

ML & NLP a brief account

- The roots
- Embeddings
- Recurrent Neural Networks
- Attention is all you need
- Applications

In []:

In []:

In []:

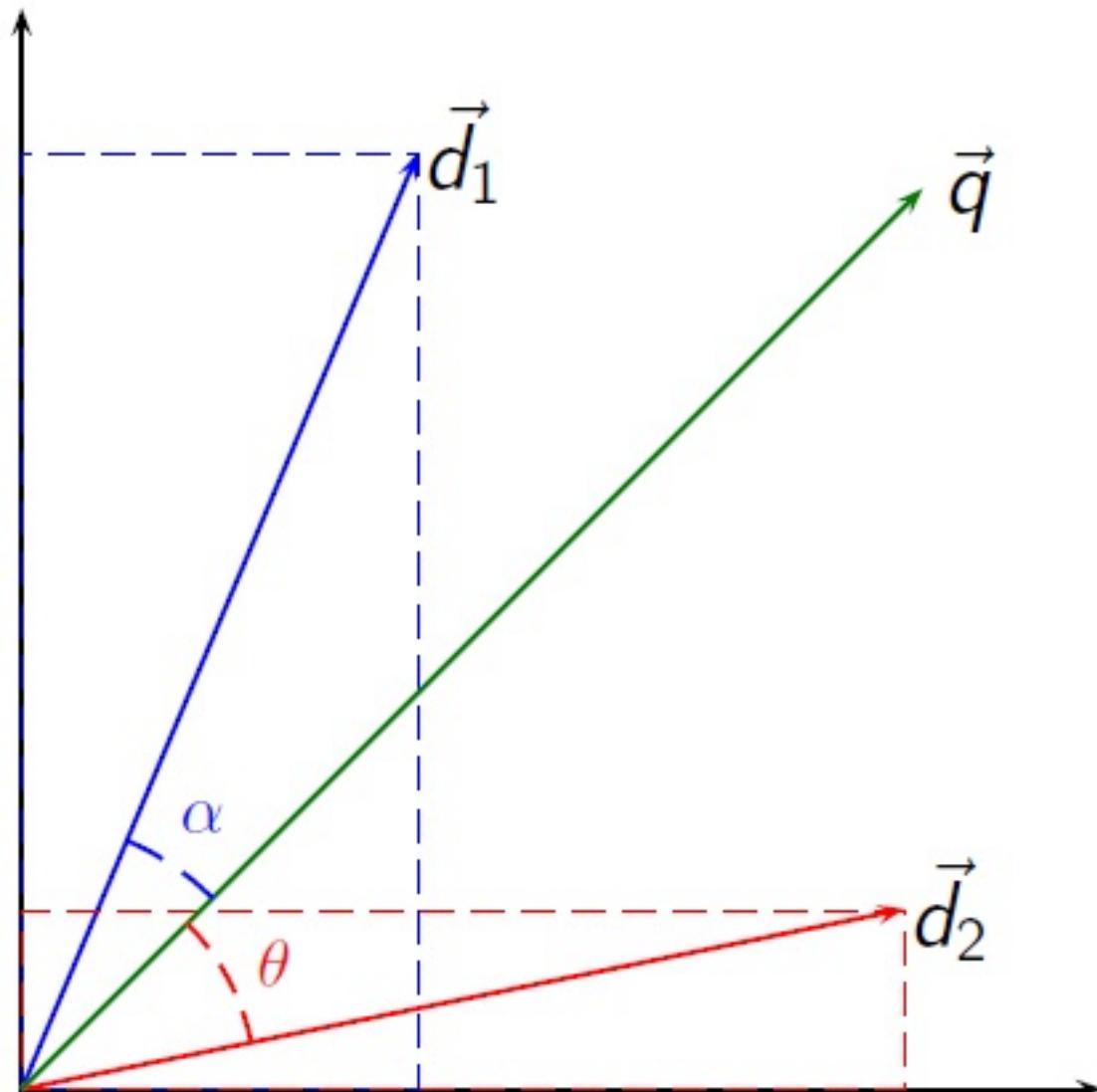
In []:

The roots

- Vector space model
- Latent Semantic analysis / Latent Dirichlet Analysis

The vector space model

It all started with Salton in 1975 and his vector space model



LSA & LDA

The main limitation of the Vector Space Model lies in its incapacity to detect close documents which do not have common terms. For instance : "this is a superb kitty" and "it's a beautiful cat"

LSA

$$\begin{matrix} \text{documents} \\ \boxed{C} \\ \text{words} \end{matrix} = \begin{matrix} \text{words} \\ \boxed{U} \\ \text{dims} \end{matrix} \begin{matrix} \text{words} \\ \boxed{D} \\ \text{dims} \end{matrix} \begin{matrix} \text{documents} \\ \boxed{V^T} \\ \text{dims} \end{matrix}$$

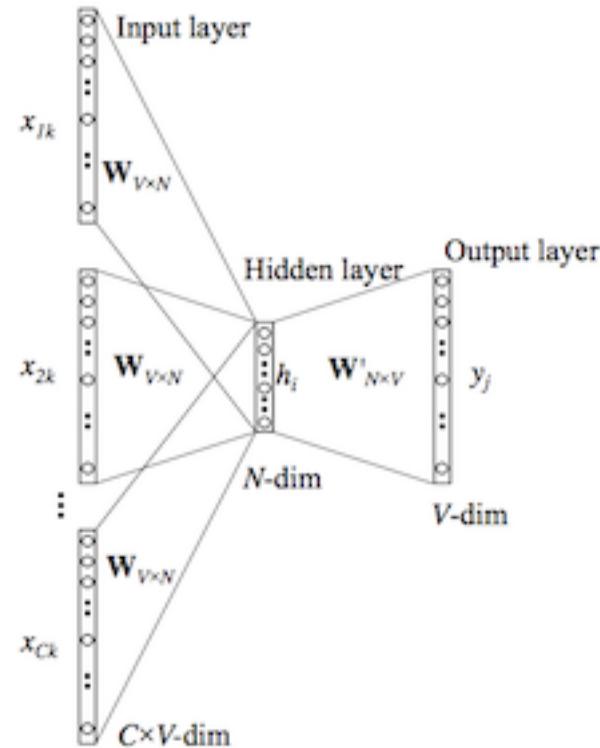
TOPIC MODEL

$$\begin{matrix} \text{documents} \\ \boxed{C} \\ \text{words} \\ \text{normalized co-occurrence matrix} \end{matrix} = \begin{matrix} \text{topics} \\ \boxed{\Phi} \\ \text{words} \\ \text{mixture components} \end{matrix} \begin{matrix} \text{topics} \\ \boxed{\Theta} \\ \text{documents} \\ \text{mixture weights} \end{matrix}$$

Embeddings

One can see them as the idea of dimension reduction of the vector space model to highlight semantic proximities by other means. [A good Medium article \(<https://towardsdatascience.com/introduction-to-word-embedding-and-word2vec-652d0c2060fa>\)](https://towardsdatascience.com/introduction-to-word-embedding-and-word2vec-652d0c2060fa)

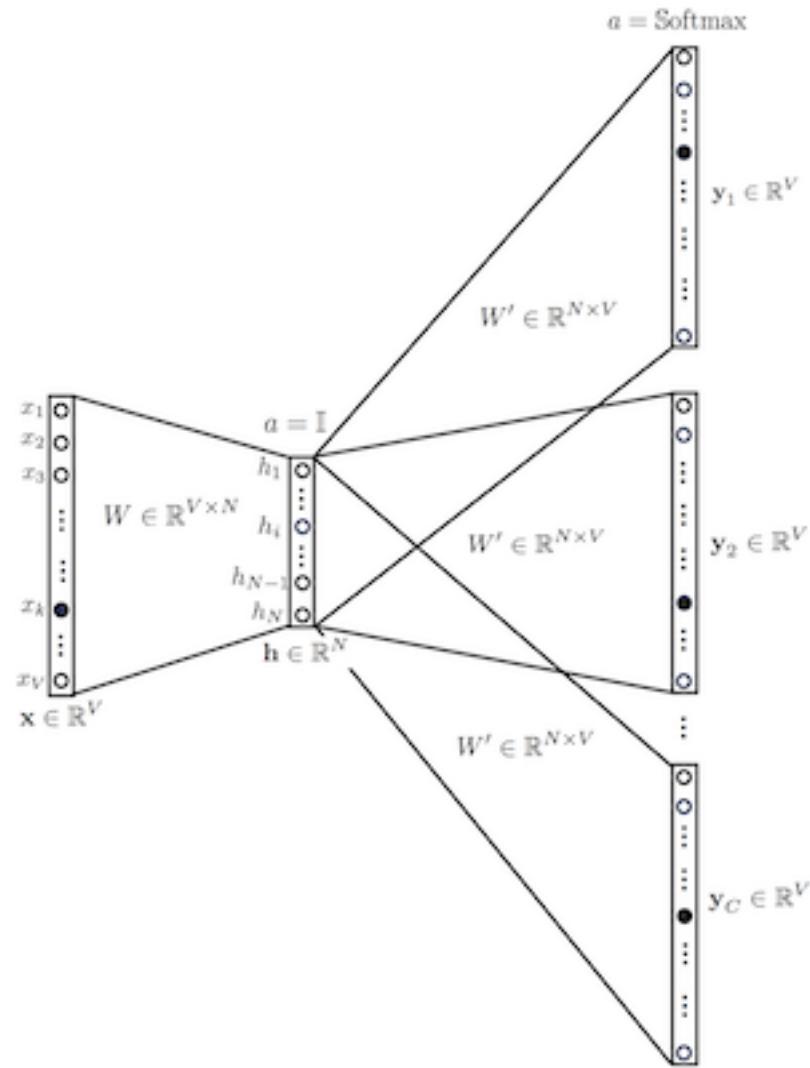
CBOW



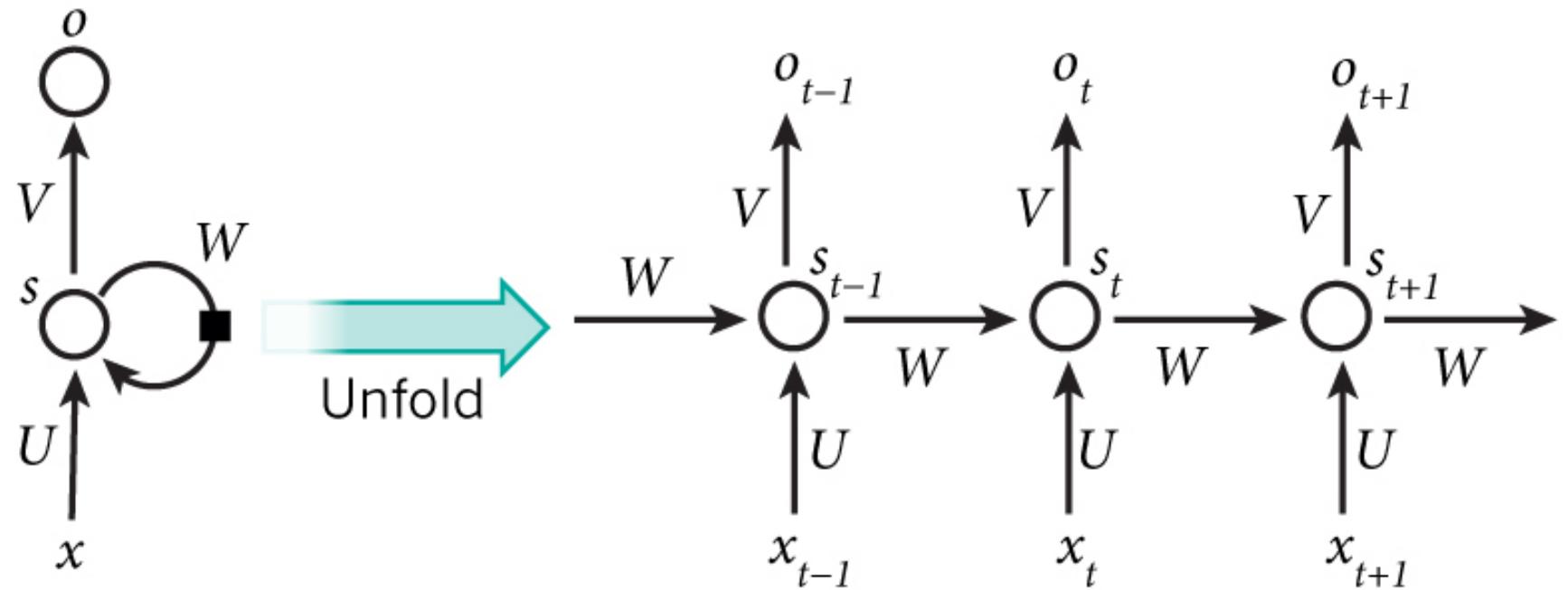
Embeddings

SkipGram

Almost the reverse model



Recurrent Neural Networks



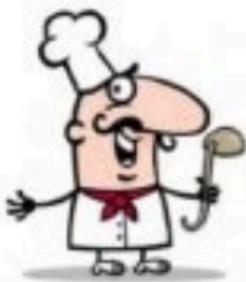
Recurrent Neural Networks



Recurrent Neural Networks

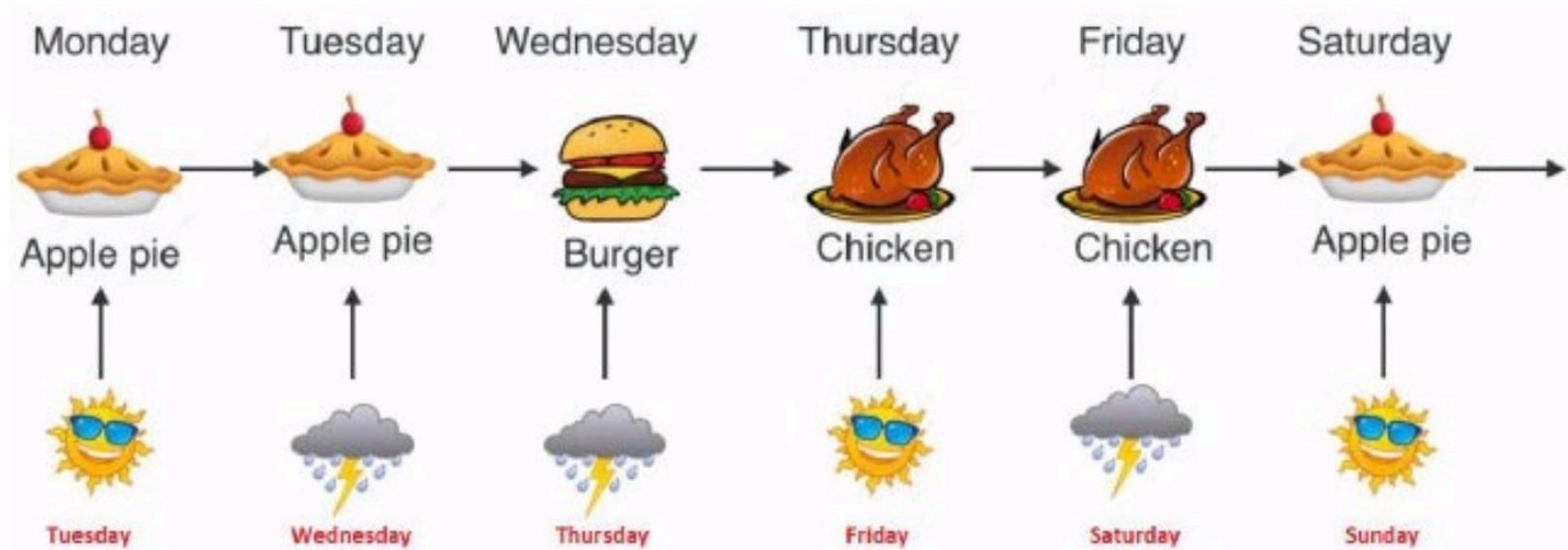


Sunny
Same as yesterday



Rain
Next dish

Recurrent Neural Networks



Recurrent Neural Networks



$$\begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}$$



$$\begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}$$



$$\begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}$$



$$\begin{bmatrix} 1 \\ 0 \end{bmatrix}$$



$$\begin{bmatrix} 0 \\ 1 \end{bmatrix}$$

Recurrent Neural Networks

$$\begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}$$
A pie emoji with a cherry on top.

$$\begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}$$
A hamburger emoji with lettuce, cheese, and meat.

$$\begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}$$
A roasted chicken emoji on a plate.

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ \hline 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} =$$

Food

$$\begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \\ 1 \\ 0 \end{bmatrix}$$
A pie emoji with a cherry on top.A hamburger emoji with lettuce, cheese, and meat.

Same

A hamburger emoji with lettuce, cheese, and meat.

Next day

Recurrent Neural Networks

$$\begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} \text{Pie} \quad \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ \hline 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \\ 0 \\ 1 \end{bmatrix} \quad \begin{array}{l} \text{Same} \\ \text{Next day} \end{array}$$

$\begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \\ 1 \end{bmatrix} \text{Food}$

Recurrent Neural Networks

$$\begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} \text{Pie} \quad \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ \hline 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 1 \\ 1 \\ 0 \\ 0 \end{bmatrix} \quad \begin{array}{l} \text{Same} \\ \text{Next day} \end{array}$$

$\begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \\ 1 \end{bmatrix} \text{Food}$

Recurrent Neural Networks

Weather

Same

Next day

Recurrent Neural Networks

Weather

Same

Next day

Recurrent Neural Networks

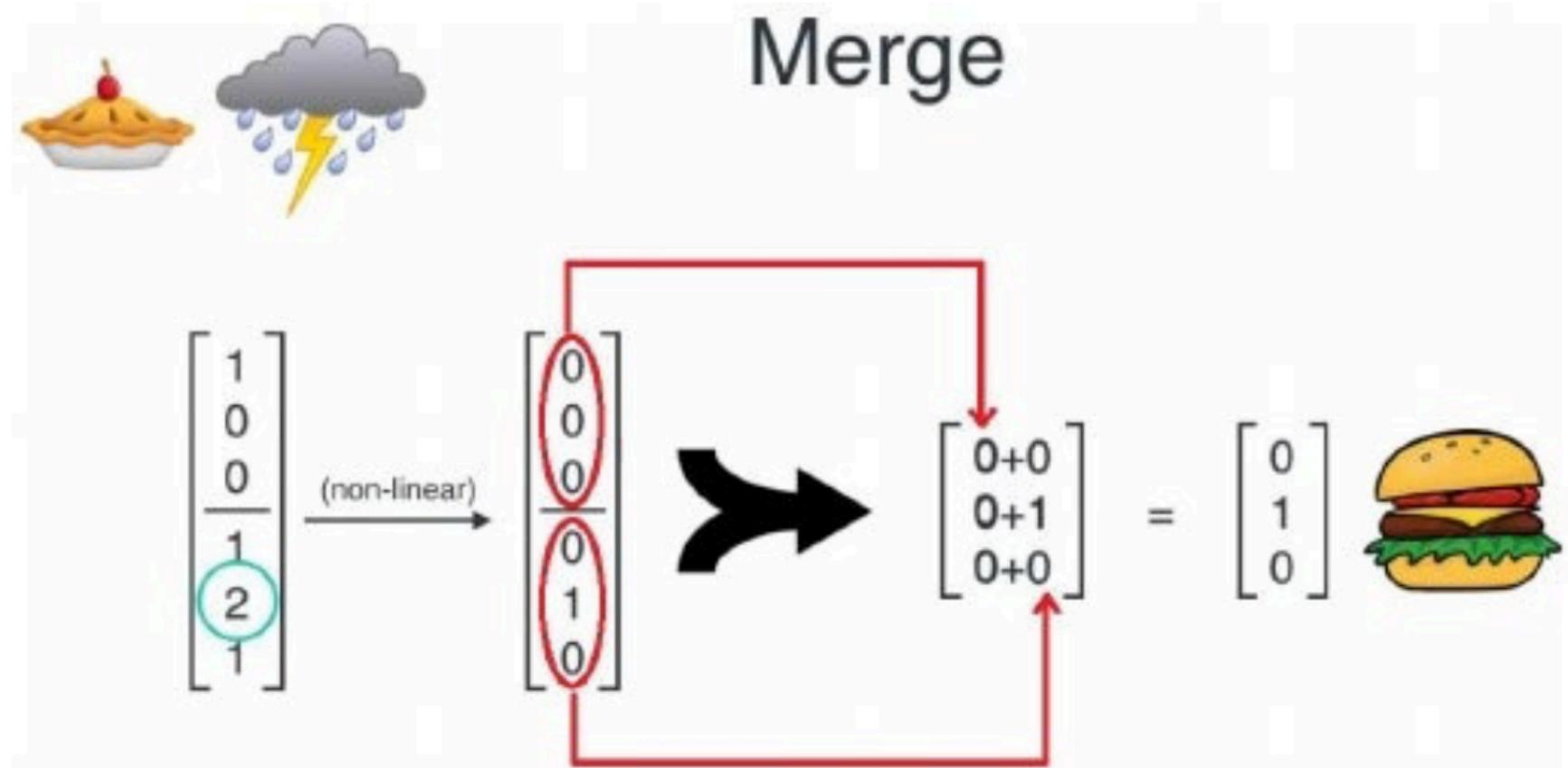
Same

Next day

Same

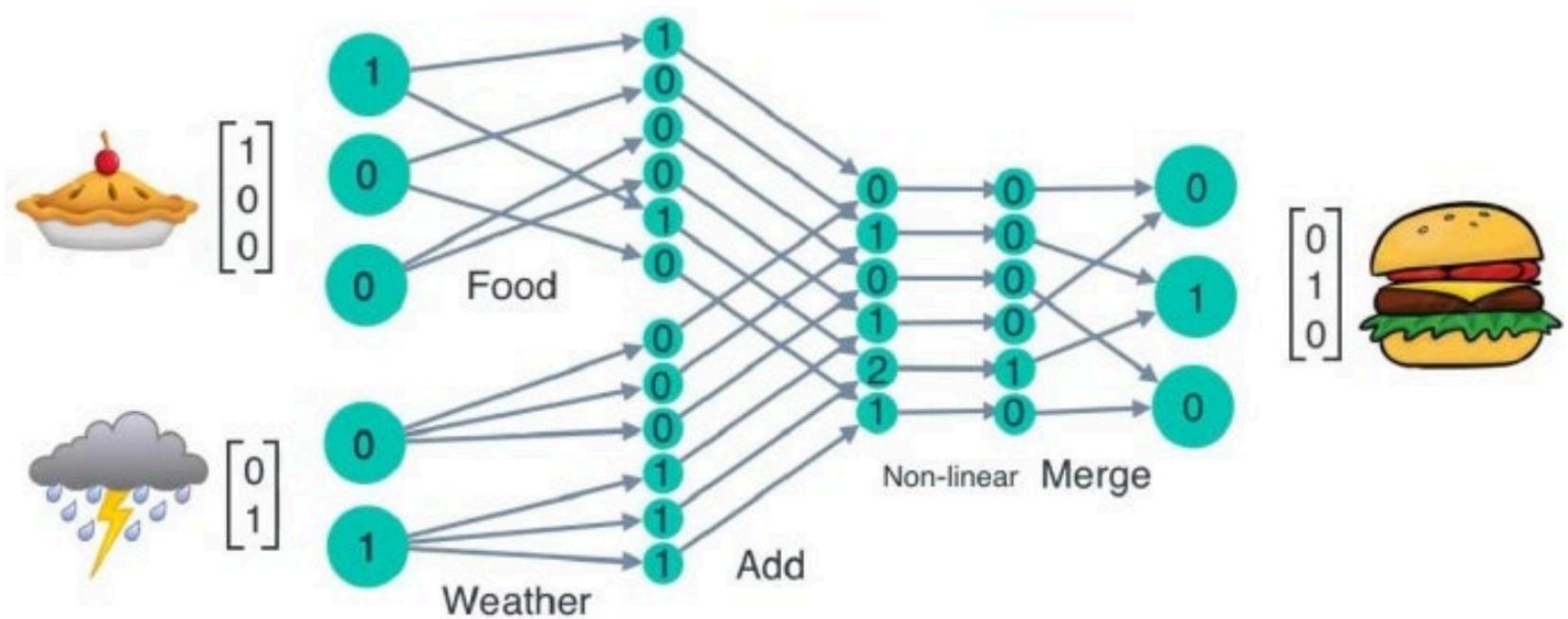
Next day

Recurrent Neural Networks



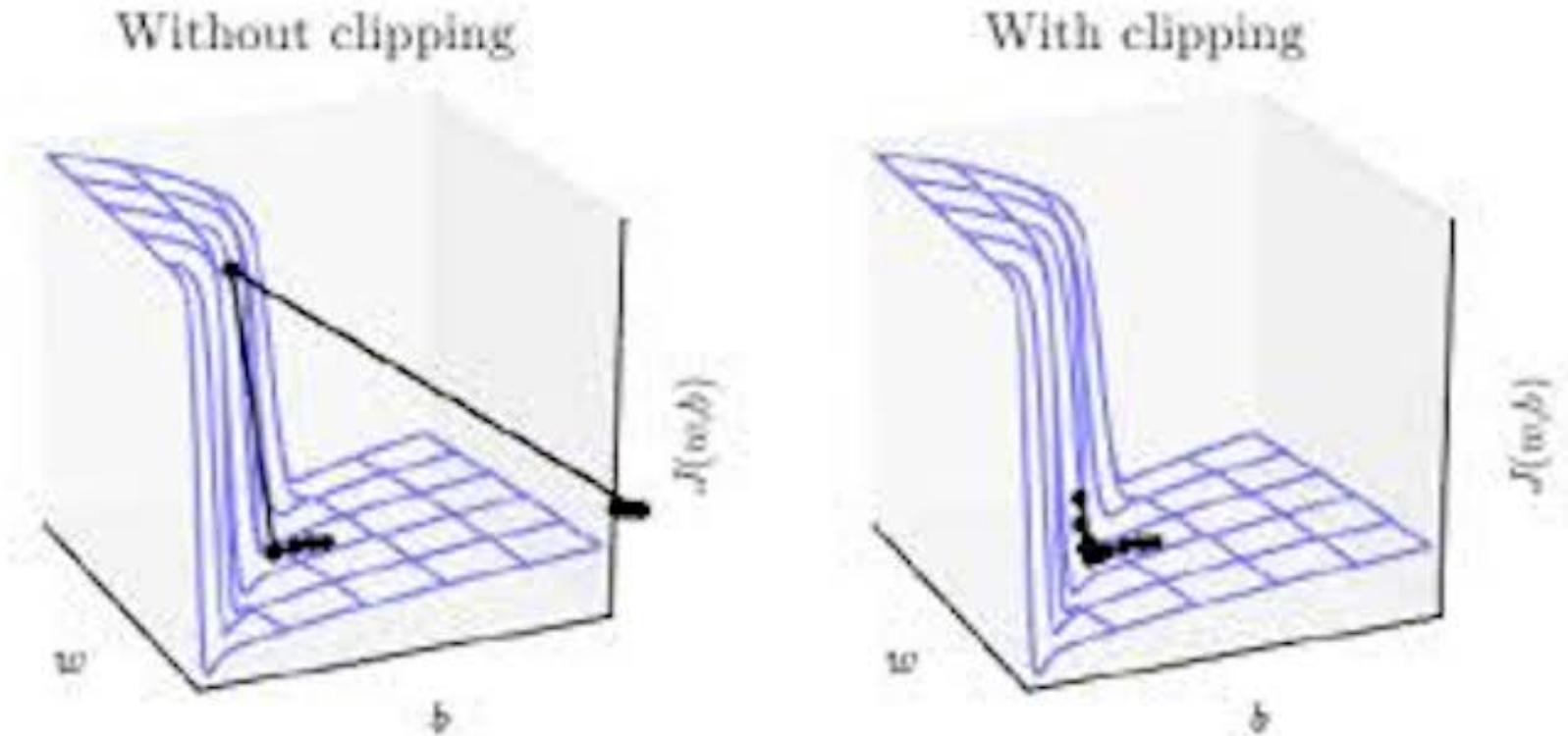
Recurrent Neural Networks

Food output is reinjected in the network as a state vector

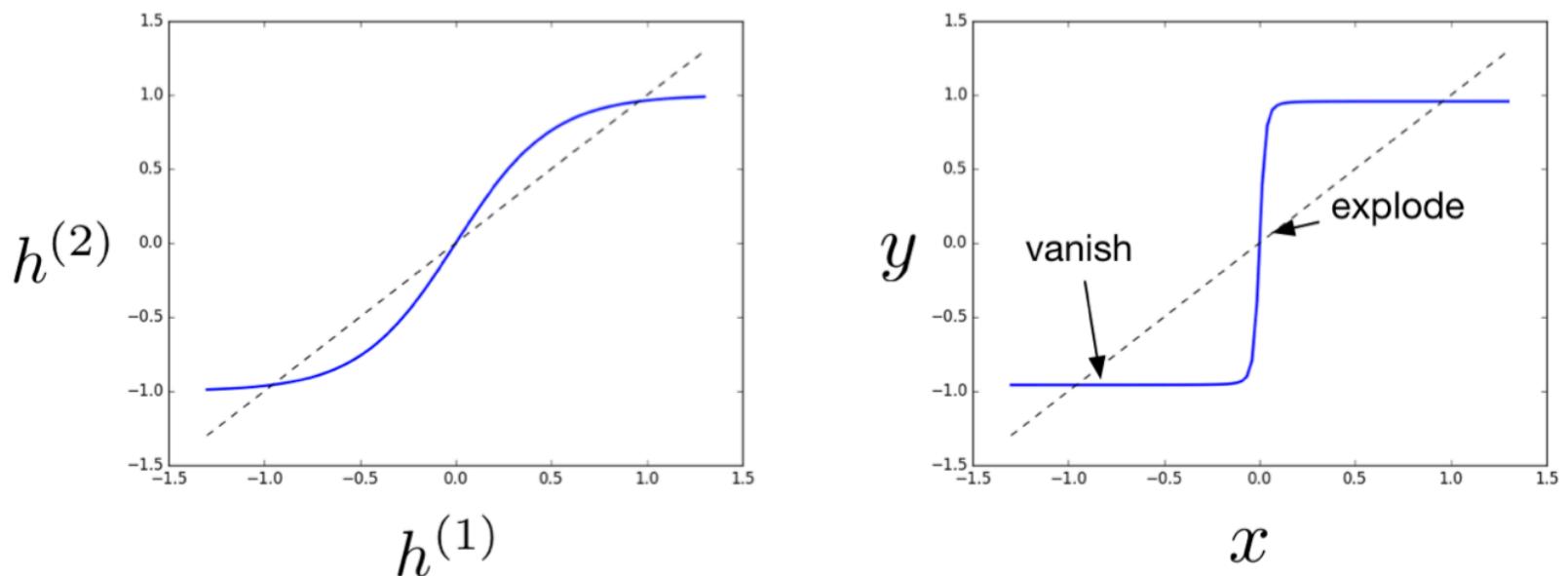


These slides on the perfect roommate are borrowed from this [Luis Serrano's Youtube video](https://www.youtube.com/watch?v=BR9h47Jtqyw)
(<https://www.youtube.com/watch?v=BR9h47Jtqyw>)

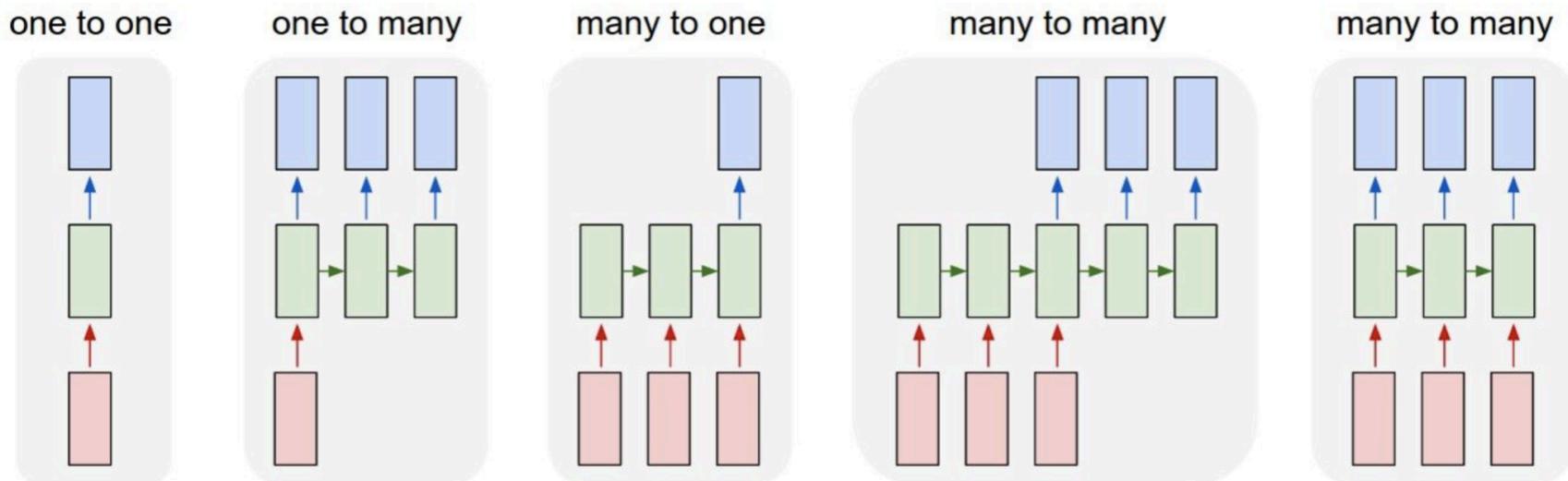
Vanishing and exploding gradient



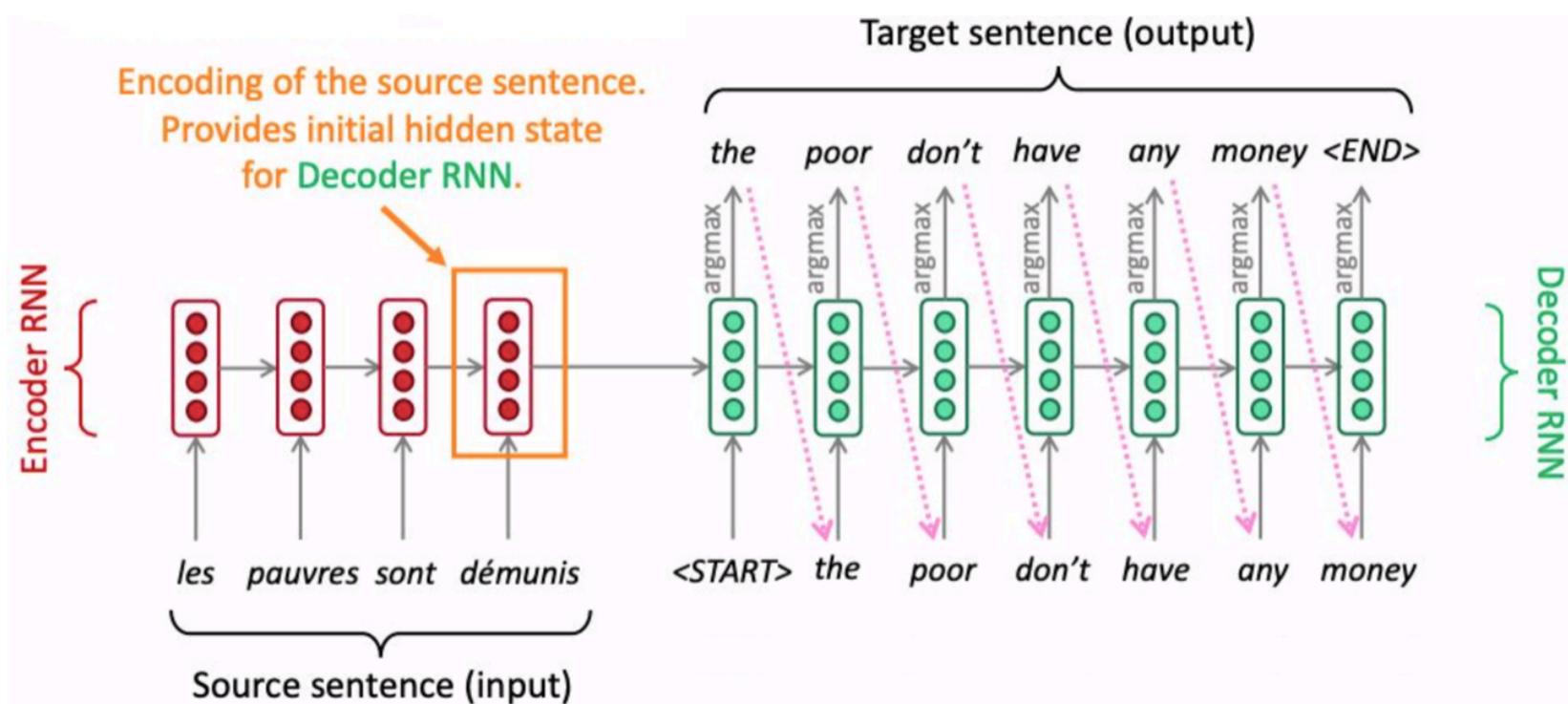
Vanishing and exploding gradient



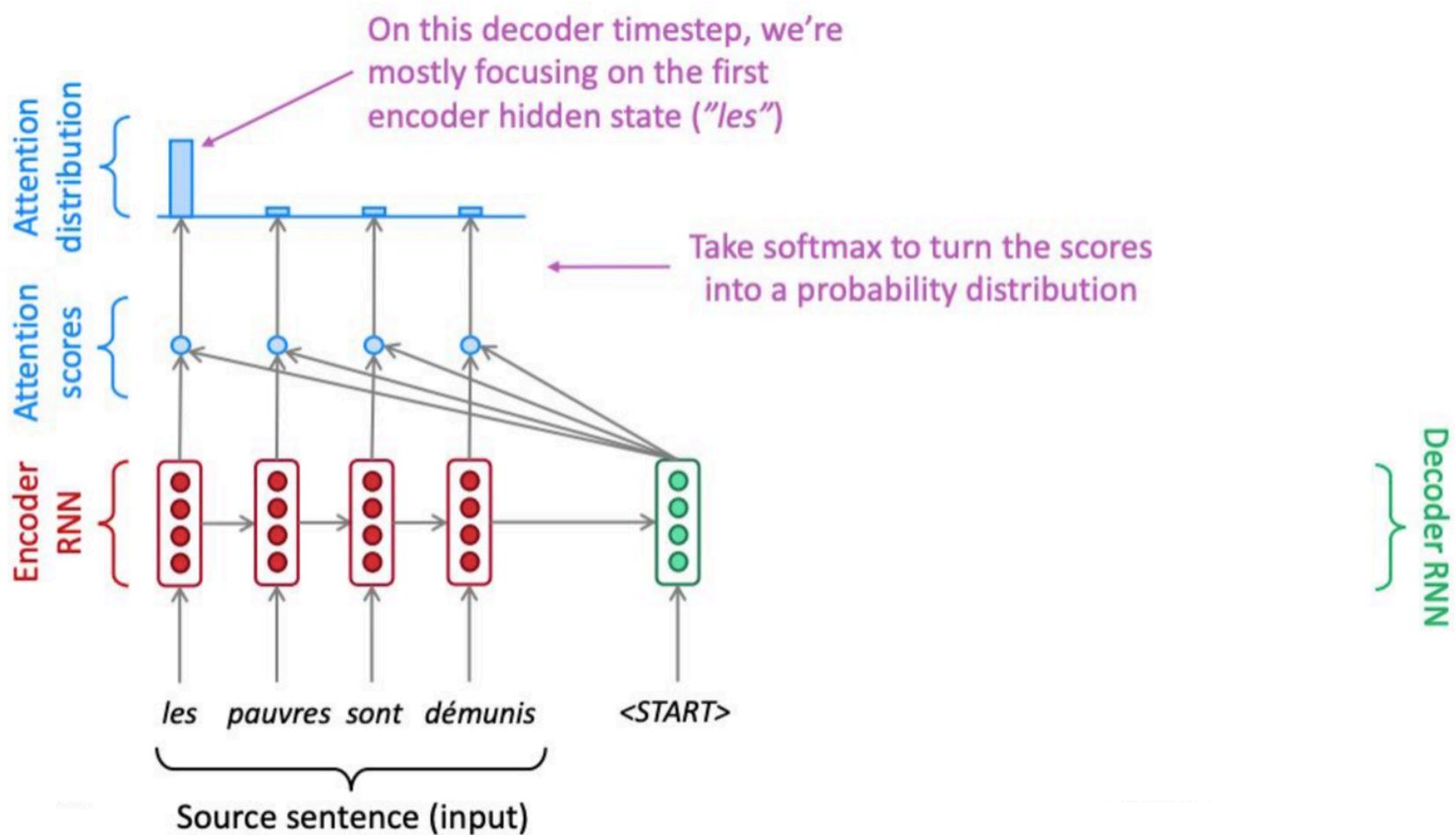
Different types of RNN



Attention is all you need (the origin)

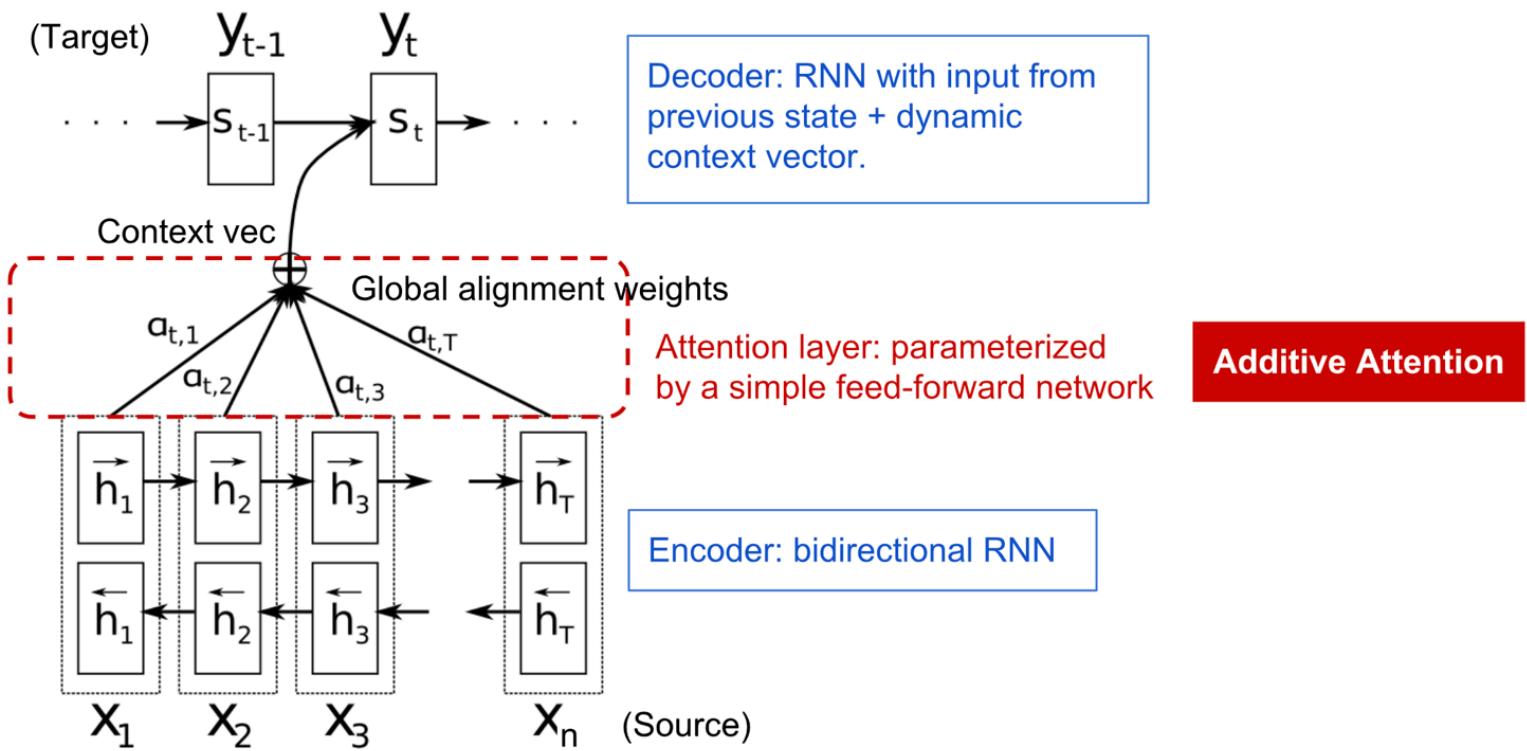


Attention is all you need (the origin)

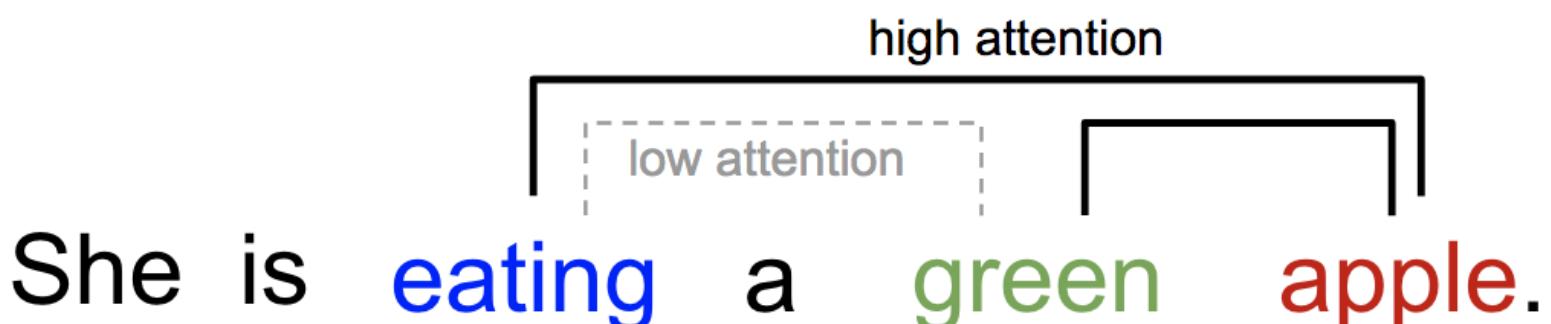


These slides come from this [Ömer Sakarya paper](https://towardsdatascience.com/introduction-to-rnns-sequence-to-sequence-language-translation-and-attention-fc43ef2cc3fd) (<https://towardsdatascience.com/introduction-to-rnns-sequence-to-sequence-language-translation-and-attention-fc43ef2cc3fd>)

Attention is all you need (the origin)



Attention is all you need (the origin)



Attention is all you need (the origin)

The FBI is chasing a criminal on the run .

The FBI is chasing a criminal on the run .

The FBI is chasing a criminal on the run .

The FBI is chasing a criminal on the run .

The FBI is chasing a criminal on the run .

The FBI is chasing a criminal on the run .

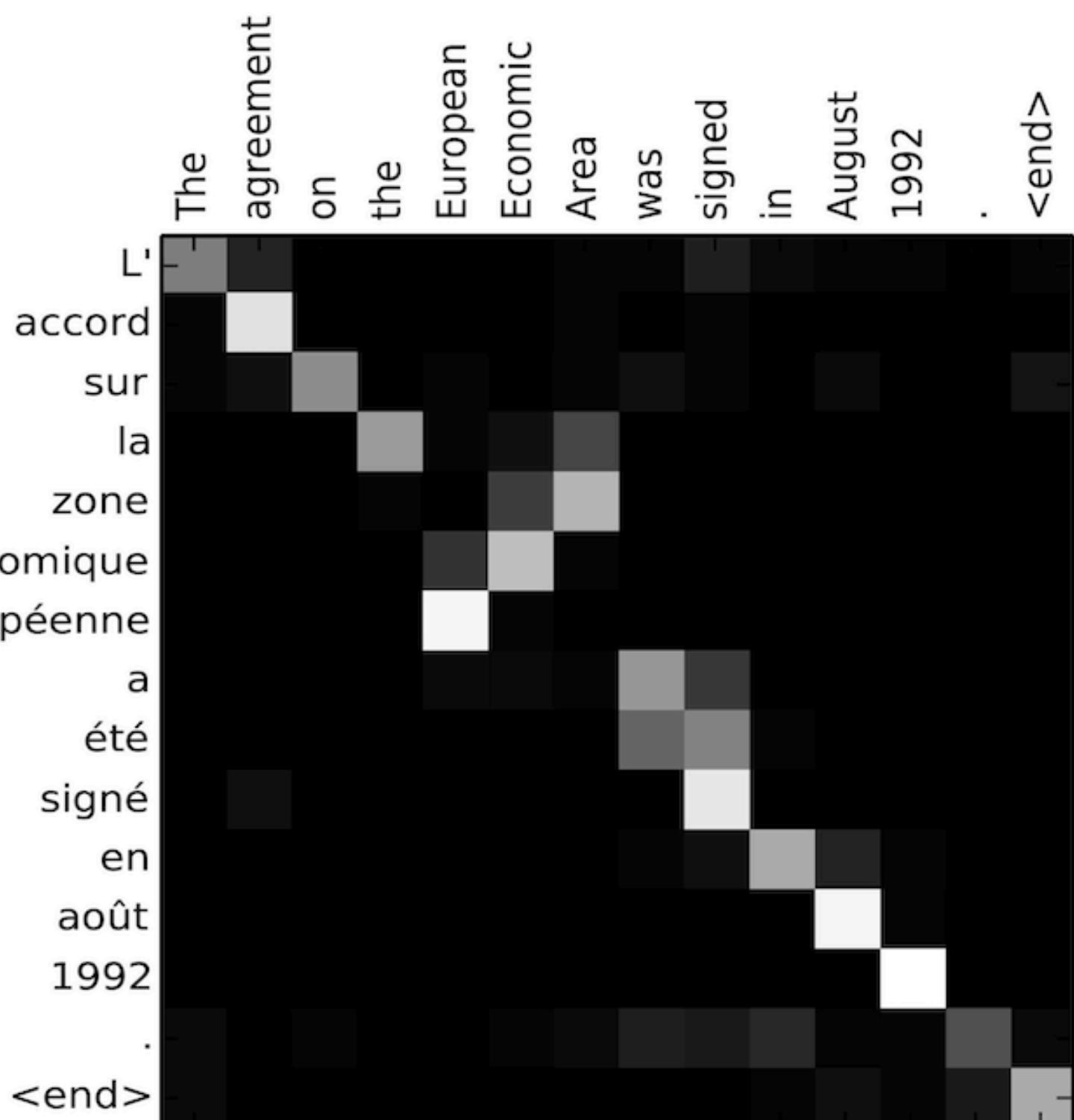
The FBI is chasing a criminal on the run .

The FBI is chasing a criminal on the run .

The FBI is chasing a criminal on the run .

The FBI is chasing a criminal on the run .

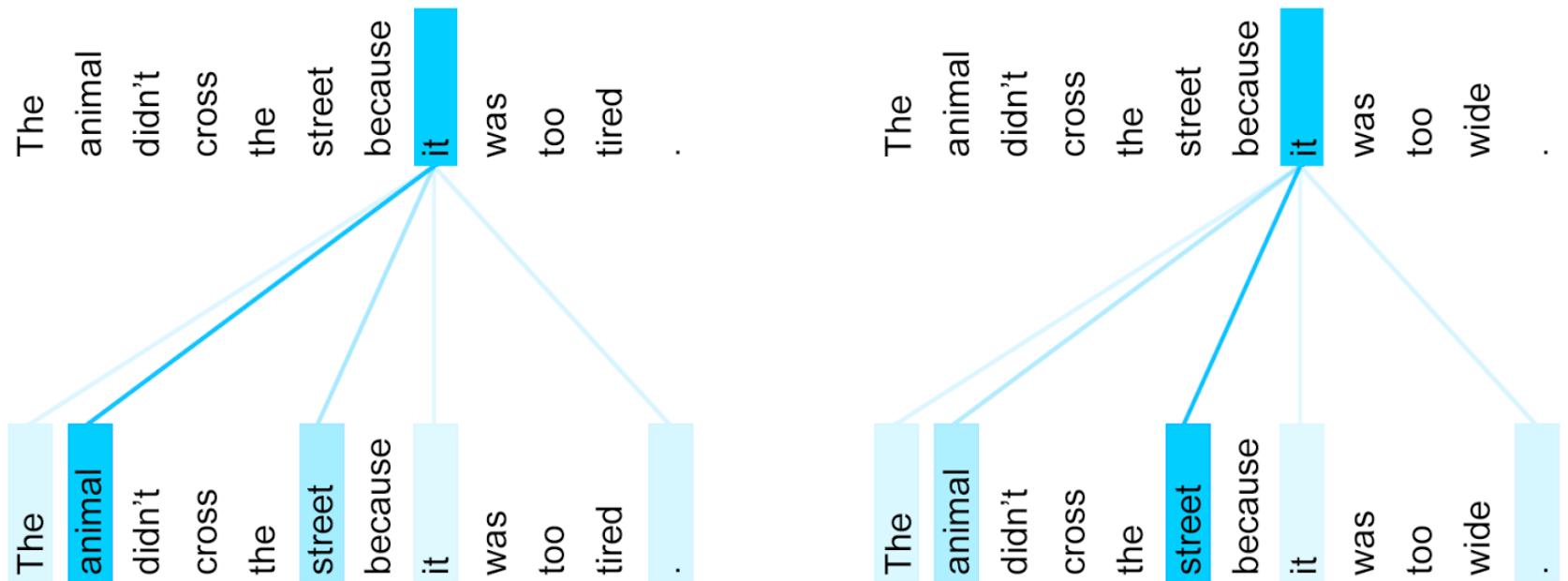
Attention is all you need (the origin)



These last eight slides are borrowed from this [Lilian Weng's article](https://lilianweng.github.io/lil-log/2018/06/24/attention-attention.html) (<https://lilianweng.github.io/lil-log/2018/06/24/attention-attention.html>)

Attention is all you need (the origin)

Attention is all you need (the origin)

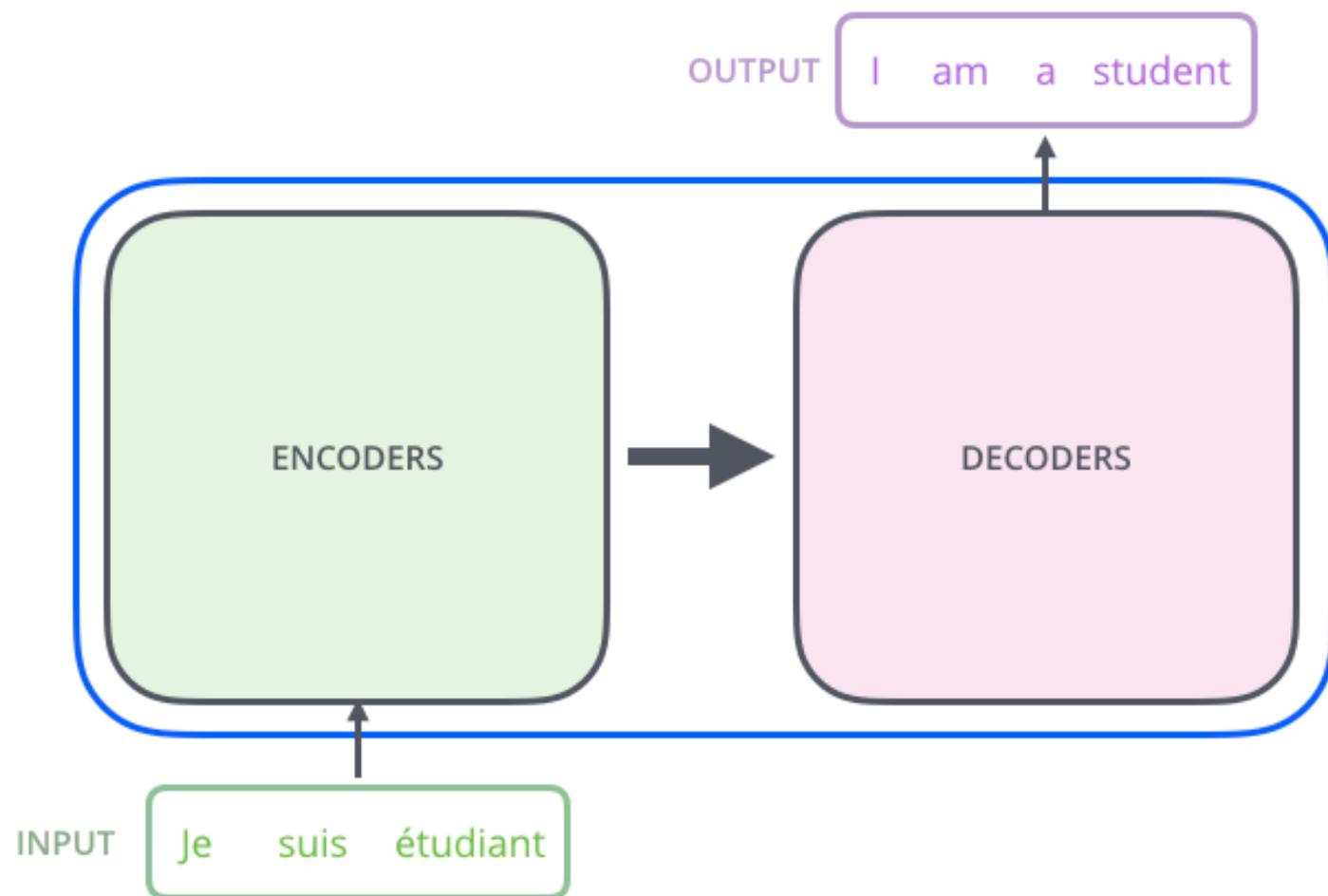


These last two slides are borrowed from this [Google AI's blog](https://ai.googleblog.com/2017/08/transformer-novel-neural-network.html)
(<https://ai.googleblog.com/2017/08/transformer-novel-neural-network.html>)

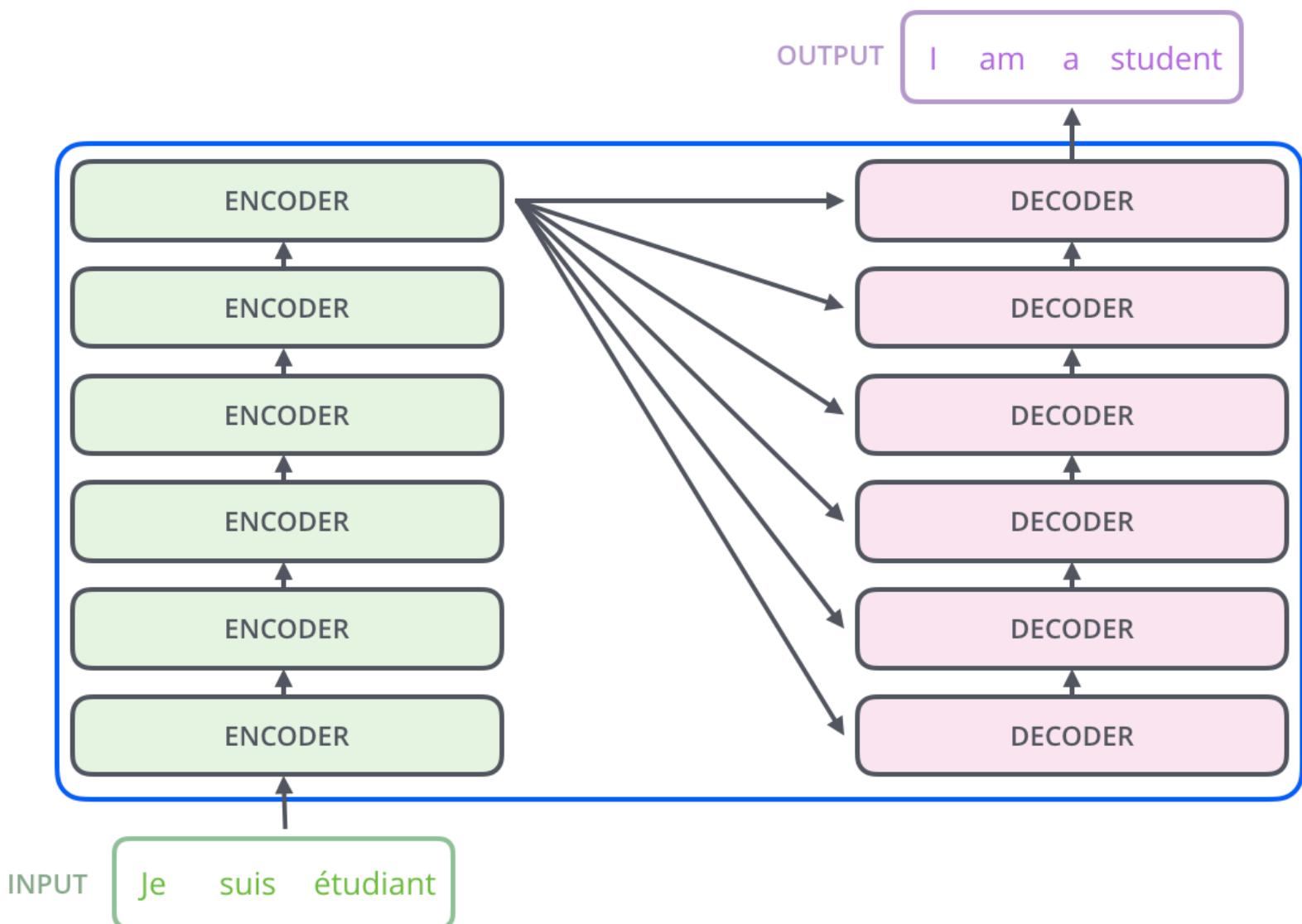
Attention is all you need (the mighty transformer)



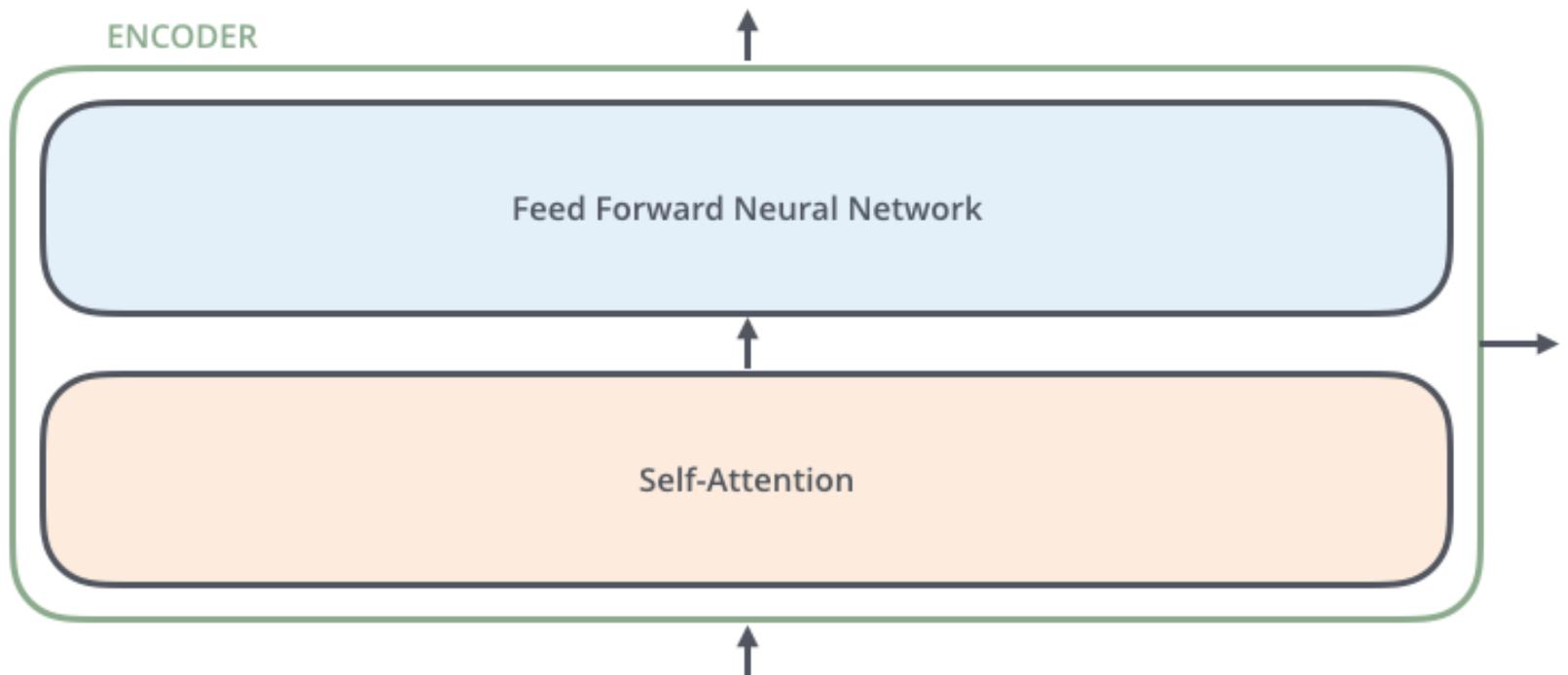
Attention is all you need (the mighty transformer)



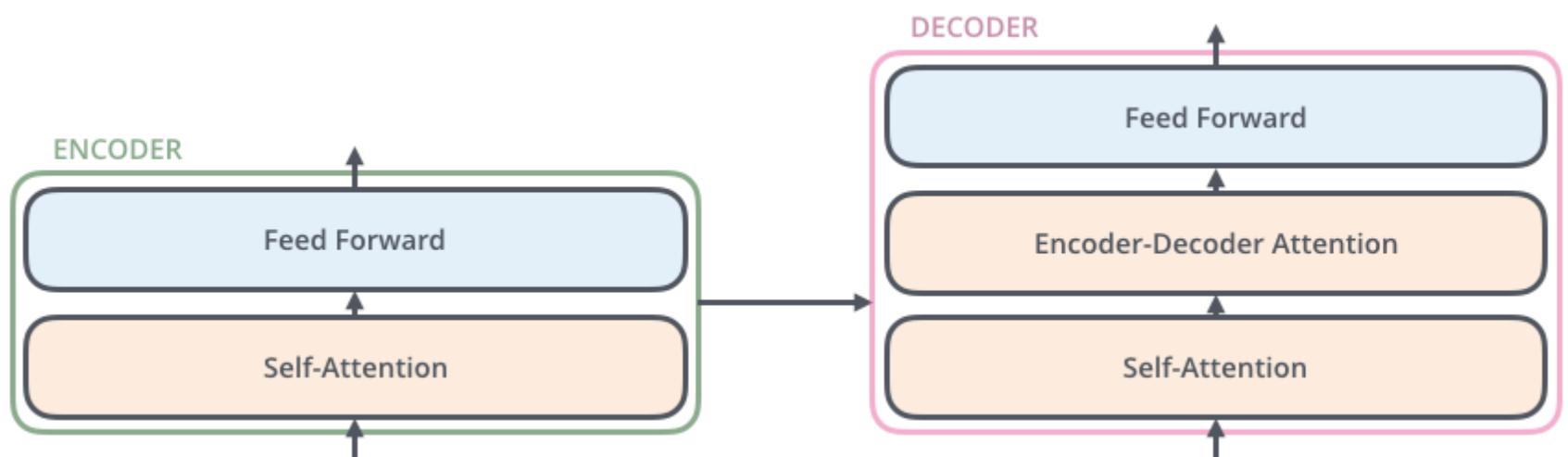
Attention is all you need (the mighty transformer)



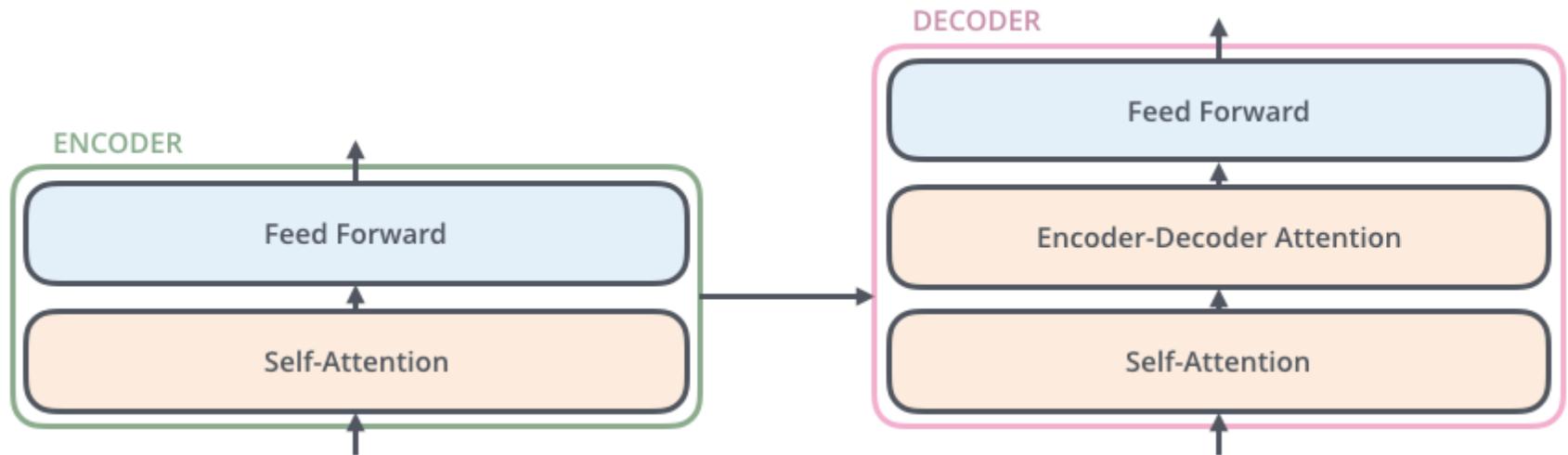
Attention is all you need (the mighty transformer)



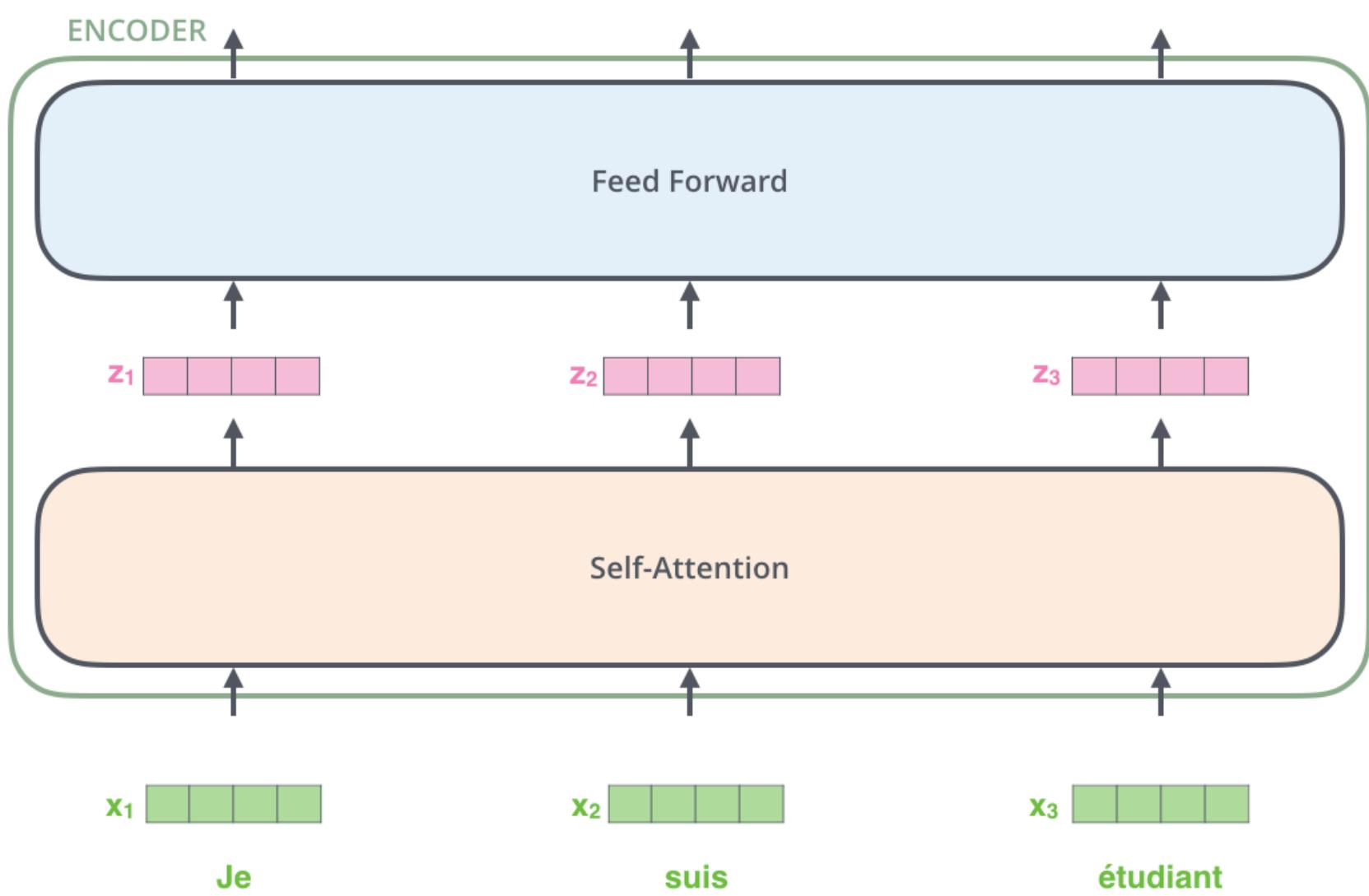
Attention is all you need (the mighty transformer)



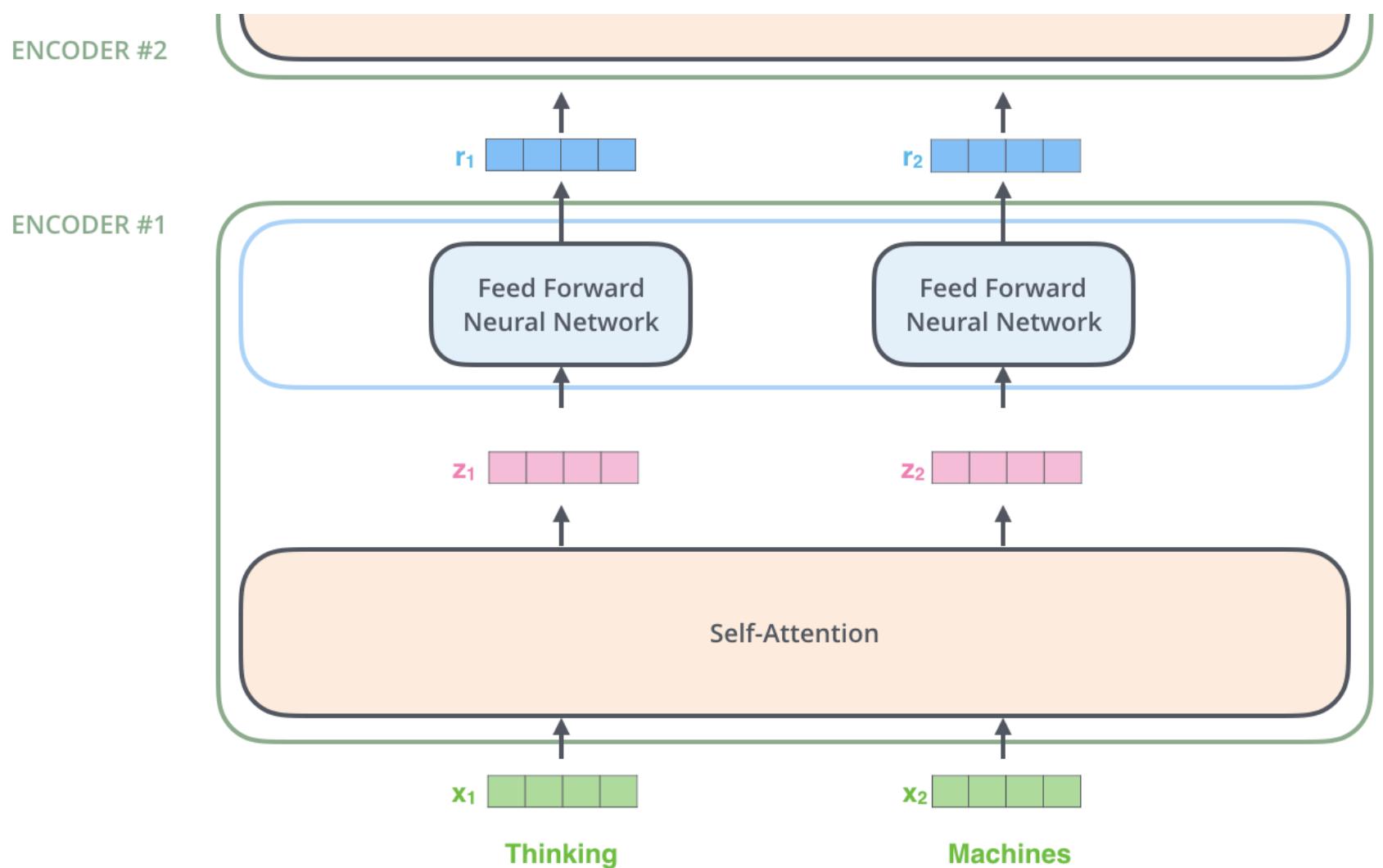
Attention is all you need (the mighty transformer)



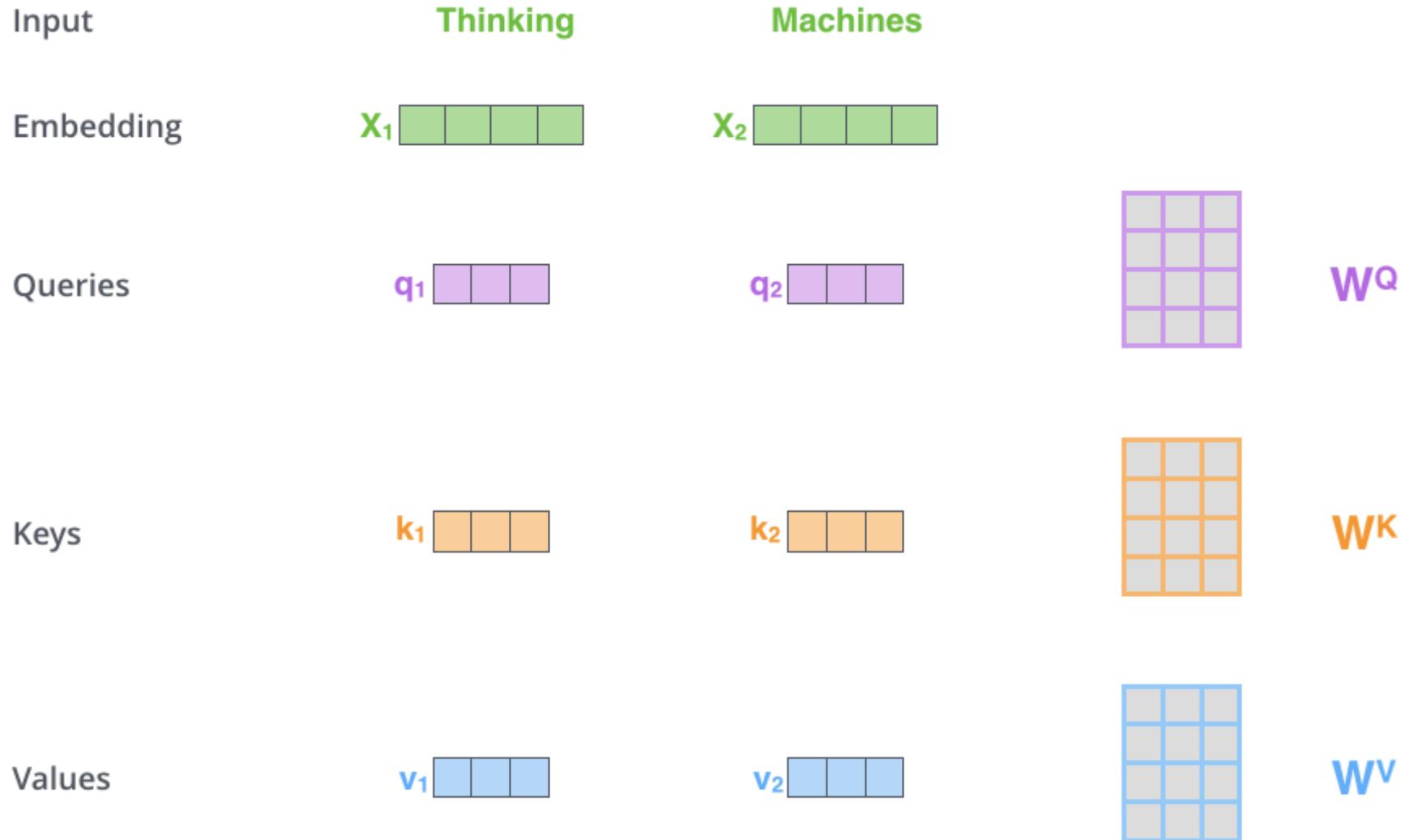
Attention is all you need (the mighty transformer)



Attention is all you need (the mighty transformer)



Attention is all you need (the mighty transformer)



Attention is all you need (the mighty transformer)

$$\mathbf{X} \times \mathbf{W}^Q = \mathbf{Q}$$

A diagram illustrating the computation of the Query matrix (\mathbf{Q}). It shows the input matrix \mathbf{X} (2 rows, 4 columns) multiplied by the weight matrix \mathbf{W}^Q (4 rows, 4 columns). The result is the Query matrix \mathbf{Q} (2 rows, 2 columns). The matrices are represented as grids of colored squares.

$$\mathbf{X} \times \mathbf{W}^K = \mathbf{K}$$

A diagram illustrating the computation of the Key matrix (\mathbf{K}). It shows the input matrix \mathbf{X} (2 rows, 4 columns) multiplied by the weight matrix \mathbf{W}^K (4 rows, 4 columns). The result is the Key matrix \mathbf{K} (2 rows, 2 columns). The matrices are represented as grids of colored squares.

$$\mathbf{X} \times \mathbf{W}^V = \mathbf{V}$$

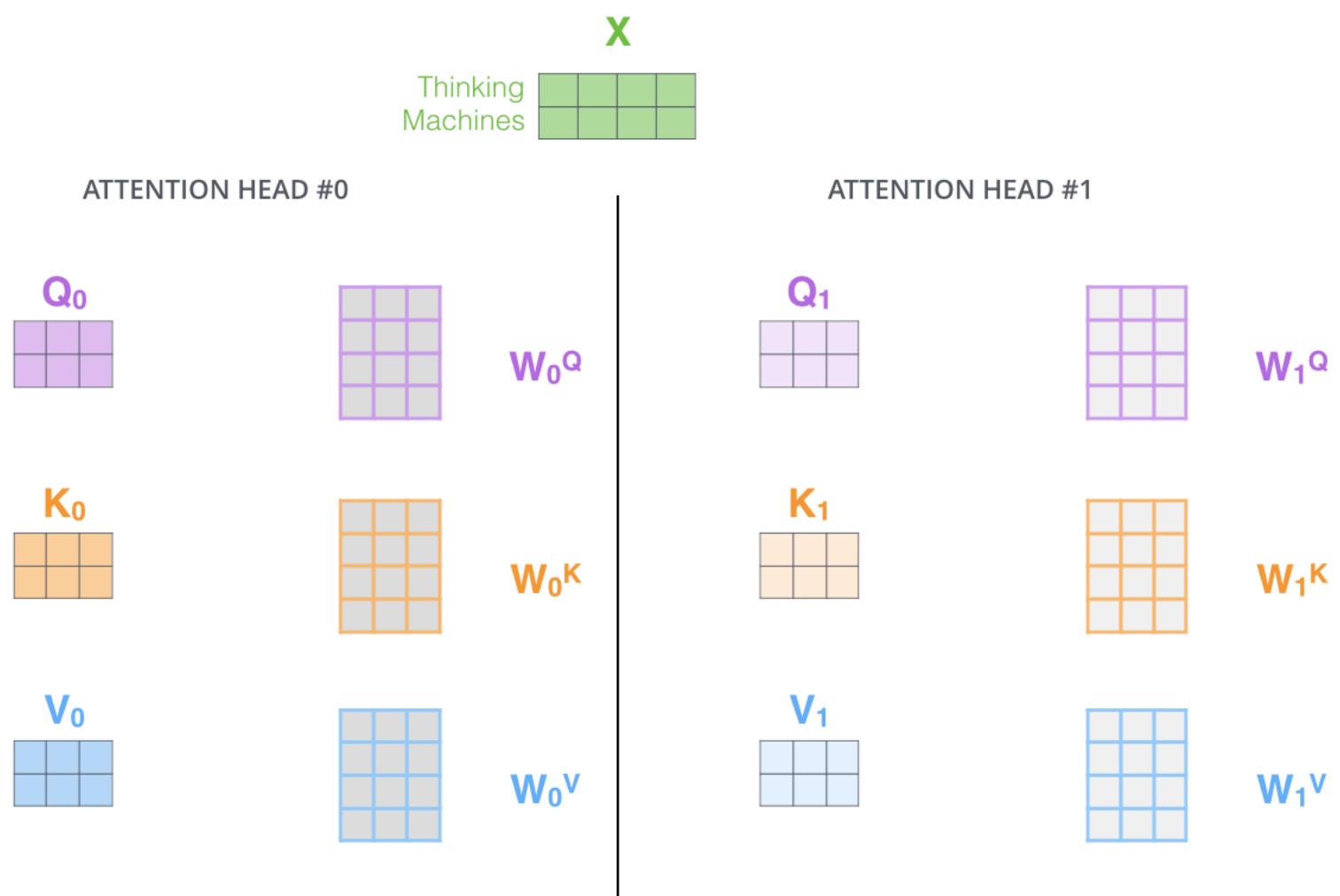
A diagram illustrating the computation of the Value matrix (\mathbf{V}). It shows the input matrix \mathbf{X} (2 rows, 4 columns) multiplied by the weight matrix \mathbf{W}^V (4 rows, 4 columns). The result is the Value matrix \mathbf{V} (2 rows, 2 columns). The matrices are represented as grids of colored squares.

Attention is all you need (the mighty transformer)

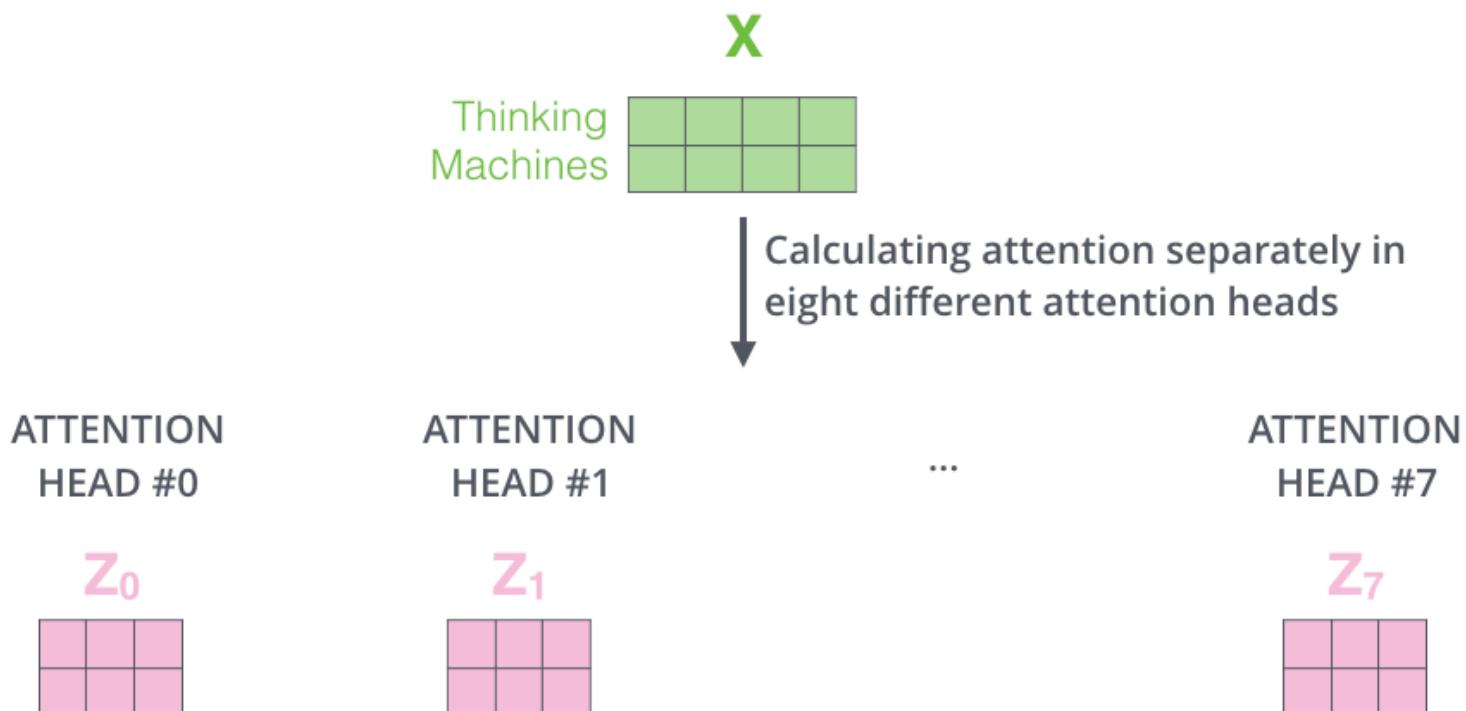
$$\text{softmax} \left(\frac{\begin{matrix} \mathbf{Q} & \mathbf{K}^T \\ \begin{matrix} \text{---} & \times \\ \hline \end{matrix} & \begin{matrix} \mathbf{V} \\ \hline \end{matrix} \end{matrix}}{\sqrt{d_k}} \right) = \mathbf{Z}$$

The diagram illustrates the computation of attention scores. It shows the multiplication of query matrix \mathbf{Q} (purple) and key matrix \mathbf{K}^T (orange) divided by $\sqrt{d_k}$, followed by a softmax operation to produce the attention probability matrix \mathbf{Z} (pink).

Attention is all you need (the mighty transformer)

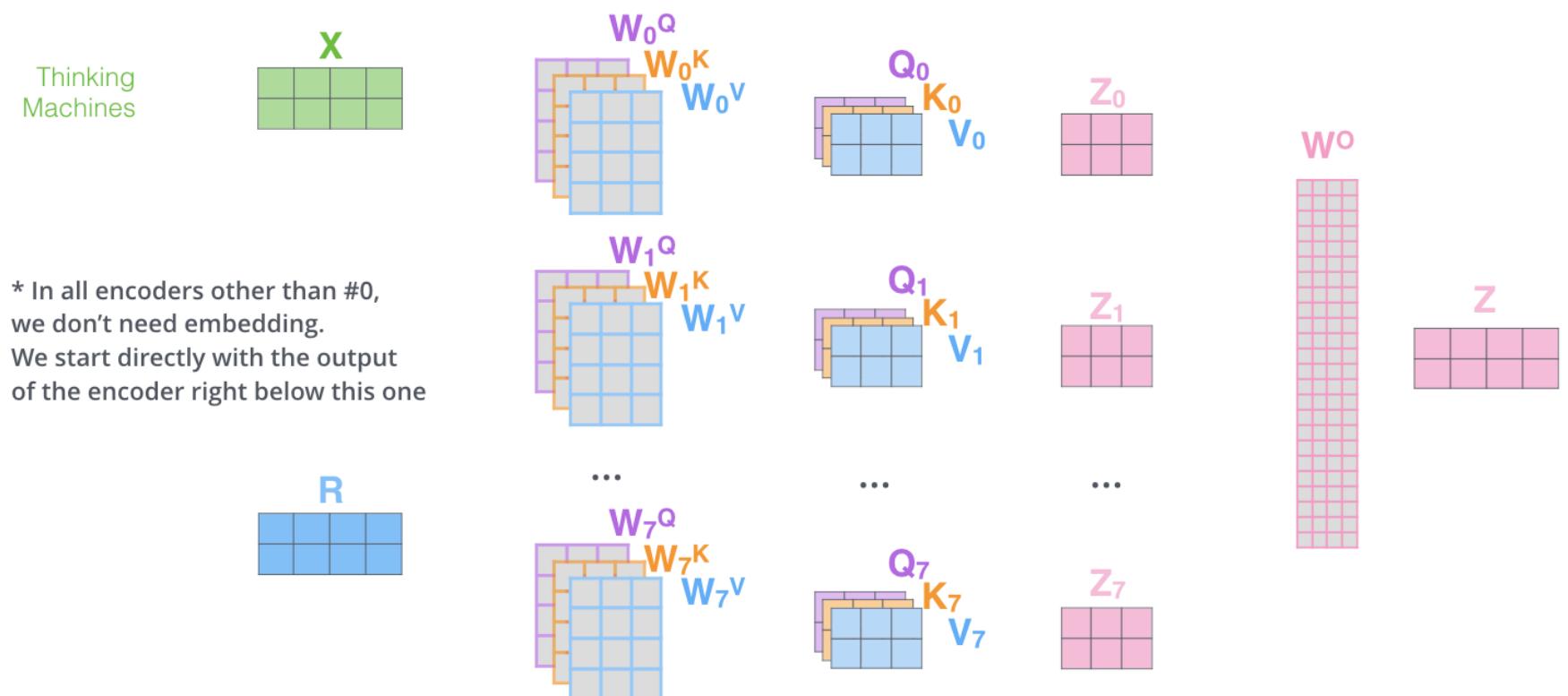


Attention is all you need (the mighty transformer)

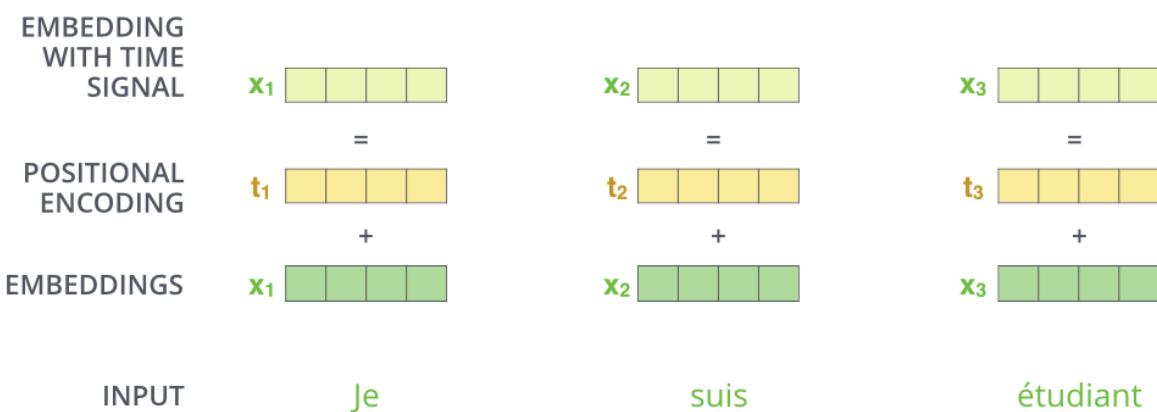
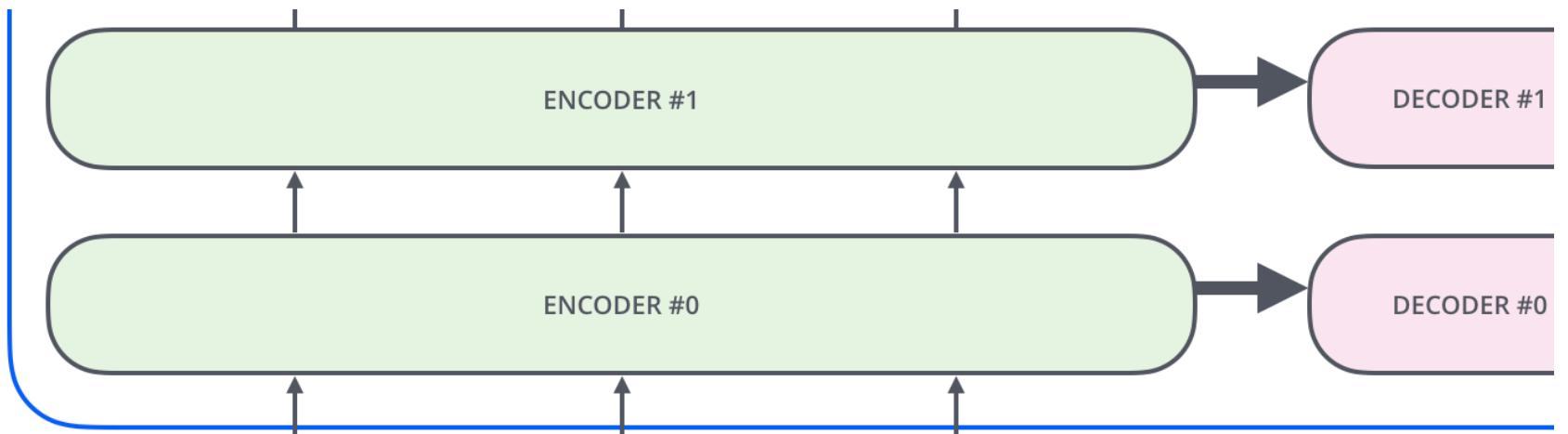


Attention is all you need (the mighty transformer)

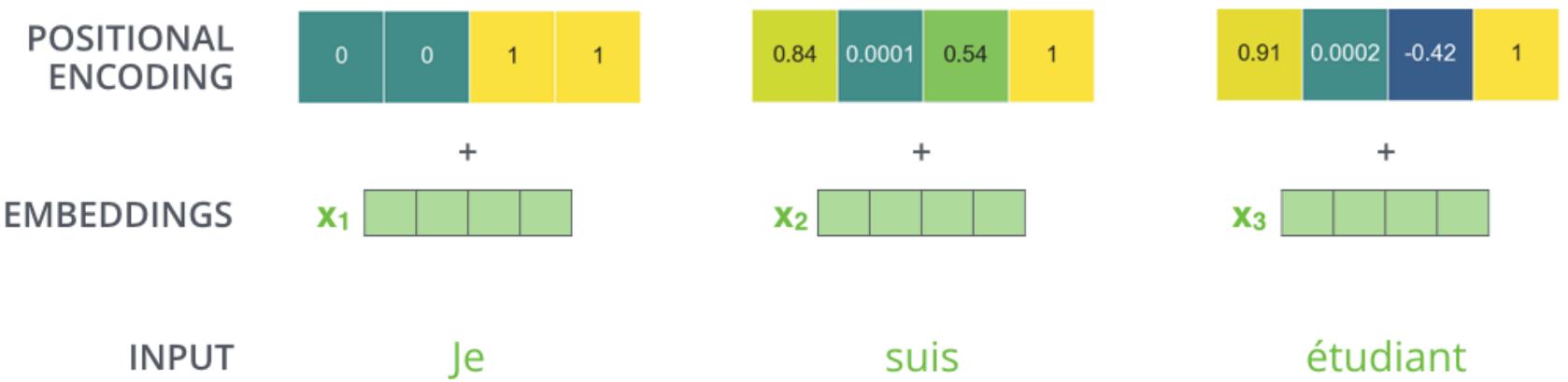
- 1) This is our input sentence* X
- 2) We embed each word*
- 3) Split into 8 heads. We multiply X or R with weight matrices
- 4) Calculate attention using the resulting $Q/K/V$ matrices
- 5) Concatenate the resulting Z matrices, then multiply with weight matrix W^O to produce the output of the layer



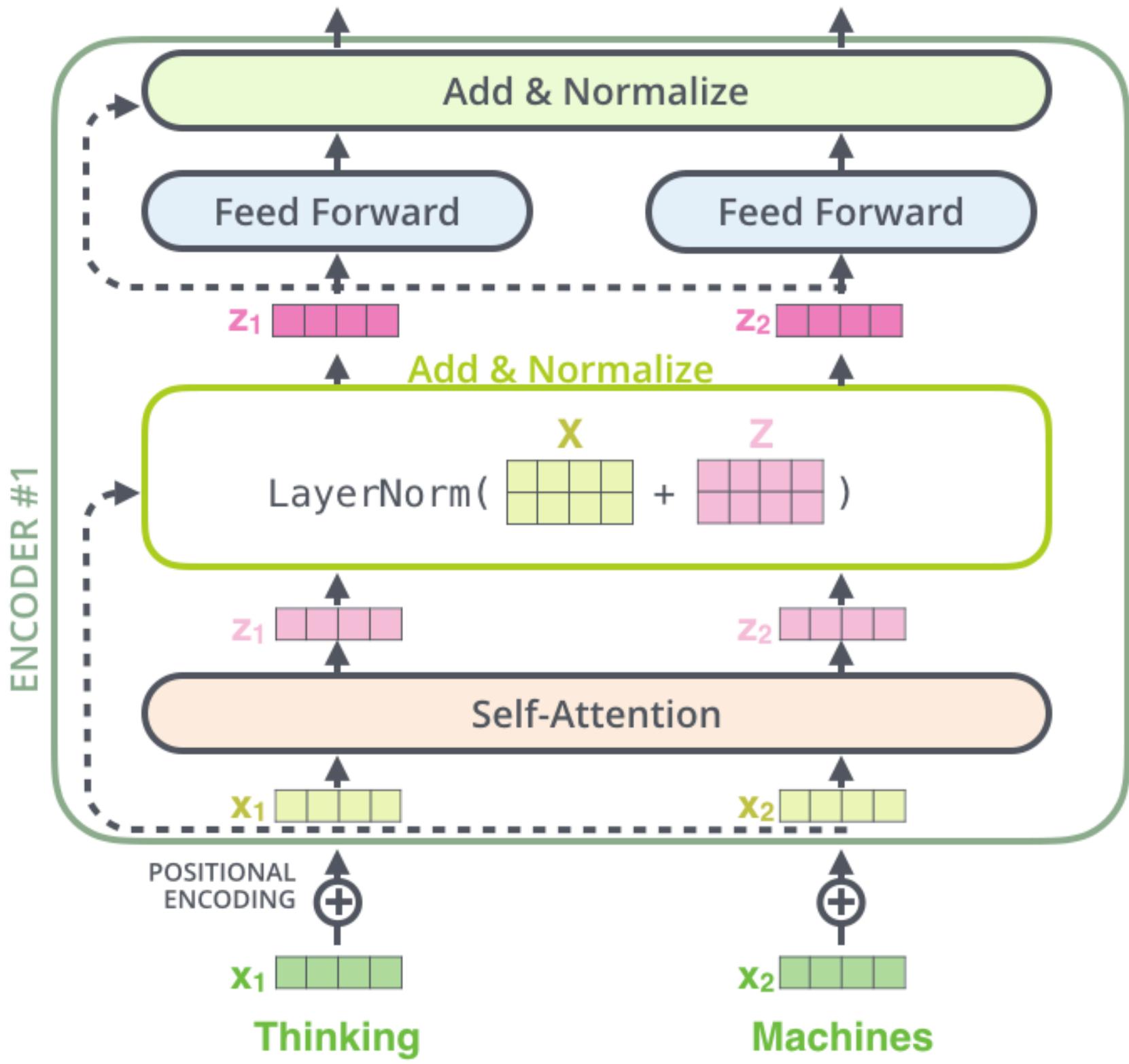
Attention is all you need (the mighty transformer)



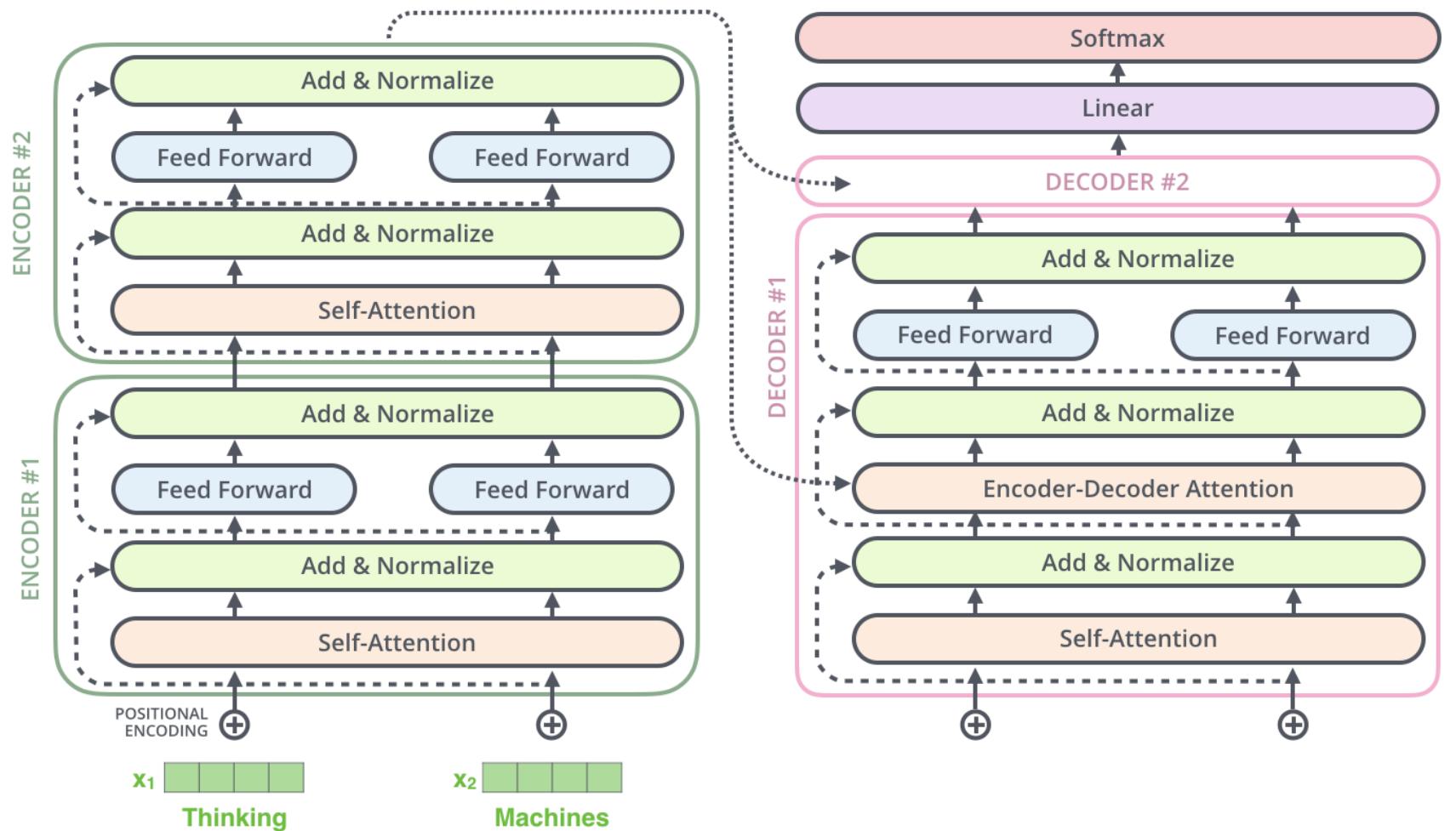
Attention is all you need (the mighty transformer)



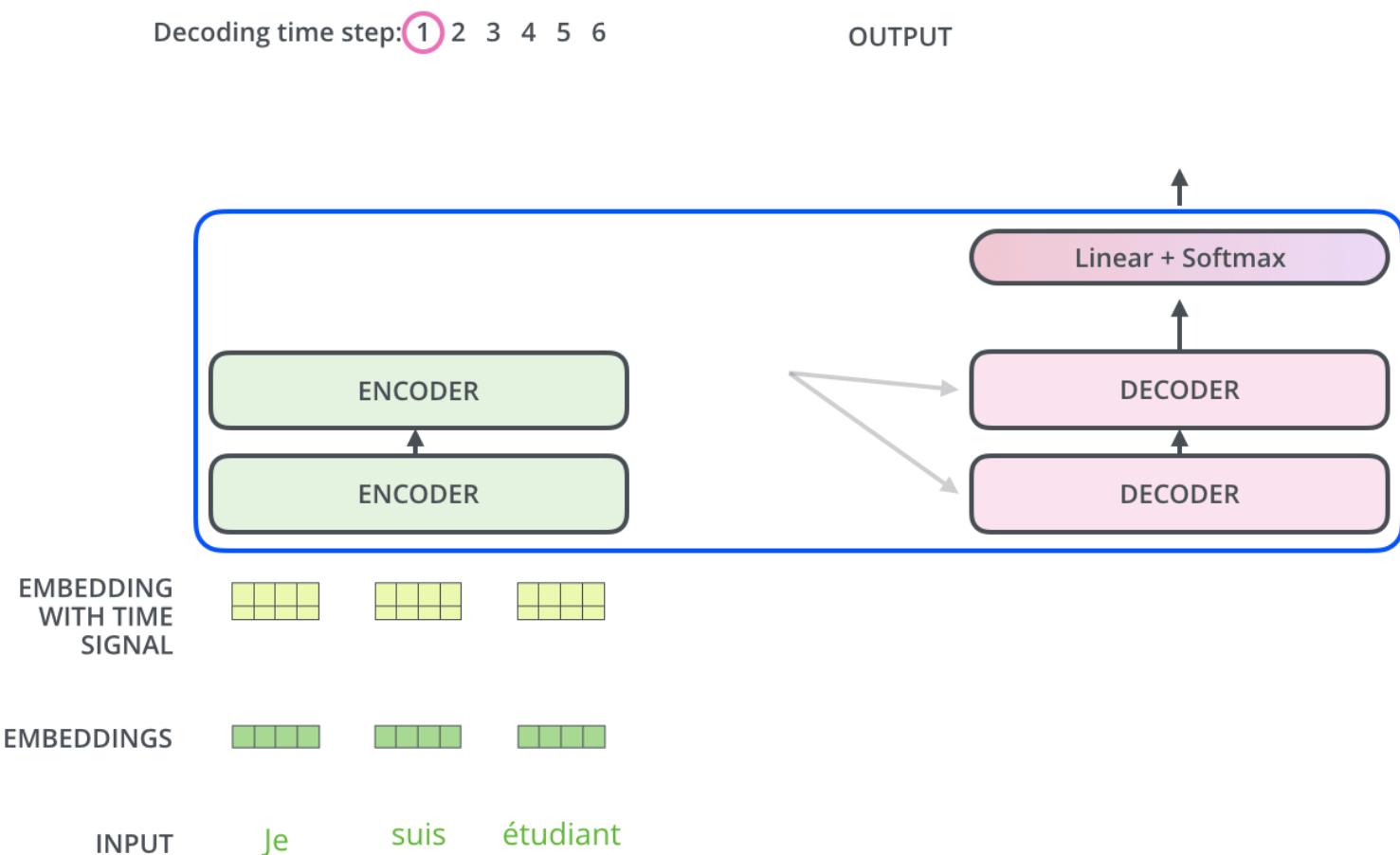
Attention is all you need (the mighty transformer)



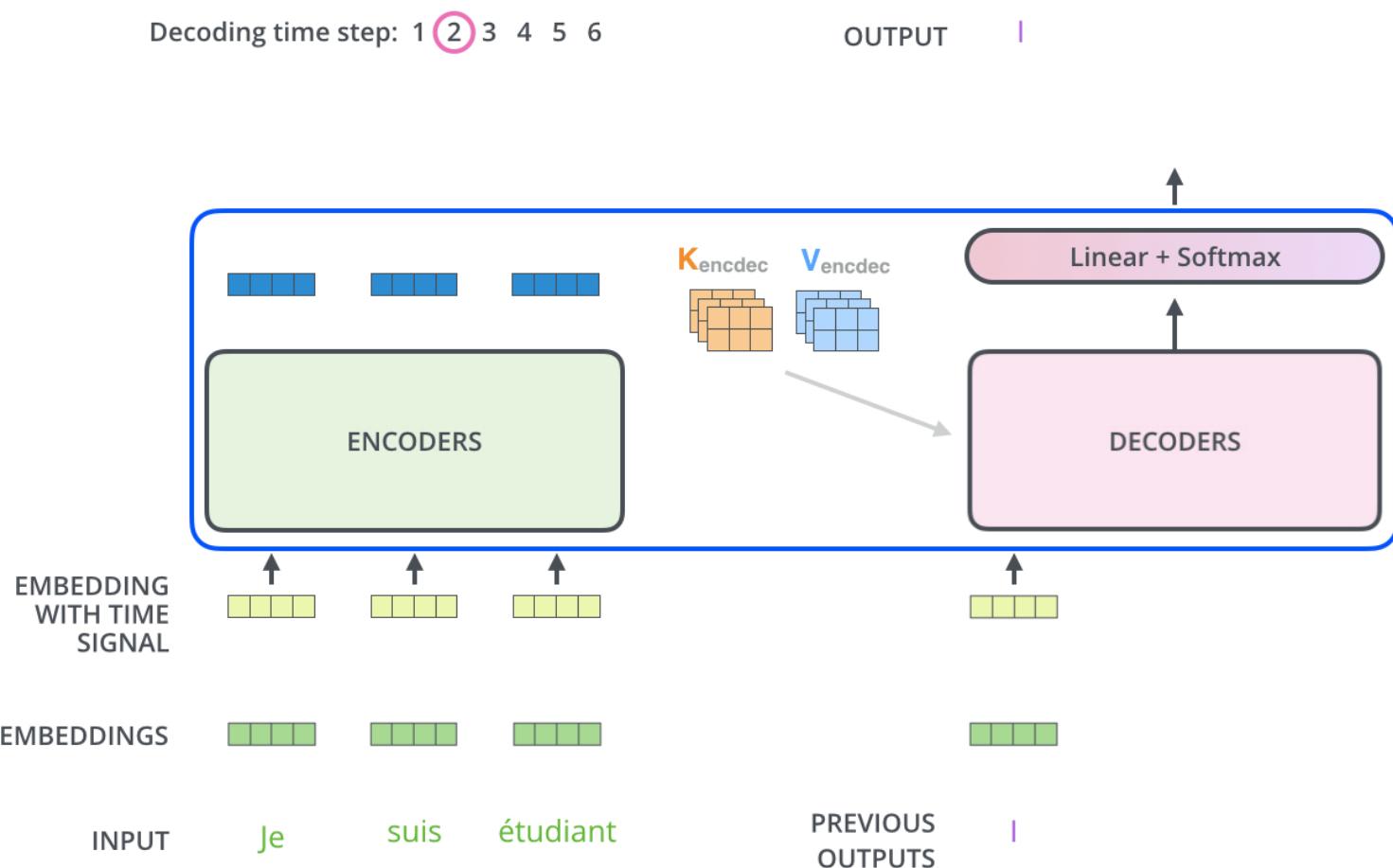
Attention is all you need (the mighty transformer)



Attention is all you need (the mighty transformer)



Attention is all you need (the mighty transformer)



Attention is all you need (the mighty transformer)

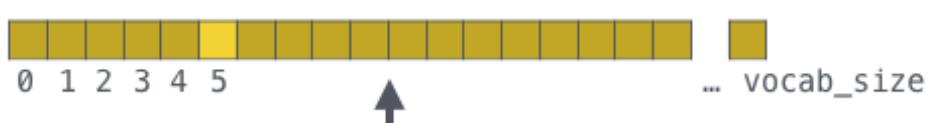
Which word in our vocabulary
is associated with this index?

am

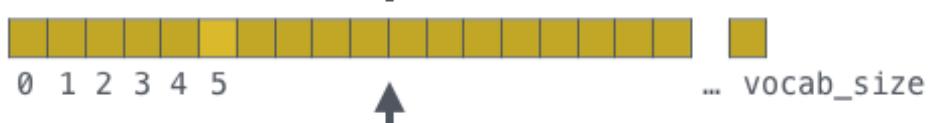
Get the index of the cell
with the highest value
(**argmax**)

5

log_probs



logits



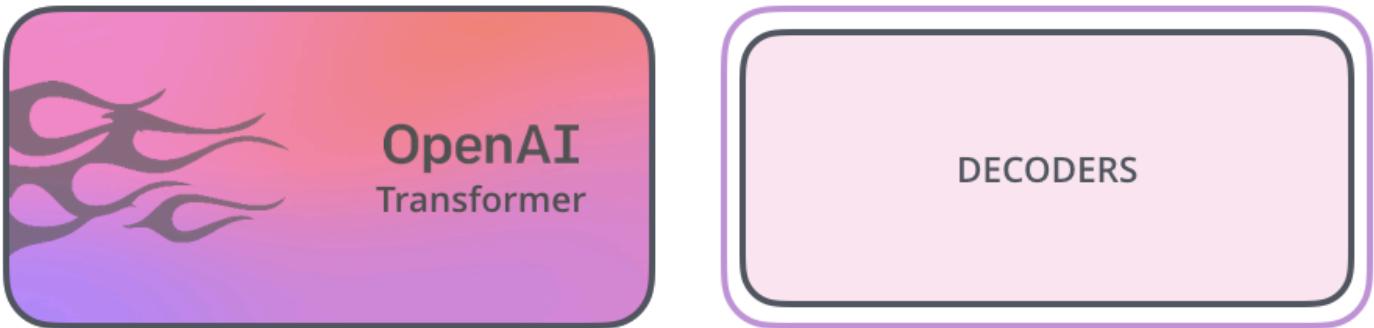
Linear



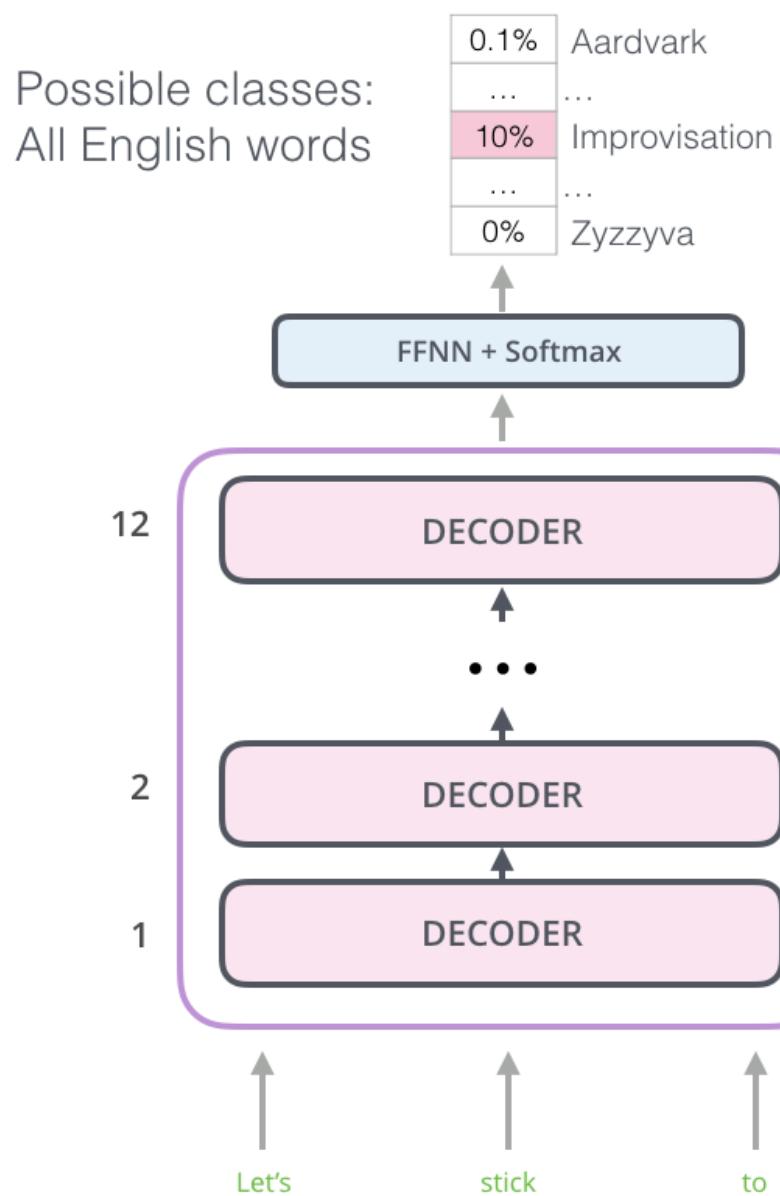
Decoder stack output

All the slides on the transformer above are borrowed from this [Jay Alammar's brilliant article](http://jalammar.github.io/illustrated-transformer/)
(<http://jalammar.github.io/illustrated-transformer/>)

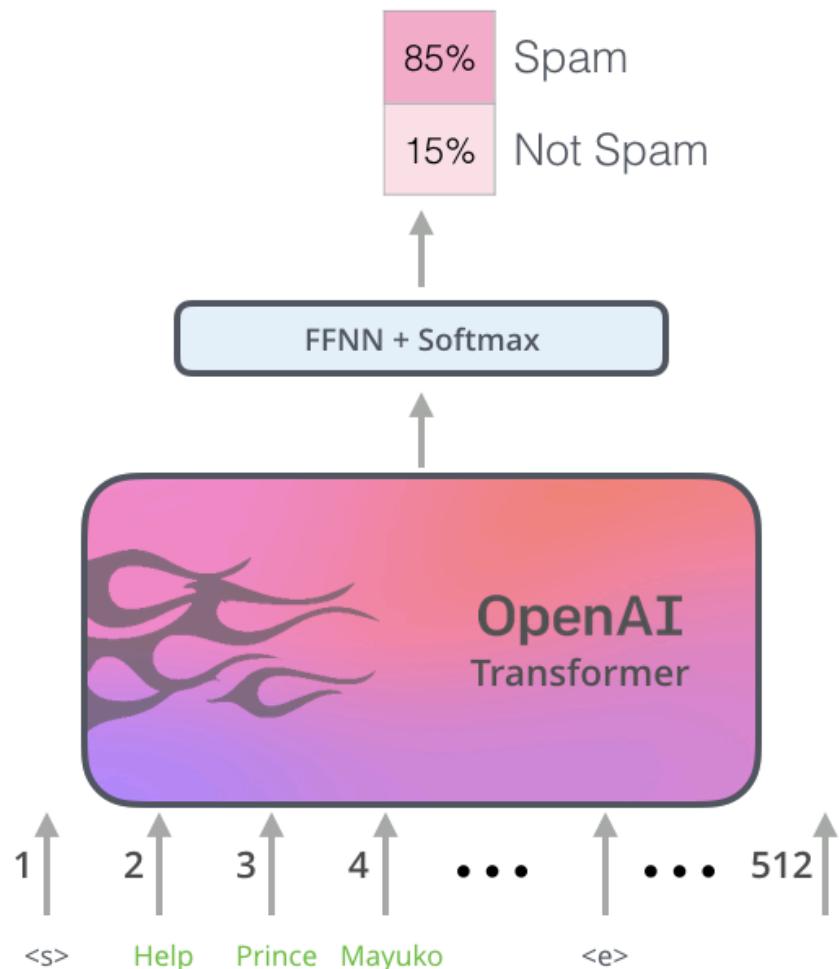
Applications



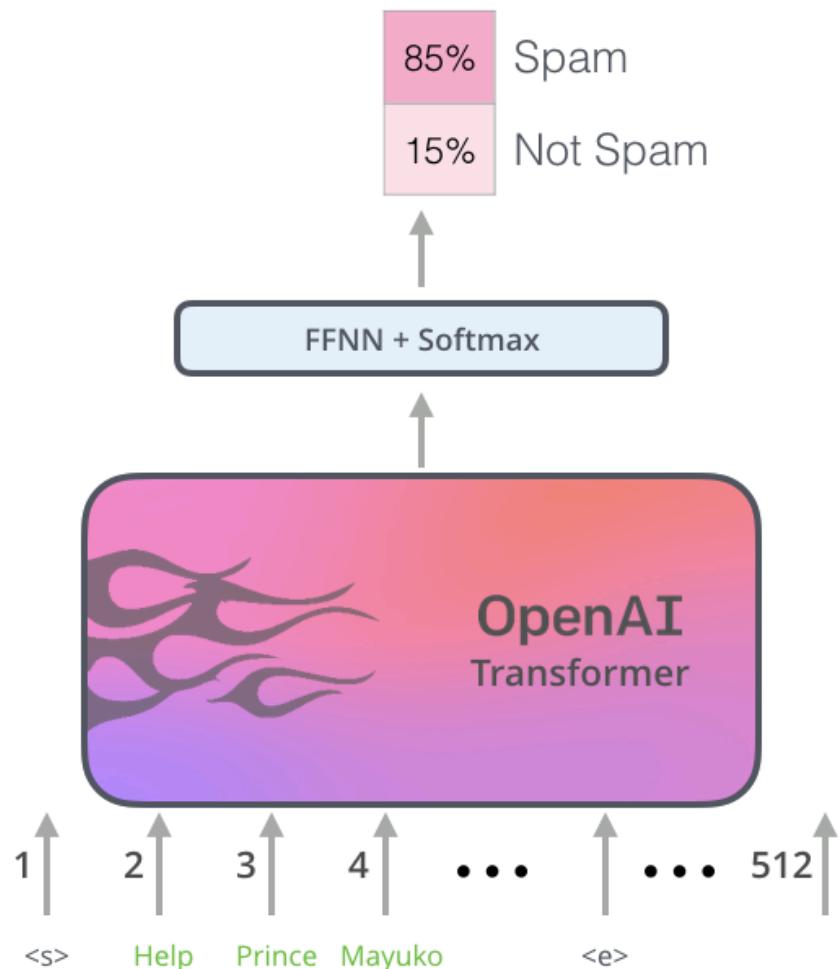
Applications



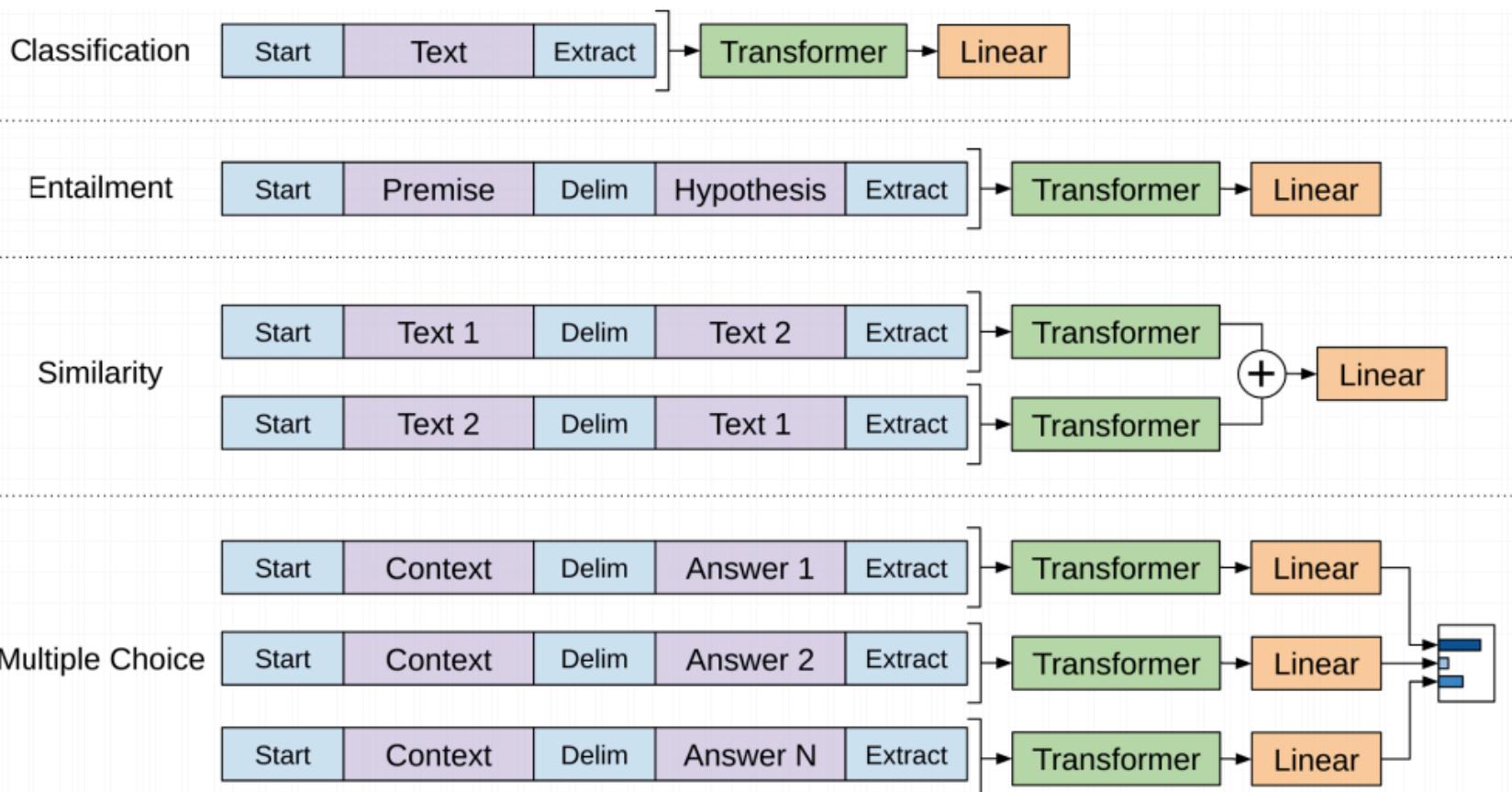
Applications



Applications



Applications

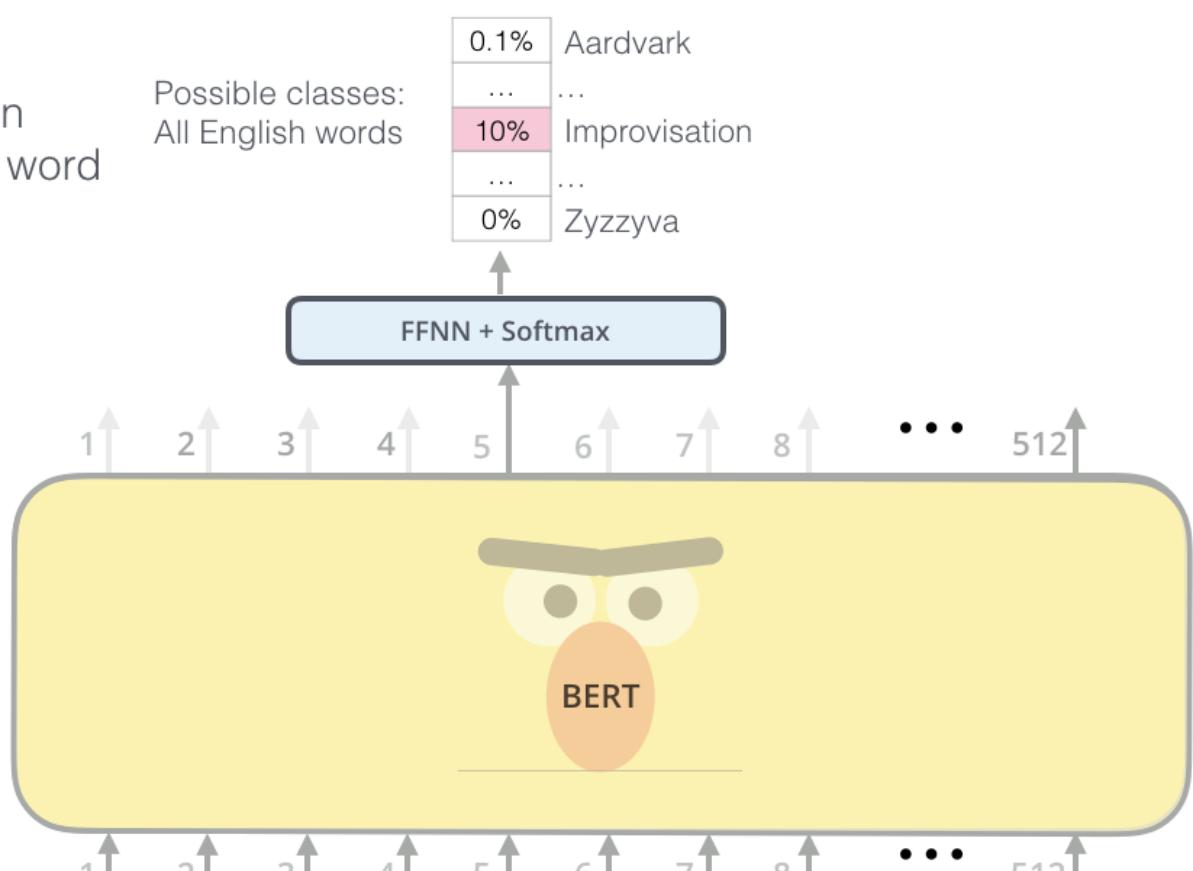


Applications

Use the output of the masked word's position to predict the masked word

Possible classes:
All English words

0.1%	Aardvark
...	...
10%	Improvisation
...	...
0%	Zyzyva



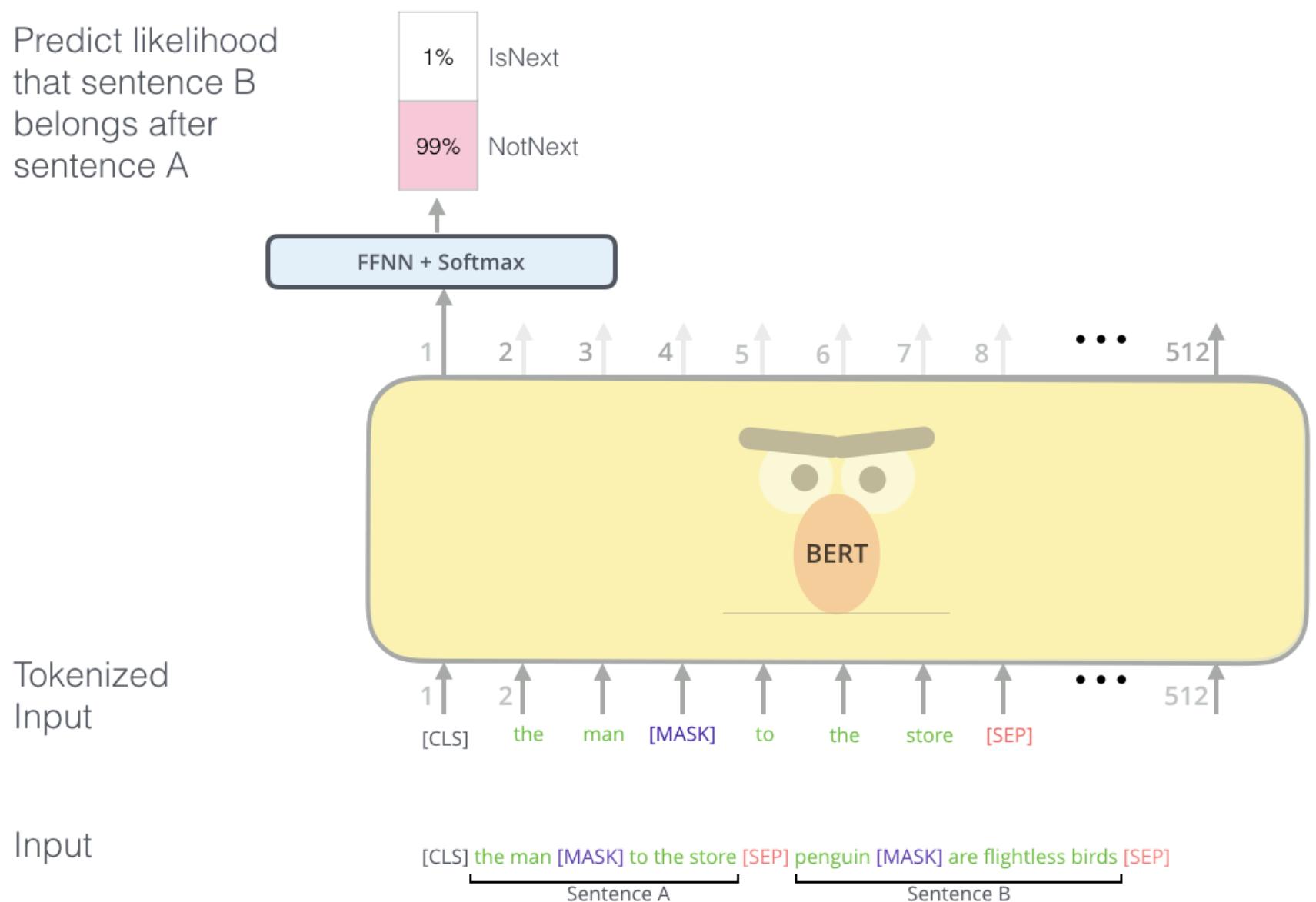
Randomly mask 15% of tokens

Input

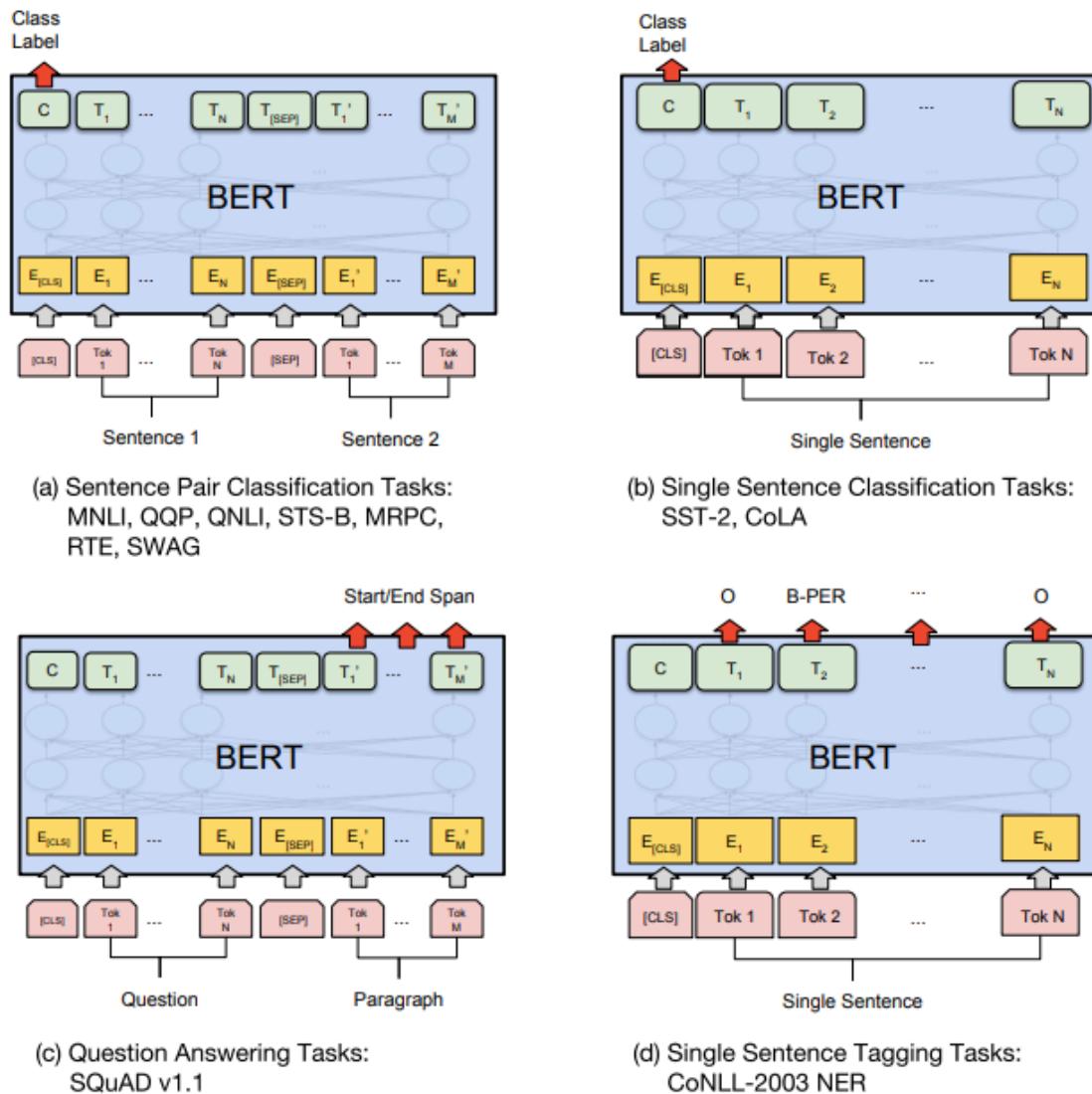
[CLS] ↑ Let's ↑ stick ↑ to ↑ [MASK] ↑ in ↑ this ↑ skit ↑

Applications

Predict likelihood
that sentence B
belongs after
sentence A



Applications



All the slides on the applications of transformer above are borrowed from this [Jay Alammar's article](http://jalammar.github.io/illustrated-bert/) (<http://jalammar.github.io/illustrated-bert/>).

GPT-2

SYSTEM PROMPT (HUMAN-WRITTEN) In a shocking finding, scientist discovered a herd of unicorns living in a remote, previously unexplored valley, in the Andes Mountains. Even more surprising to the researchers was the fact that the unicorns spoke perfect English.

MODEL COMPLETION (MACHINE-WRITTEN, 10 TRIES) The scientist named the population, after their distinctive horn, Ovid's Unicorn. These four-horned, silver-white unicorns were previously unknown to science.

Now, after almost two centuries, the mystery of what sparked this odd phenomenon is finally solved.

Dr. Jorge Pérez, an evolutionary biologist from the University of La Paz, and several companions, were exploring the Andes Mountains when they found a small valley, with no other animals or humans. Pérez noticed that the valley had what appeared to be a natural fountain, surrounded by two peaks of rock and silver snow.

Pérez and the others then ventured further into the valley. “By the time we reached the top of one peak, the water looked blue, with some crystals on top,” said Pérez.

Pérez and his friends were astonished to see the unicorn herd. These creatures could be seen from the air without having to move too much to see them – they were so close they could touch their horns.

While examining these bizarre creatures the scientists discovered that the creatures also spoke some fairly regular English. Pérez stated, “We can see, for example, that they have a common ‘language,’ something like a dialect or dialectic.” Dr. Pérez believes that the unicorns may have originated in Argentina, where the animals were believed to be descendants of a lost race of people who lived there before the arrival of humans in those parts of South America.

While their origins are still unclear, some believe that perhaps the creatures were created when a human and a unicorn met each other in a time before human civilization. According to Pérez, “In South America, such incidents seem to be quite common.”

However, Pérez also pointed out that it is likely that the only way of knowing for sure if unicorns are indeed the descendants of a lost alien race is through DNA. “But they seem to be able to communicate in English quite well, which I believe is a sign of evolution, or at least a change in social organization,” said the scientist.

SQuAD 2.0

Let's explore [SQuAD 2.0 \(<https://rajpurkar.github.io/SQuAD-explorer/explore/v2.0/dev/>\)](https://rajpurkar.github.io/SQuAD-explorer/explore/v2.0/dev/).

