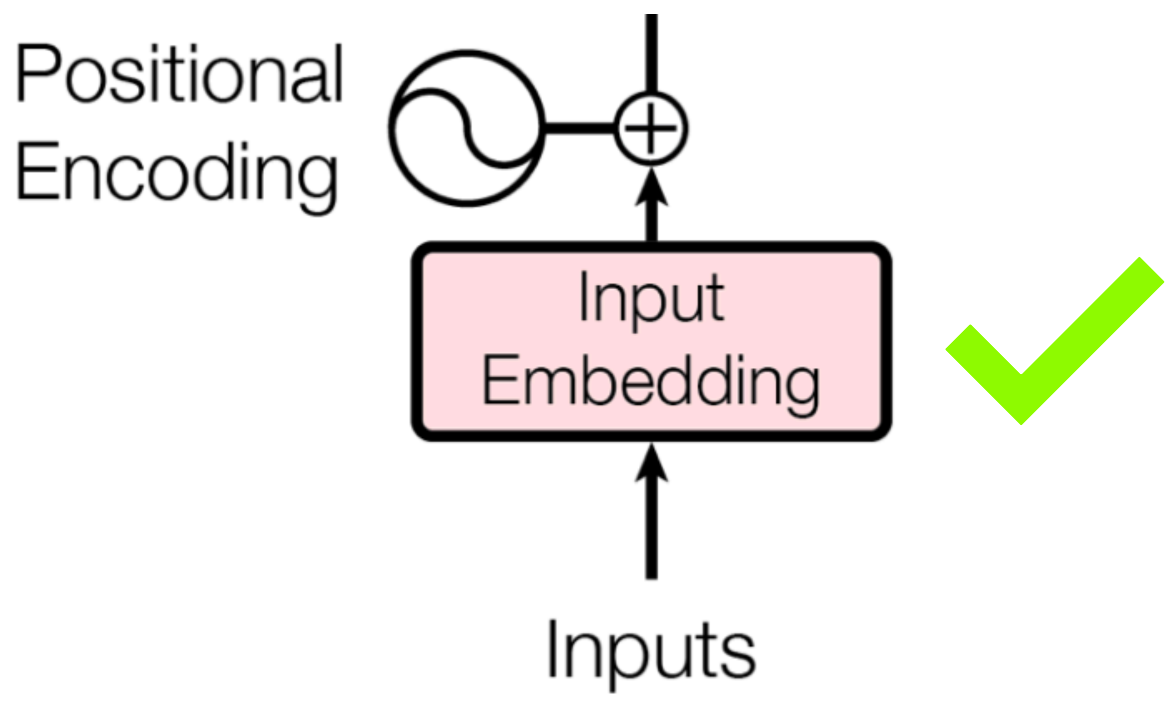
🇮🇹 Il gatto è sotto il tavolo

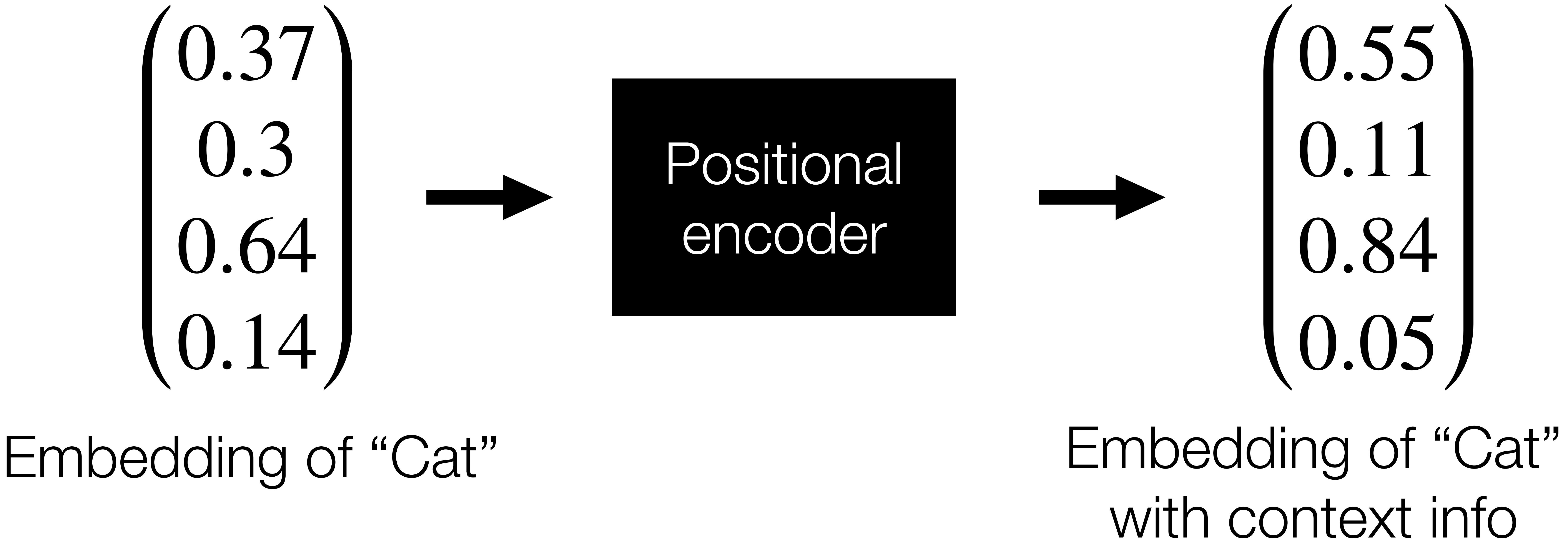
Input Embedding

Input Embedding
Inputs
Cat
Sun
Red
Dog
Horse
Cat
Human
Son
Happy
Sad
Love
Guitar
Violin
0.37
0.3
0.64
0.14

Embedding of the word cat

Positional Encoding





$$\begin{pmatrix}0.37\\0.3\\0.64\\0.14\end{pmatrix} \rightarrow \boxed{\text{Positional encoder}} \rightarrow \begin{pmatrix}0.55\\0.11\\0.84\\0.05\end{pmatrix}$$

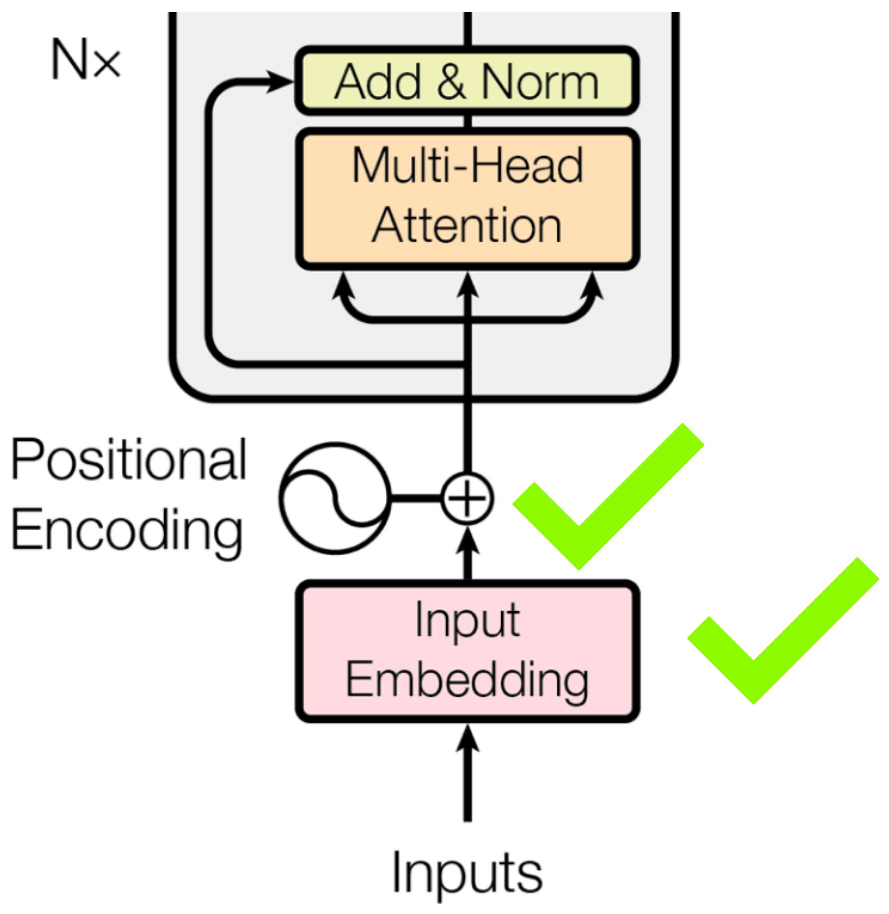
Embedding of “Cat”

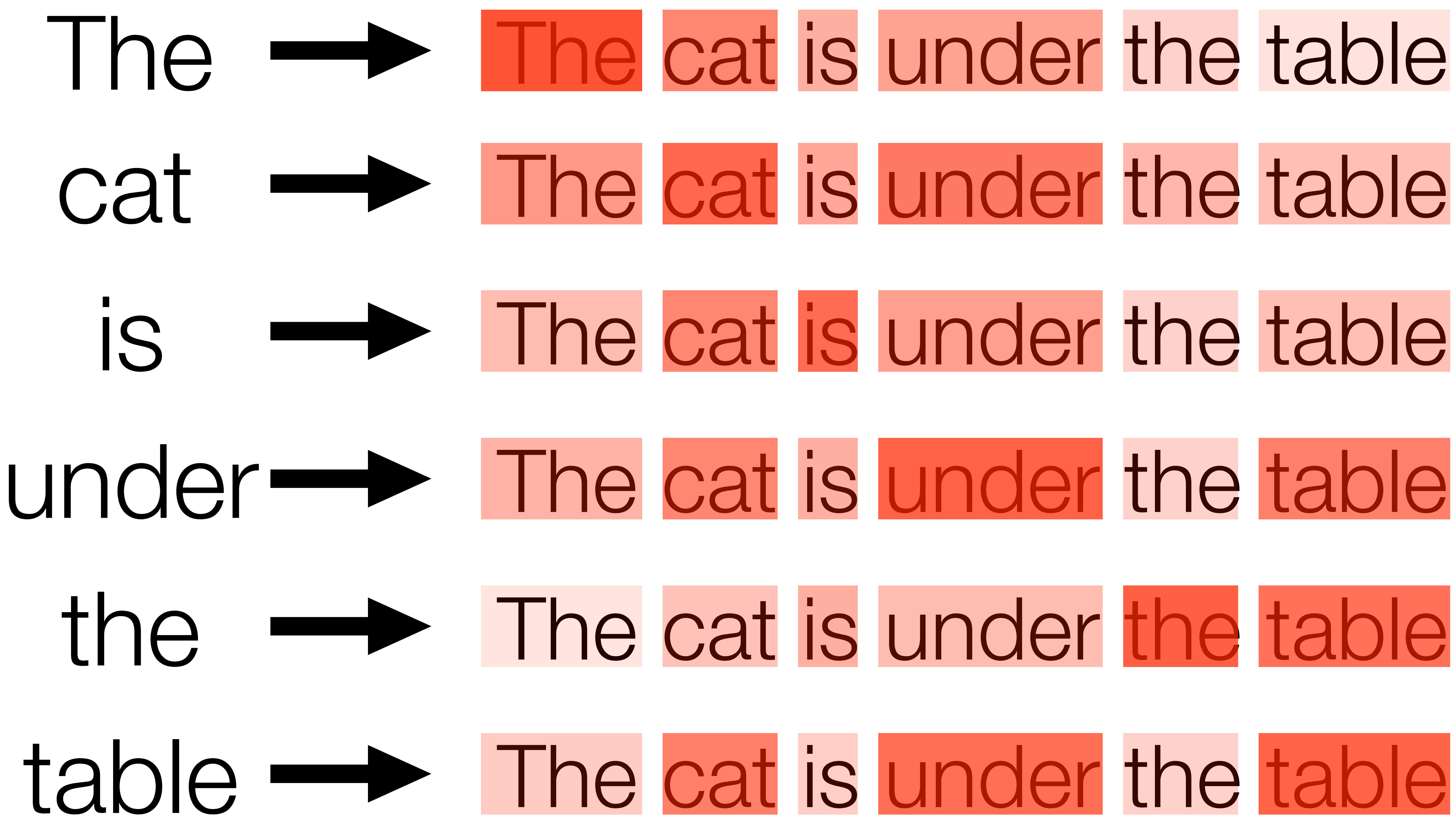
Embedding of “Cat” with context info

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

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

Multi-Head Attention (1)





The → The cat is under the table

cat → The cat is under the table

is → The cat is under the table

under → The cat is under the table

the → The cat is under the table

table → The cat is under the table

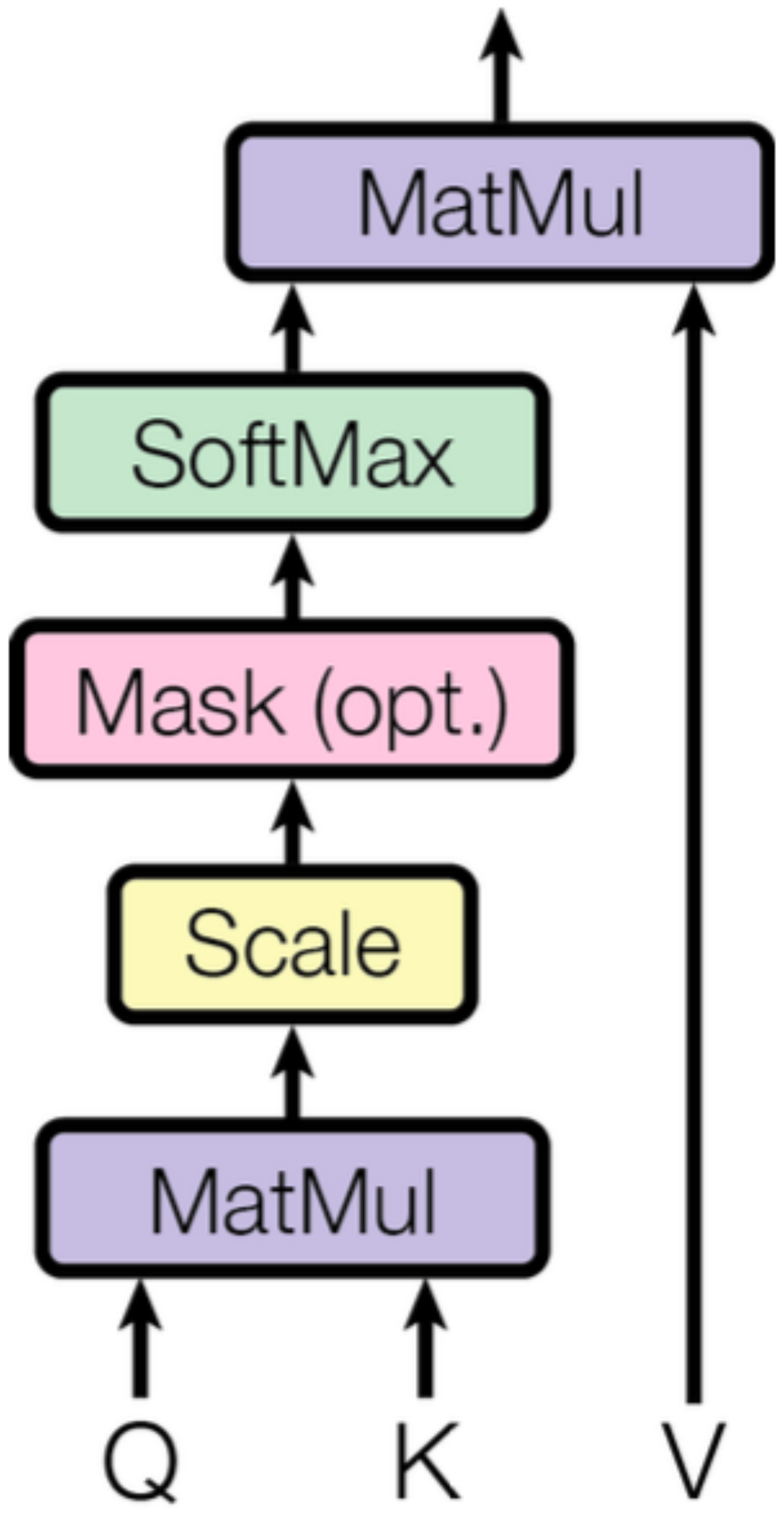
Attention Vectors

[0.70 0.10 0.05 0.09 0.2 0.04]

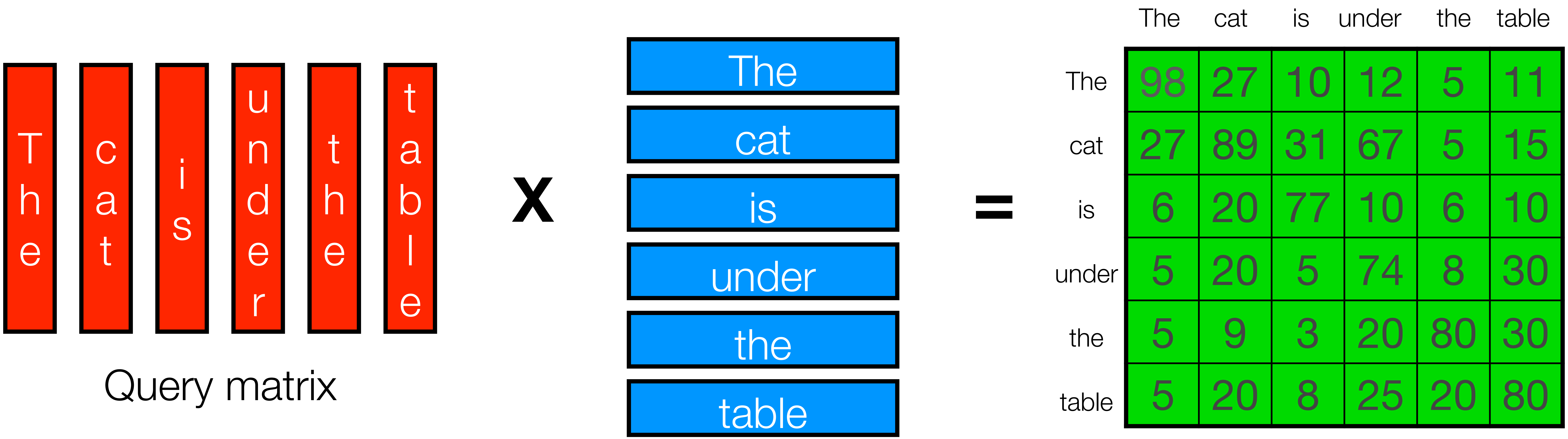
[0.13 0.64 0.02 0.10 0.01 0.10]

.
.
.

Multi-Head Attention (4)



Multi-Head Attention (3)



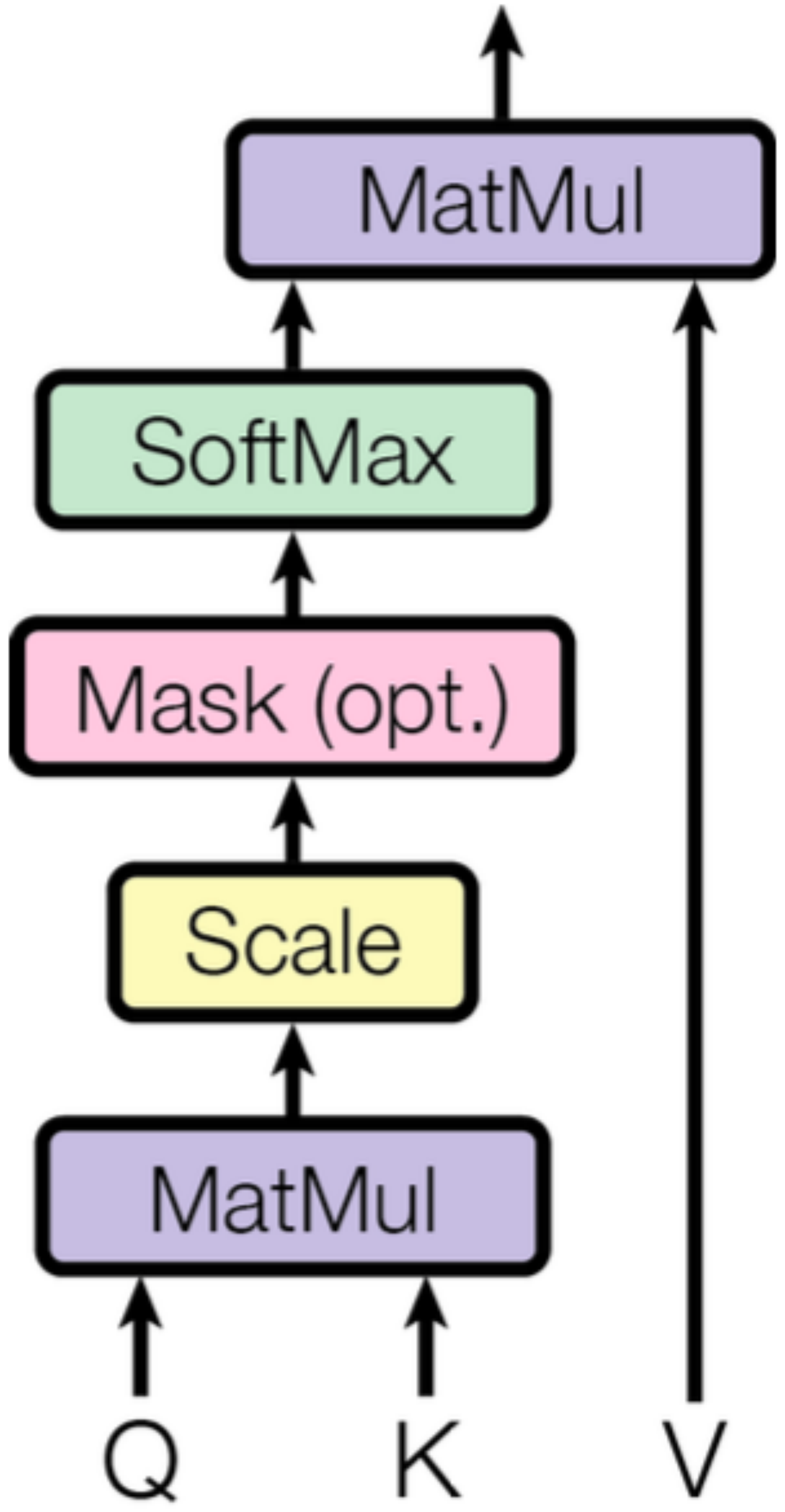
Query matrix

Key matrix

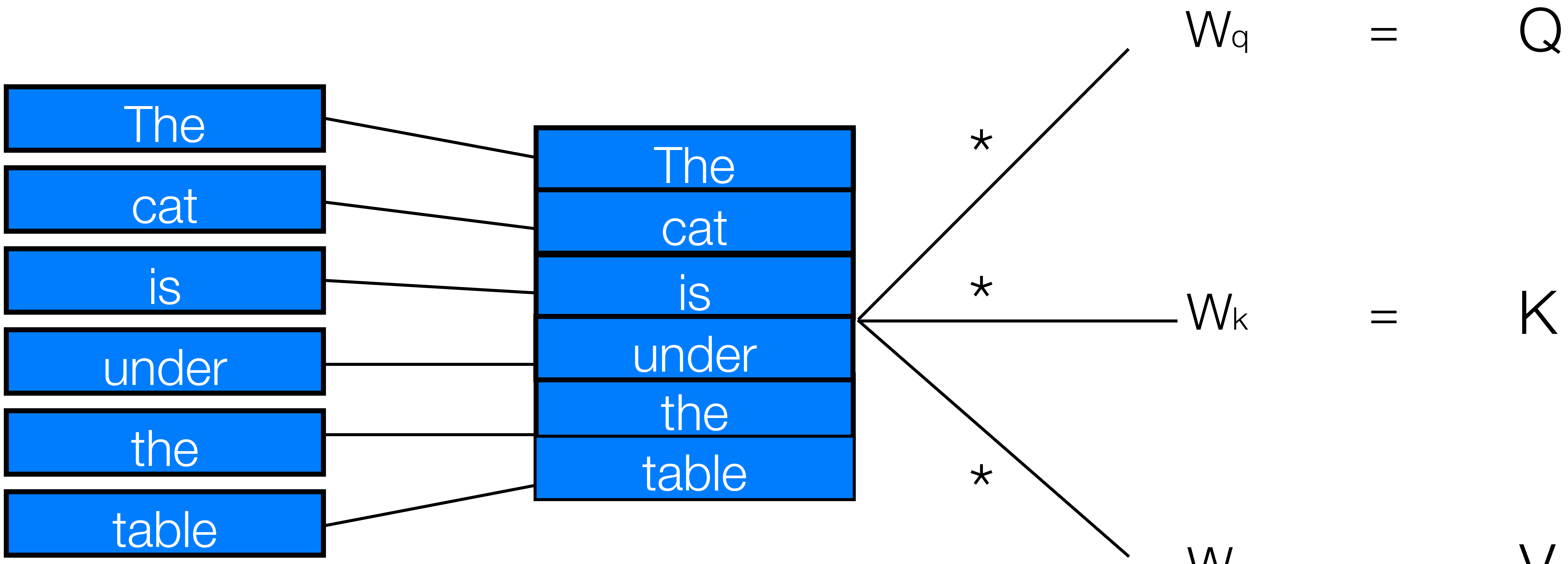
	The	cat	is	under	the	table
The	98	27	10	12	5	11
cat	27	89	31	67	5	15
is	6	20	77	10	6	10
under	5	20	5	74	8	30
the	5	9	3	20	80	30
table	5	20	8	25	20	80

Score matrix

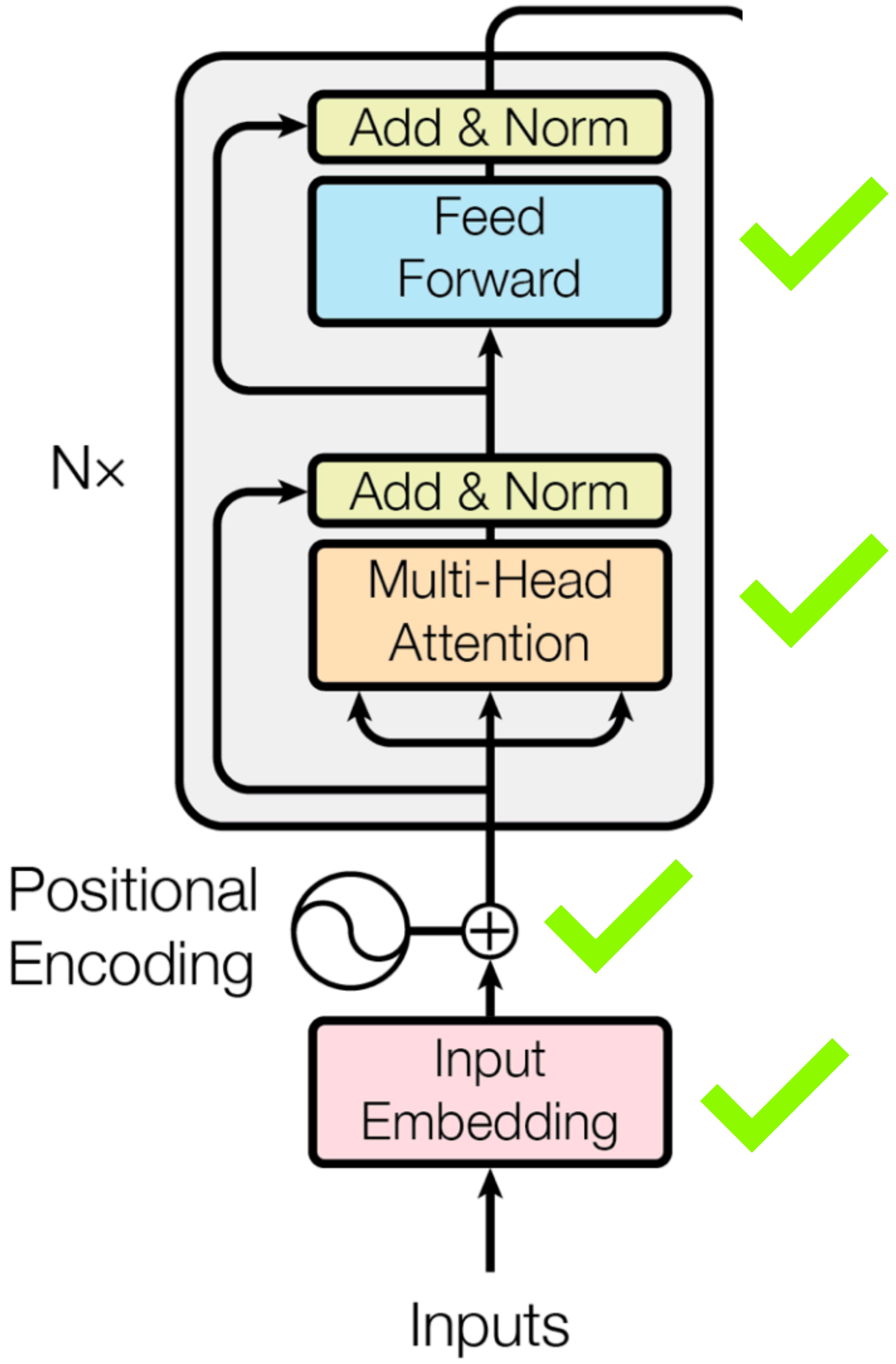
Multi-Head Attention (4)



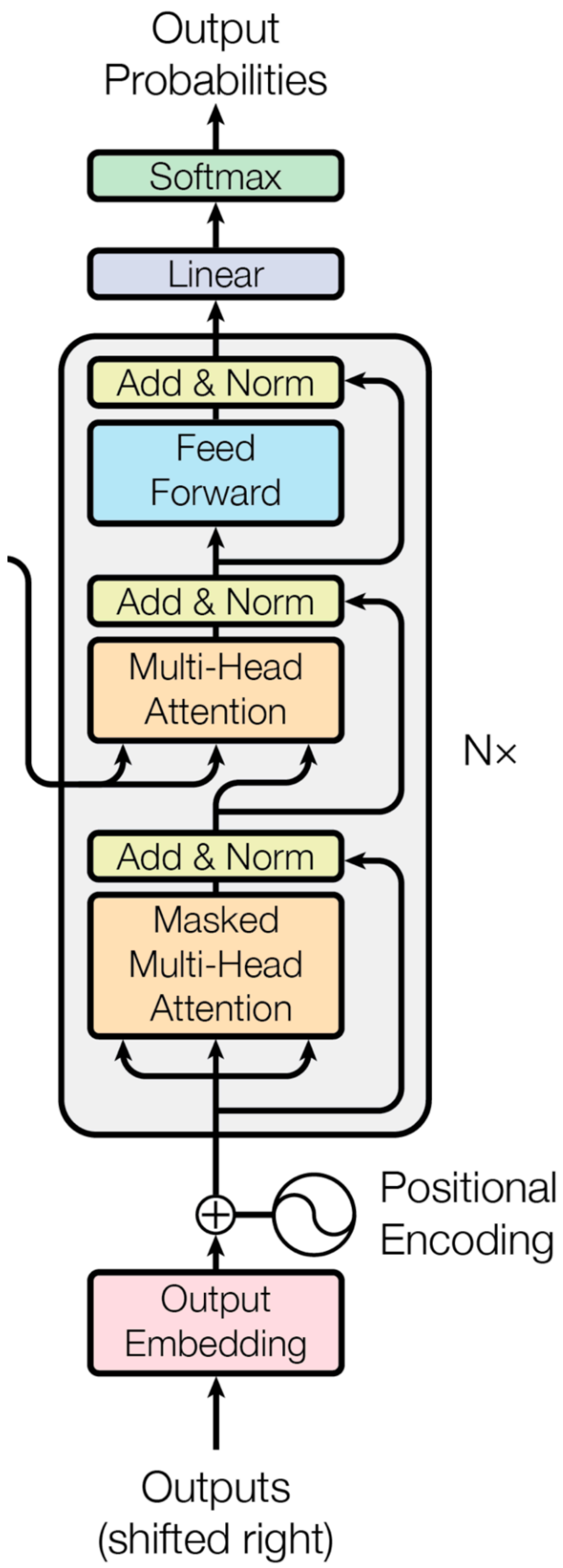
Multi-Head Attention (5)



Feed Forward



Decoder


Output
Probabilities
Softmax
Linear
Add & Norm
Feed
Forward
Add & Norm
Multi-Head
Attention
Nx
Add & Norm
Masked
Multi-Head
Attention
Positional
Encoding
Output
Embedding
Outputs
(shifted right)

Masked Multi-Head Attention

	Il	gatto	è	sotto	il	tavolo
Il	98	~~27~~	~~10~~	~~12~~	~~5~~	~~11~~
gatto	27	89	~~31~~	~~67~~	~~5~~	~~15~~
è	6	20	77	~~10~~	~~6~~	~~10~~
sotto	5	20	5	74	~~8~~	~~30~~
il	5	9	3	20	80	~~30~~
tavolo	5	20	8	25	20	80

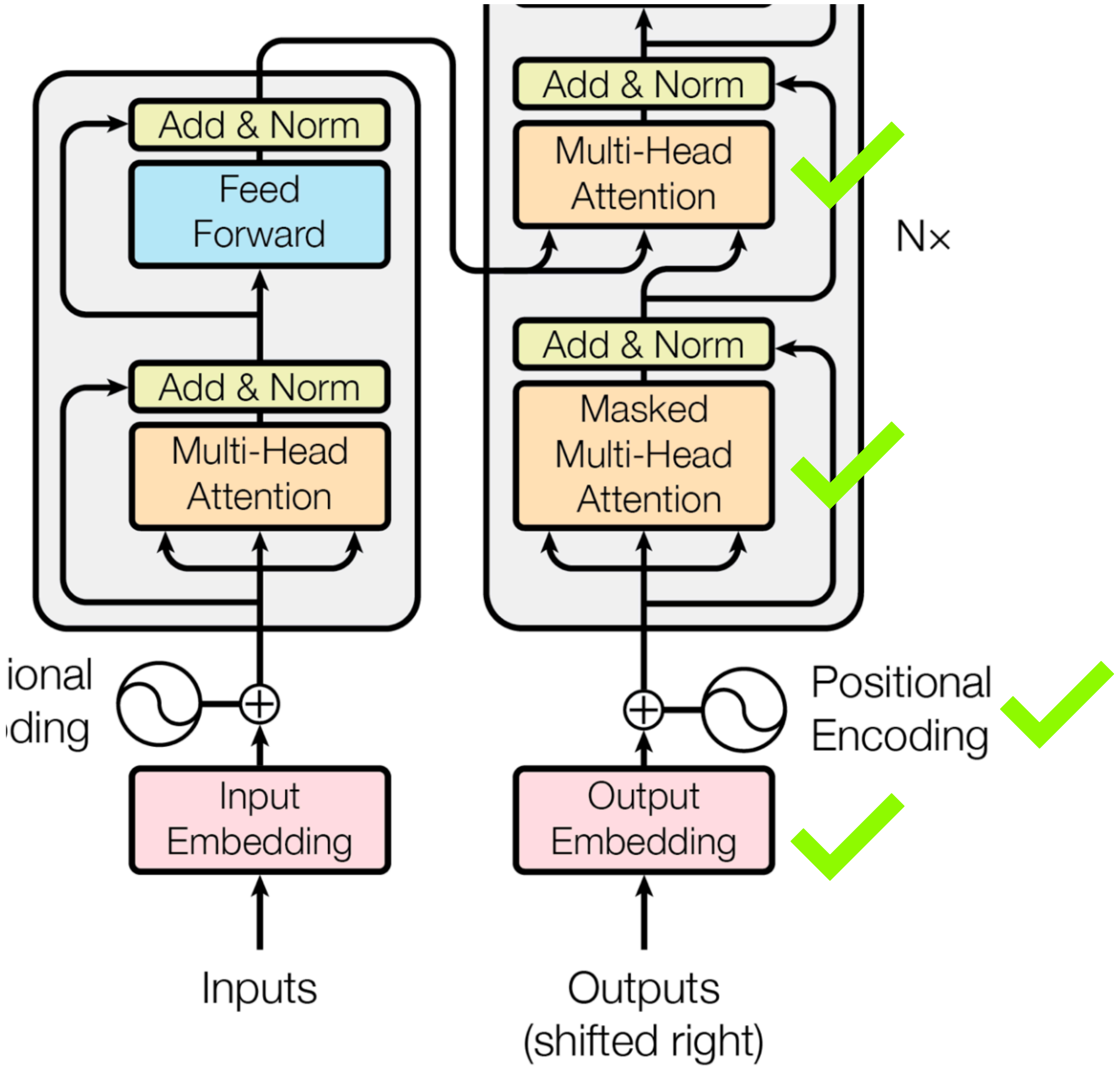
Score matrix

+

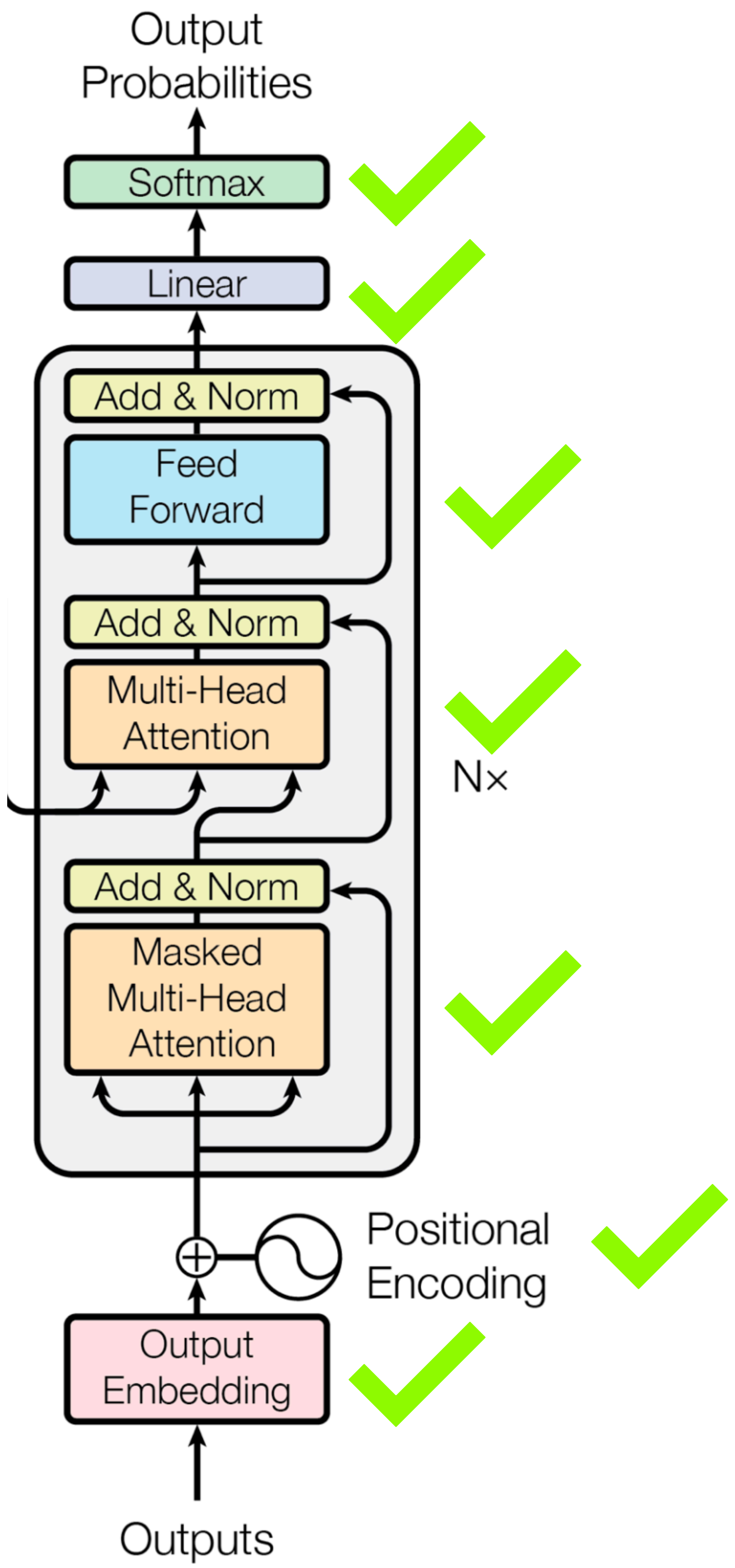
0	-inf	-inf	-inf	-inf	-inf
0	0	-inf	-inf	-inf	-inf
0	0	0	-inf	-inf	-inf
0	0	0	0	-inf	-inf
0	0	0	0	0	-inf
0	0	0	0	0	0

Mask matrix

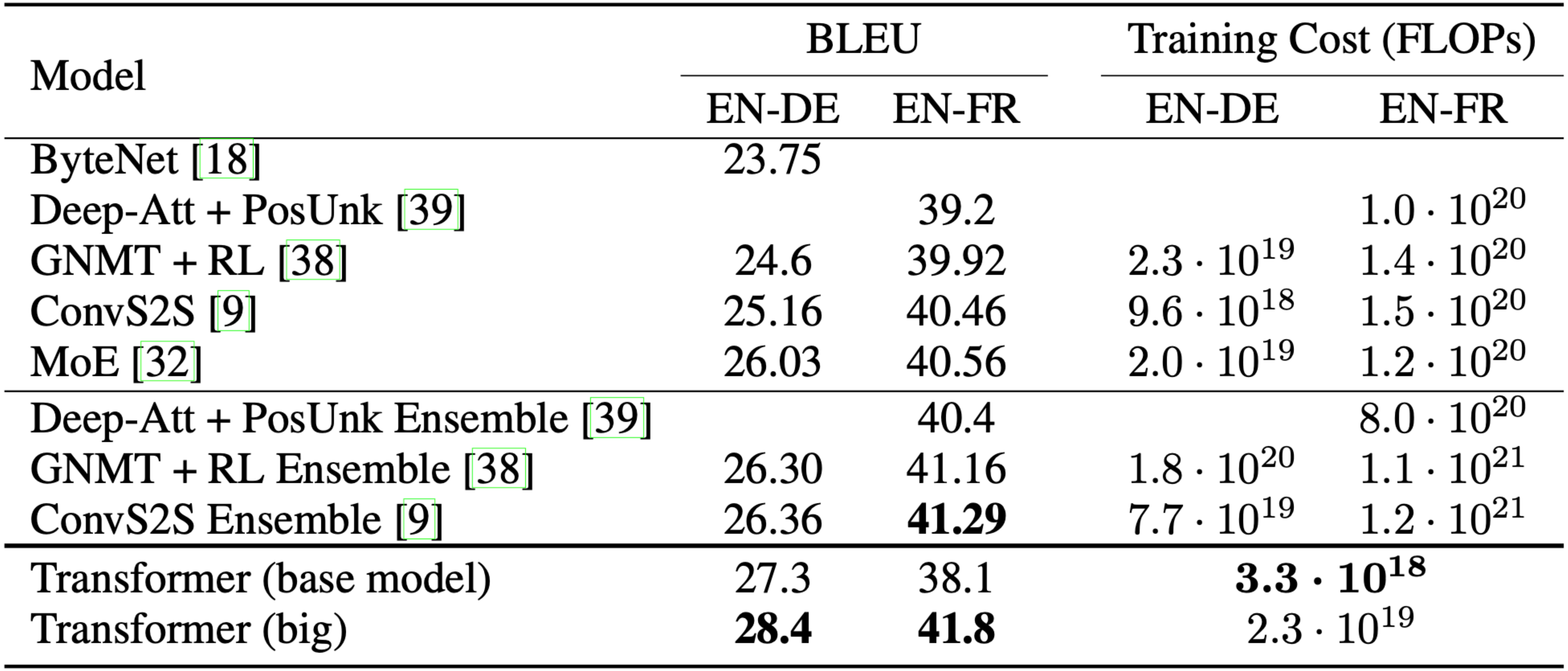
Multi-Head Attention



Multi-Head Attention



Results



Model	BLEU		Training Cost (FLOPs)	
	EN-DE	EN-FR	EN-DE	EN-FR
ByteNet [18]	23.75			
Deep-Att + PosUnk [39]		39.2		$1.0 \cdot 10^{20}$
GNMT + RL [38]	24.6	39.92	$2.3 \cdot 10^{19}$	$1.4 \cdot 10^{20}$
ConvS2S [9]	25.16	40.46	$9.6 \cdot 10^{18}$	$1.5 \cdot 10^{20}$
MoE [32]	26.03	40.56	$2.0 \cdot 10^{19}$	$1.2 \cdot 10^{20}$
Deep-Att + PosUnk Ensemble [39]		40.4		$8.0 \cdot 10^{20}$
GNMT + RL Ensemble [38]	26.30	41.16	$1.8 \cdot 10^{20}$	$1.1 \cdot 10^{21}$
ConvS2S Ensemble [9]	26.36	**41.29**	$7.7 \cdot 10^{19}$	$1.2 \cdot 10^{21}$
Transformer (base model)	27.3	38.1	$\mathbf{3.3 \cdot 10^{18}}$	
Transformer (big)	**28.4**	**41.8**	$2.3 \cdot 10^{19}$	

Thanks for your attention

Q&A