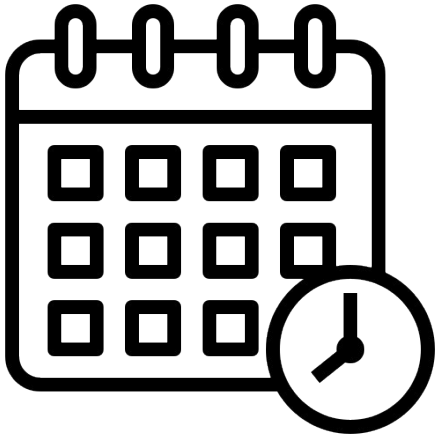
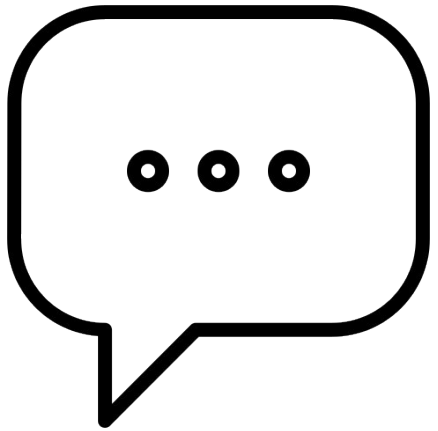
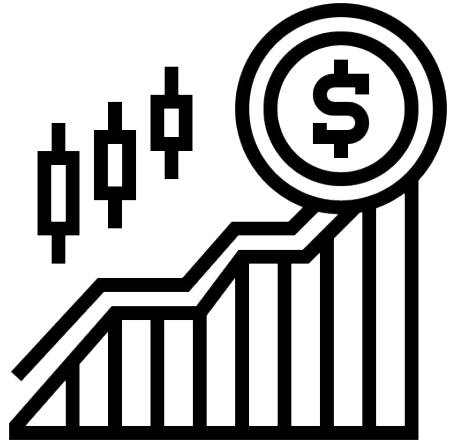
Hands-on Introduction to Deep Learning

Sequences

X_0 → X_1 → ... → X_t → ... → X_{n-1} → X_n

Sequential Data

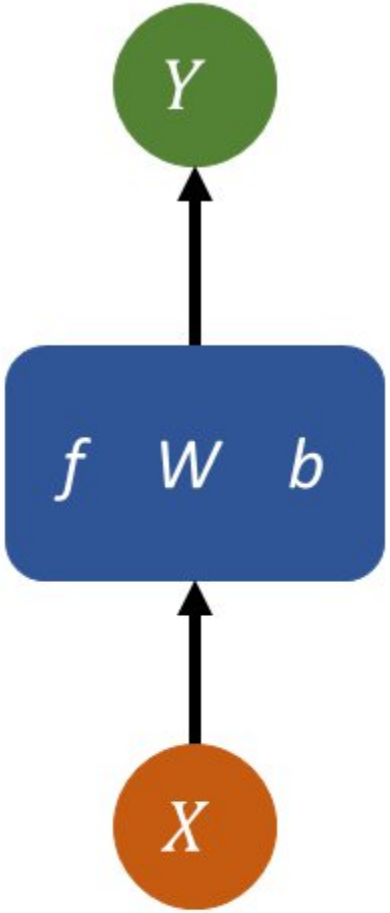
Stock market

	day 1	day 2	day 3
asset 1	9.77	79.94	64.13
asset 2	47.66	74.07	70.90
asset 3	94.25	76.34	99.95
asset 4	41.19	9.99	89.50
asset 5	65.44	63.79	67.14

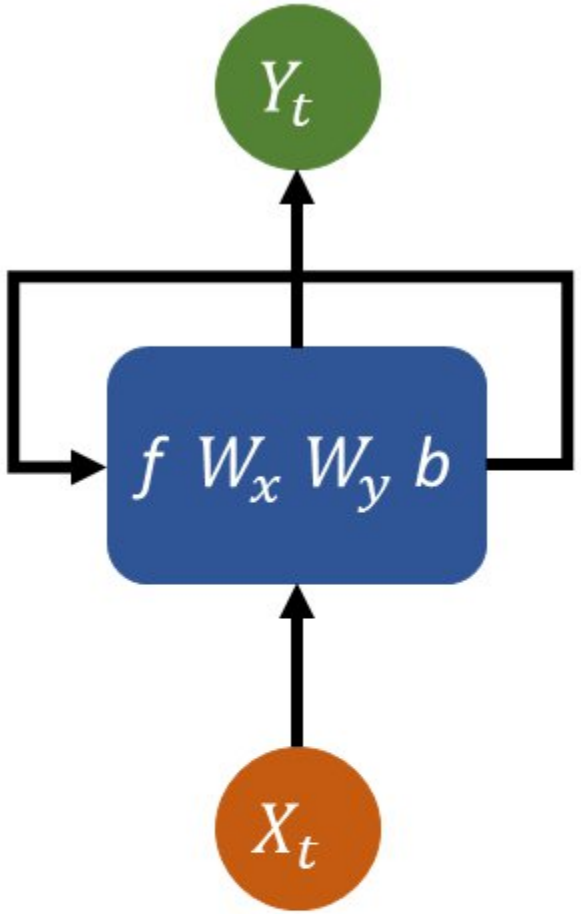
Text

I	am	learning	.
0,83	0,65	-0,90	-0,04
-0,53	0,81	-0,61	-0,12
0,24	-0,14	0,58	0,66
-0,31	0,32	0,37	-0,11
-0,53	0,50	-0,96	0,48
-0,34	-0,85	0,19	-0,78
-0,79	0,53	-0,31	-0,28
-0,23	-0,13	0,33	0,45
0,95	0,53	0,74	-0,24
-0,60	0,04	-0,96	-0,96

Sequential Data - Examples

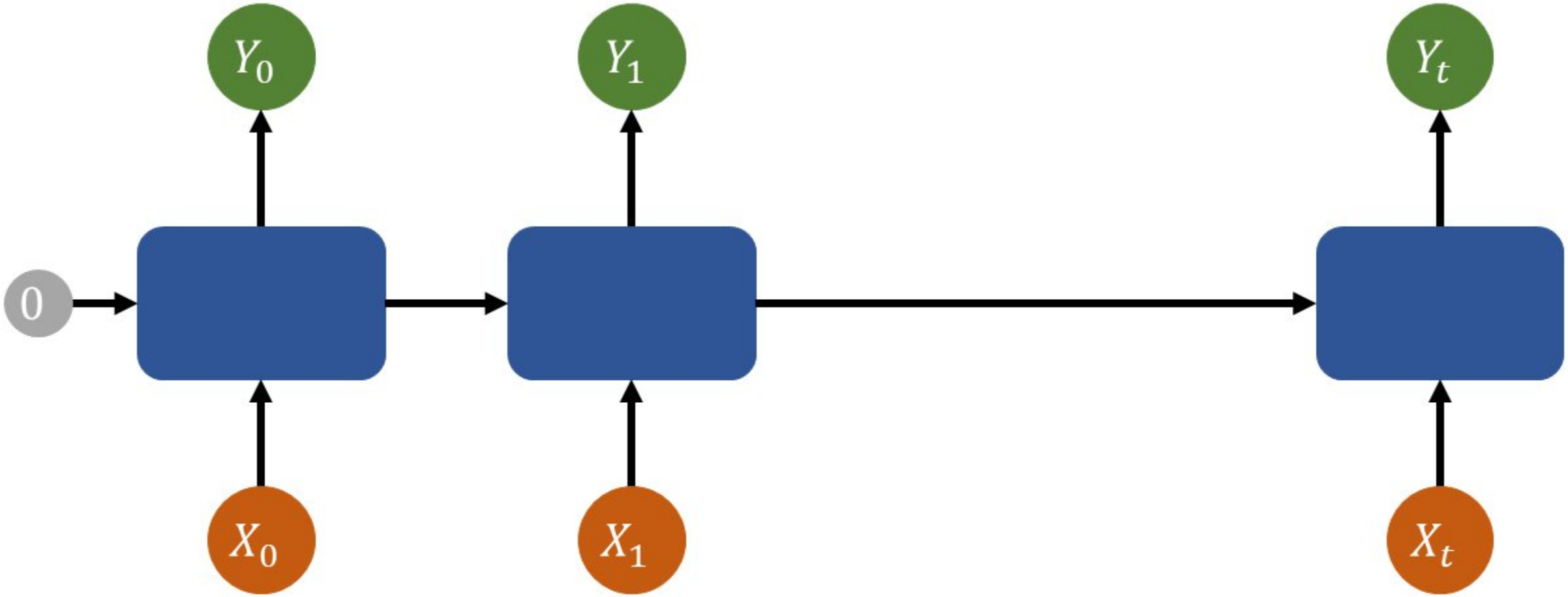


$$Y = f(W.X + b)$$

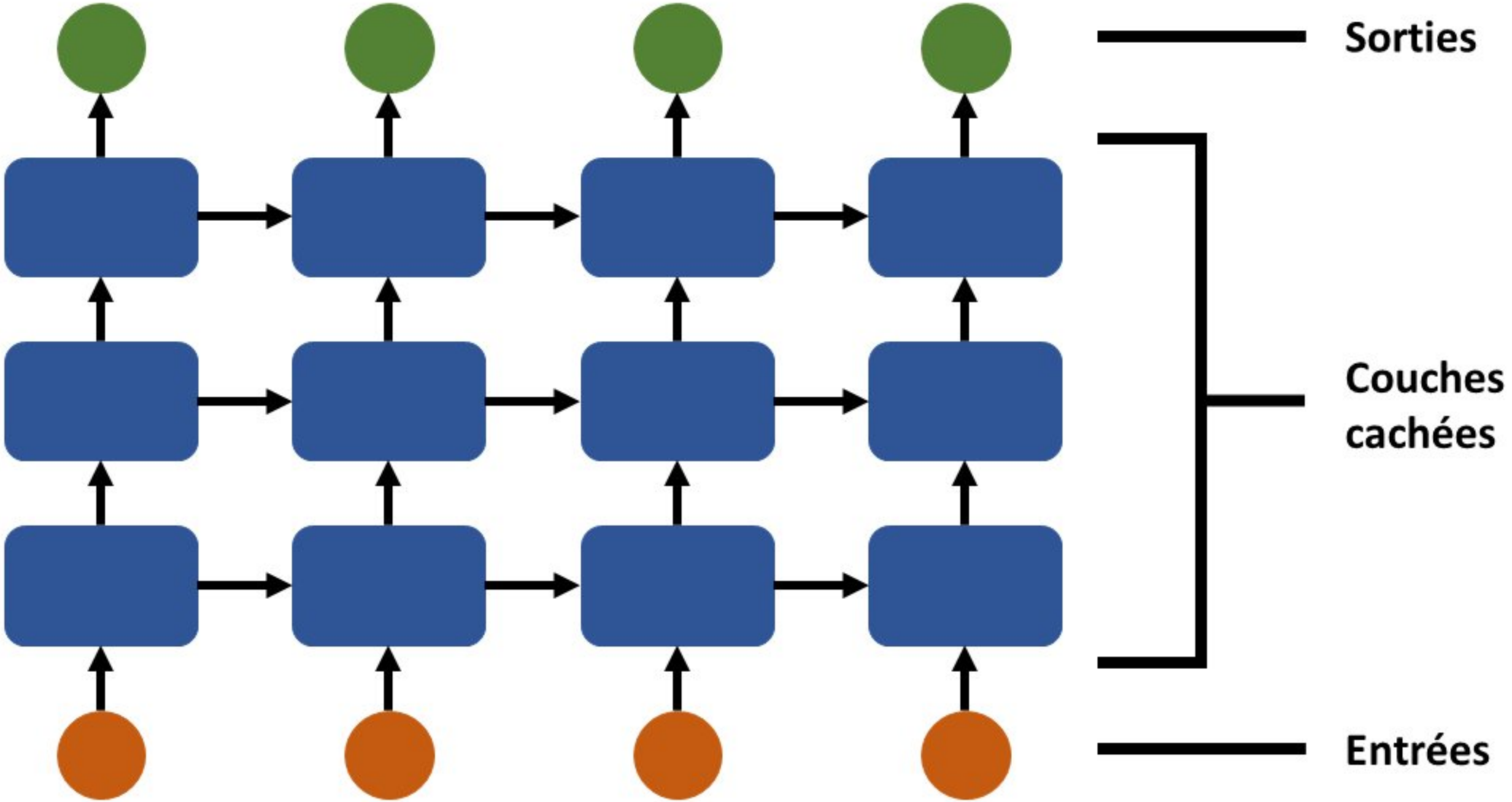


$$Y_t = f(W_x.X_t + W_y Y_{t-1} + b)$$

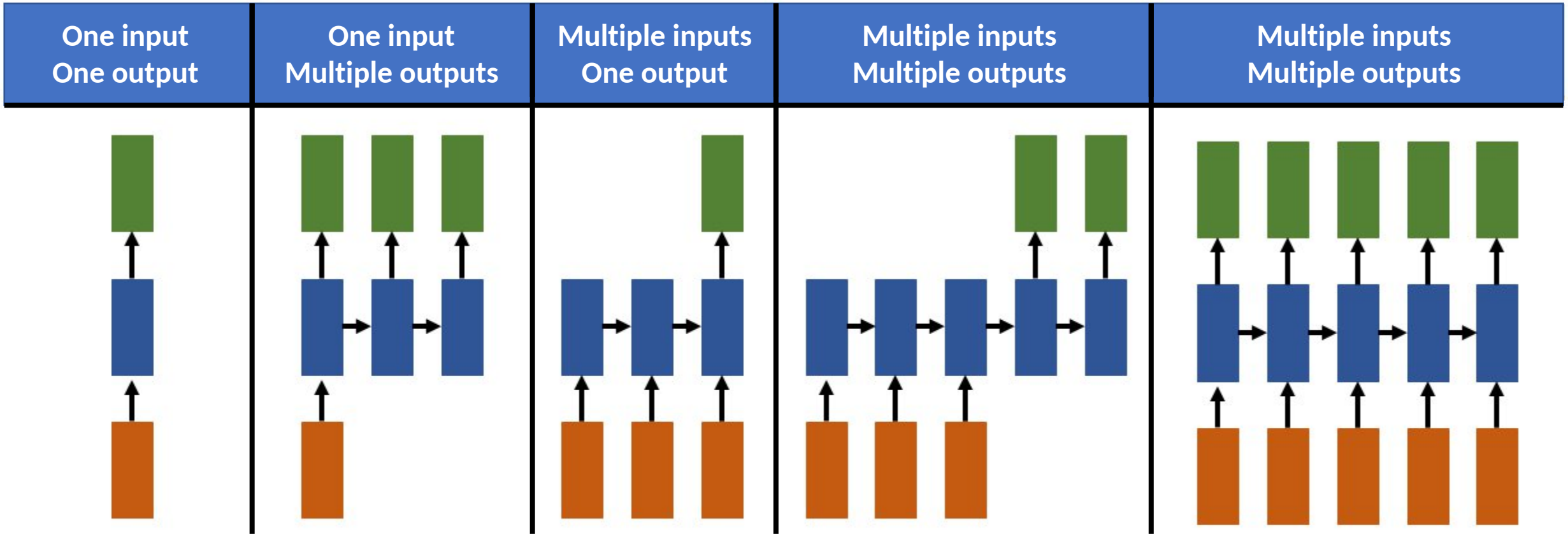
Simple recurrent neuron



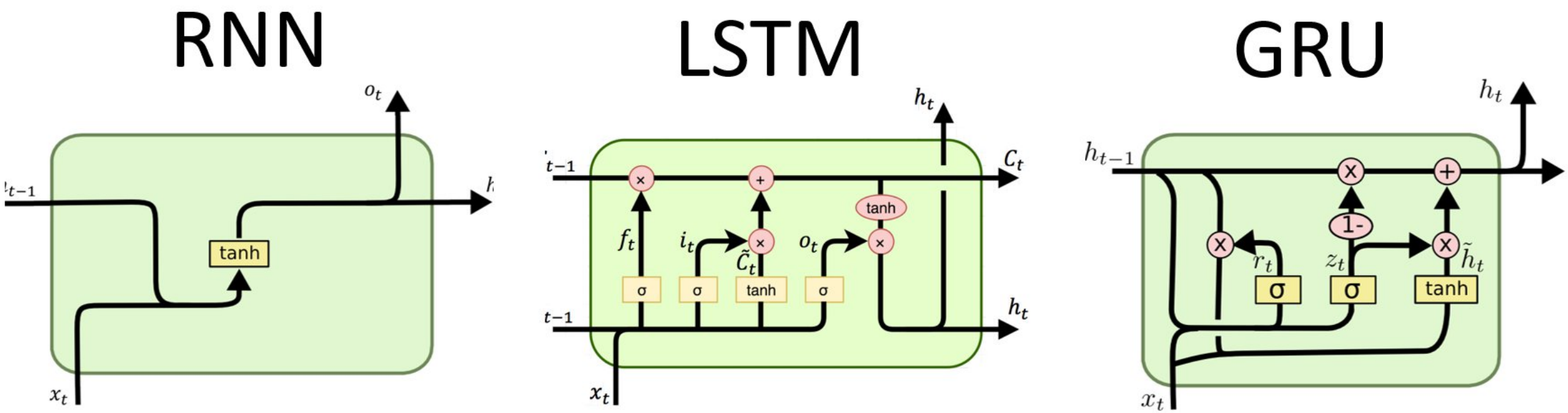
Recurrent neuron unfolded



Recurrent Neural Network

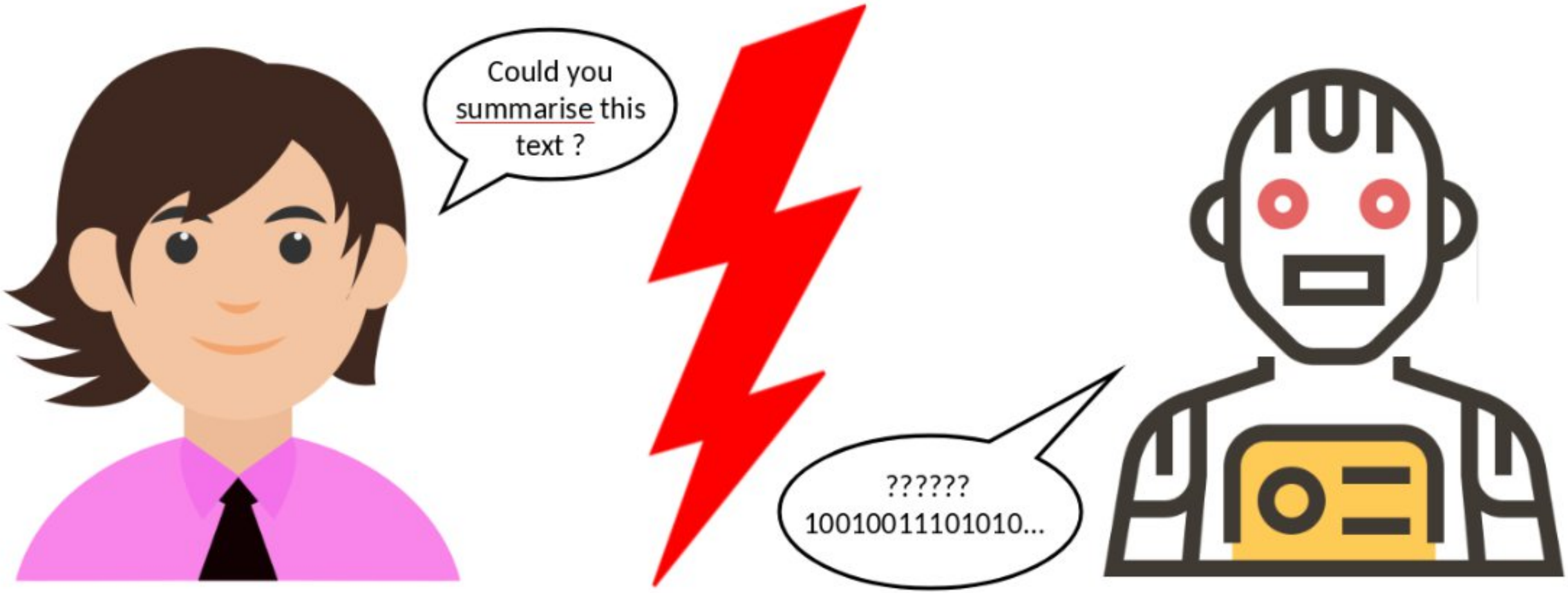


A flexible model type

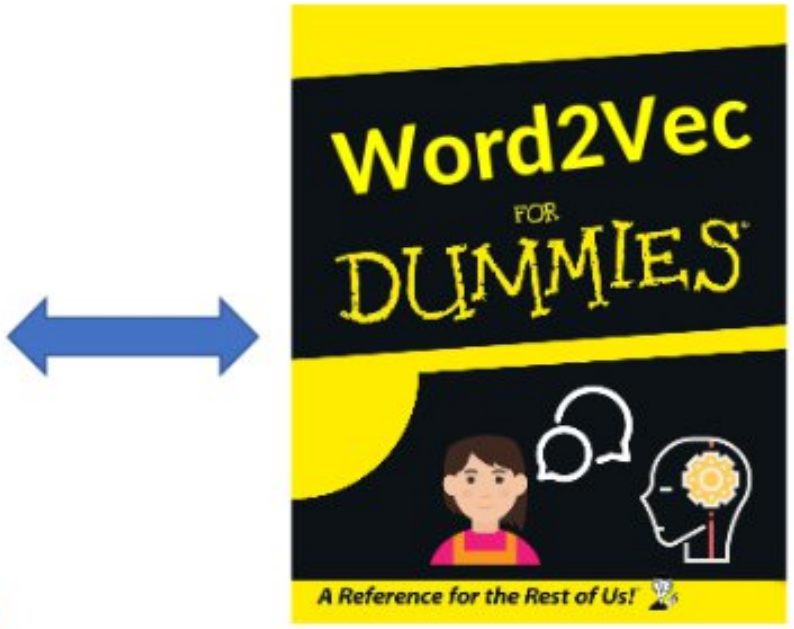


Tembhurne, Jitendra V., and Tausif Diwan. « Sentiment analysis in textual, visual and multimodal inputs using recurrent neural networks. » *Multimedia Tools and Applications* 80.5 (2021) : 6871-6910.

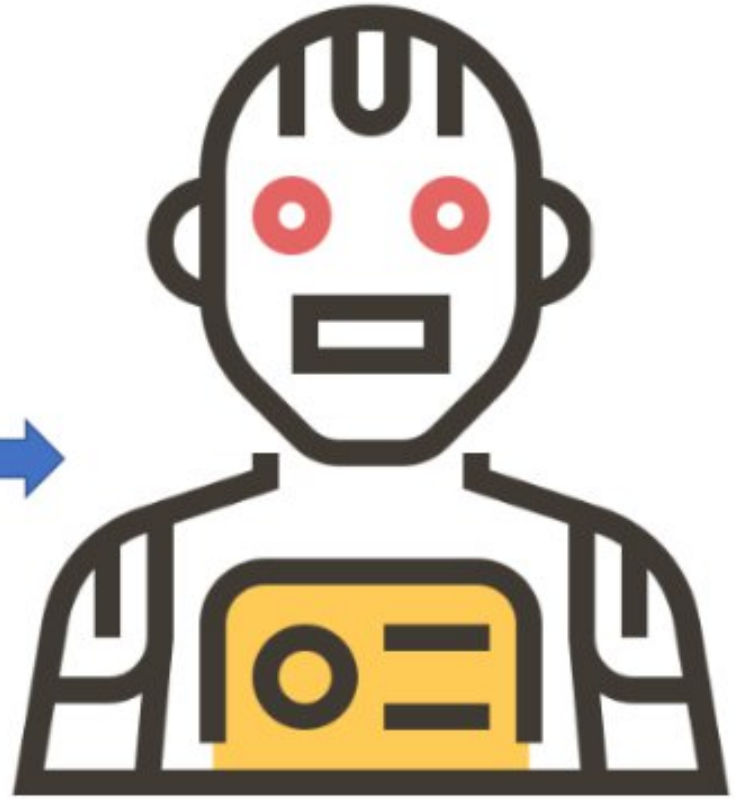
Simple RNN vs LSTM vs GRU



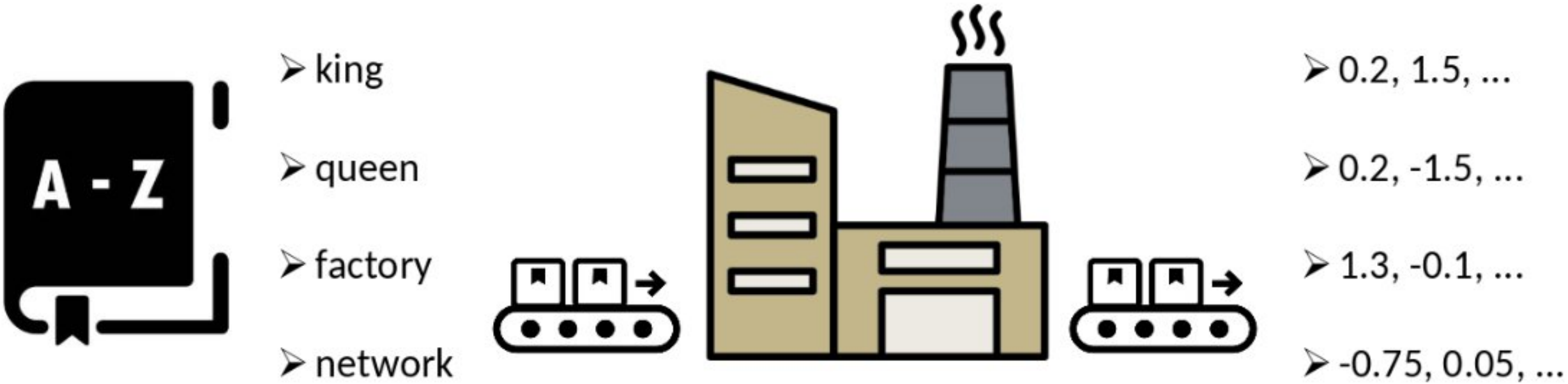
The language barrier between Men and Machines



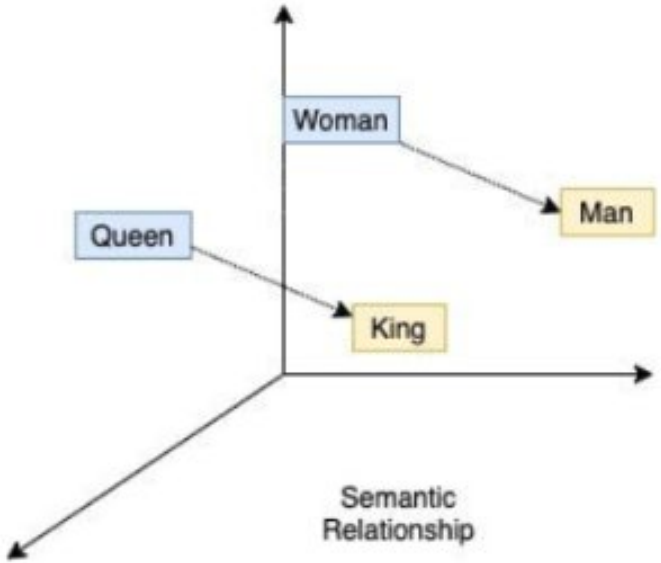
Embedding layer

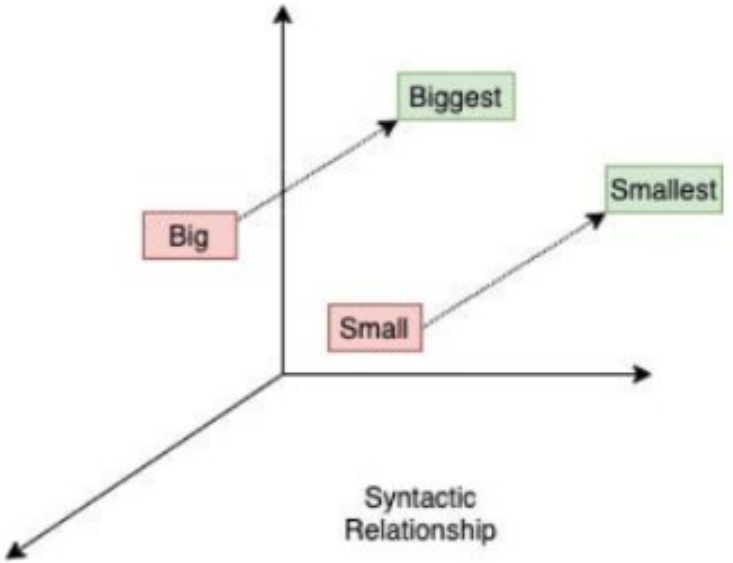


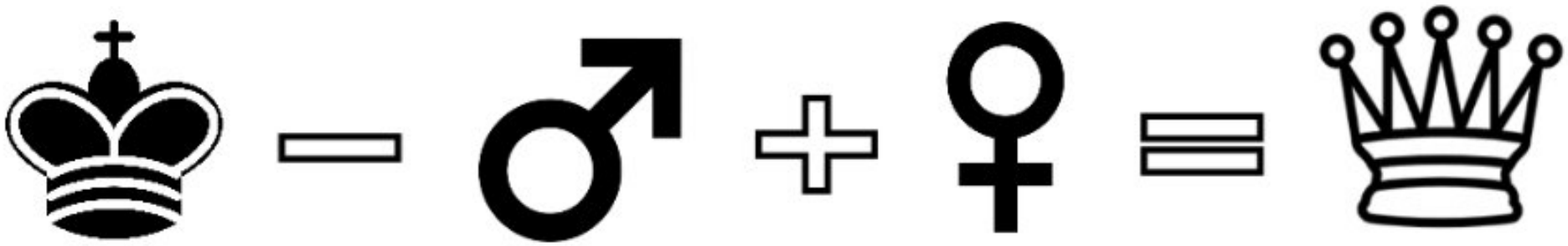
Embeddings



Embeddings


Woman
Man
Queen
King
Semantic
Relationship


Biggest
Smallest
Big
Small
Syntactic
Relationship



Embeddings

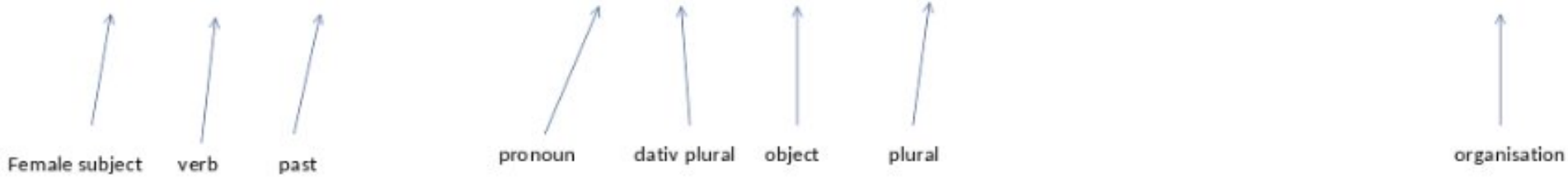
Anti - *chocs*, ***constitution***
et *actu* - ***elle*** - ***ment*** sont appris.

Anticonstitutionnellement sera compris.

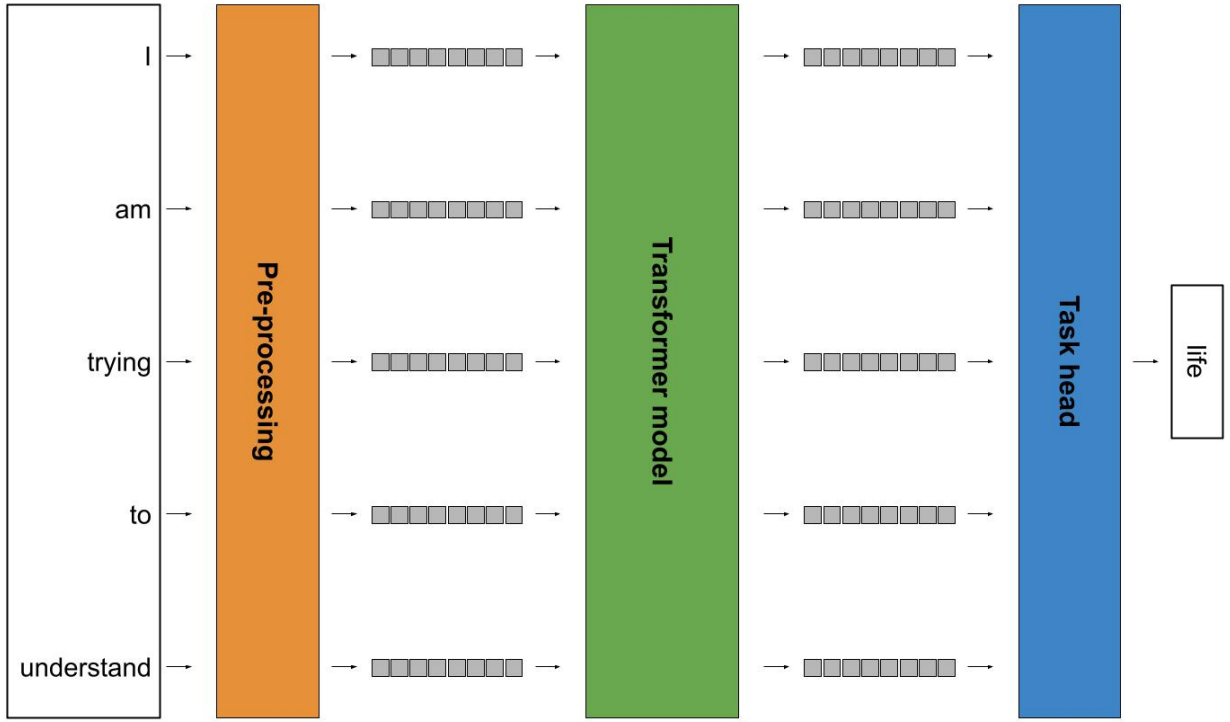
Tokens are a smaller unit of meaning. Splitting words into tokens helps understanding the meaning of the sentence, especially with agglutinative languages.

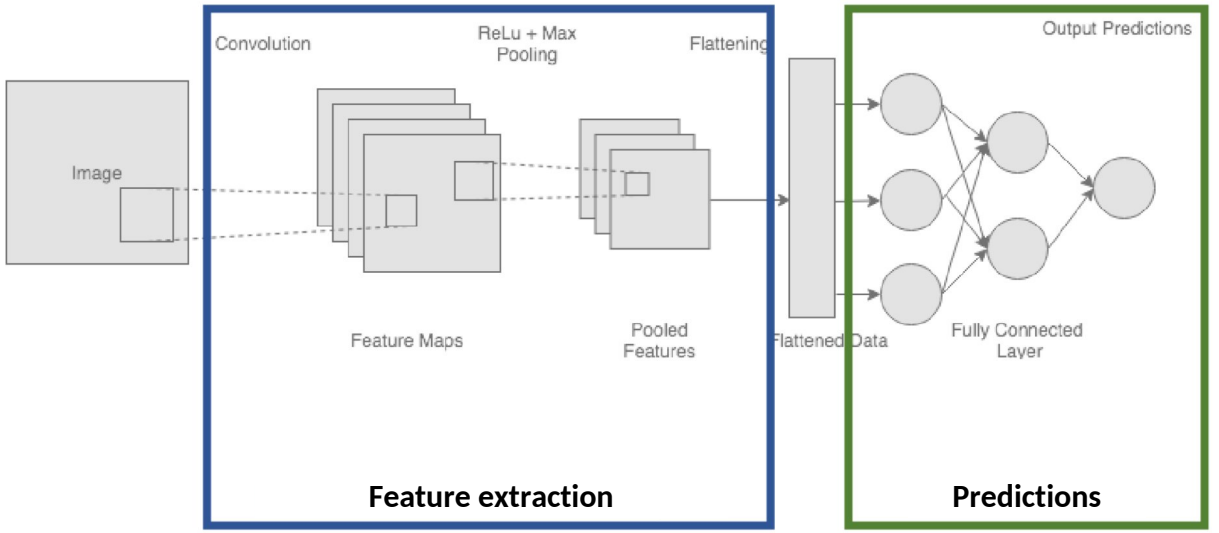
Example : Sie schaute sich mit ihren Freunden ein Spiel von Paris Saint-Germain an.

Sie / schaut / e / sich / mit / ihr / en / Freund / en / ein / Spiel / von / Paris Saint-Germain / an

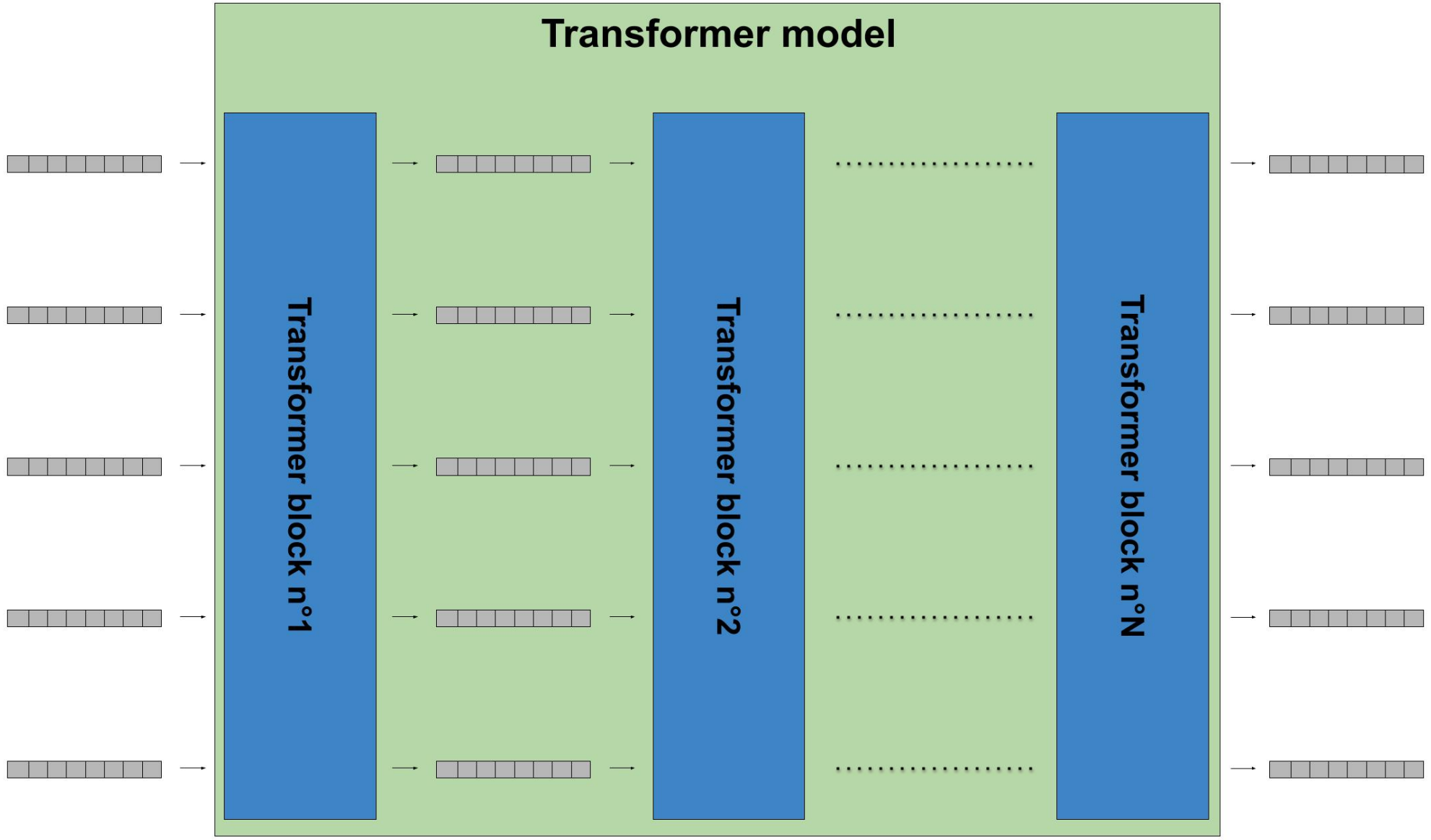


Tokenization

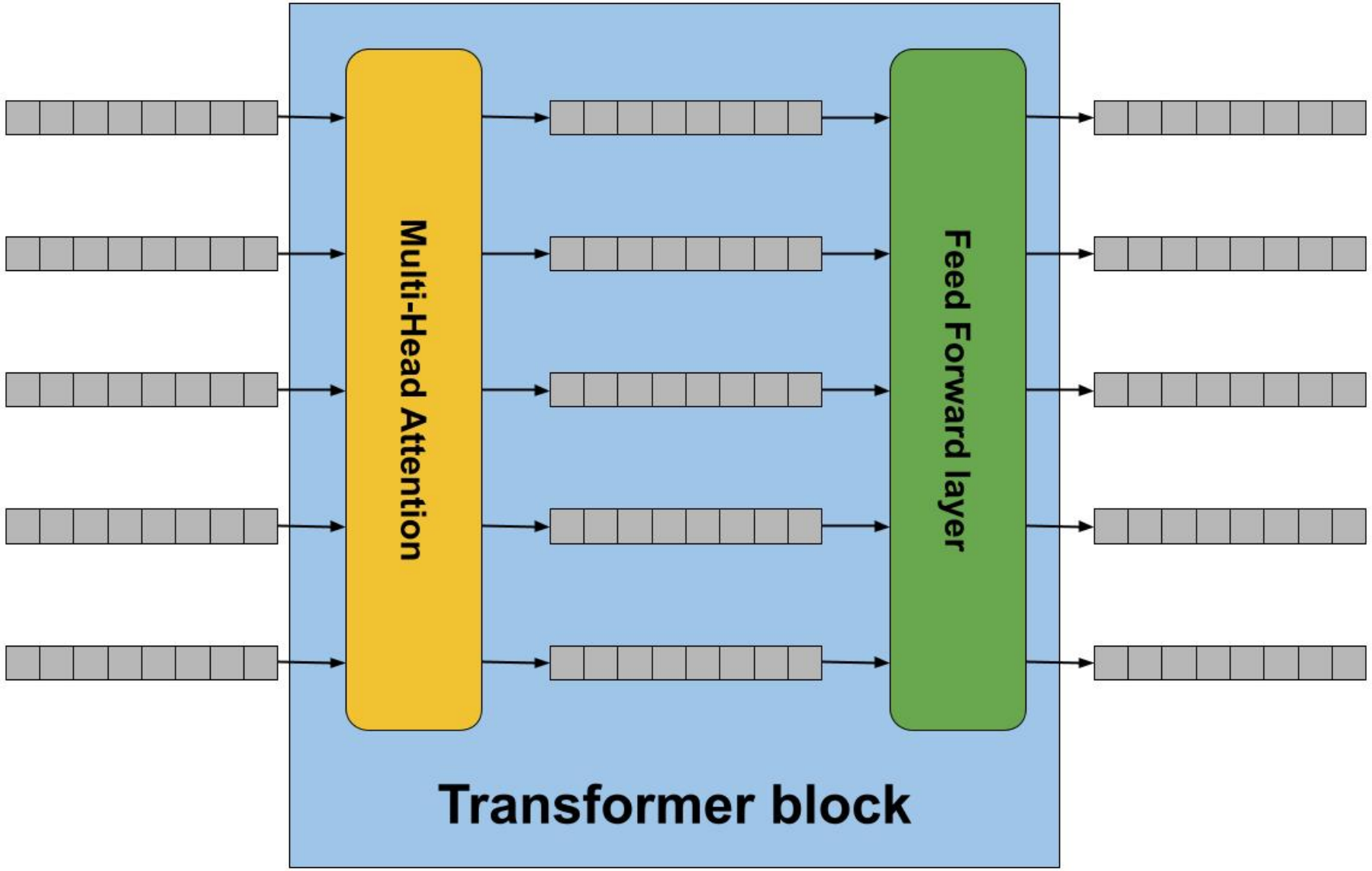




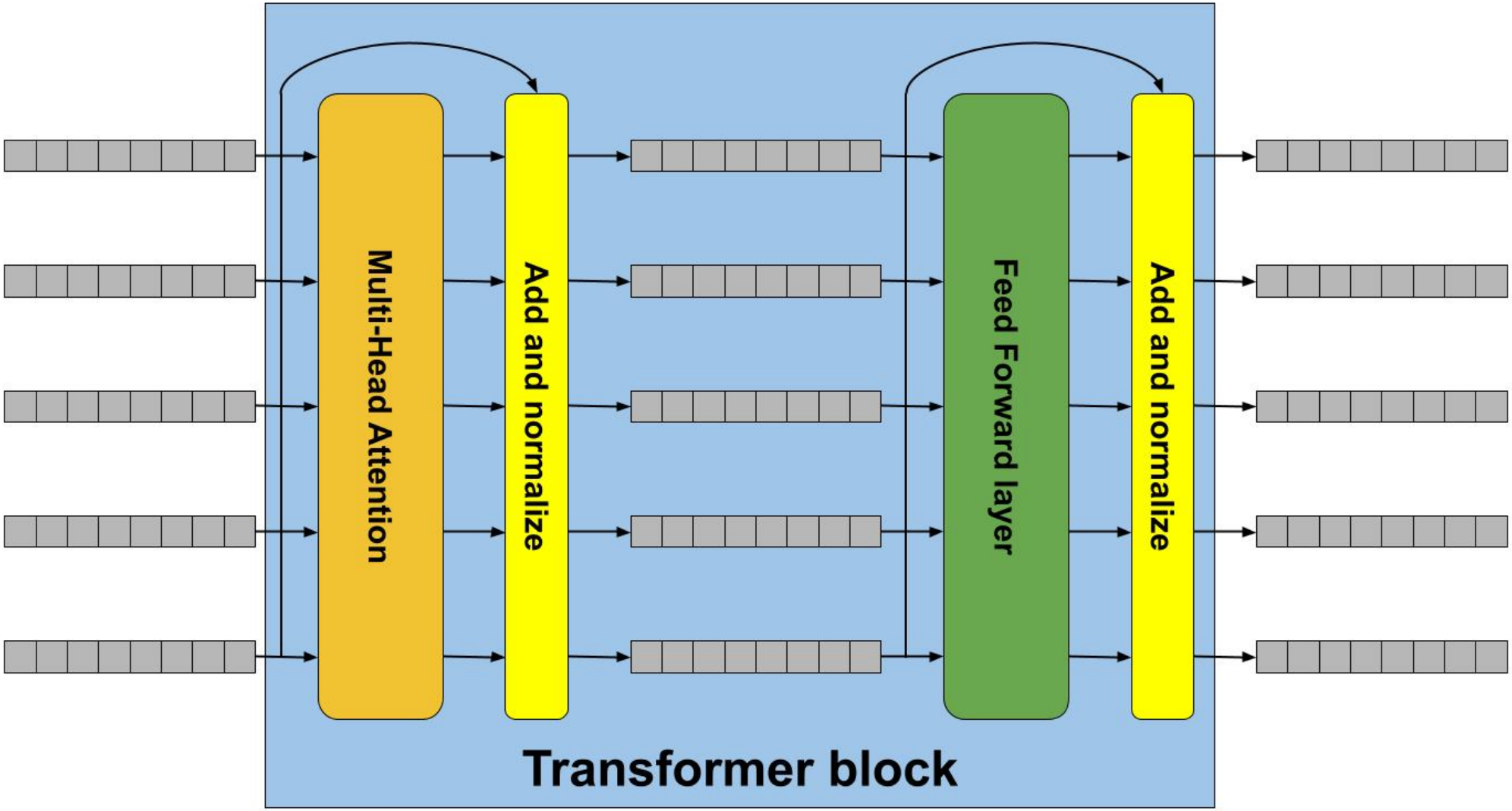
Transformer in a system



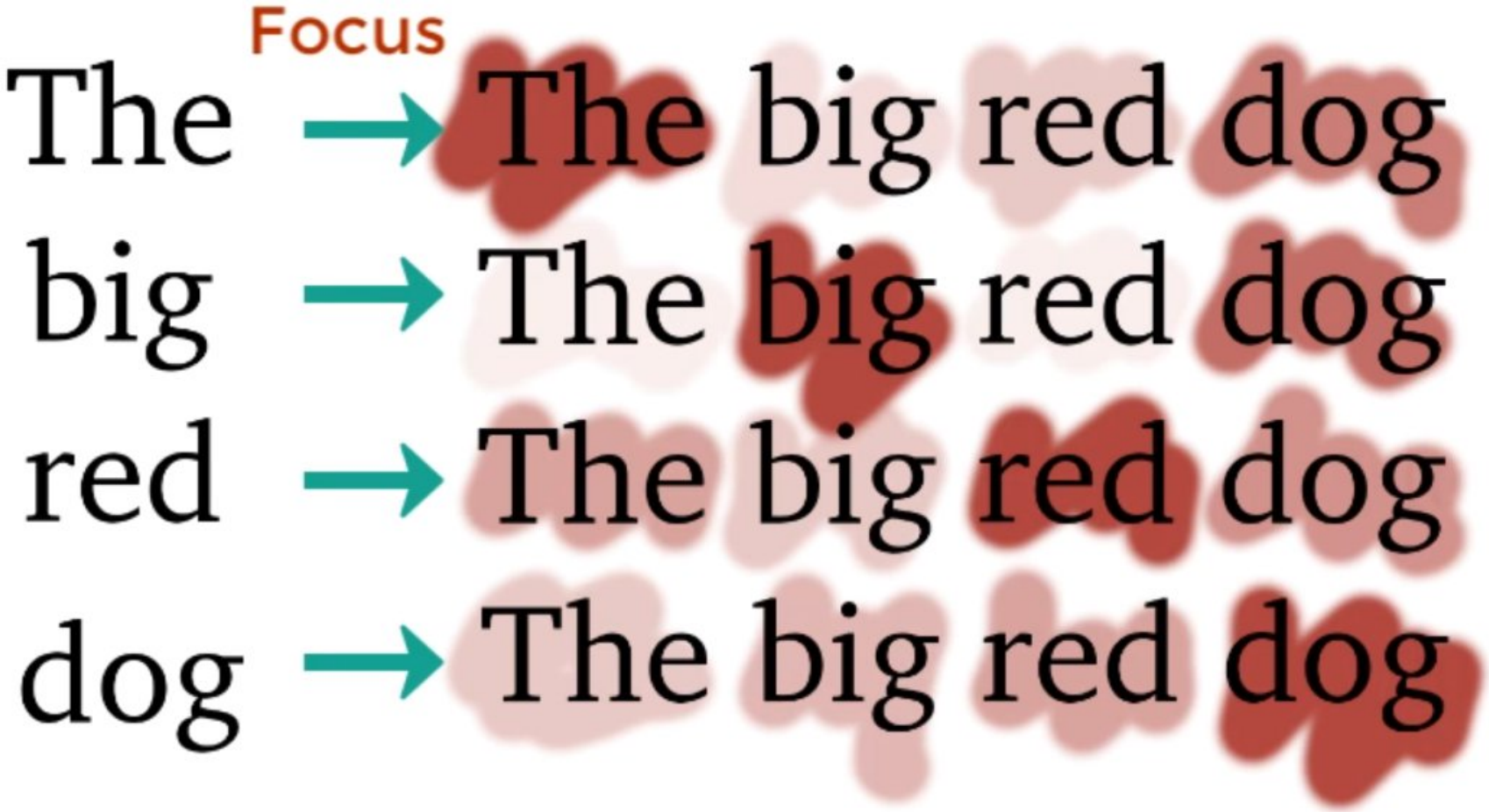
Transformer architecture (1)



Transformer architecture (2)

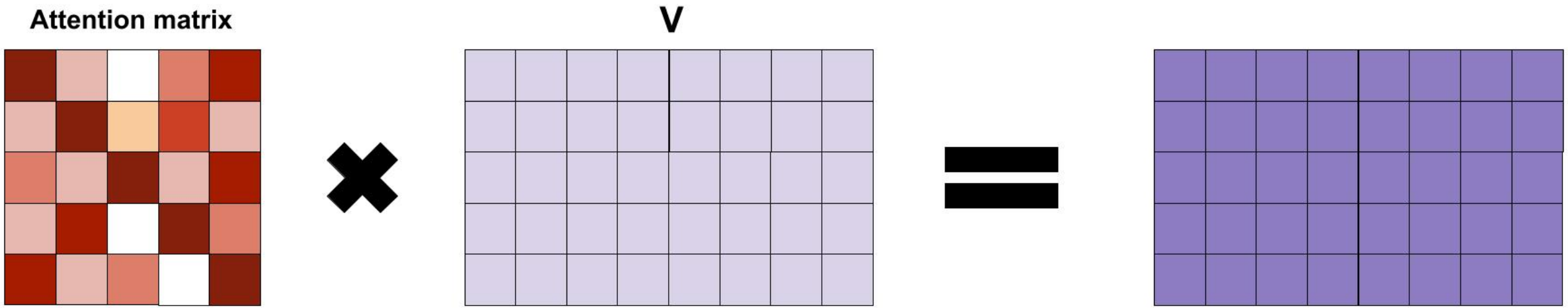


Transformer architecture (3)

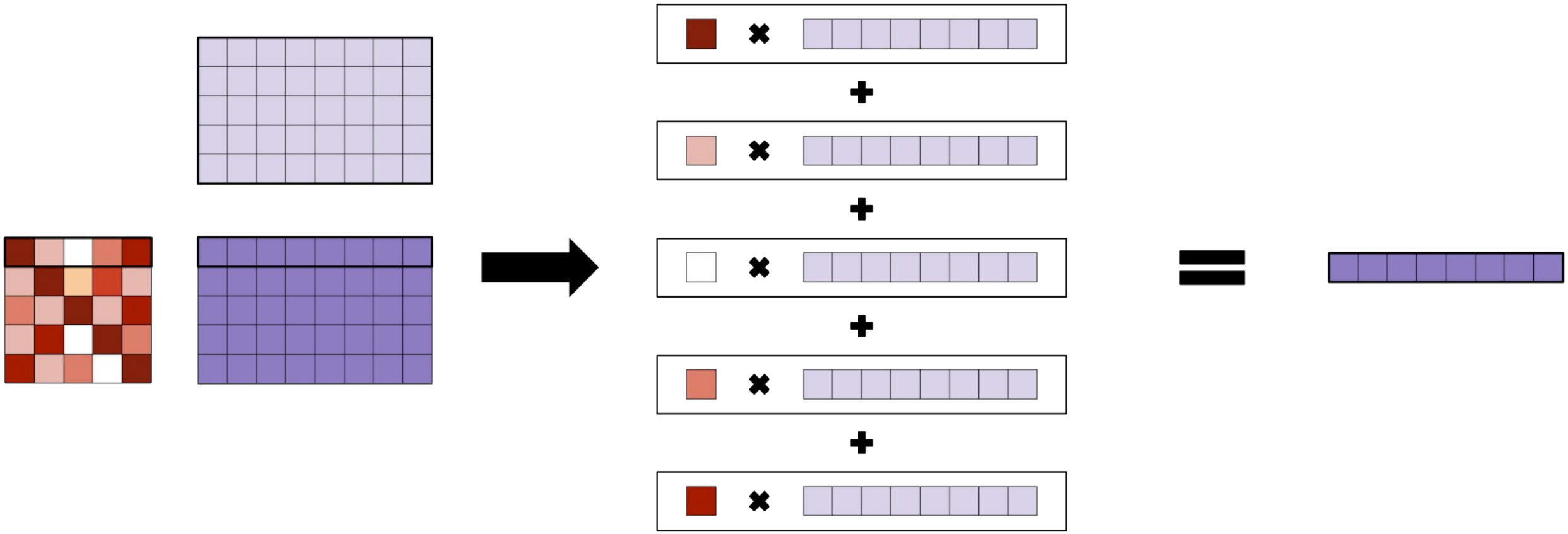


Transformer Neural Networks - EXPLAINED! (Attention is all you need) : https://www.youtube.com/watch?v=TQQlZhbC5ps

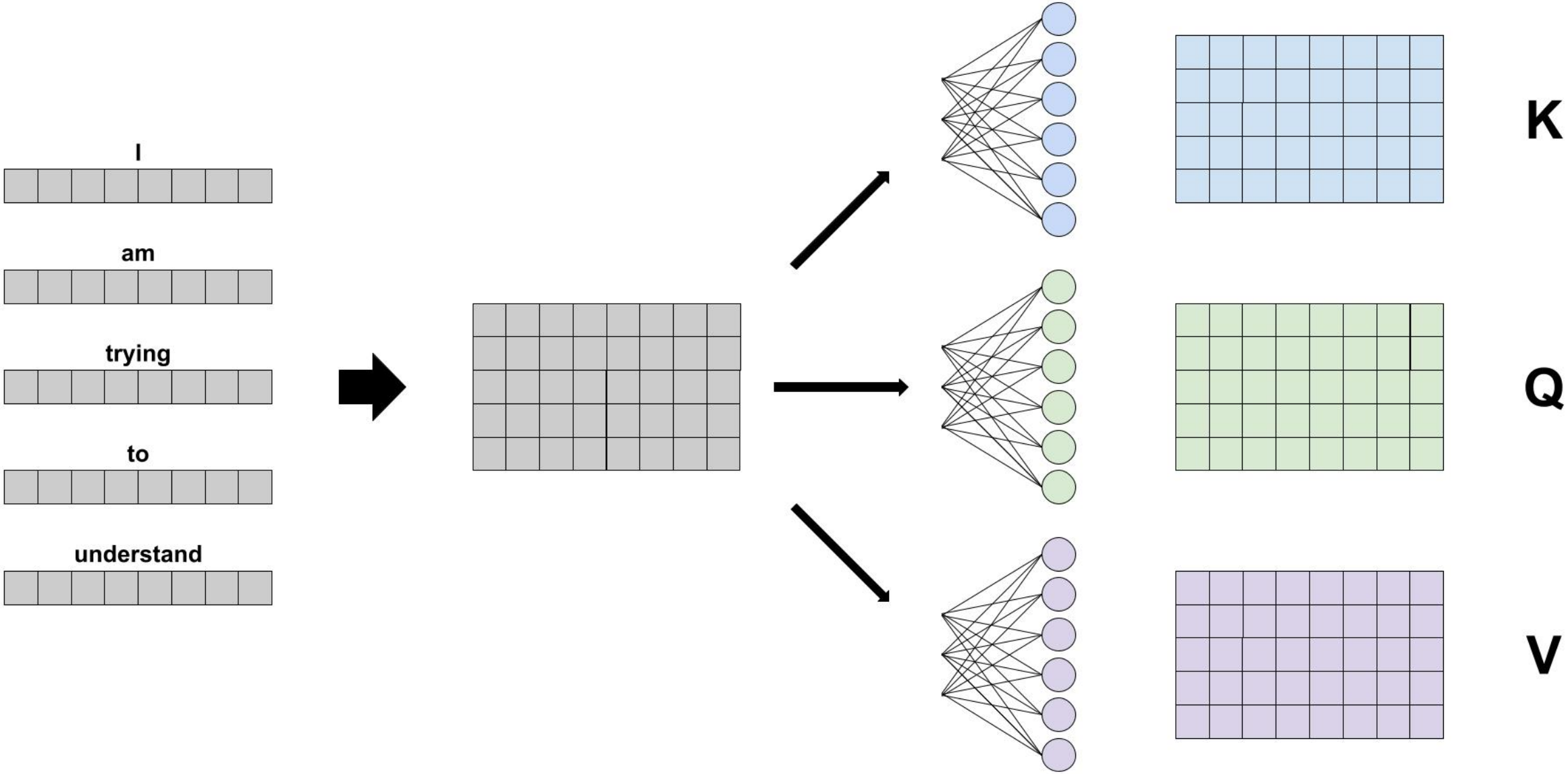
Intuition behind the Attention mechanism (1)



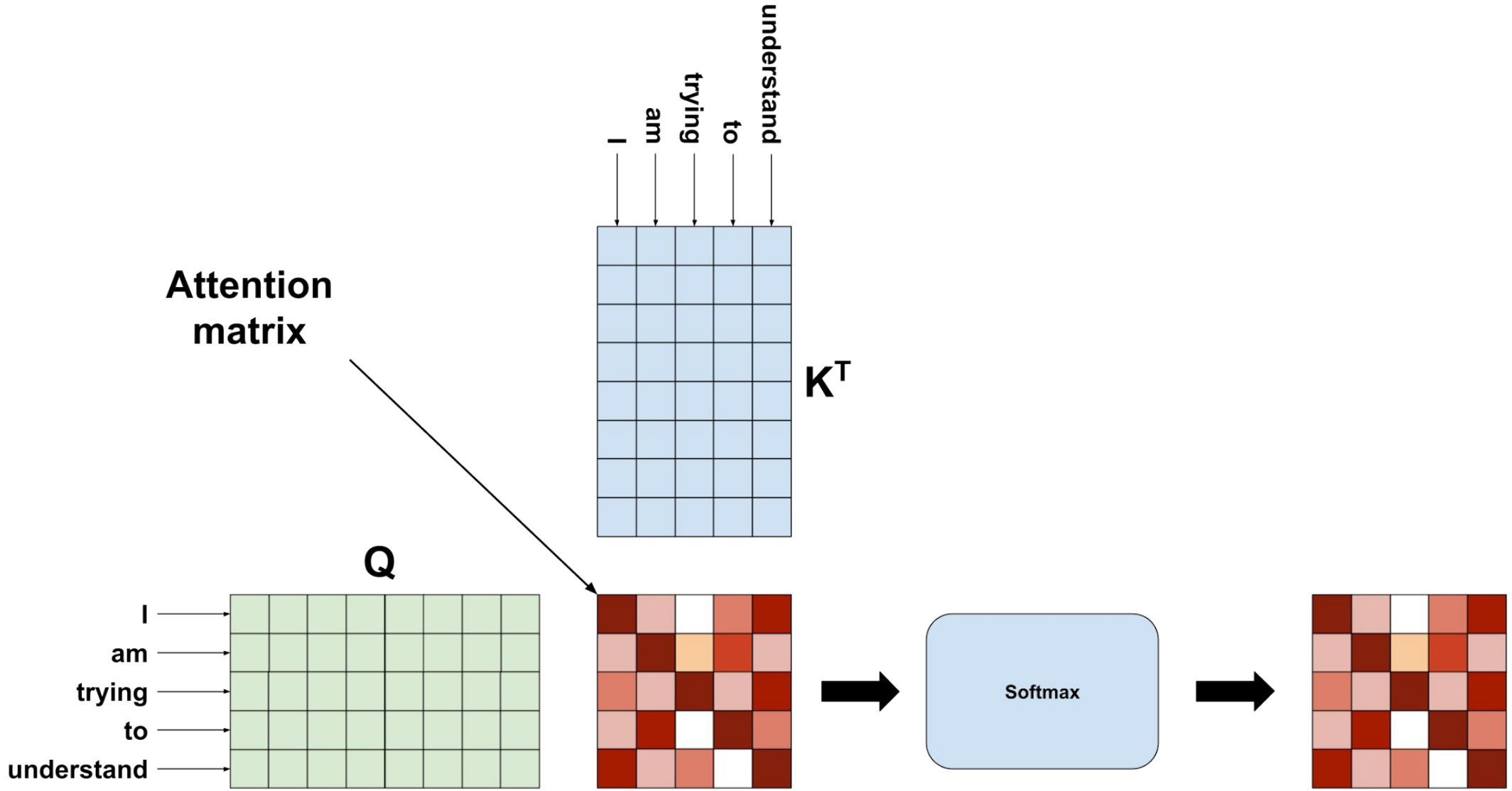
Intuition behind the Attention mechanism (2)



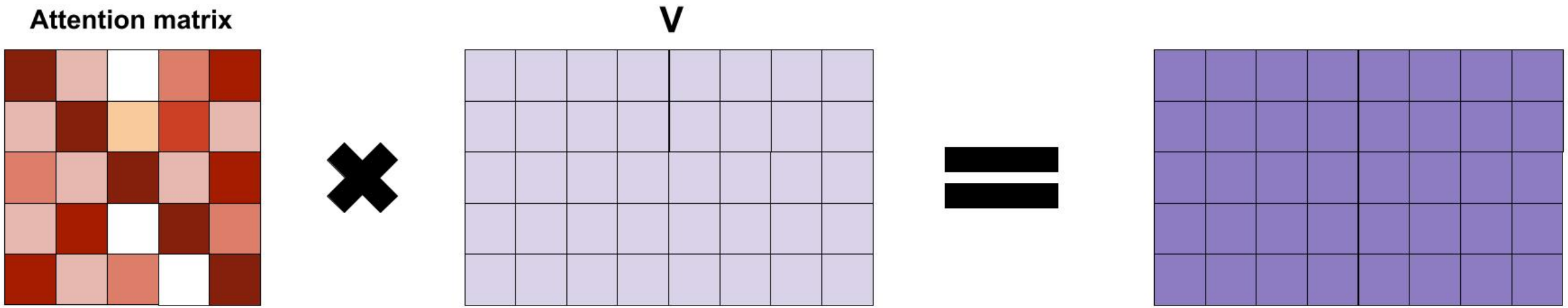
Intuition behind the Attention mechanism (3)



Attention mechanism (1)

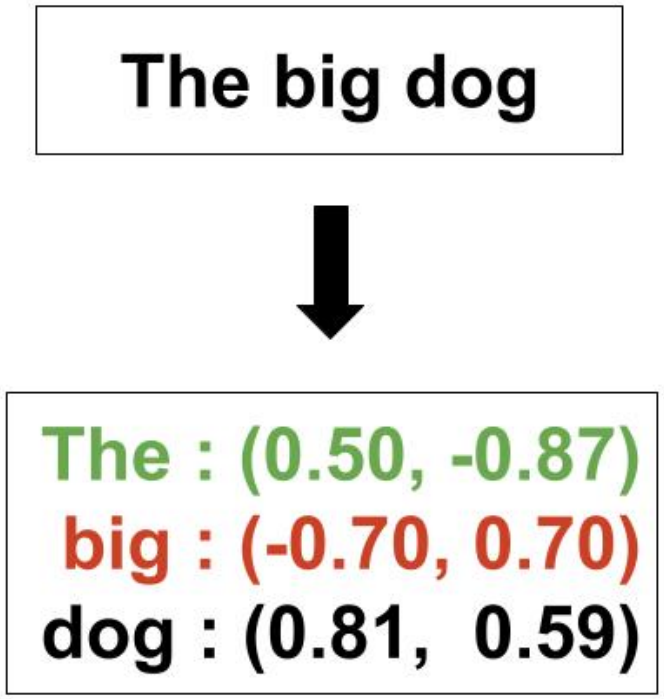


Attention mechanism (2)

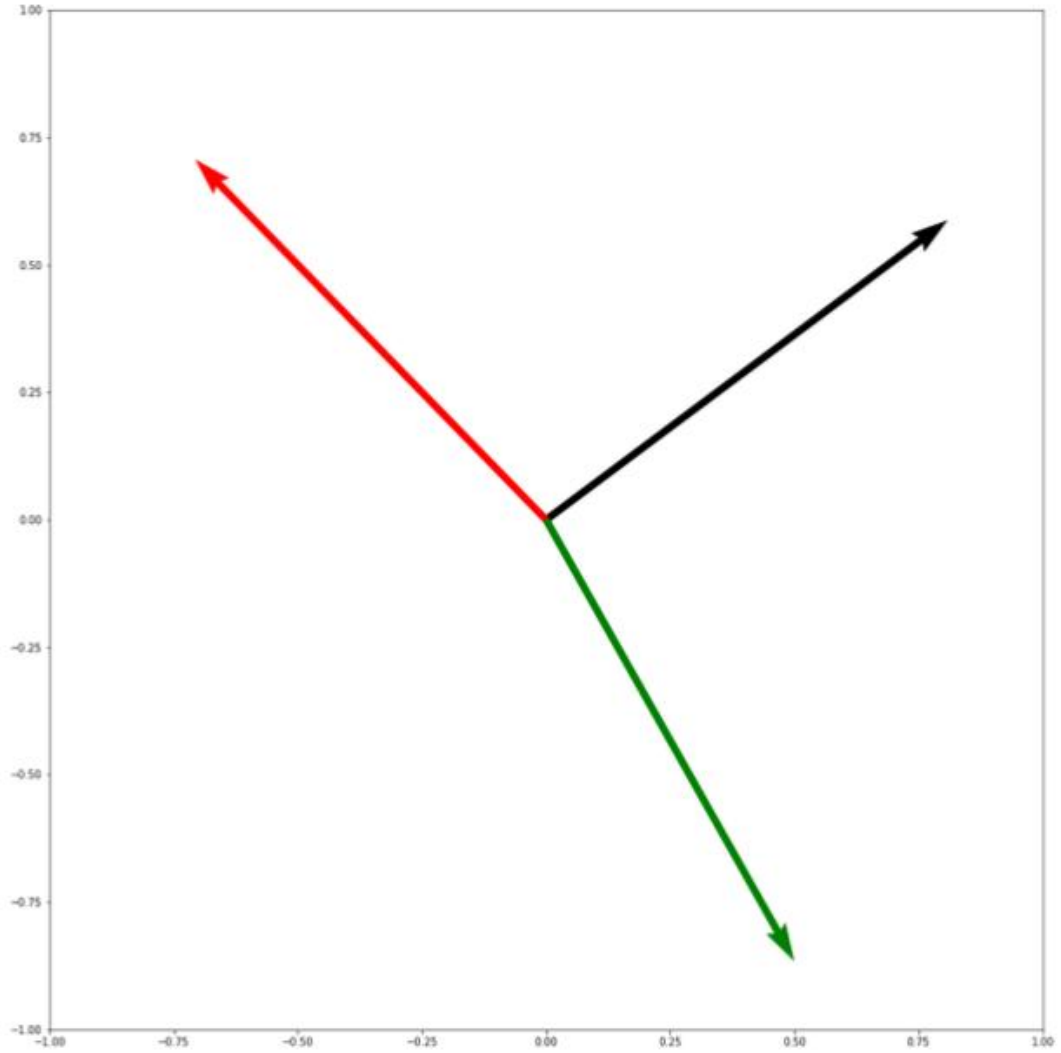


Attention mechanism (3)

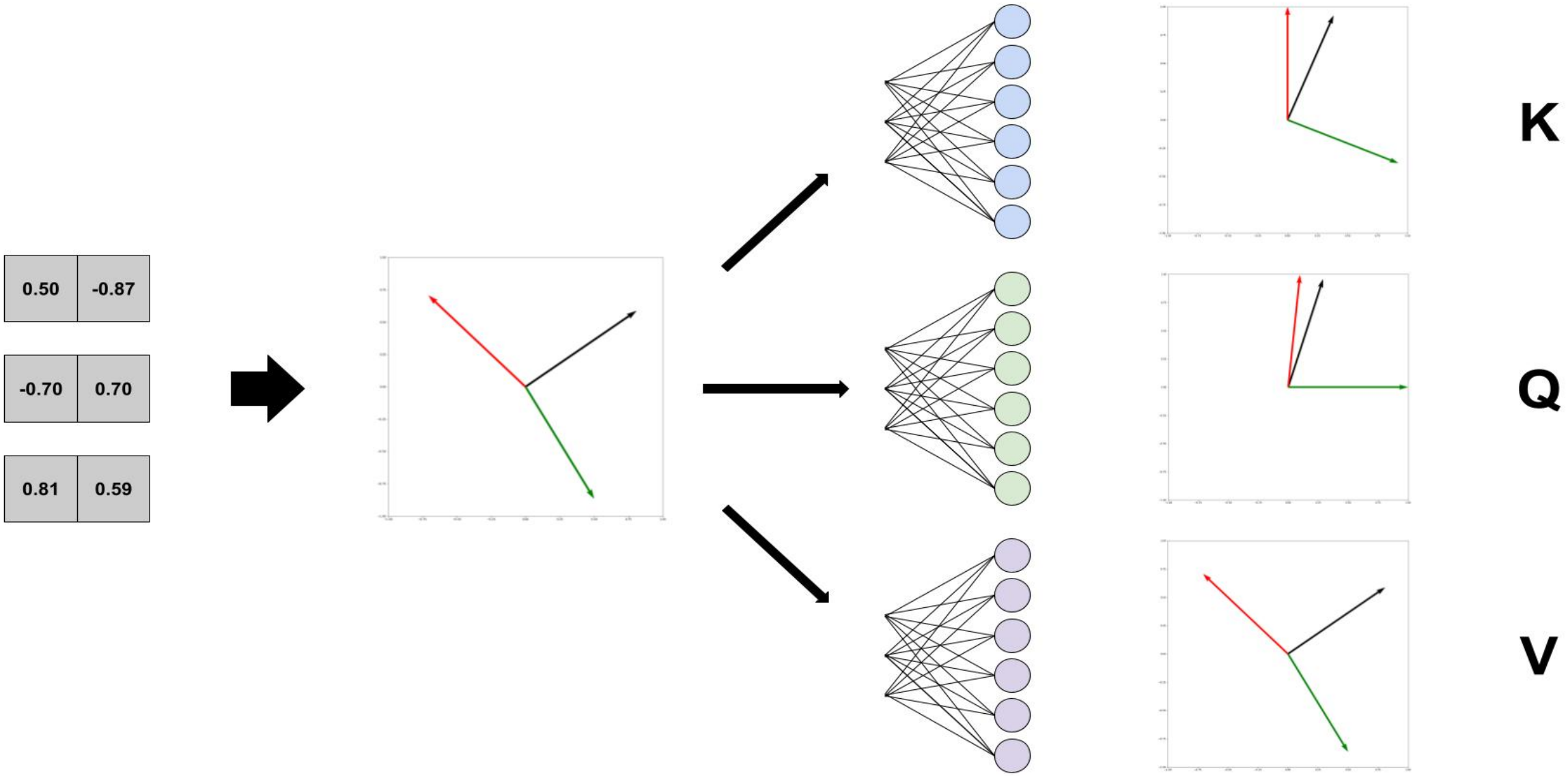
The big dog



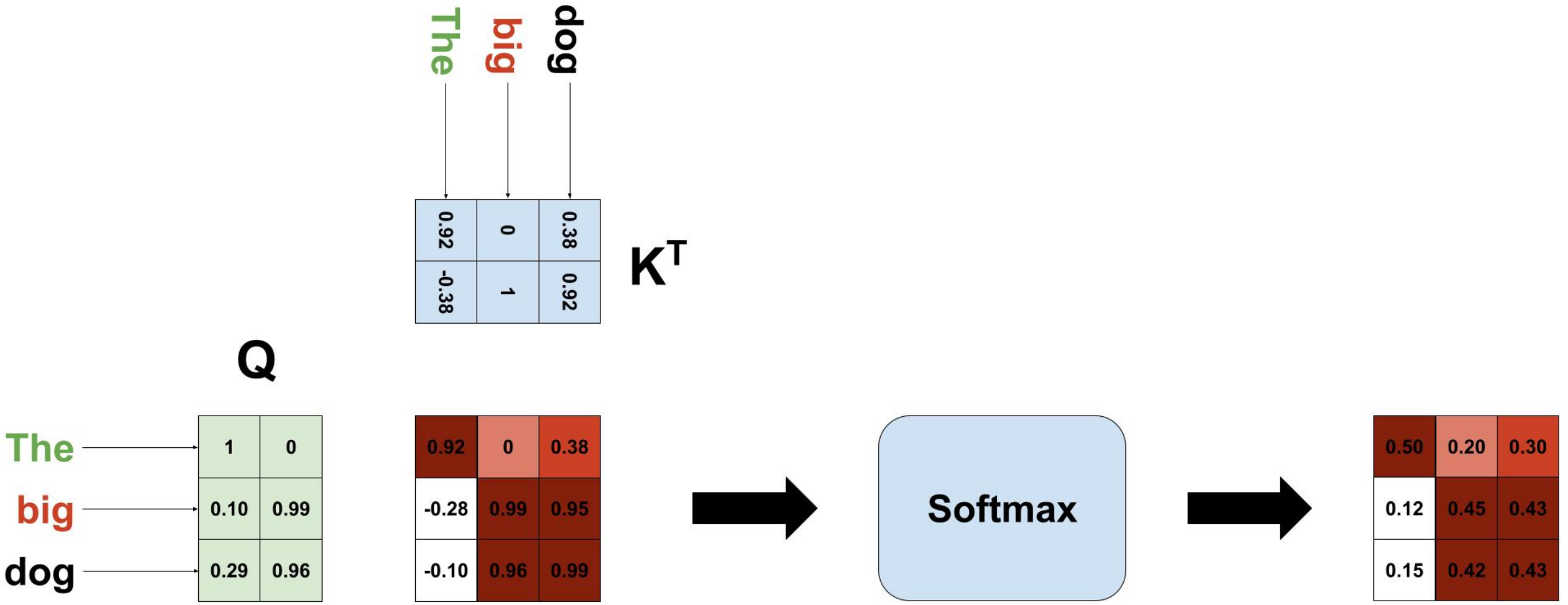
The : (0.50, -0.87)
big : (-0.70, 0.70)
dog : (0.81, 0.59)



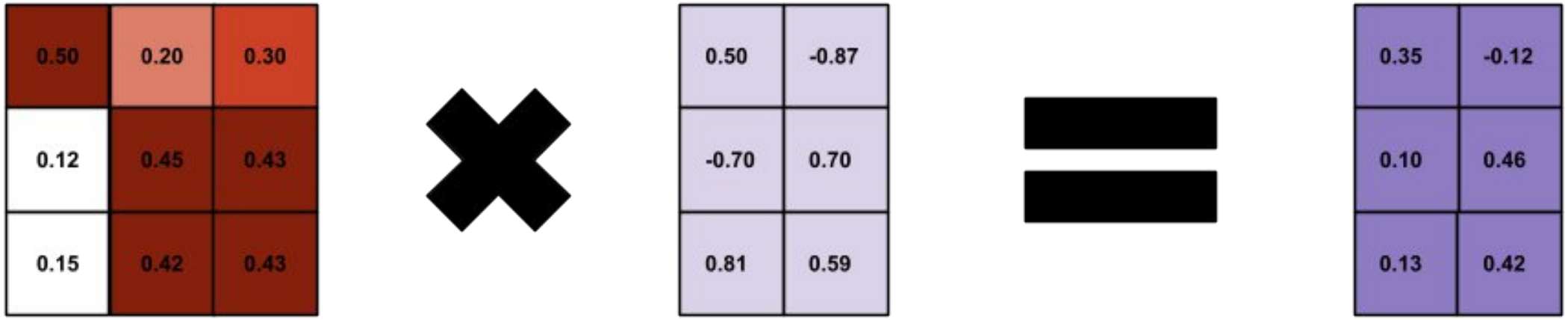
Attention mechanism - Example (1)



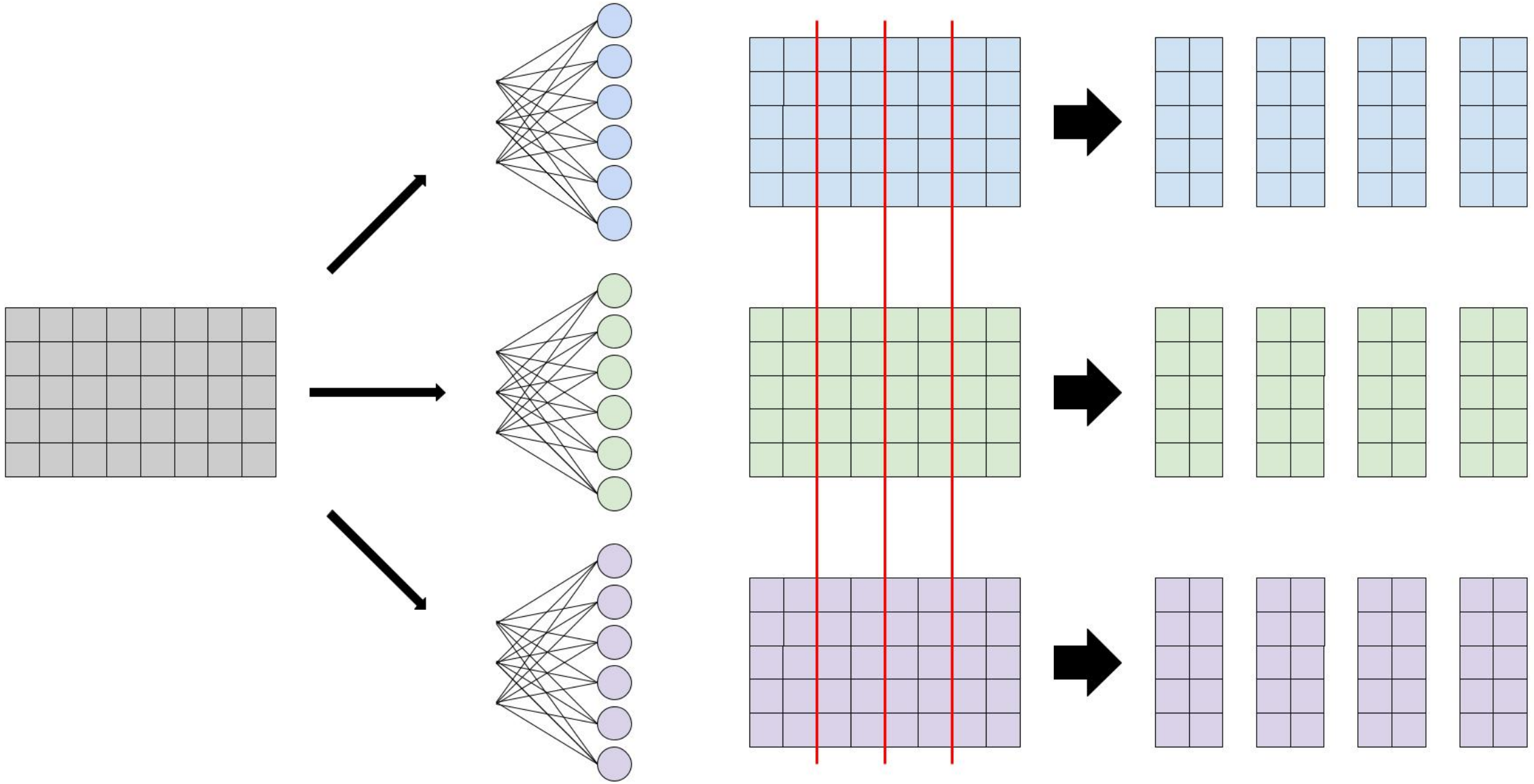
Attention mechanism - Example (2)



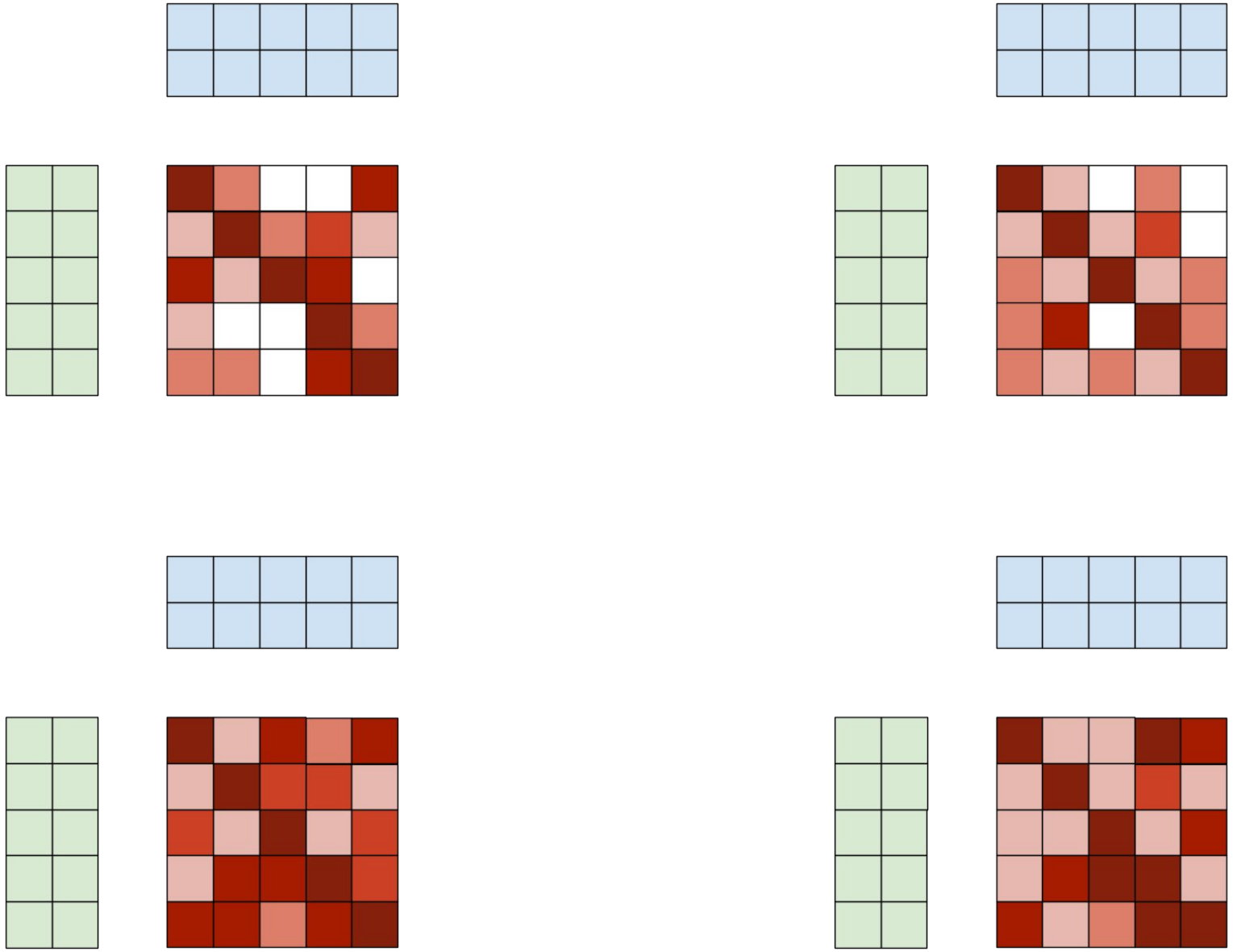
Attention mechanism - Example (3)



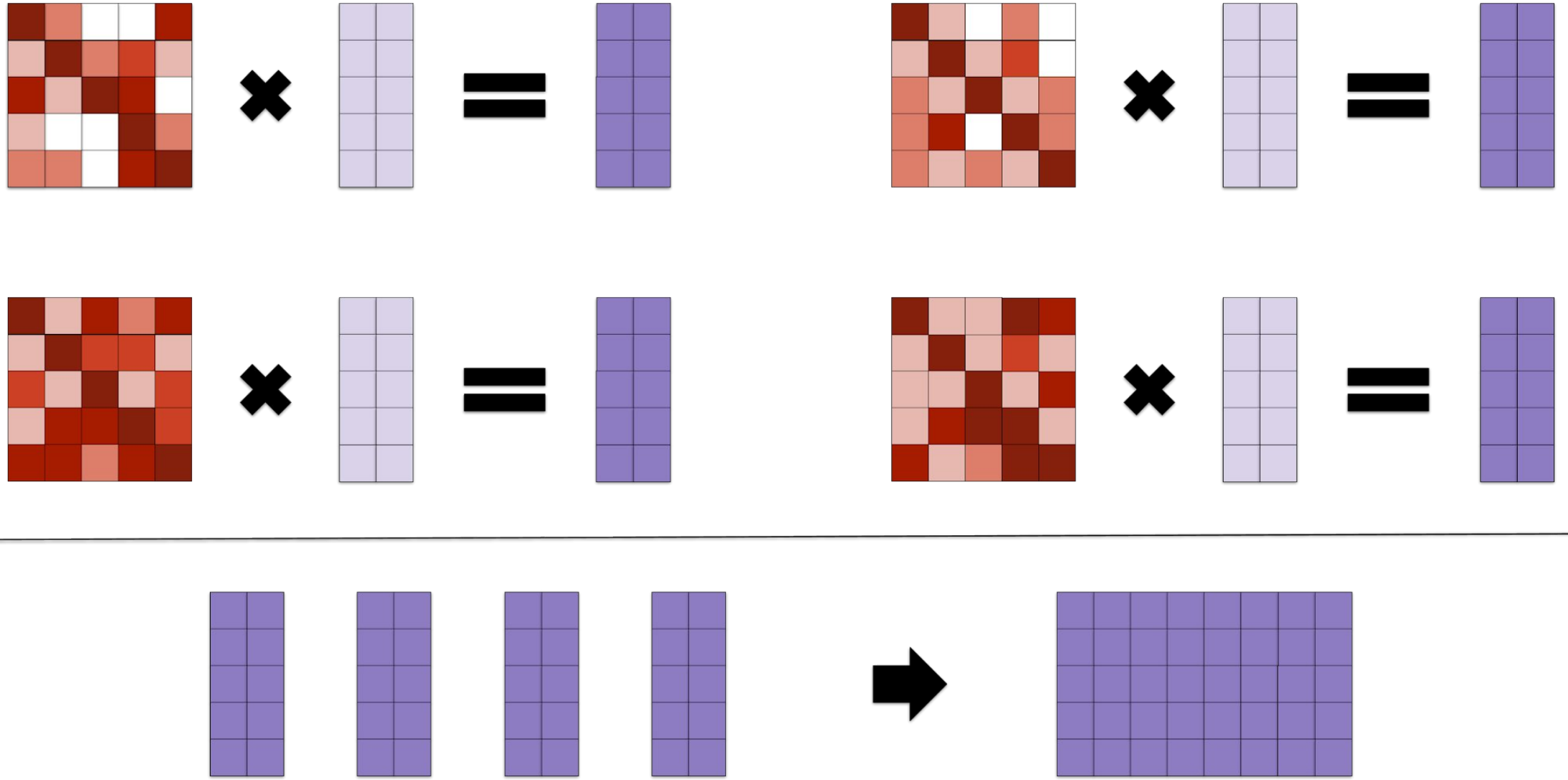
Attention mechanism - Example (4)



Multi-Head Attention (1)

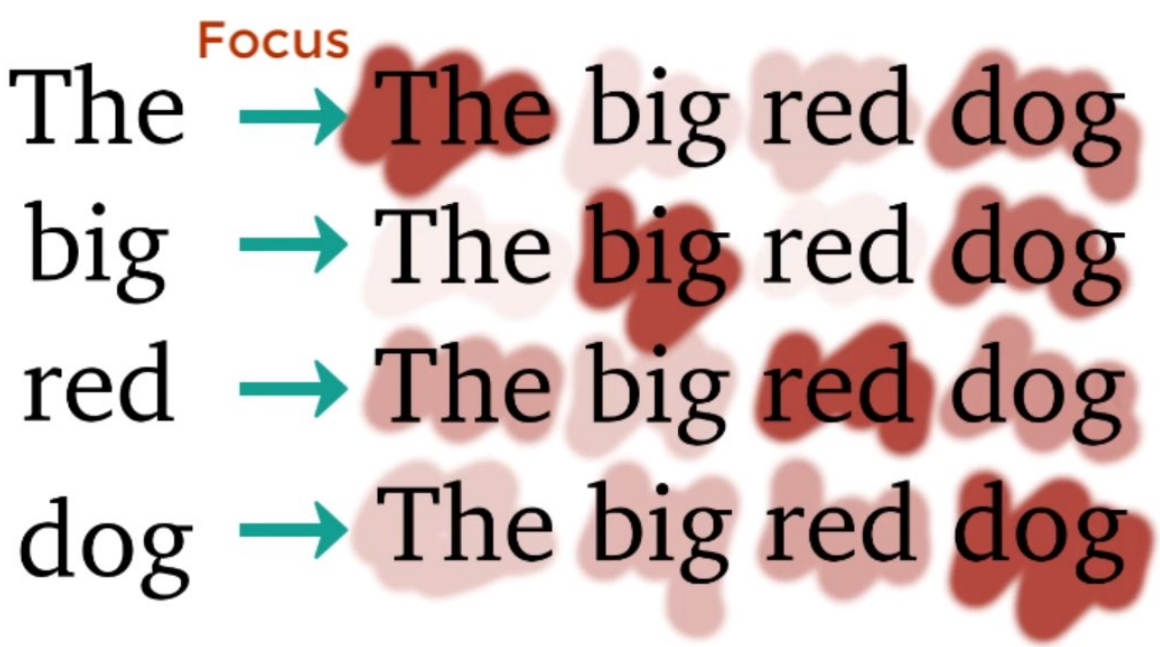


Multi-Head Attention (2)



Multi-Head Attention (3)

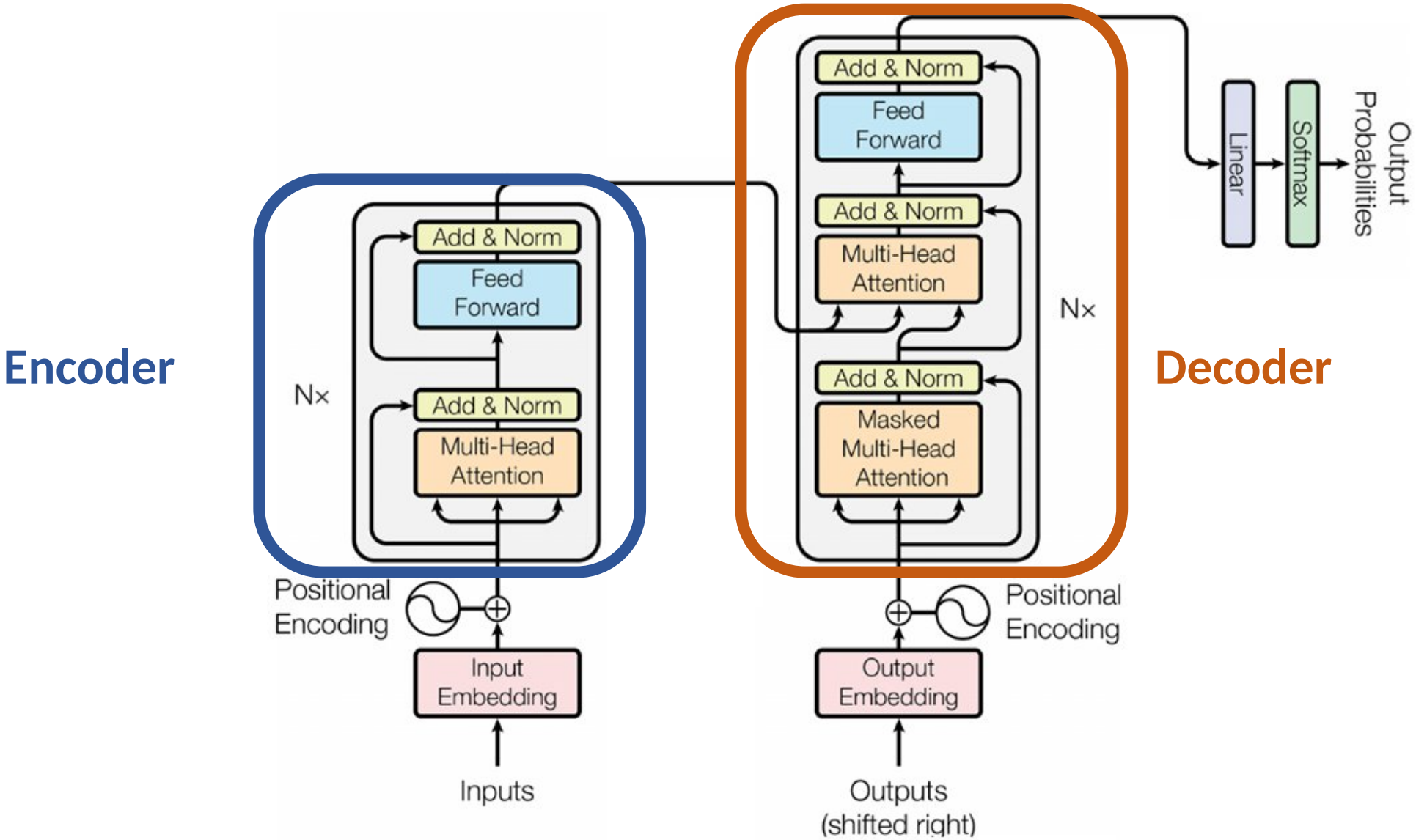
Bidirectional attention (BERT - Encoder - Auto-encoding)



Unidirectional attention (GPT - Decoder - Auto-regressive)

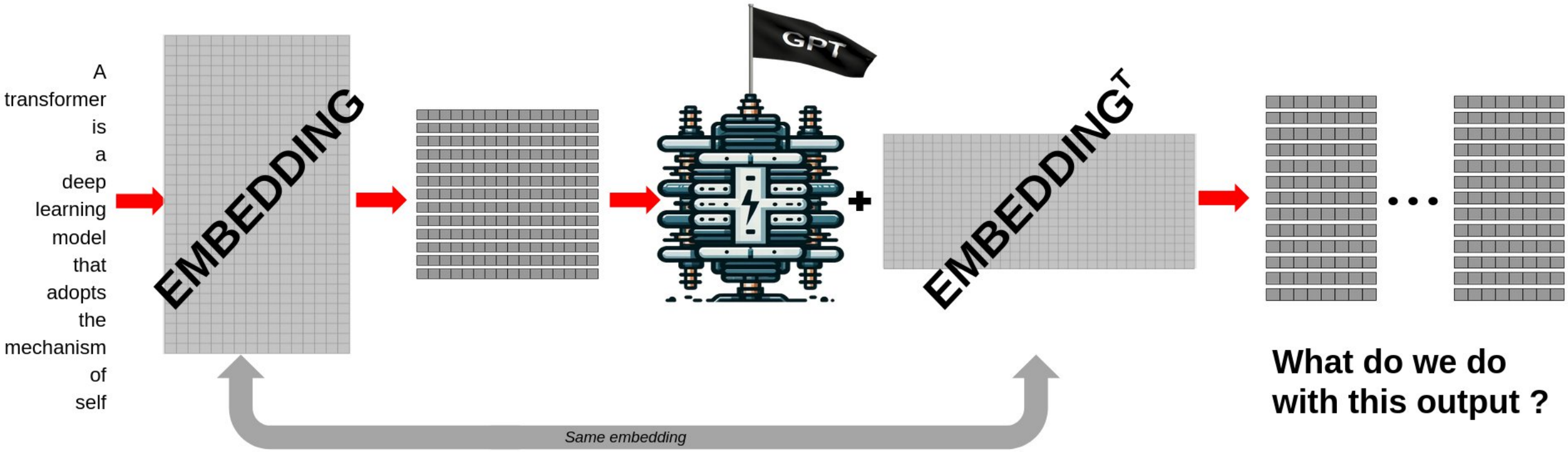
Transformer Neural Networks - EXPLAINED! (Attention is all you need) : https://www.youtube.com/watch?v=TQQlZhbC5ps

Transformer types (1)

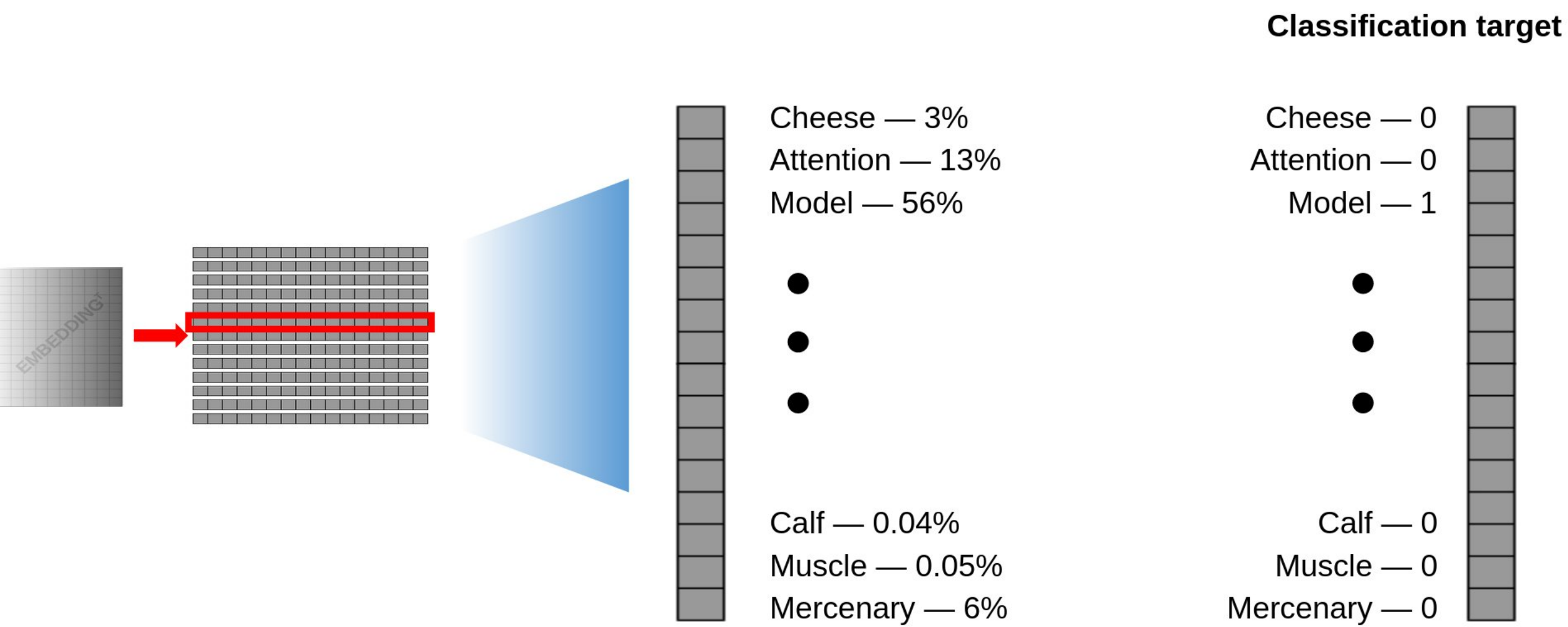


Vaswani, Ashish, et al. "Attention is all you need." Advances in neural information processing systems 30 (2017).

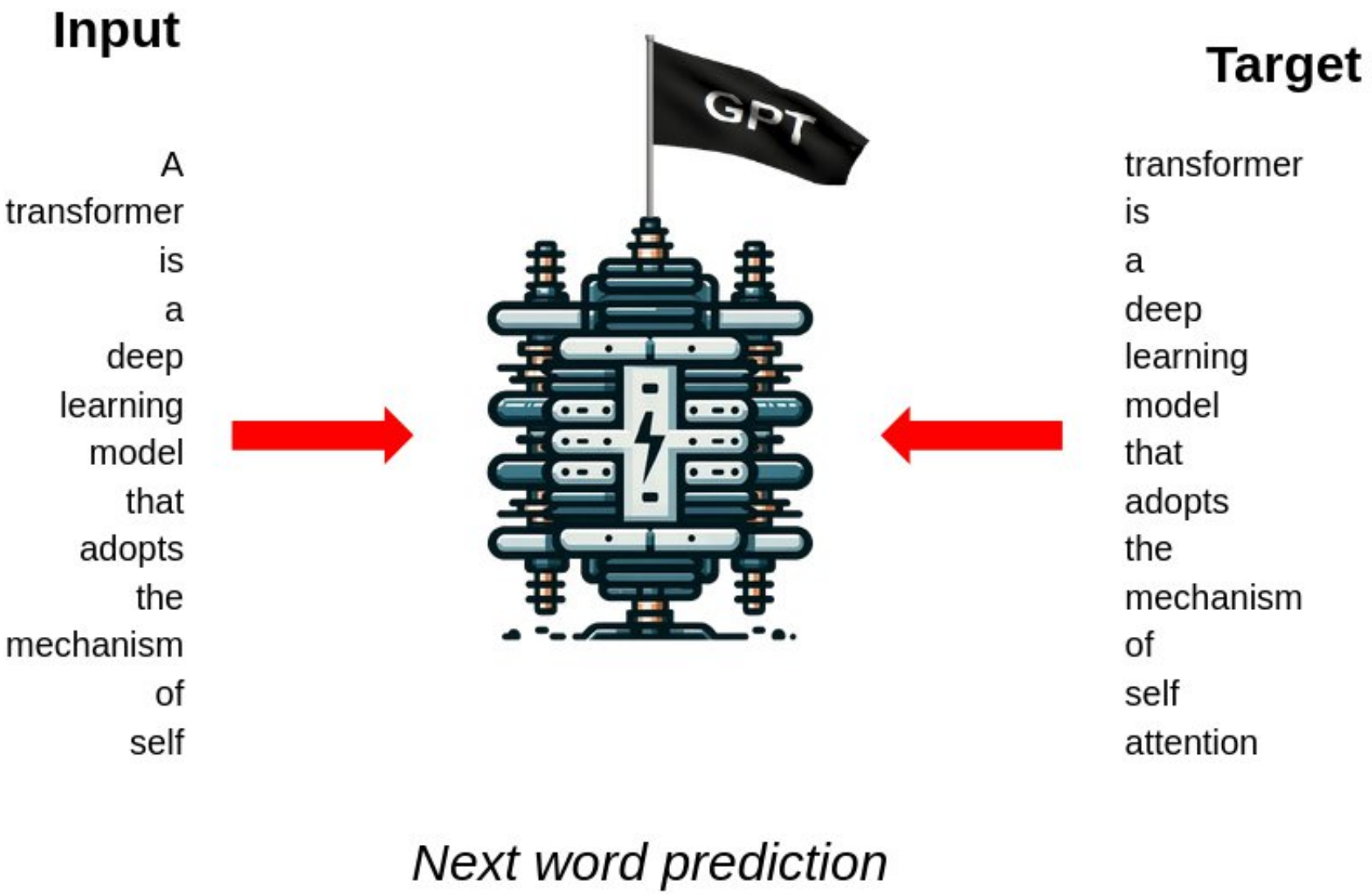
Transformer types (2)



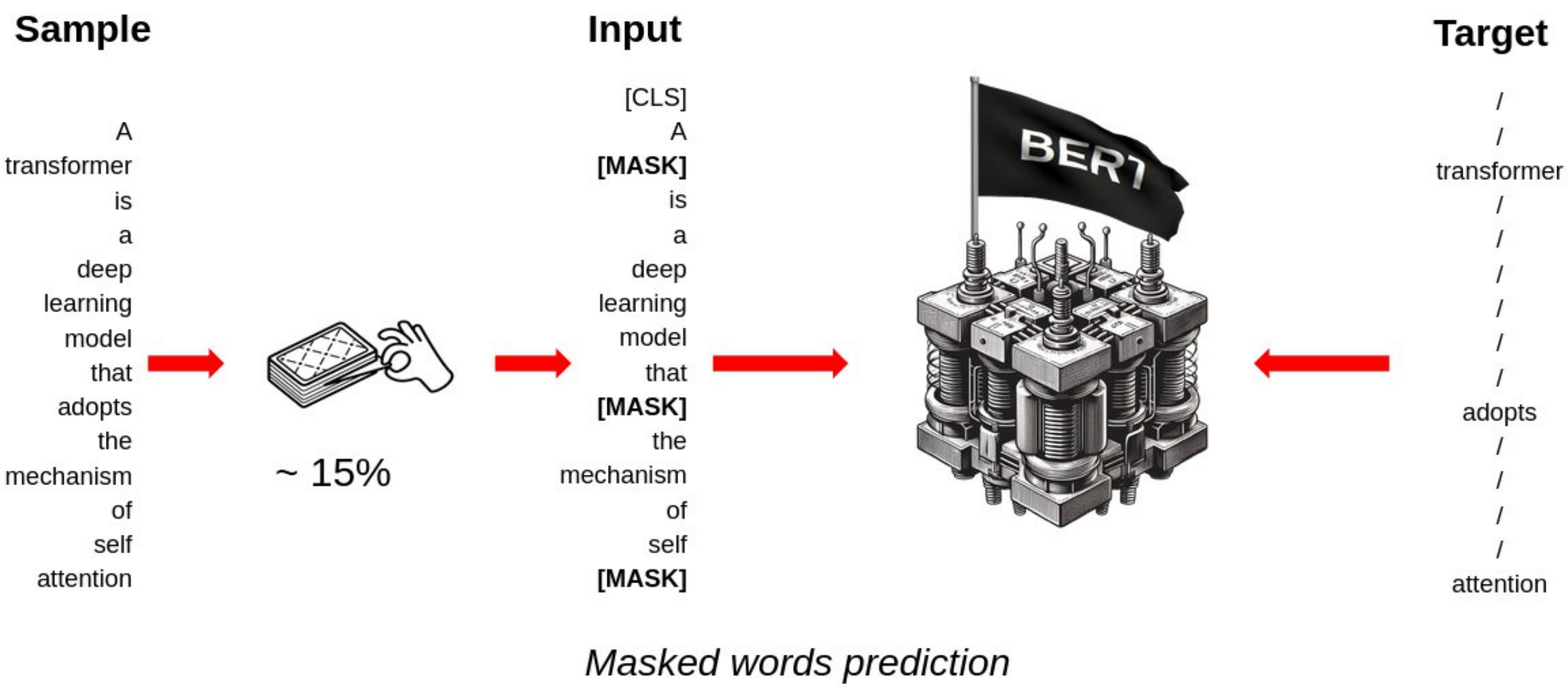
Training a transformer



Training a transformer



Training a GPT-style transformer



Training a BERT-style transformer

Which word do we choose ?

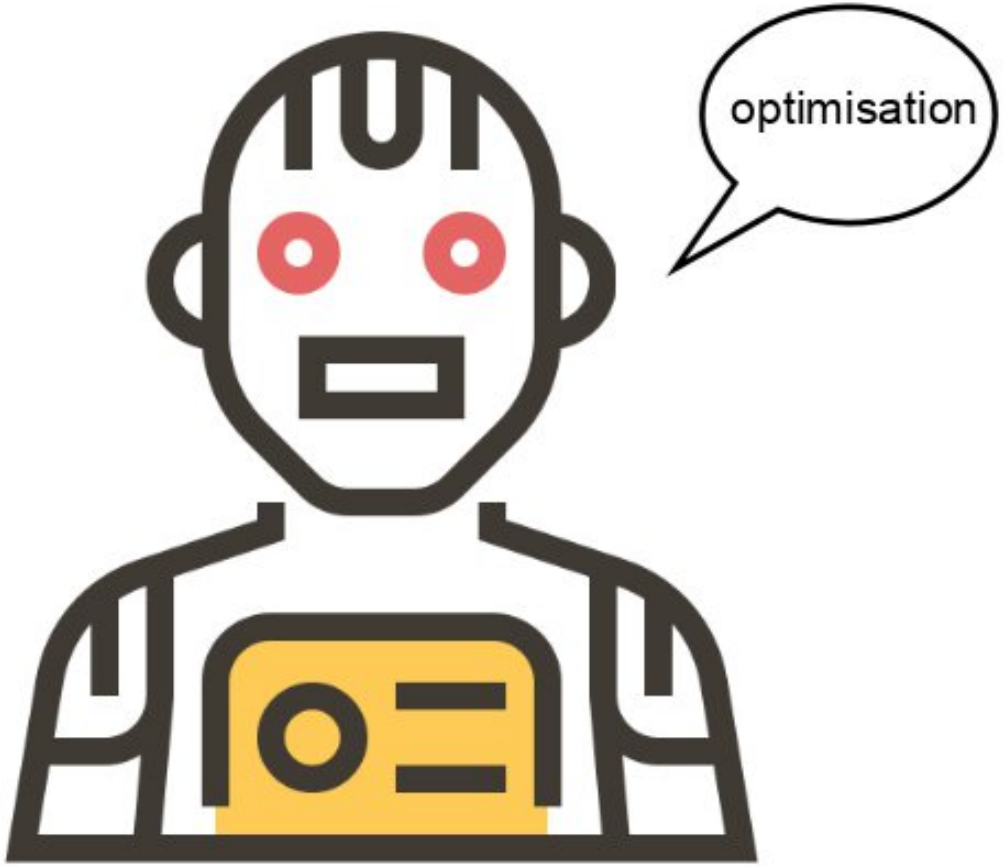
- Optimisation — 35%
- Intelligence — 22%
- Apprentissage — 18%
- Machine — 10%
- Isolement — 1.2%
- Fromage — 0.04%

• • •

Prediction

- Optimisation — 35%
- Intelligence — 22%
- Apprentissage — 18%
- Machine — 10%
- Isolement — 1.2%
- Fromage — 0.04%

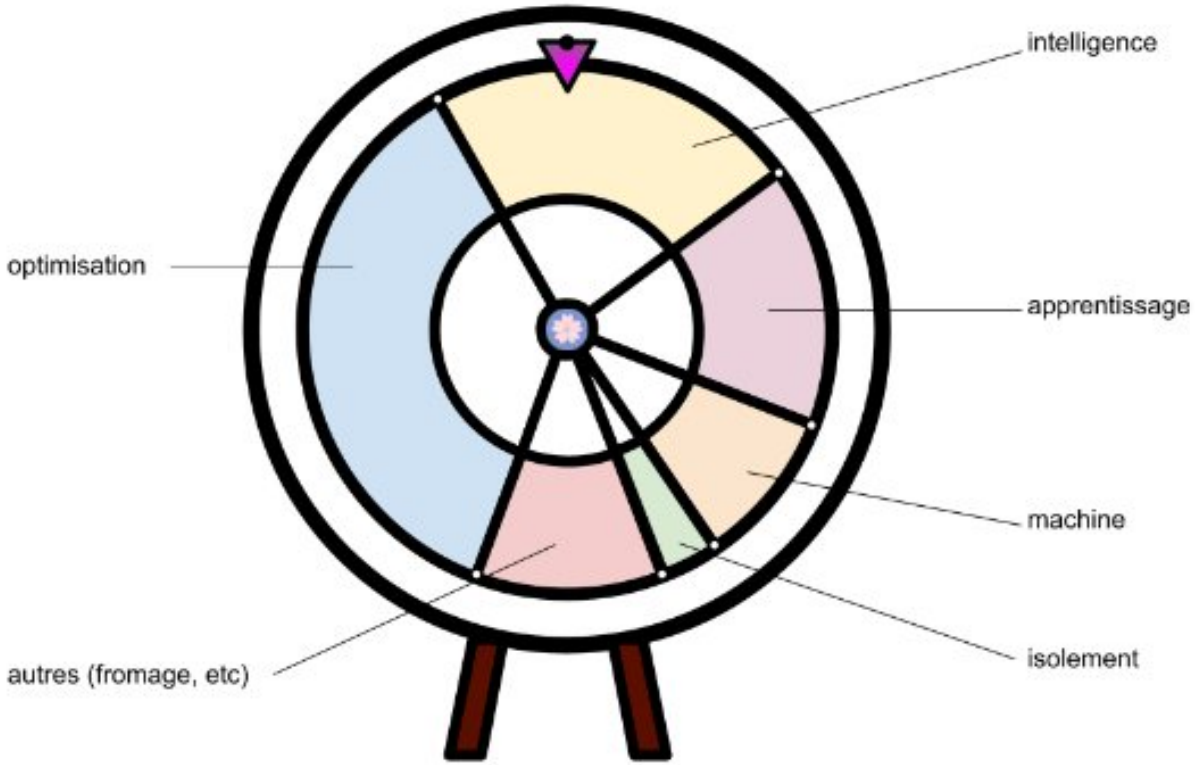
• • •

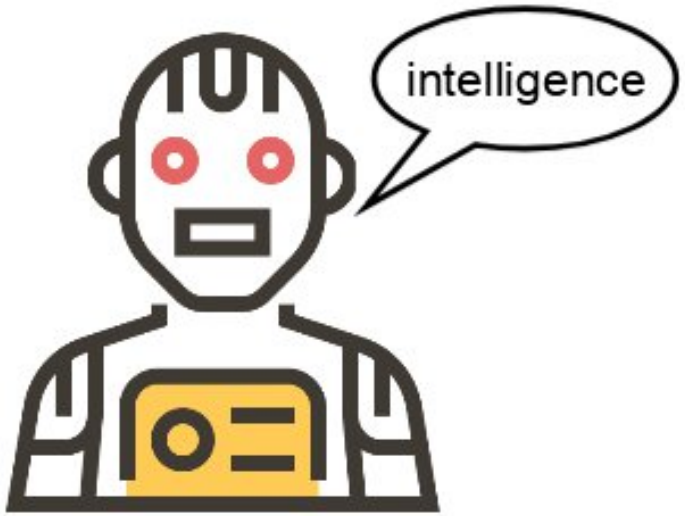


Prediction mode : Greedy

- Optimisation — 35%
- Intelligence — 22%
- Apprentissage — 18%
- Machine — 10%
- Isolement — 1.2%
- Fromage — 0.04%

• • •

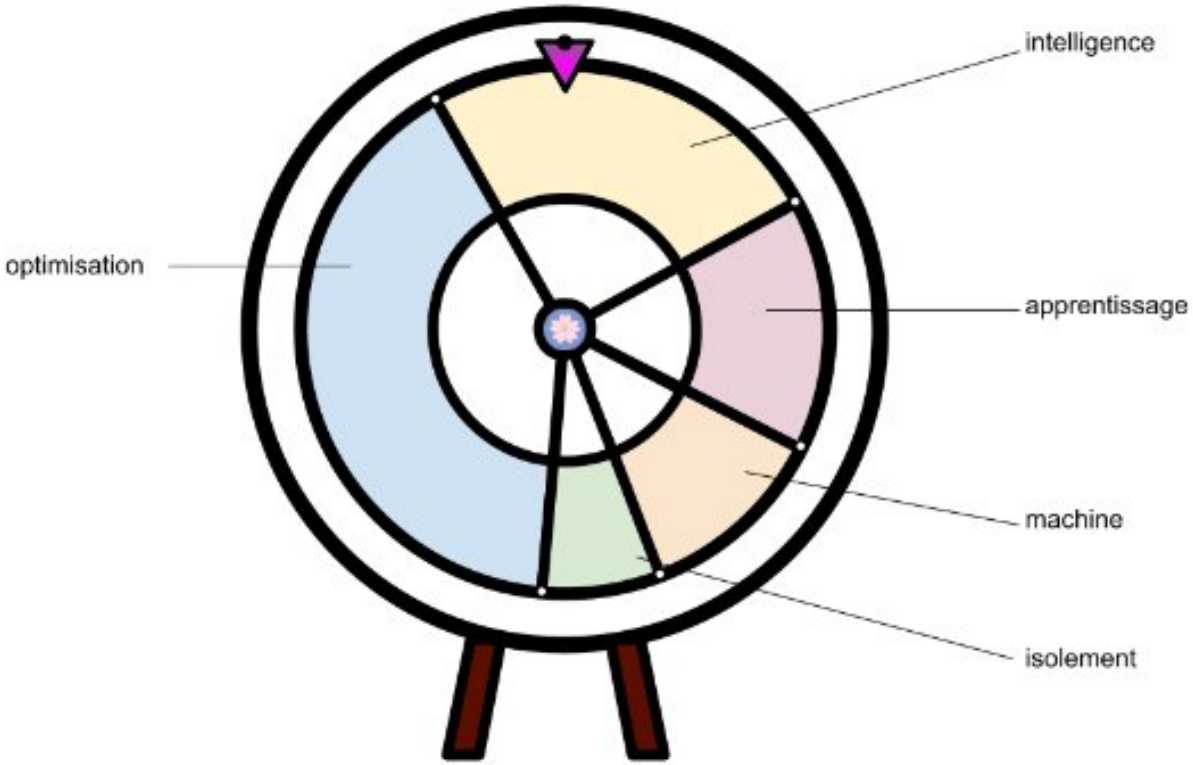


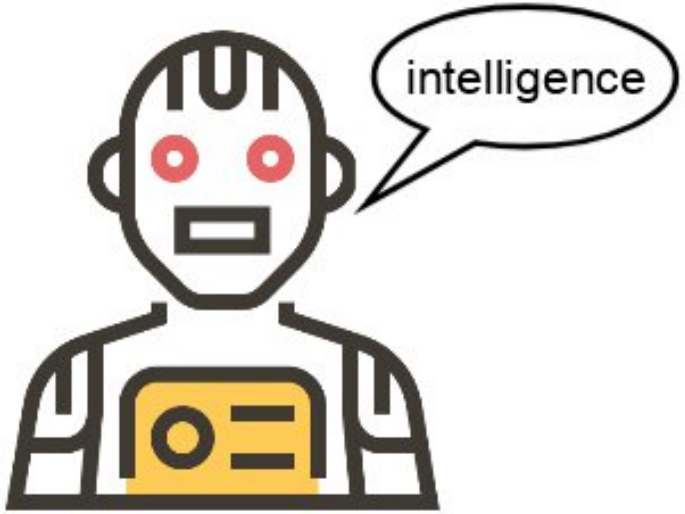


Prediction mode : Sampling

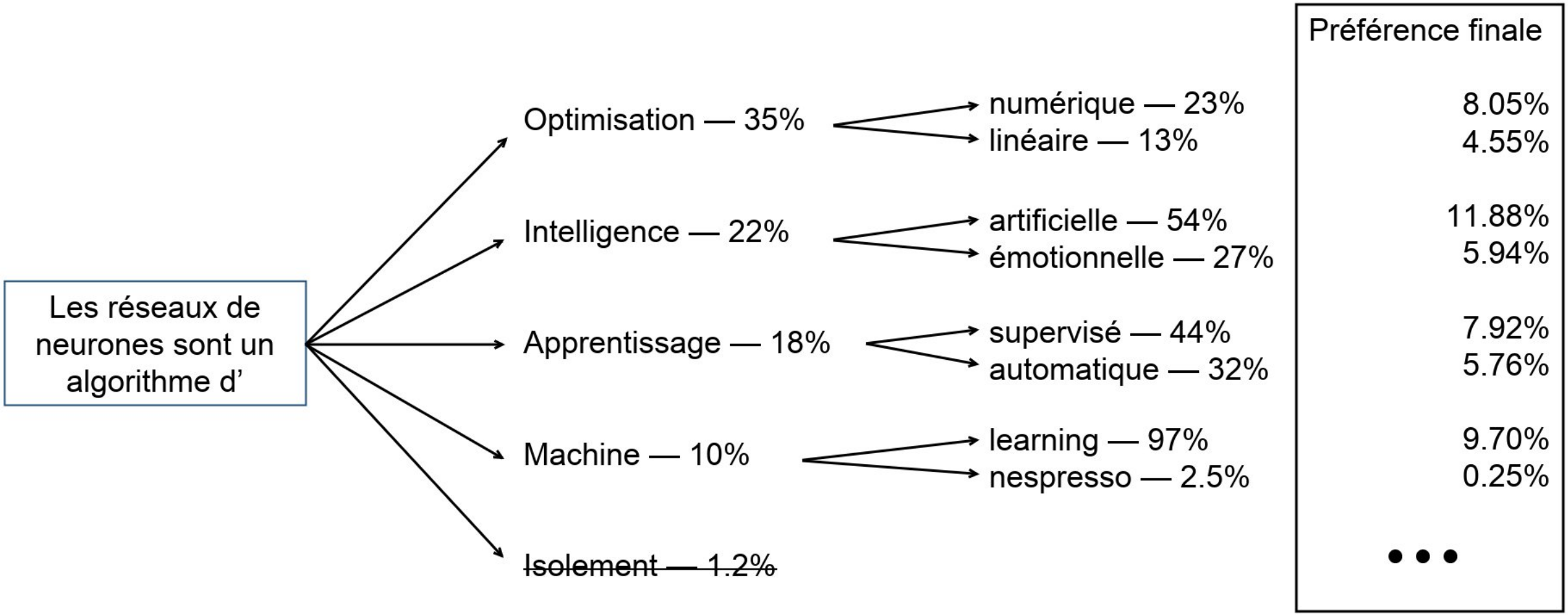
- Optimisation — 35%
- Intelligence — 22%
- Apprentissage — 18%
- Machine — 10%
- Isolement — 1.2%
- ~~Fromage — 0.04%~~

• • •

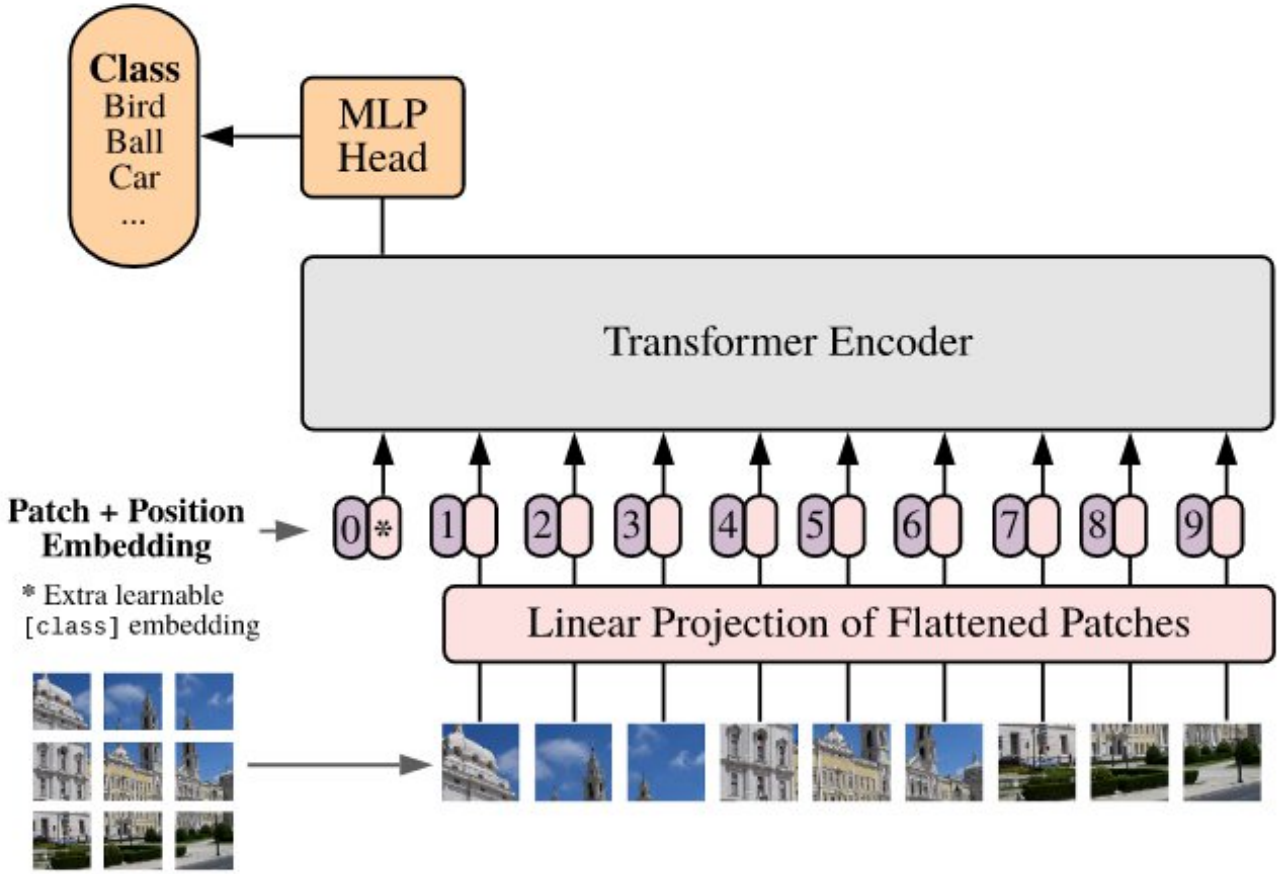


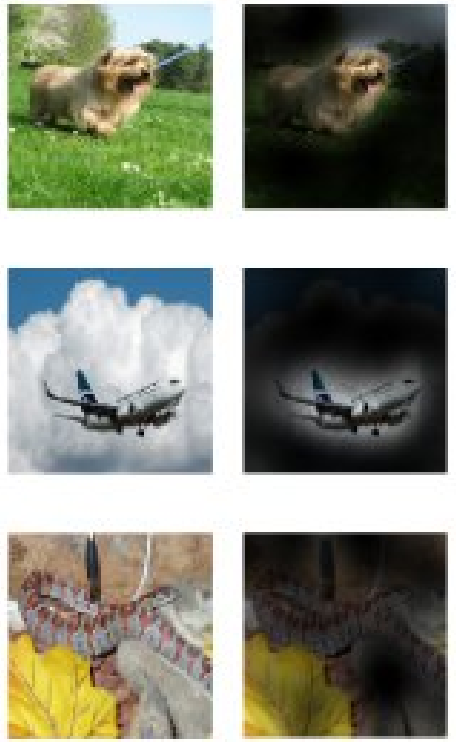


Prediction mode : Sampling (top-k variant)



Prediction mode : Beam Search





Dosovitskiy, Alexey, et al. "An image is worth 16x16 words: Transformers for image recognition at scale." arXiv preprint arXiv:2010.11929 (2020).

Vision Transformers