

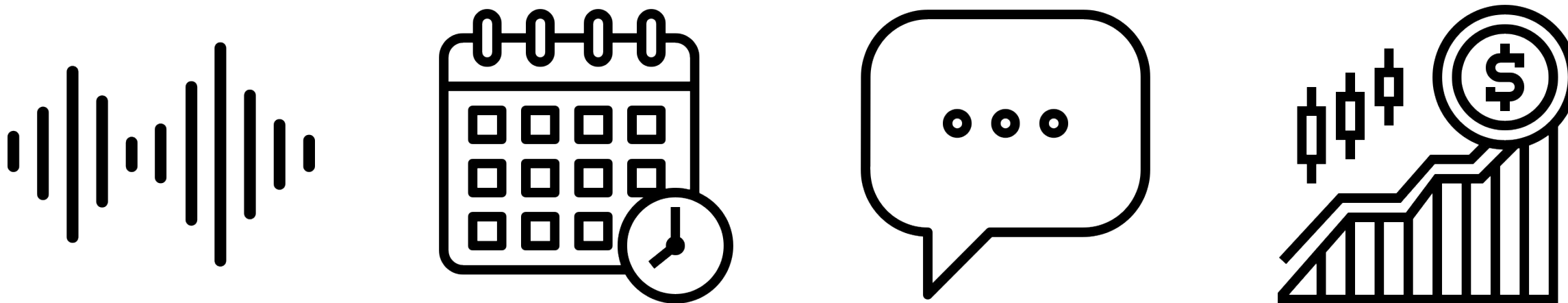


Hands-on Introduction to Deep Learning

Sequences



INSTITUT DU
DÉVELOPPEMENT ET DES
RESSOURCES EN
INFORMATIQUE
SCIENTIFIQUE

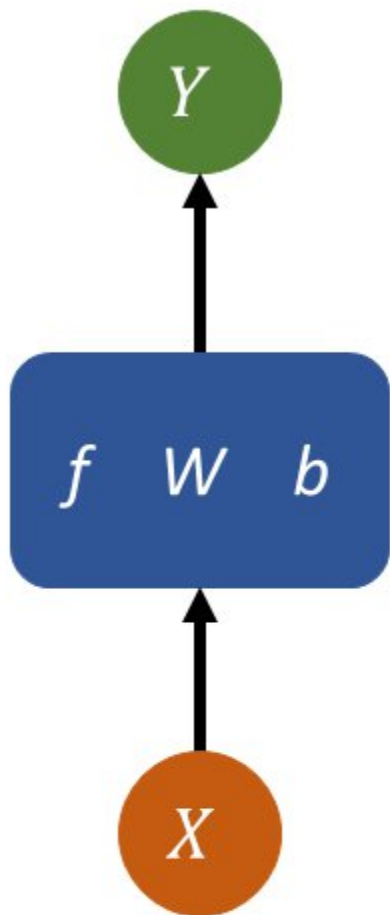


Stock market

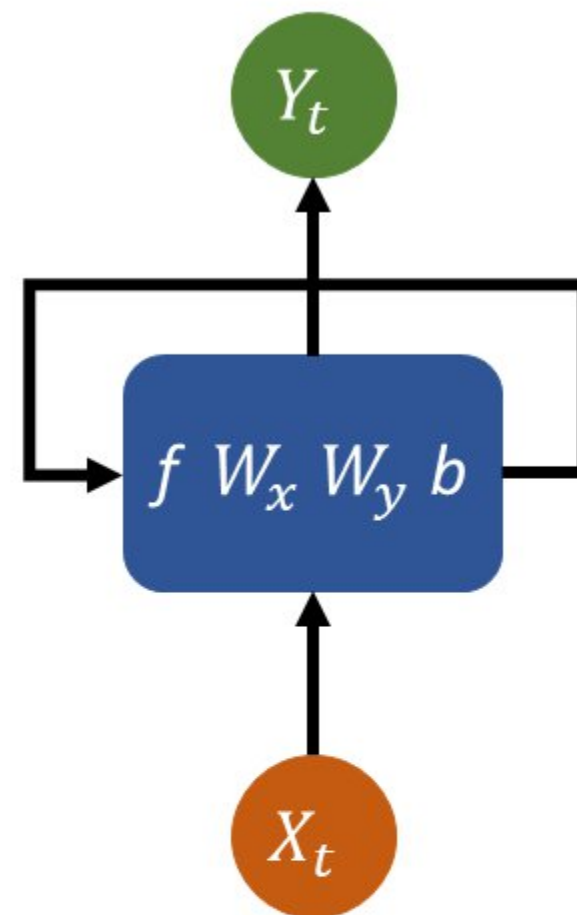
	day 1	day 2	day 3
asset 1	9.77	79.94	64.13
asset 2	47.66	74.07	70.90
asset 3	94.25	76.34	99.95
asset 4	41.19	9.99	89.50
asset 5	65.44	63.79	67.14

Text

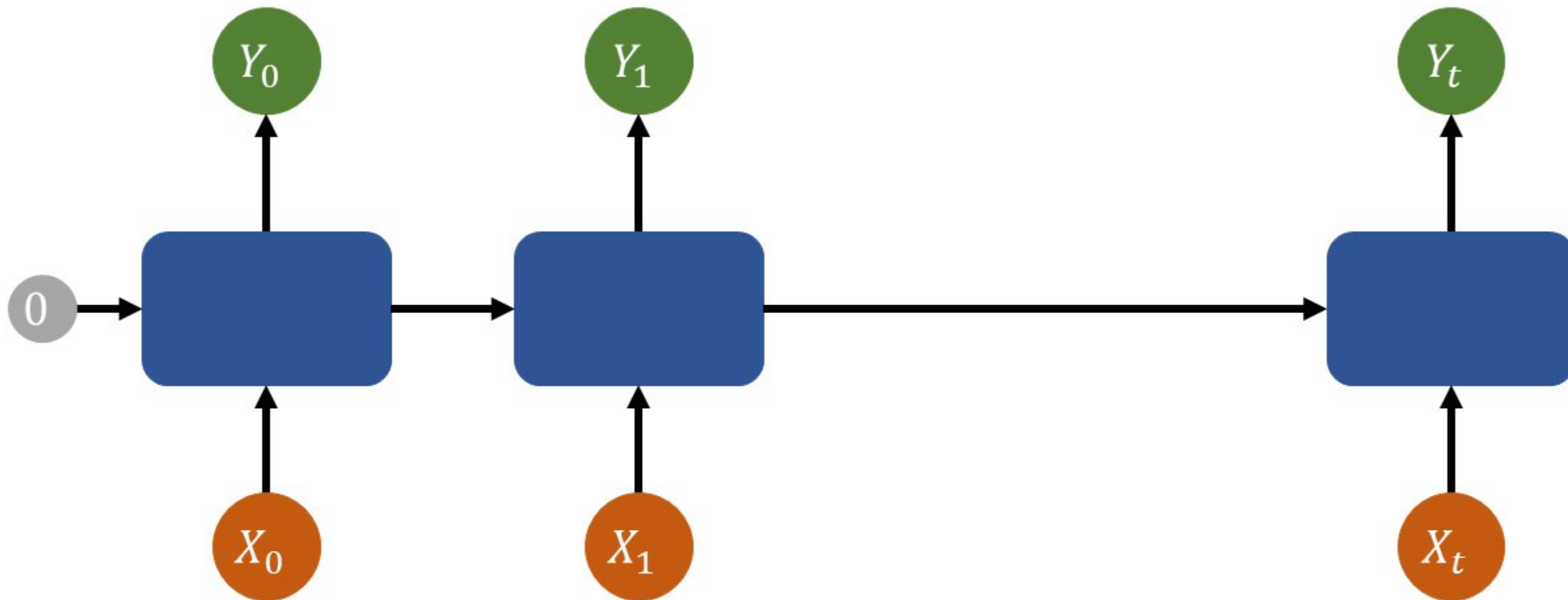
I	am	learning	.
0,83	0,65	-0,90	-0,04
-0,53	0,81	-0,61	-0,12
0,24	-0,14	0,58	0,66
-0,31	0,32	0,37	-0,11
-0,53	0,50	-0,96	0,48
-0,34	-0,85	0,19	-0,78
-0,79	0,53	-0,31	-0,28
-0,23	-0,13	0,33	0,45
0,95	0,53	0,74	-0,24
-0,60	0,04	-0,96	-0,96

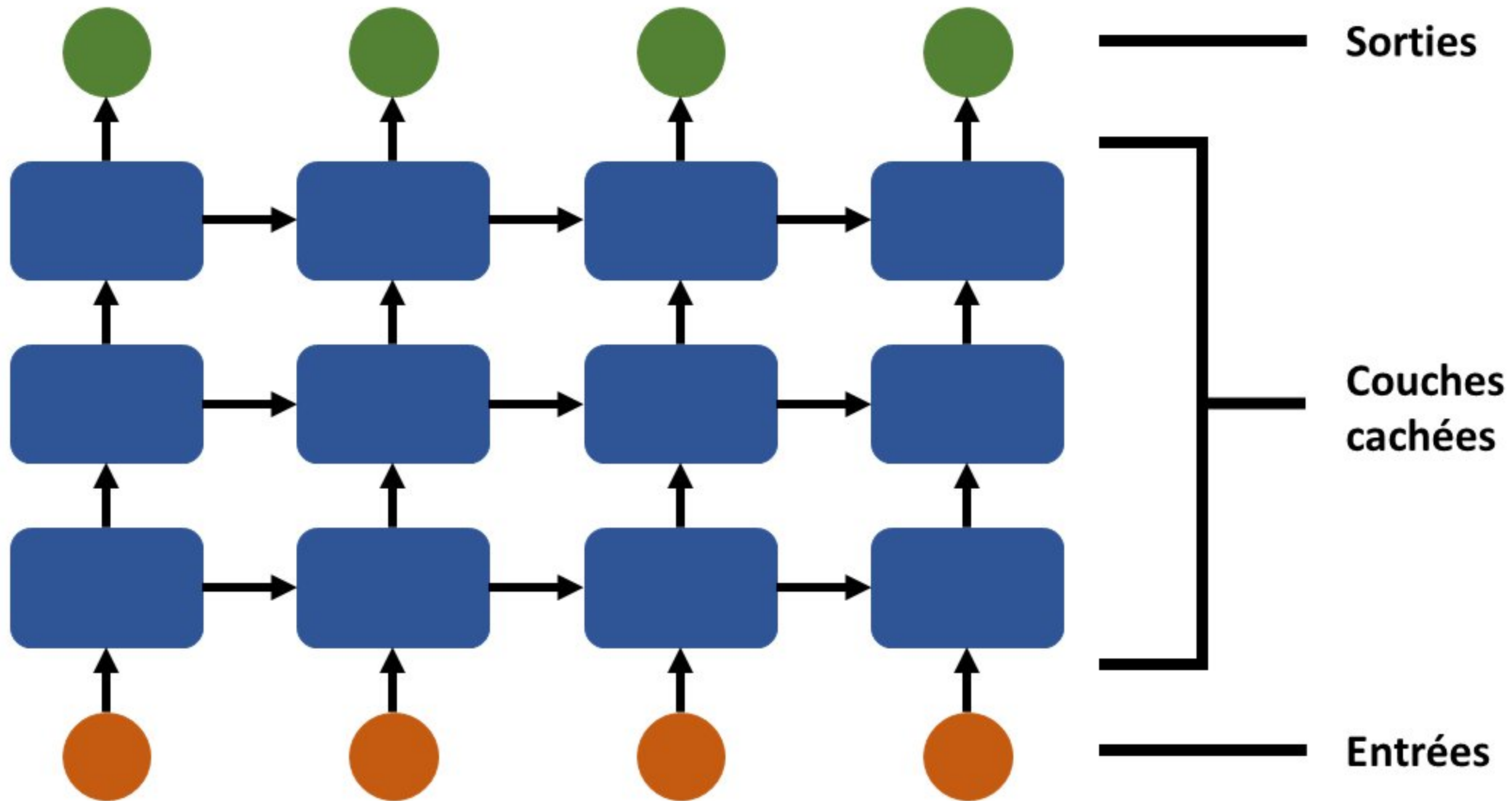


$$Y = f(W \cdot X + b)$$

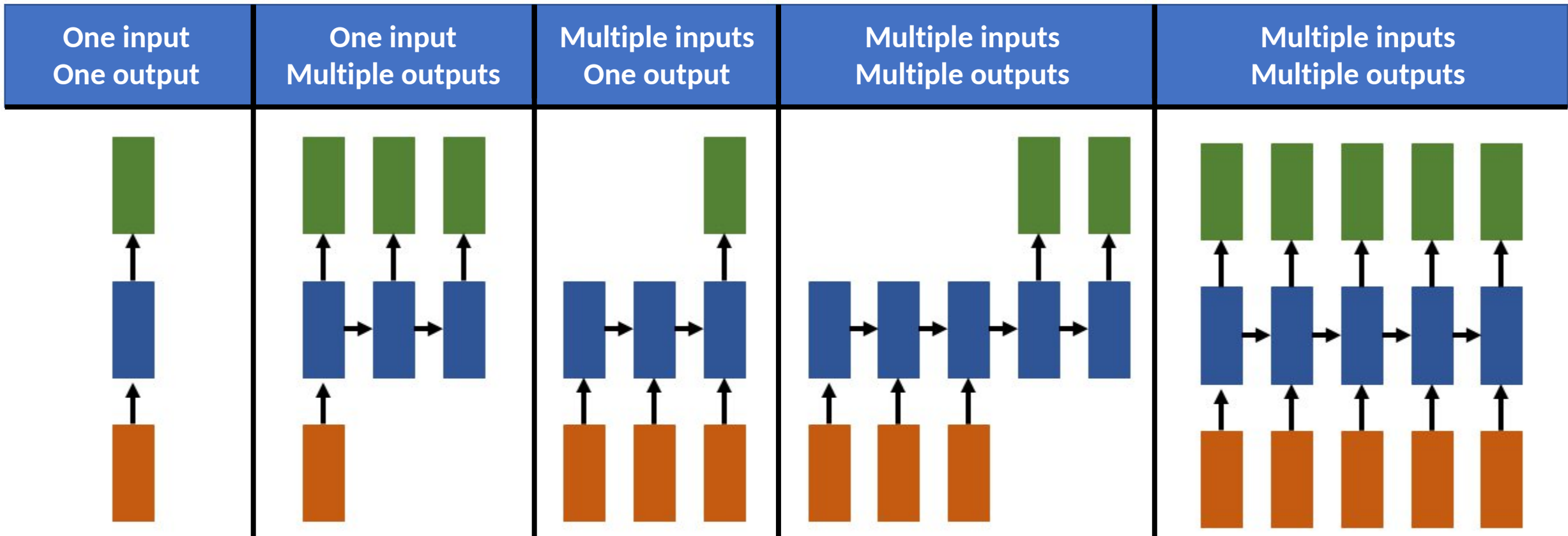


$$Y_t = f(W_x \cdot X_t + W_y Y_{t-1} + b)$$



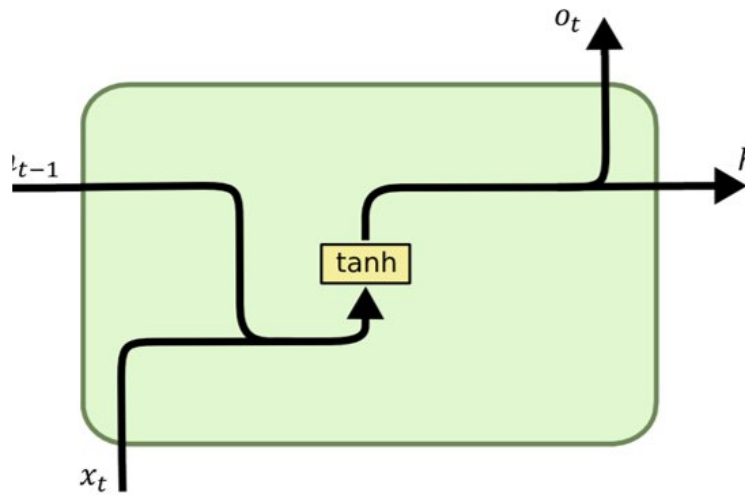


Recurrent Neural Network

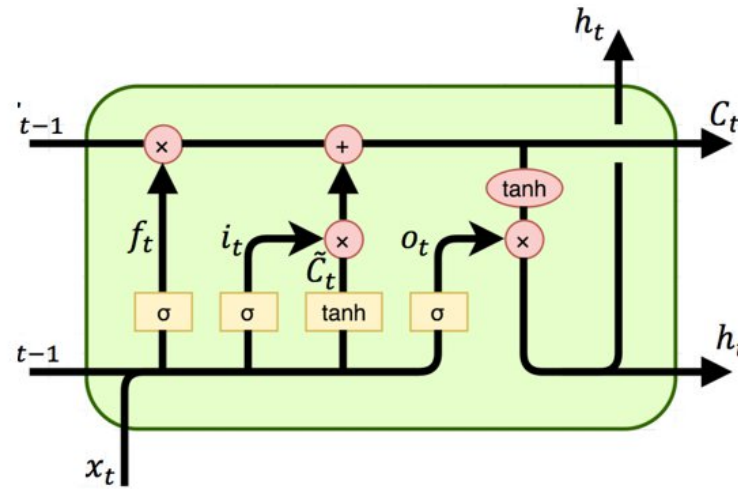


A flexible model type

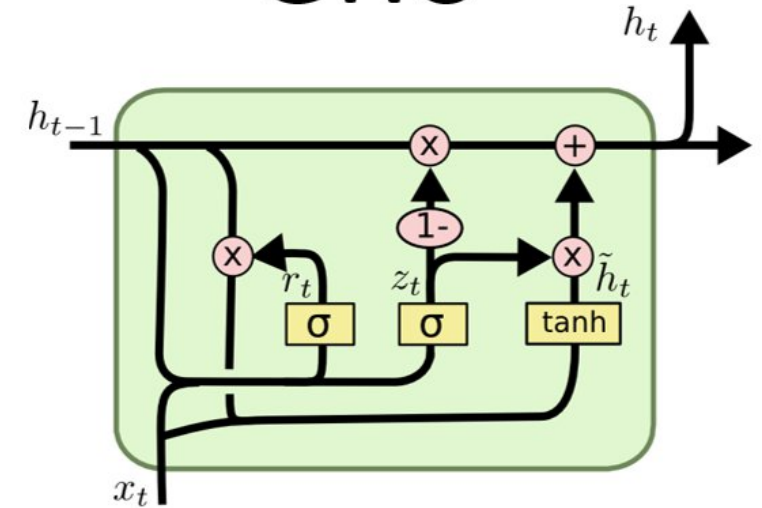
RNN



LSTM



GRU



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.

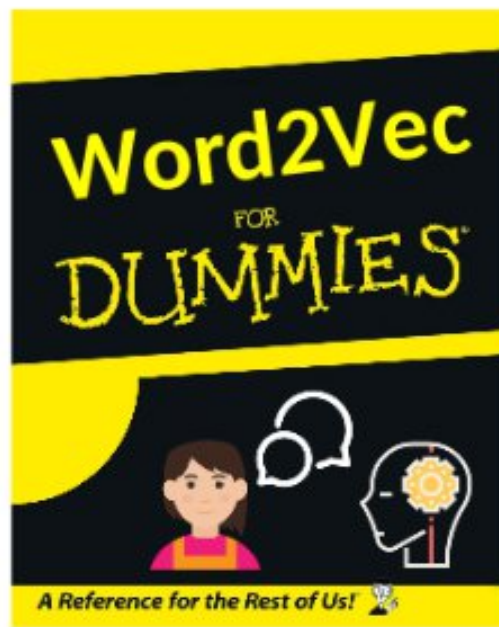


Could you summarise this text ?

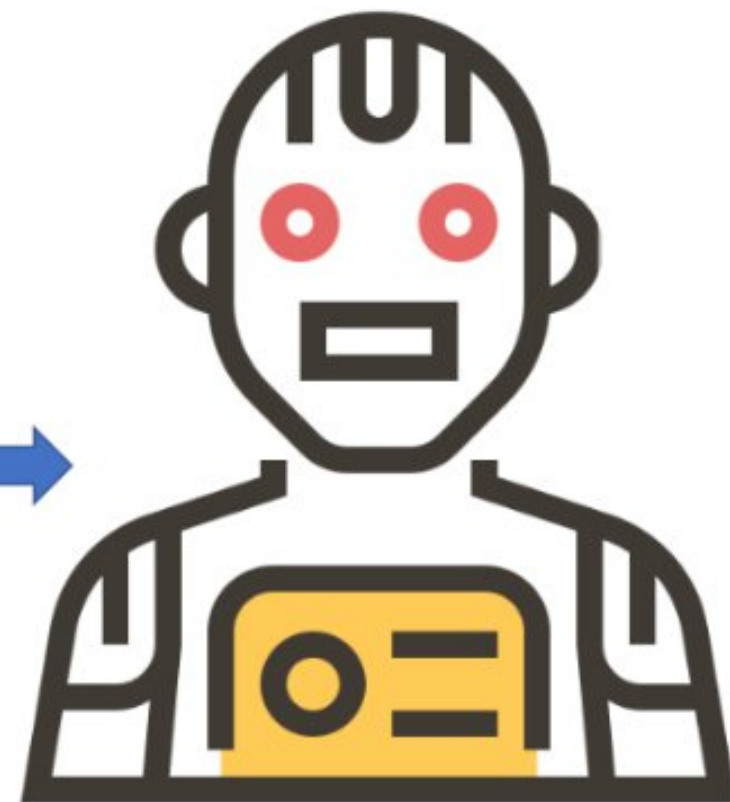


??????
10010011101010...



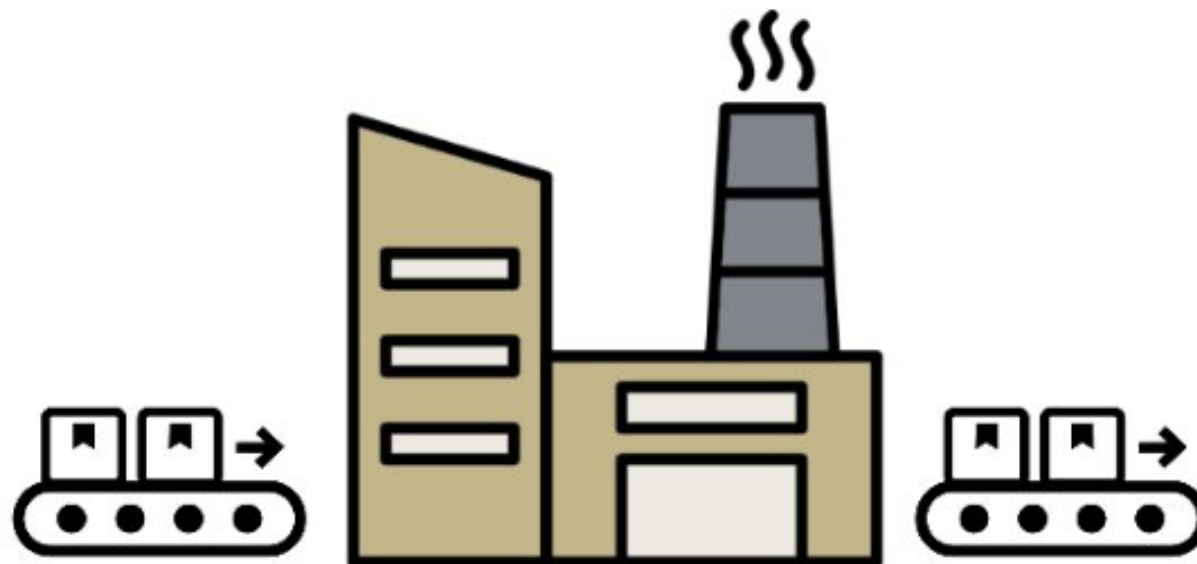


Embedding layer

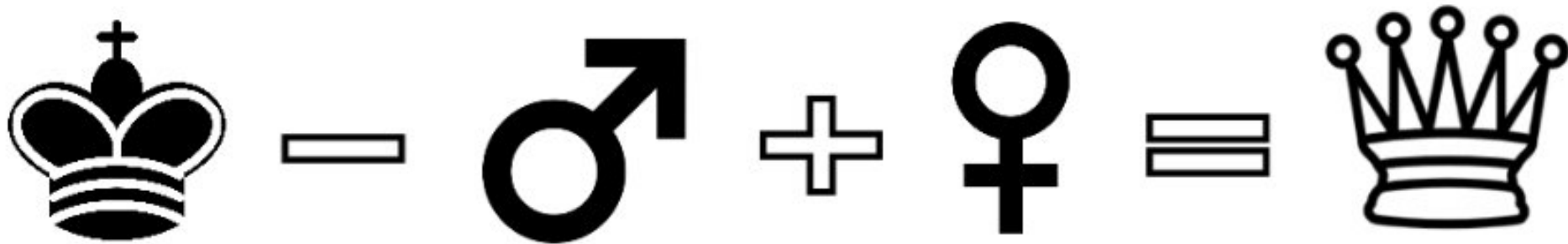
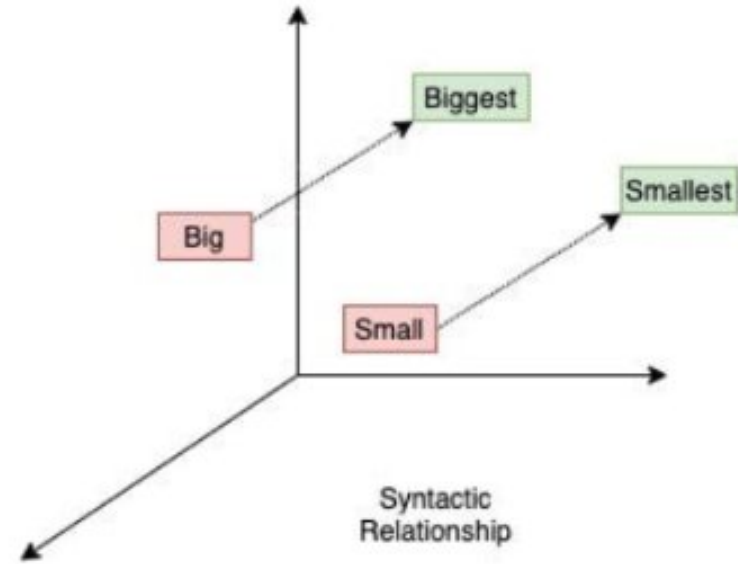
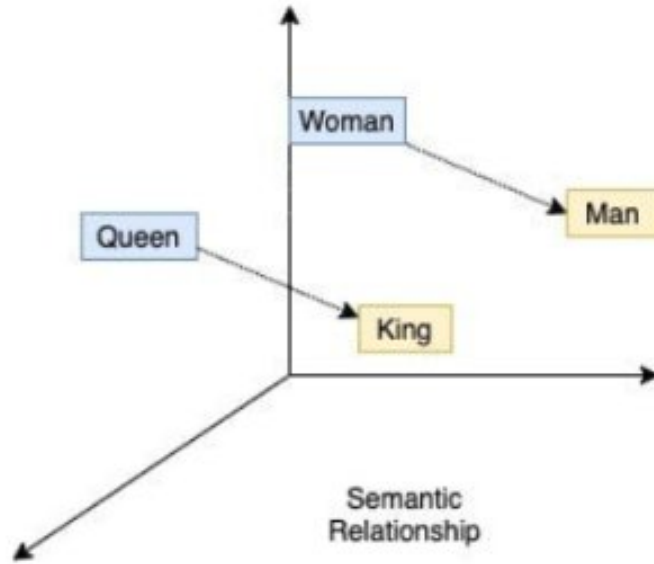




- king
- queen
- factory
- network



- 0.2, 1.5, ...
- 0.2, -1.5, ...
- 1.3, -0.1, ...
- -0.75, 0.05, ...



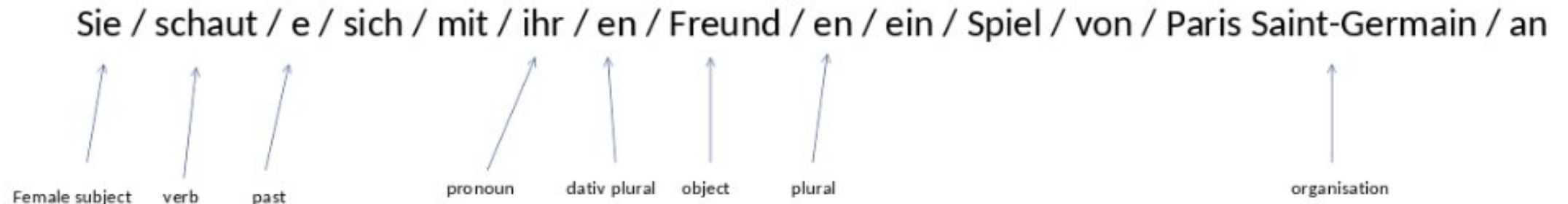
Embeddings

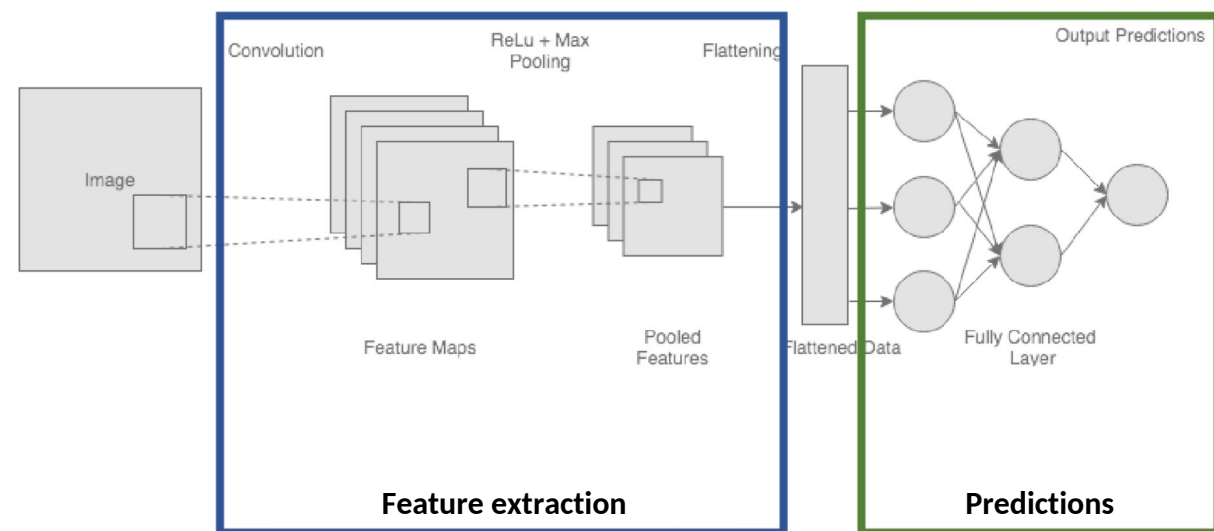
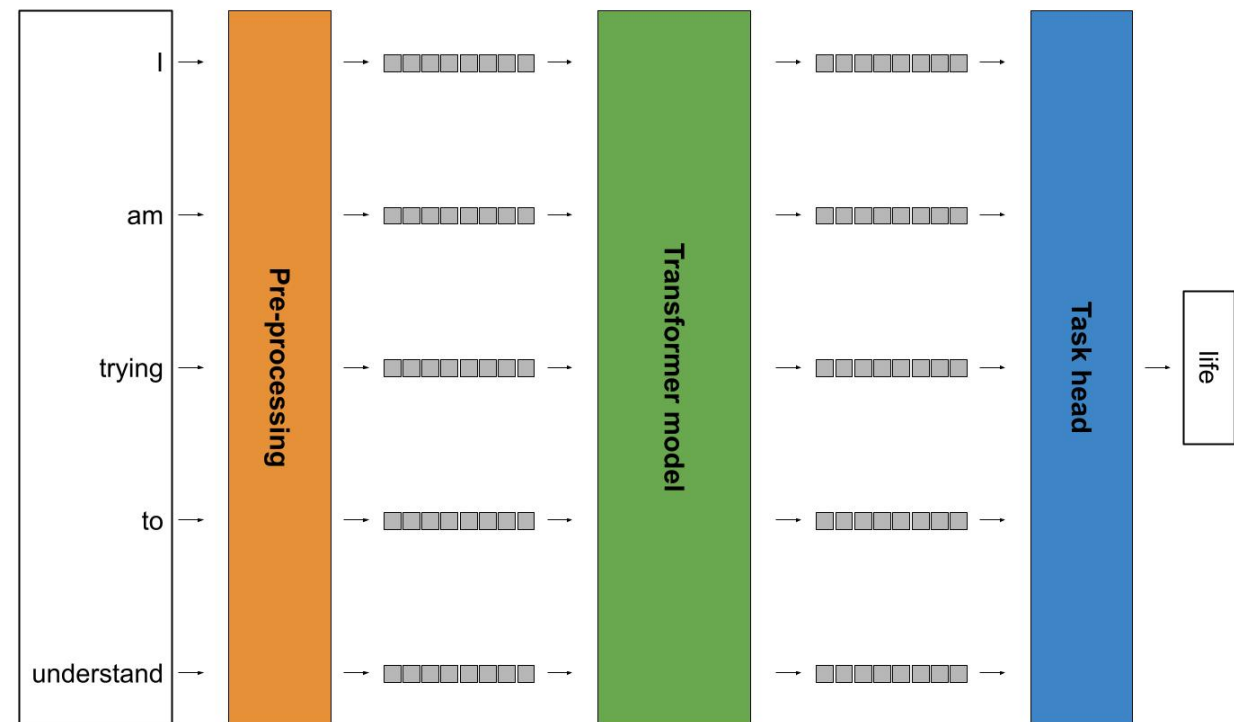
Anti - *chocs*, **constitution**
et *actu* - **elle** - **ment** sont appris.

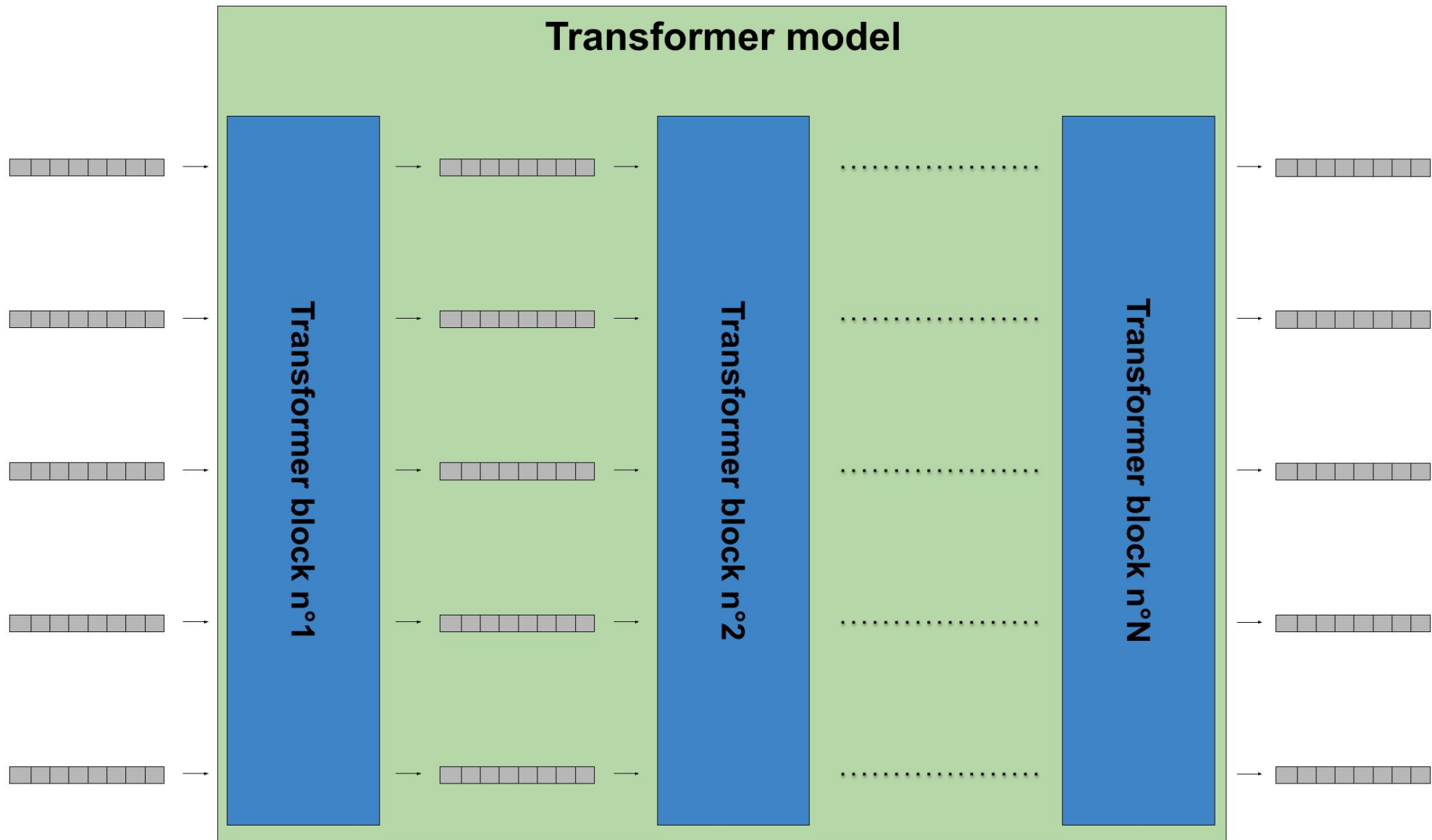
Anticonstitutionnellement sera compris.

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

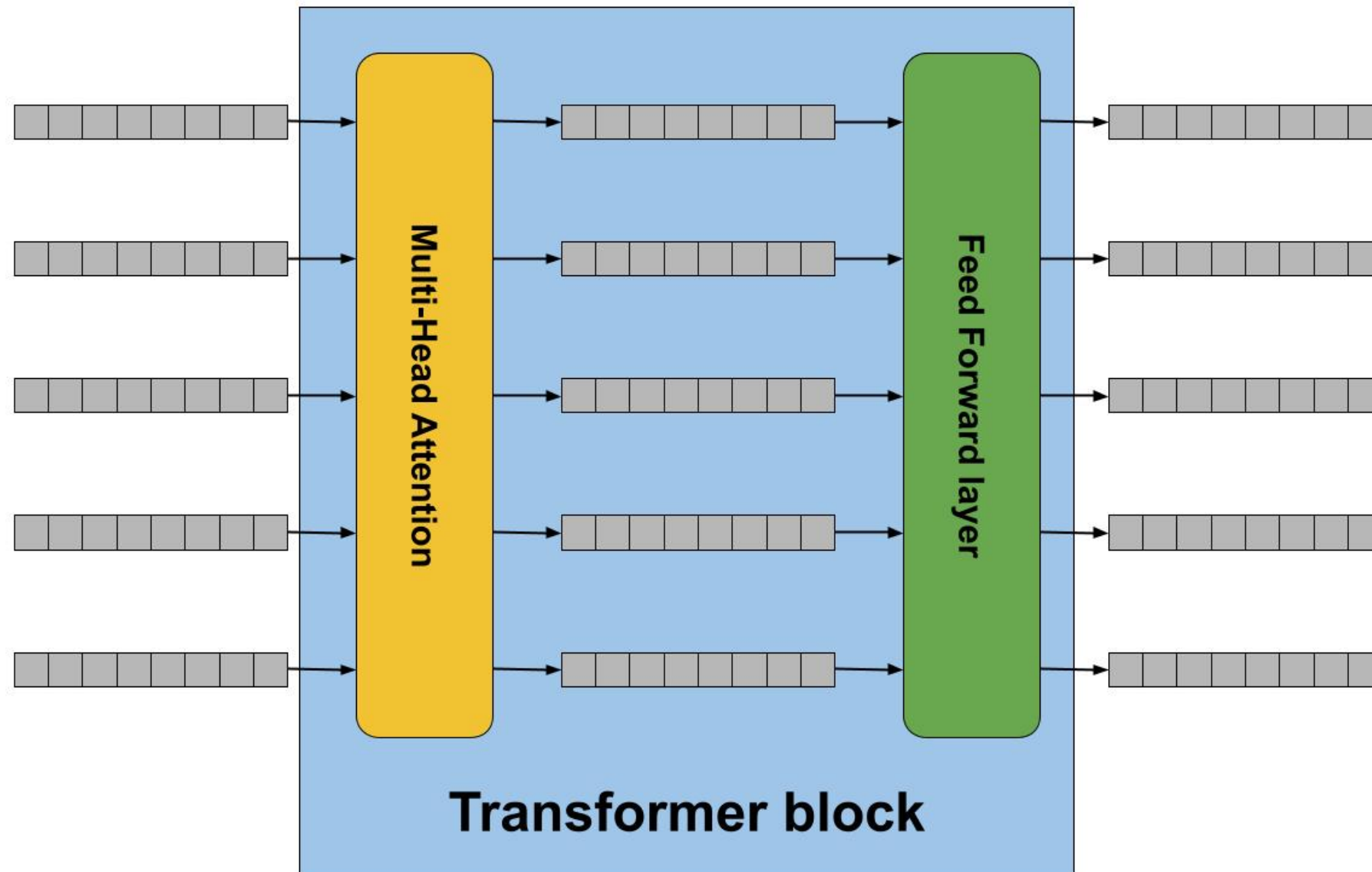
Example : Sie schaute sich mit ihren Freunden ein Spiel von Paris Saint-Germain an.



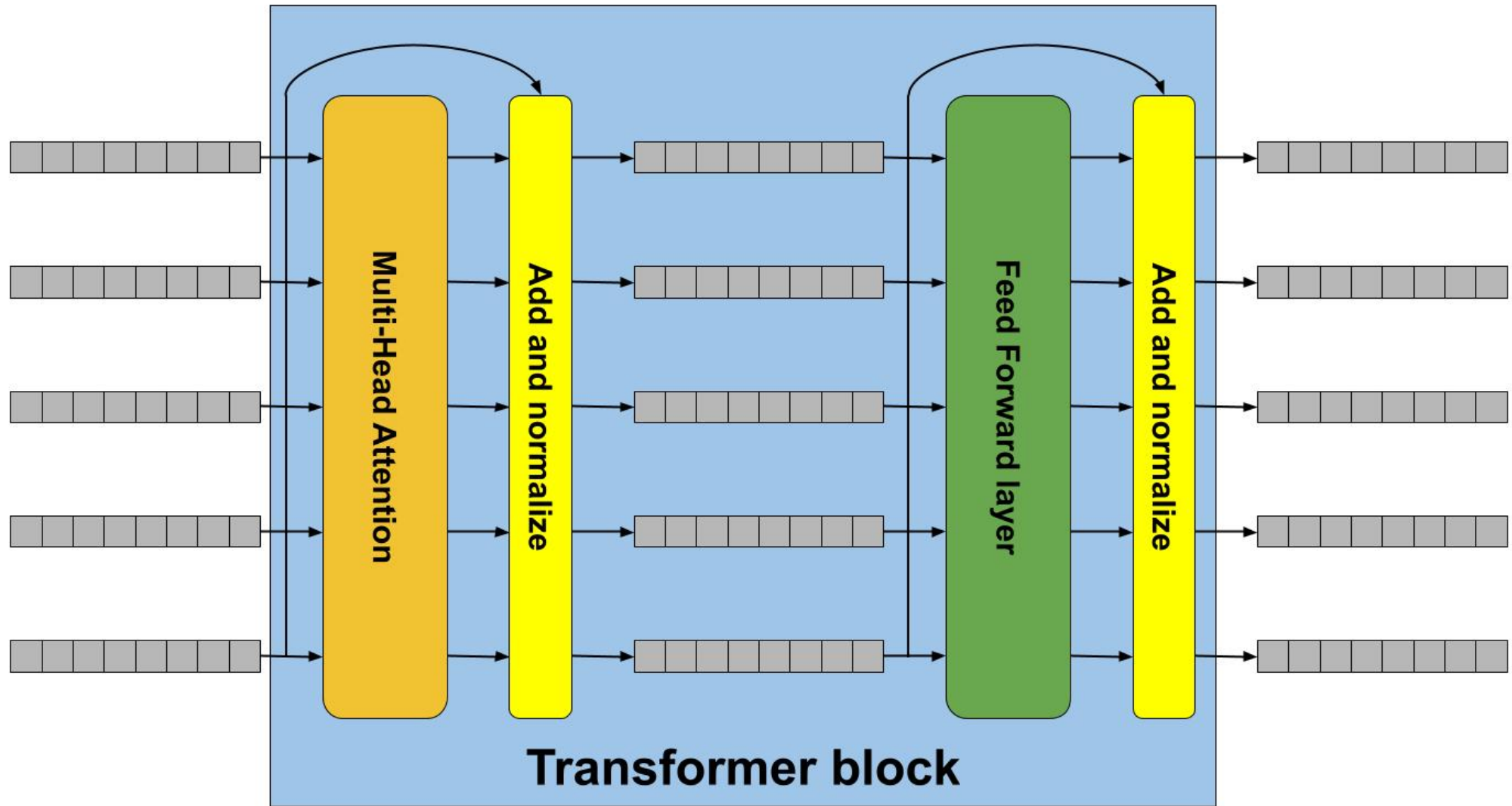




Transformer architecture (1)



Transformer architecture (2)



Transformer architecture (3)

Focus

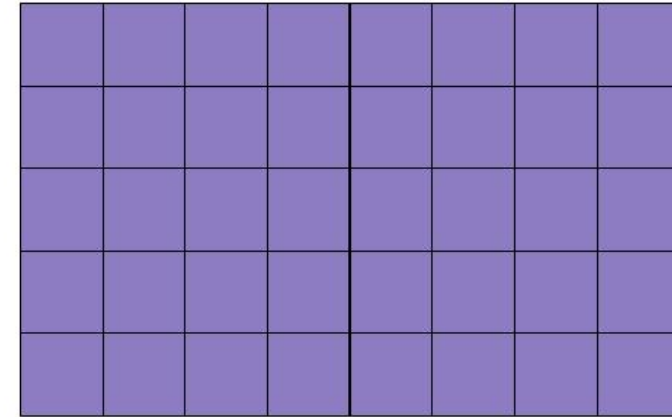
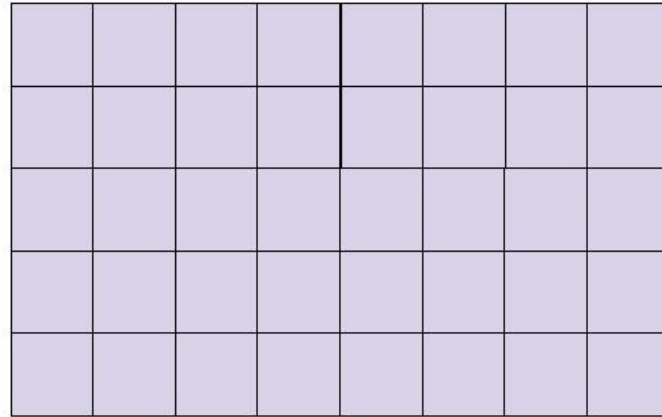
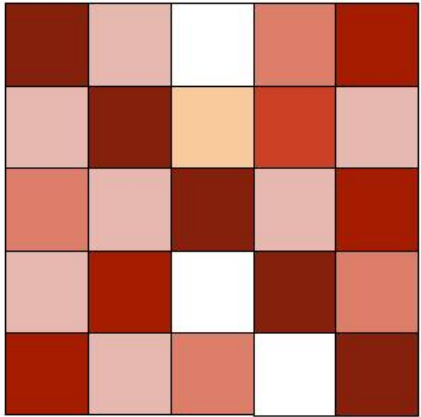
The → The big red dog
big → The big red dog
red → The big red dog
dog → The big red dog

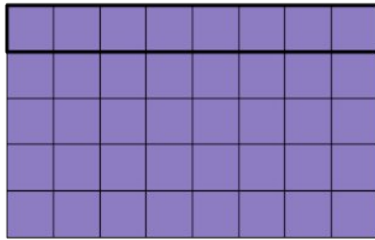
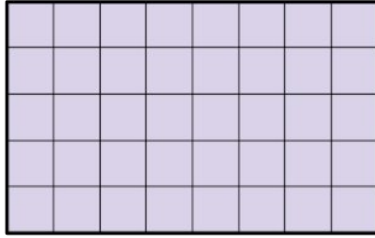
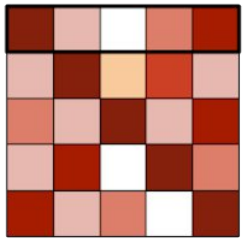
Transformer Neural Networks - EXPLAINED! (Attention is all you need) : <https://www.youtube.com/watch?v=TQQIZhbC5ps>

Intuition behind the Attention mechanism (1)

Attention matrix

V





$$\text{dark red square} \times \text{row of purple squares}$$

+

$$\text{light red square} \times \text{row of purple squares}$$

+

$$\text{white square} \times \text{row of purple squares}$$

+

$$\text{orange square} \times \text{row of purple squares}$$

+

$$\text{dark red square} \times \text{row of purple squares}$$

=



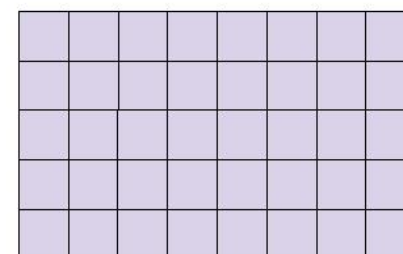
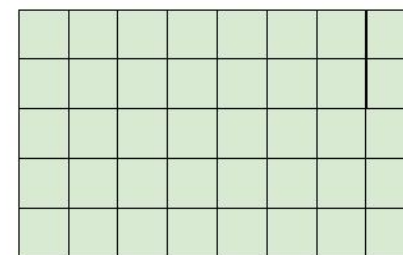
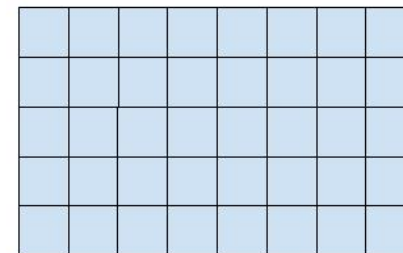
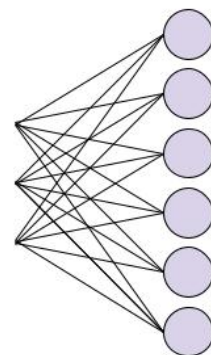
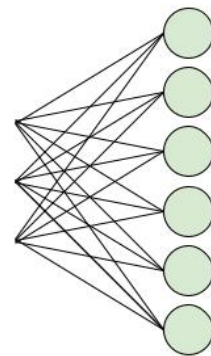
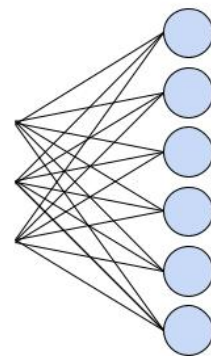
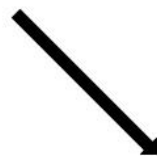
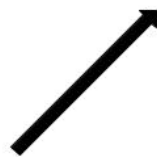
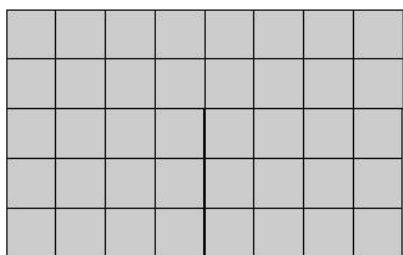
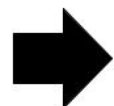
I

am

trying

to

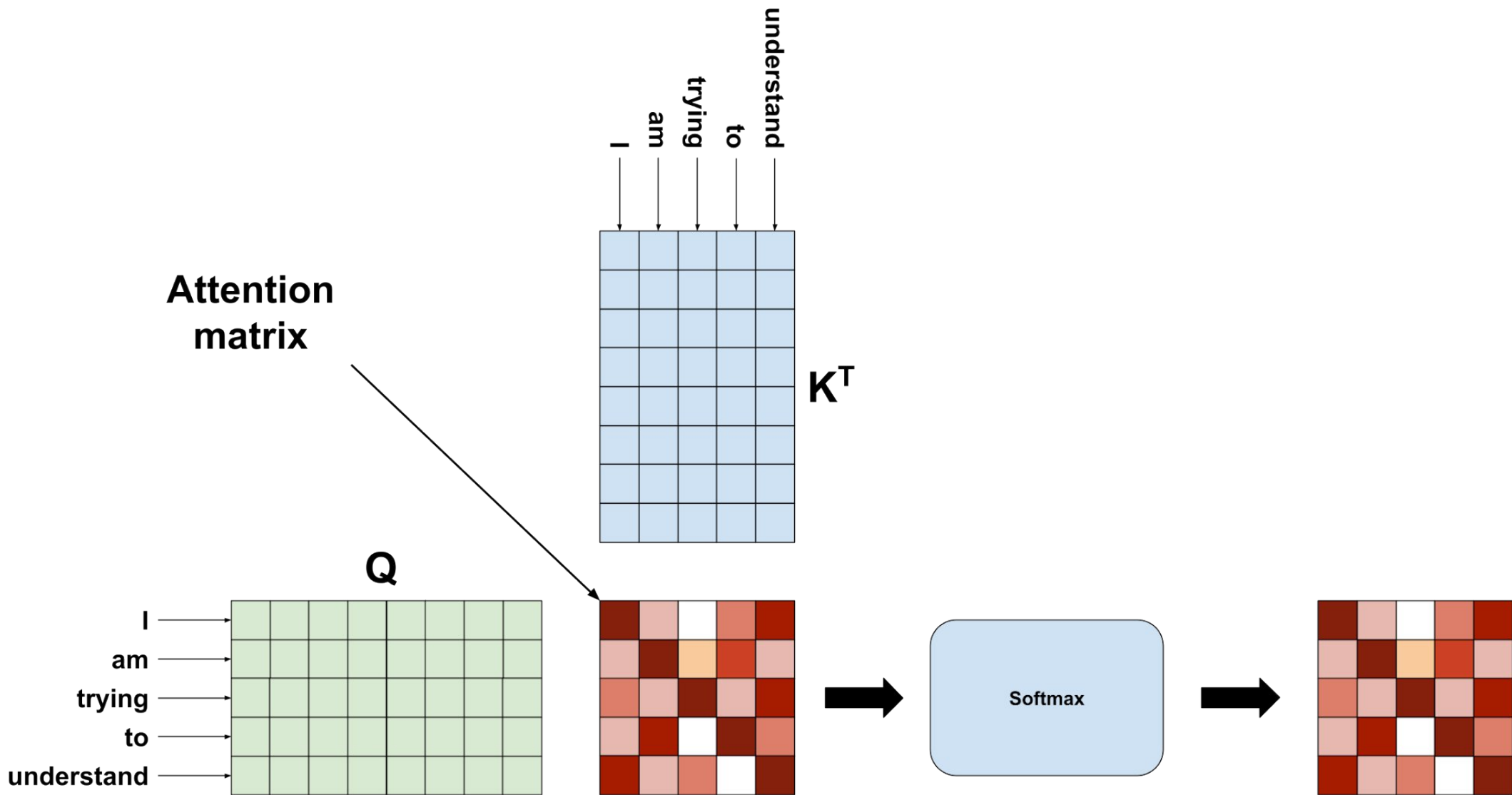
understand



K

Q

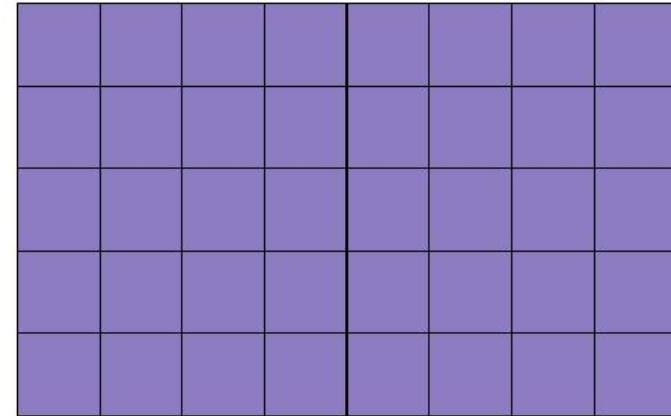
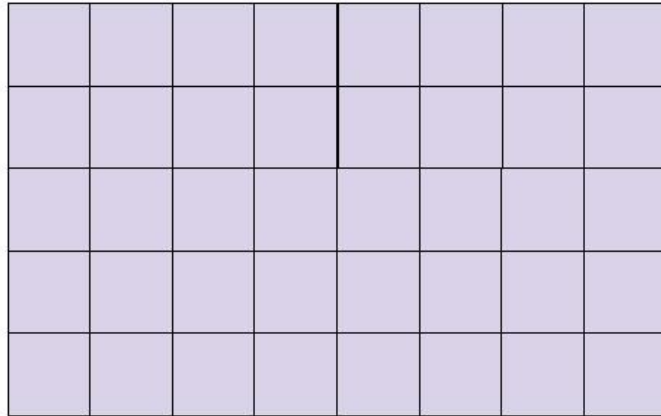
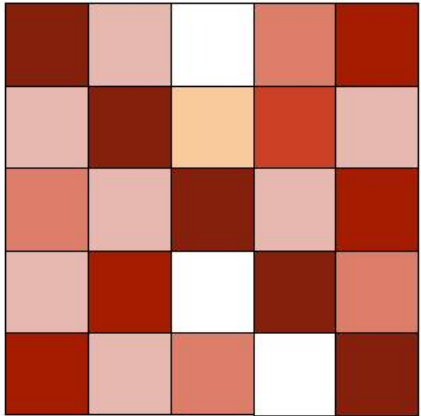
V



Attention mechanism (2)

Attention matrix

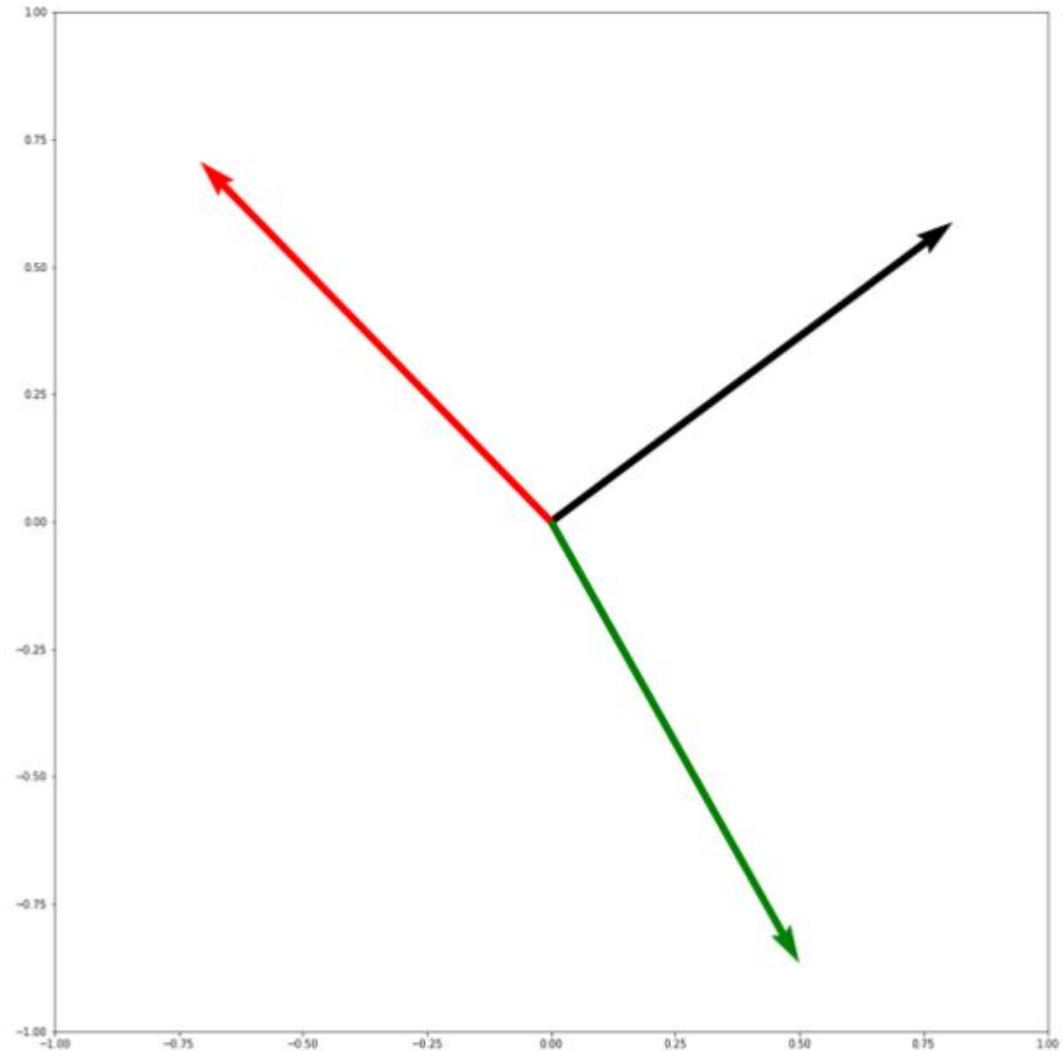
V



The big dog



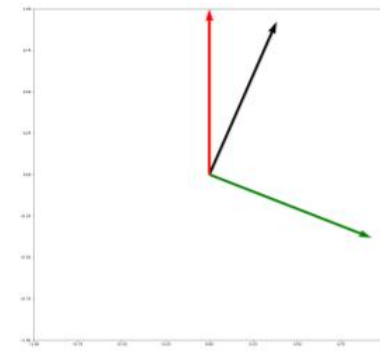
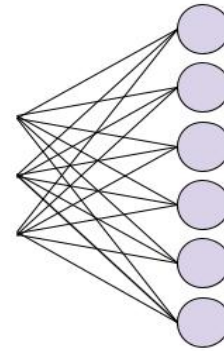
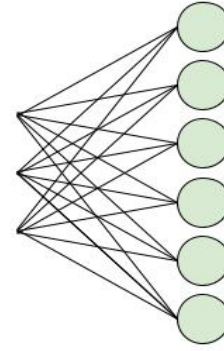
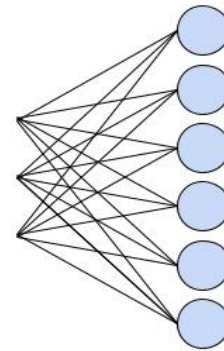
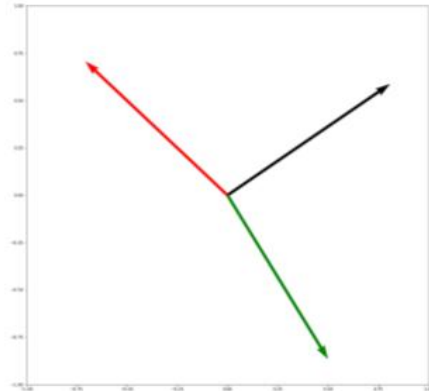
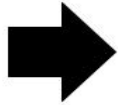
The : (0.50, -0.87)
big : (-0.70, 0.70)
dog : (0.81, 0.59)



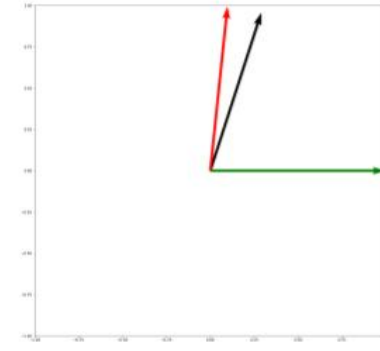
0.50	-0.87
------	-------

-0.70	0.70
-------	------

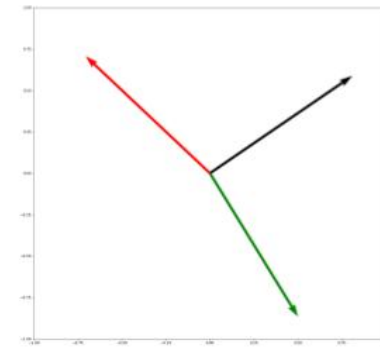
0.81	0.59
------	------



K

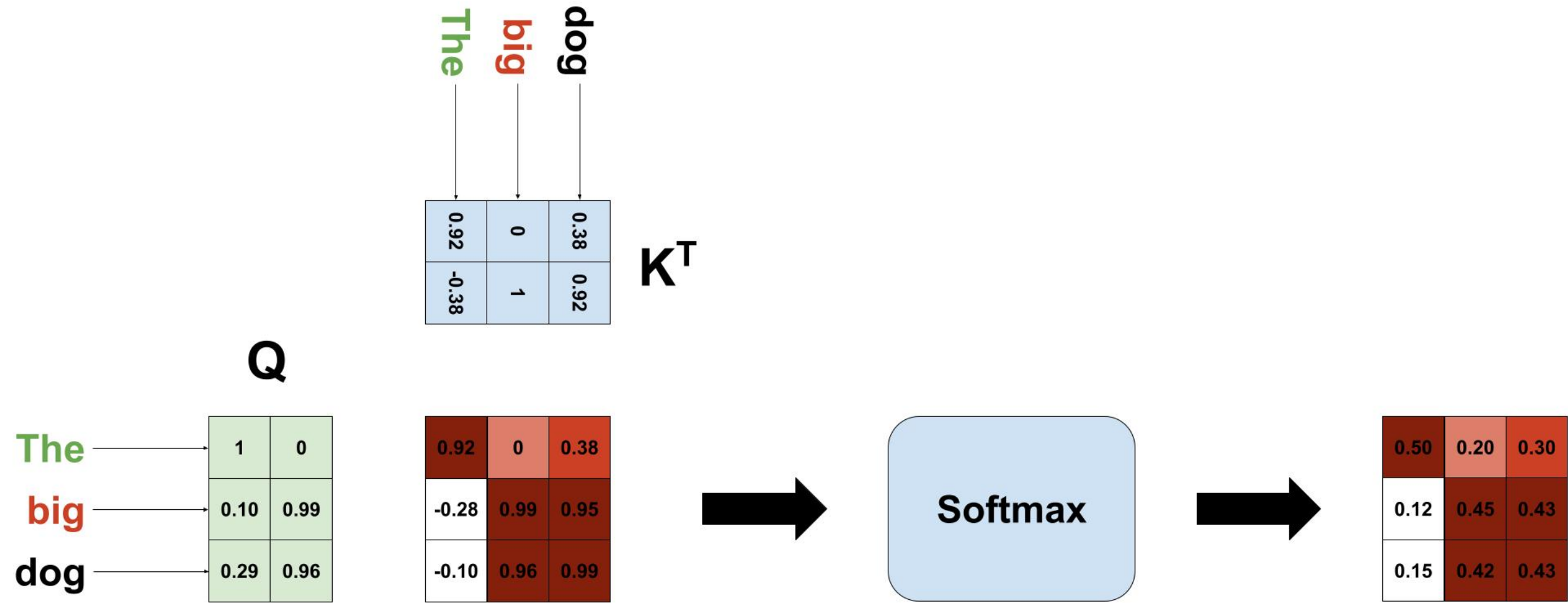


Q



V

Attention mechanism - Example (2)



Attention mechanism - Example (3)

0.50	0.20	0.30
0.12	0.45	0.43
0.15	0.42	0.43

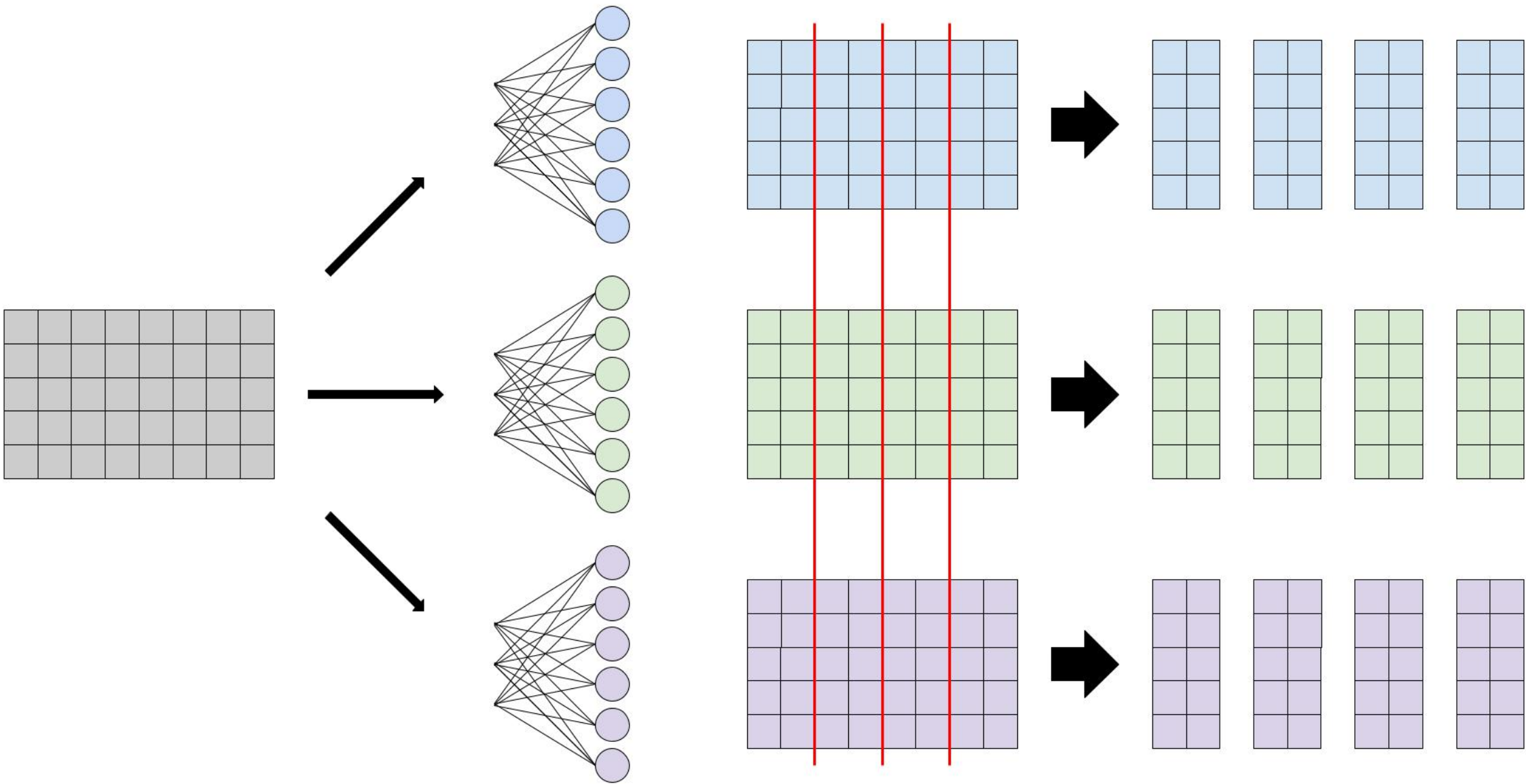


0.50	-0.87
-0.70	0.70
0.81	0.59

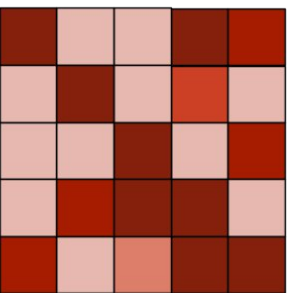
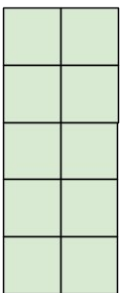
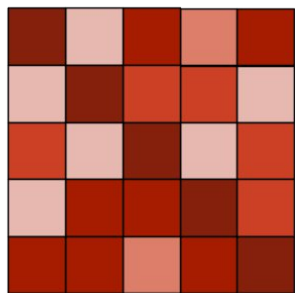
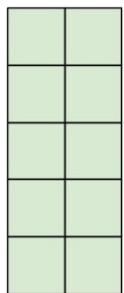
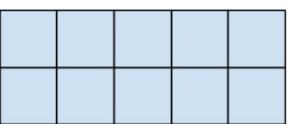
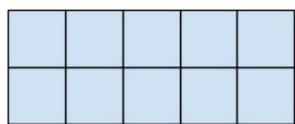
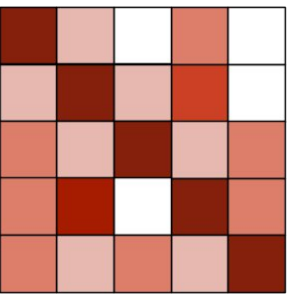
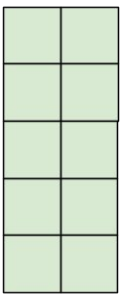
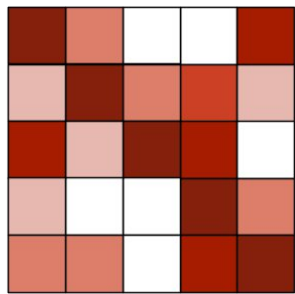
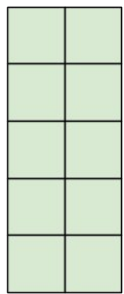
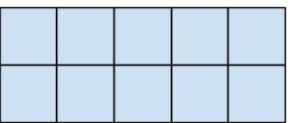
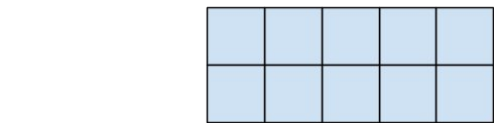


0.35	-0.12
0.10	0.46
0.13	0.42

Attention mechanism - Example (4)

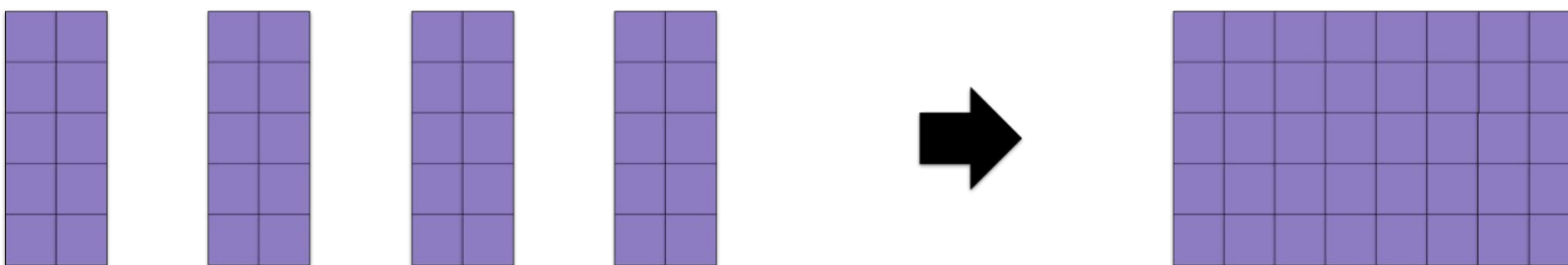
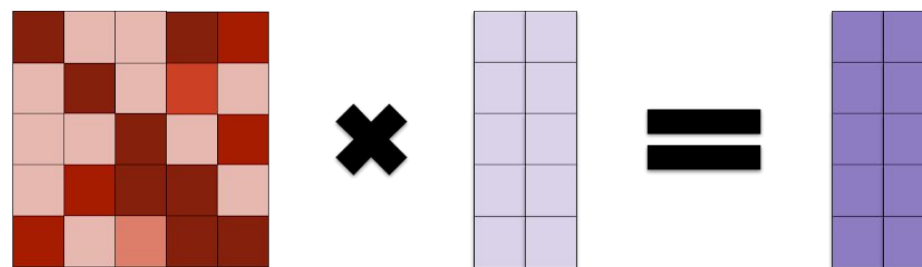
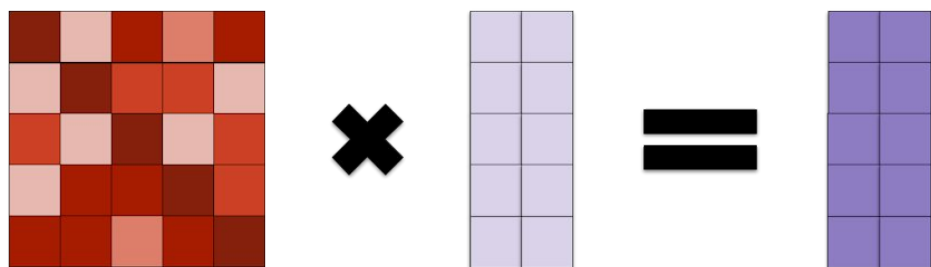
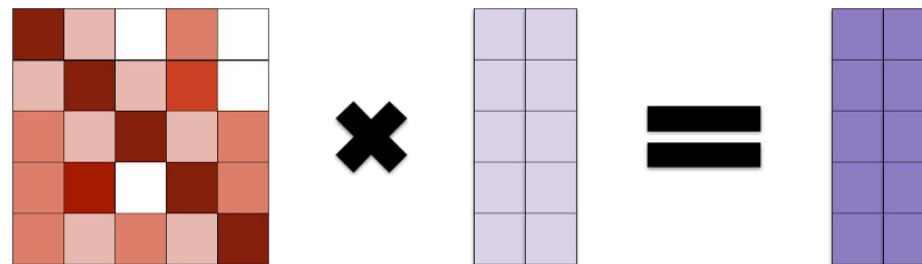
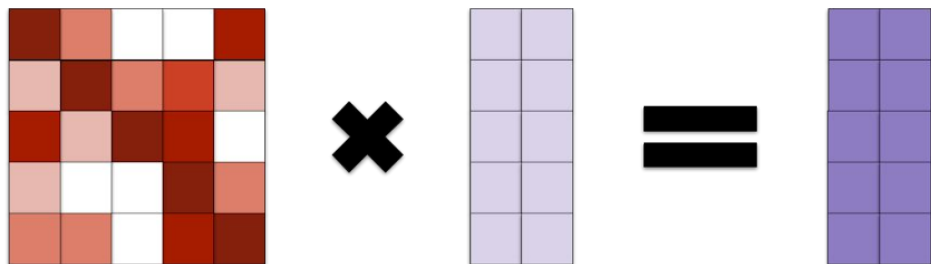


Multi-Head Attention (1)



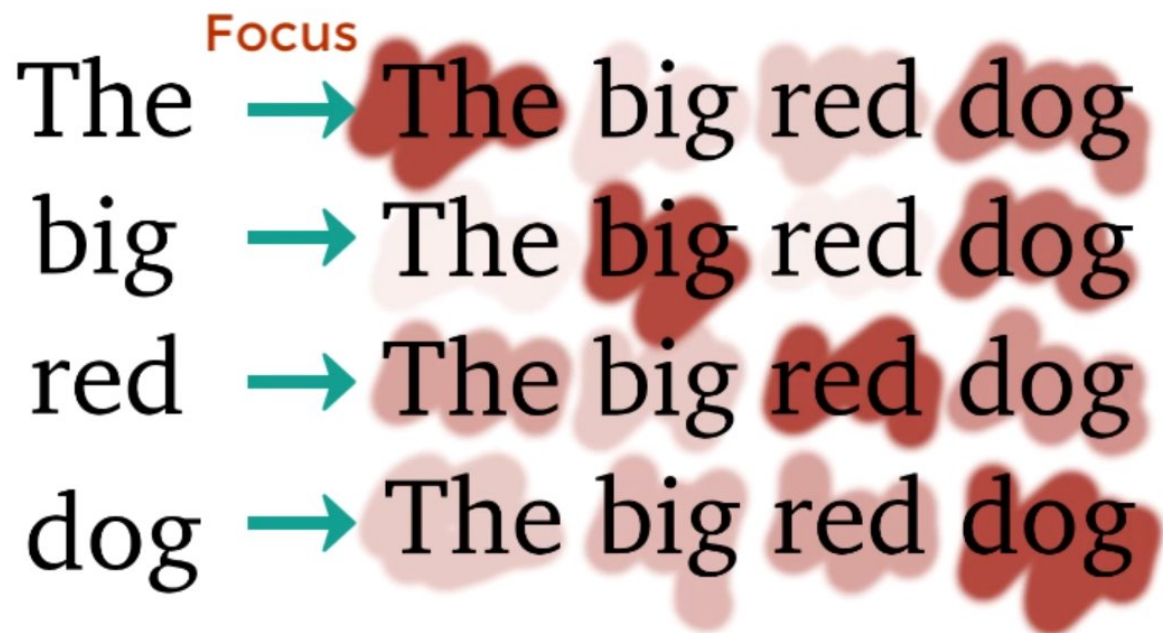
Multi-Head Attention (2)





Multi-Head Attention (3)

Bidirectional attention (BERT - Encoder - Auto-encoding)

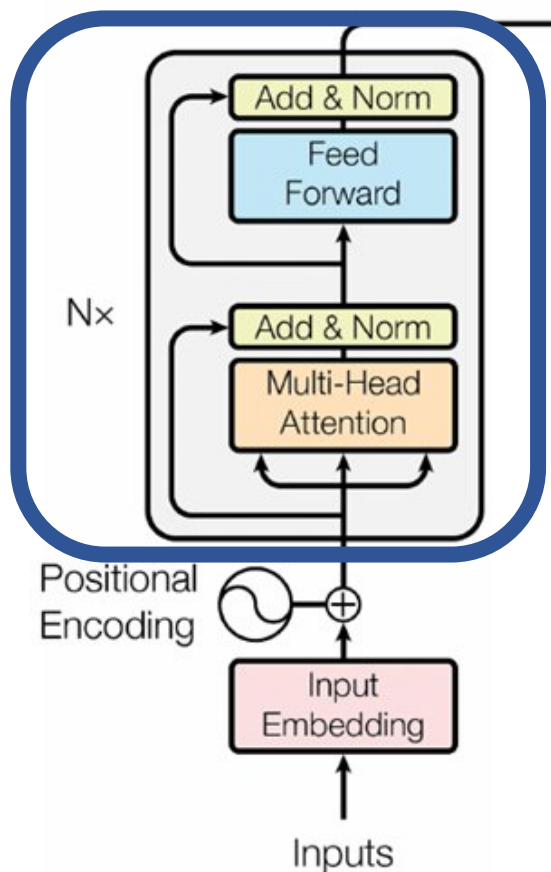


Unidirectional attention (GPT - Decoder - Auto-regressive)

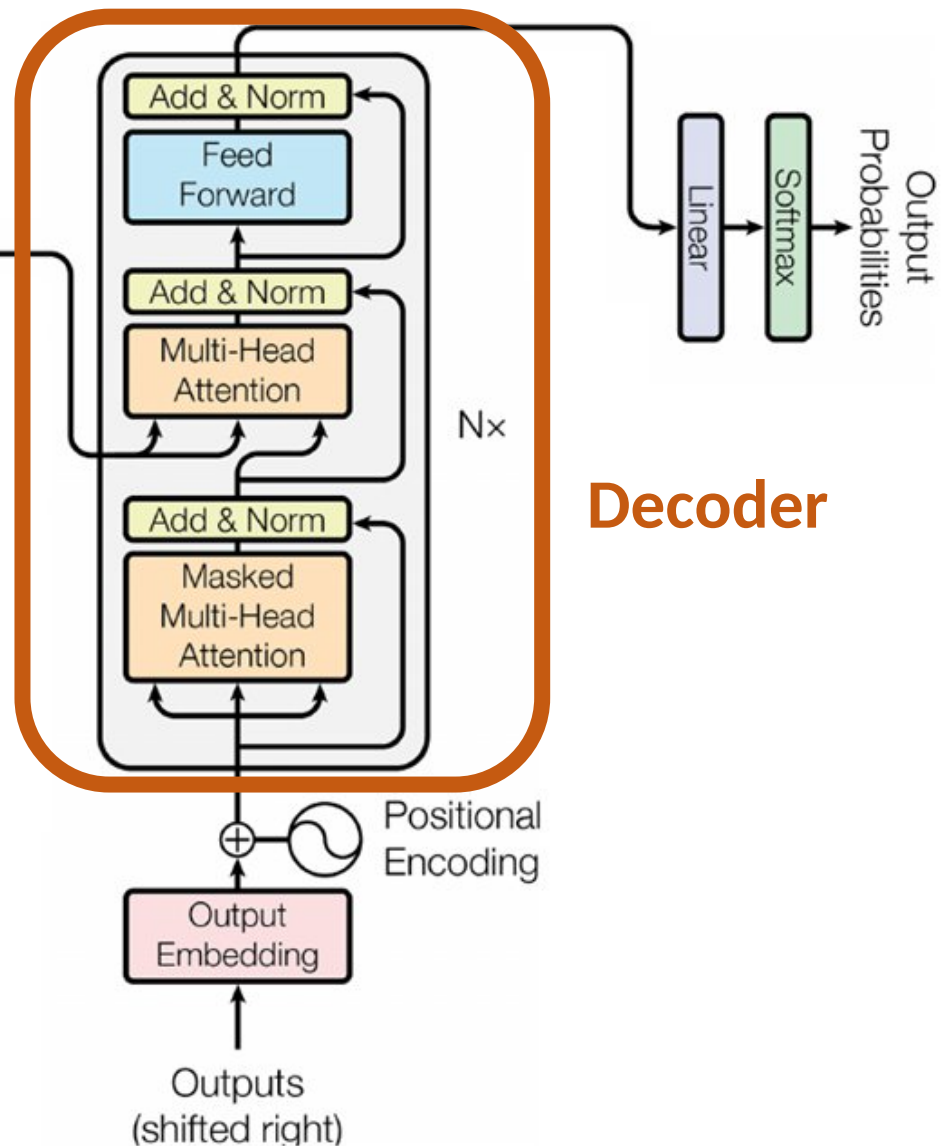


Transformer Neural Networks - EXPLAINED! (Attention is all you need) : <https://www.youtube.com/watch?v=TQQIZhbC5ps>

Encoder



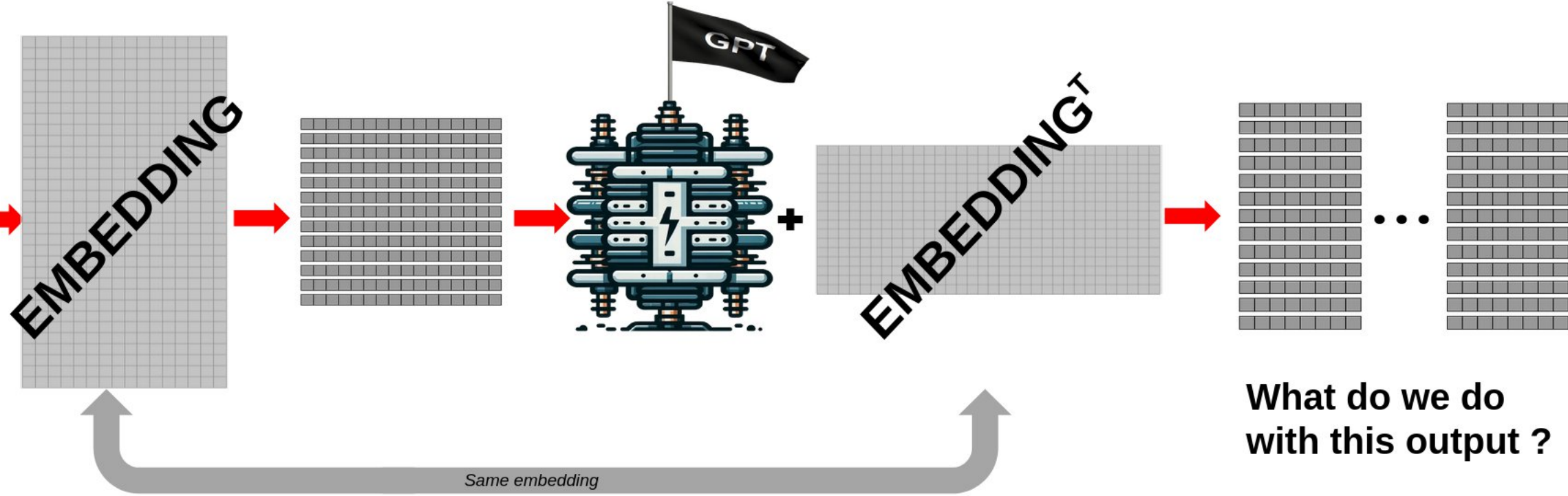
Decoder



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

Transformer types (2)

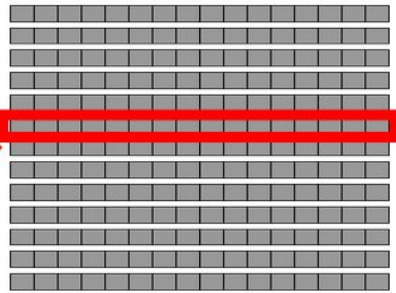
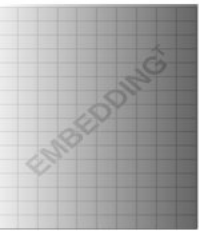
A transformer is a deep learning model that adopts the mechanism of self



What do we do with this output ?



Training a transformer



Cheese — 3%
 Attention — 13%
 Model — 56%

-
-
-

Calf — 0.04%
 Muscle — 0.05%
 Mercenary — 6%

Classification target

Cheese — 0
 Attention — 0
 Model — 1

-
-
-

Calf — 0
 Muscle — 0
 Mercenary — 0



Input

A
transformer
is
a
deep
learning
model
that
adopts
the
mechanism
of
self



Target

transformer
is
a
deep
learning
model
that
adopts
the
mechanism
of
self
attention

Next word prediction

Training a GPT-style transformer

Sample

A transformer is a deep learning model that adopts the mechanism of self attention



~ 15%



Input

[CLS]
A
[MASK] is a deep learning model that [MASK] the mechanism of self [MASK]



Target

/
/
transformer
/
/
/
/
/
/
adopts
/
/
/
/
attention

Masked words prediction

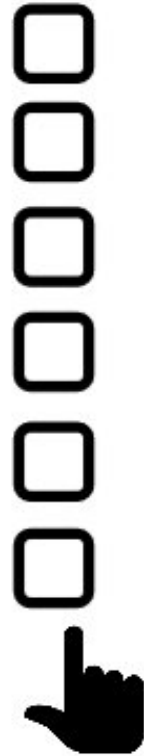
Training a BERT-style transformer



Which word do we choose ?

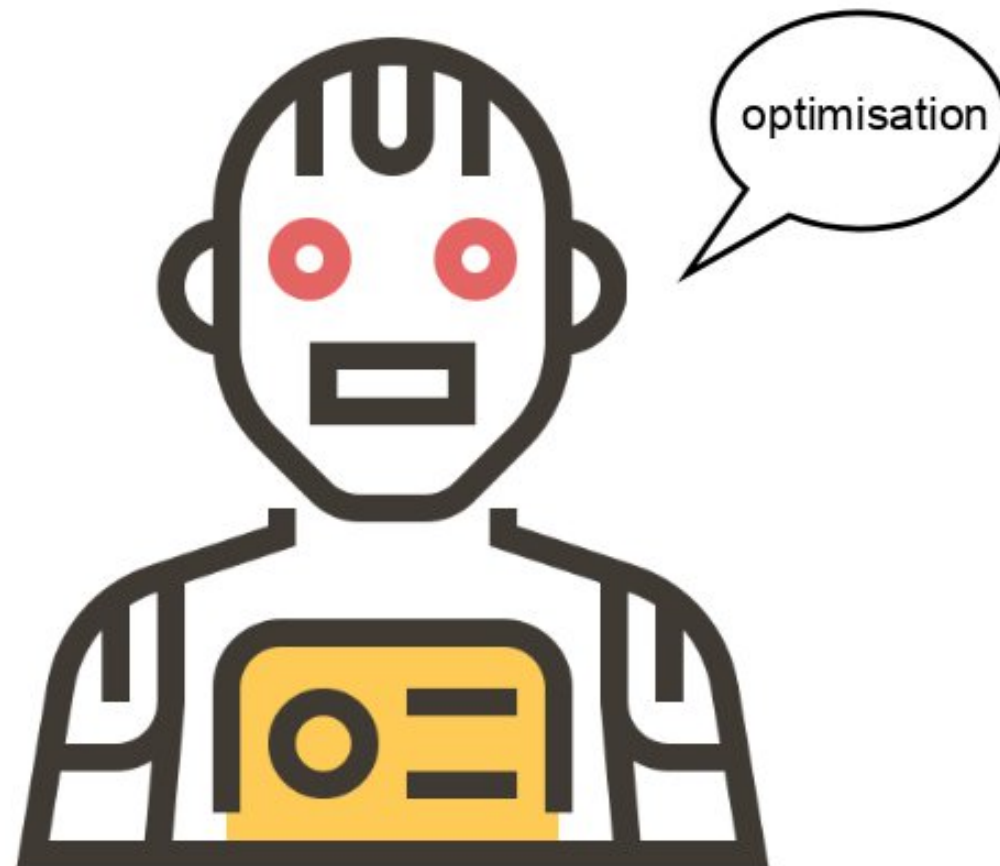
- Optimisation — 35%
- Intelligence — 22%
- Apprentissage — 18%
- Machine — 10%
- Isolement — 1.2%
- Fromage — 0.04%

...



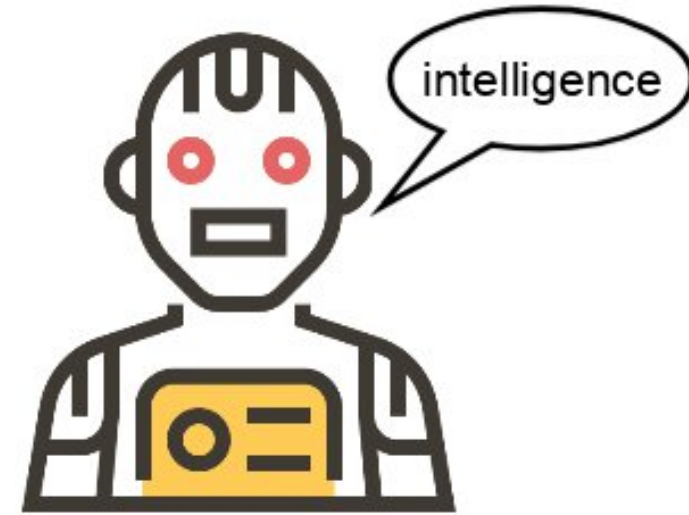
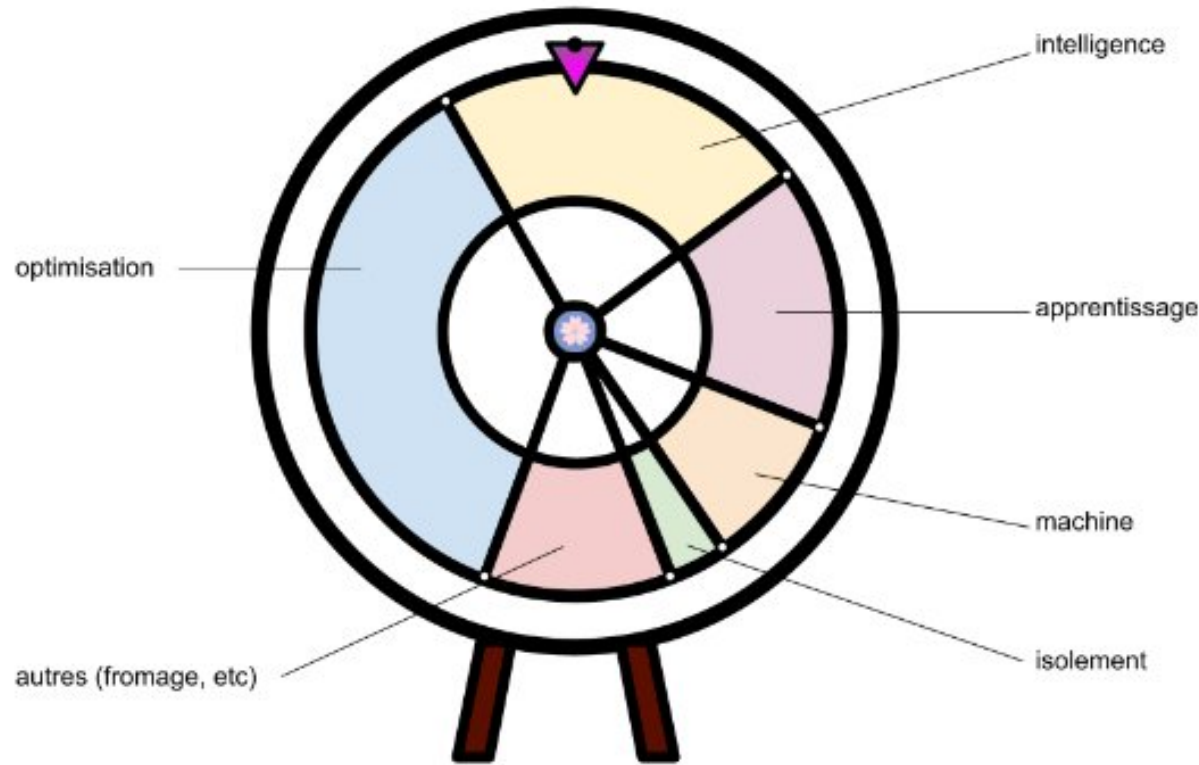
- Optimisation — 35%
- Intelligence — 22%
- Apprentissage — 18%
- Machine — 10%
- Isolement — 1.2%
- Fromage — 0.04%

...



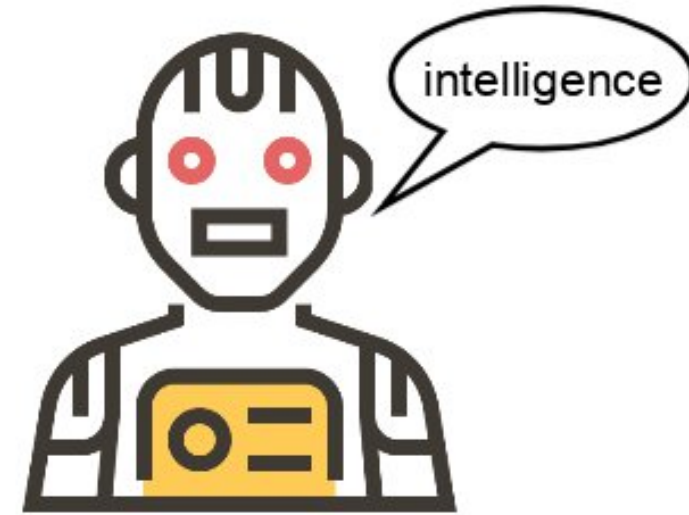
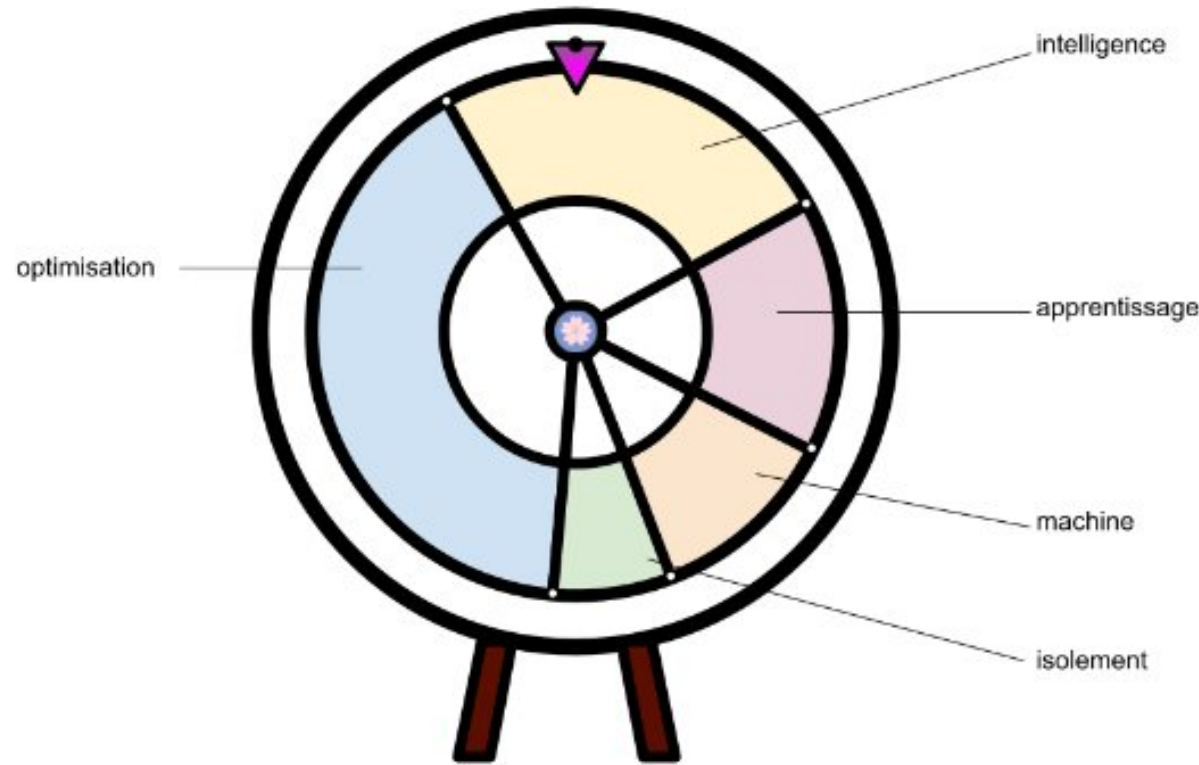
- Optimisation — 35%
- Intelligence — 22%
- Apprentissage — 18%
- Machine — 10%
- Isolement — 1.2%
- Fromage — 0.04%

...



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

...



Les réseaux de neurones sont un algorithme d'

Optimisation — 35%

numérique — 23%

linéaire — 13%

Intelligence — 22%

artificielle — 54%

émotionnelle — 27%

Apprentissage — 18%

supervisé — 44%

automatique — 32%

Machine — 10%

learning — 97%

nespresso — 2.5%

Isolement — 1.2%

Préférence finale

8.05%

4.55%

11.88%

5.94%

7.92%

5.76%

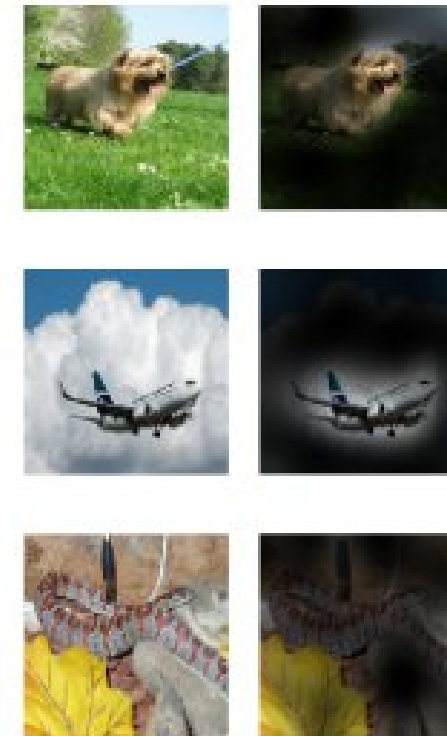
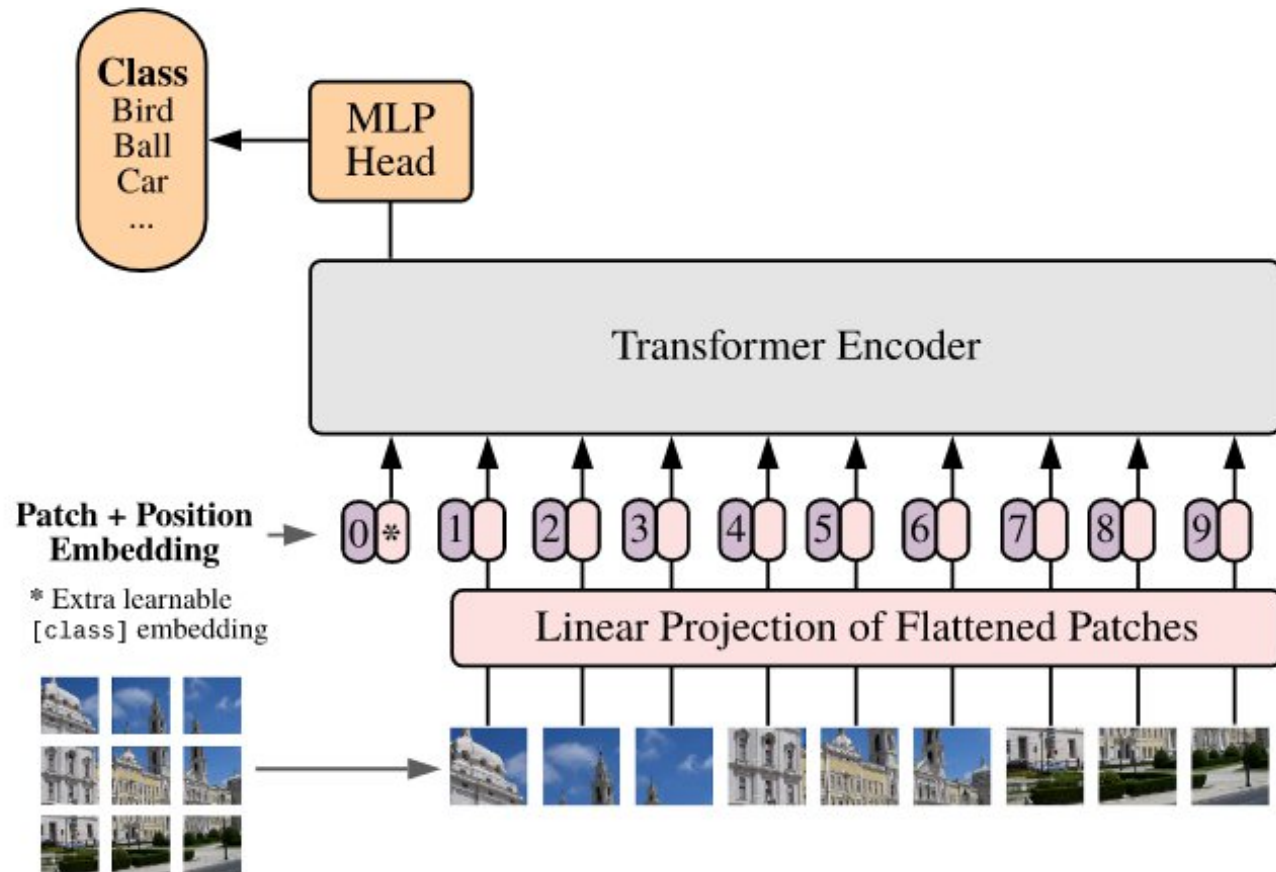
9.70%

0.25%

...

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).