AI CAMP

school of ai Algiers

Attention mechanism and Transformers architecture

- Presented by AIT SAID Azzedine Idir -

Introduction

"Generative AI is the most powerful tool for creativity that has ever been created. It has the potential to unleash a new era of human innovation." ~Elon Musk

Gemini

AI CAMP

school of ai Algiers

Introduction

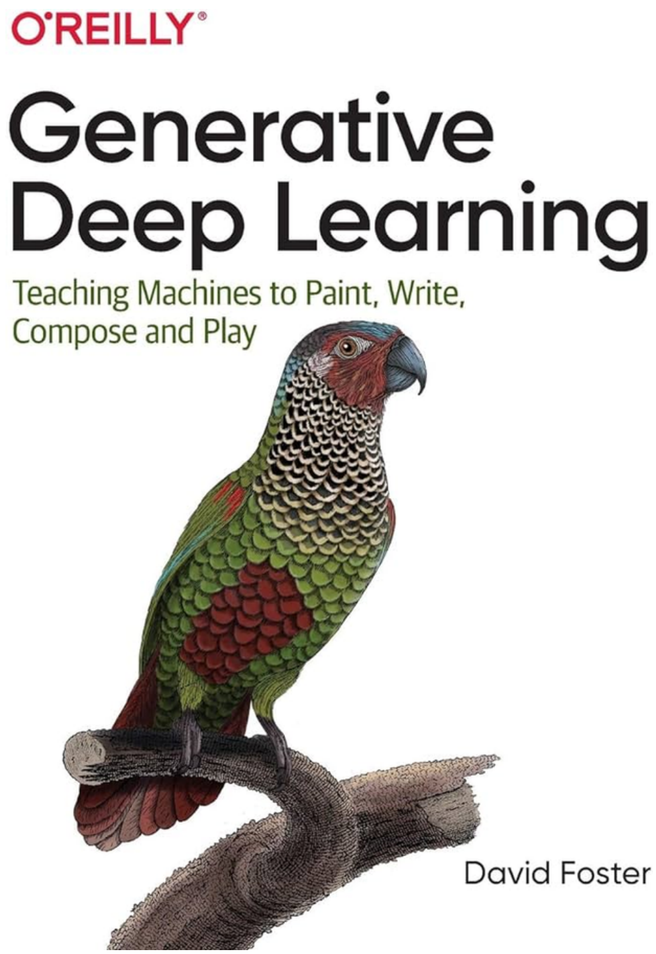
"Generative AI is the most powerful tool for creativity that has ever been created. It has the potential to unleash a new era of human innovation." ~Elon Musk

AI CAMP

school of ai Algiers

Introduction

Generative AI

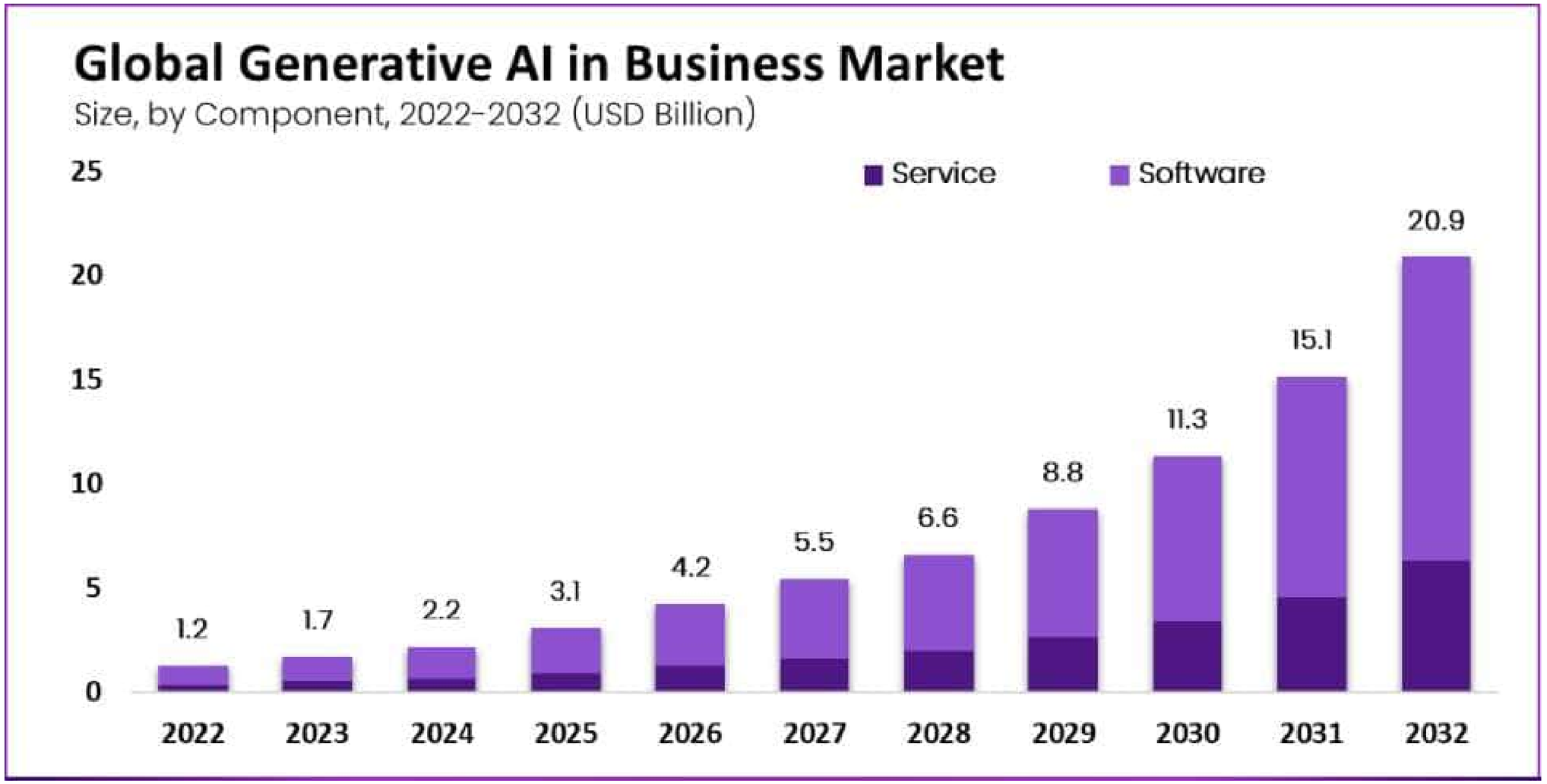


Proof of a trend

AI CAMP

school of ai Algiers

Introduction



It’s not just a trend after all

AI CAMP

school of ai Algiers

Introduction

Why is it working ?

No code for today

How ? :’(

AI CAMP

school of ai Algiers

Introduction

Why is it working ?

Embeddings

Encoder-decoder

Key Concepts

Attention

Transformers

AI CAMP

school of ai Algiers

Introduction

Why is it working ?

Embedding

Encoder-decoder

Attention

Transformers

AI CAMP

school of ai Algiers

Plan

Introduction

Reminders !

Attention is all you need

Encoder-decoder

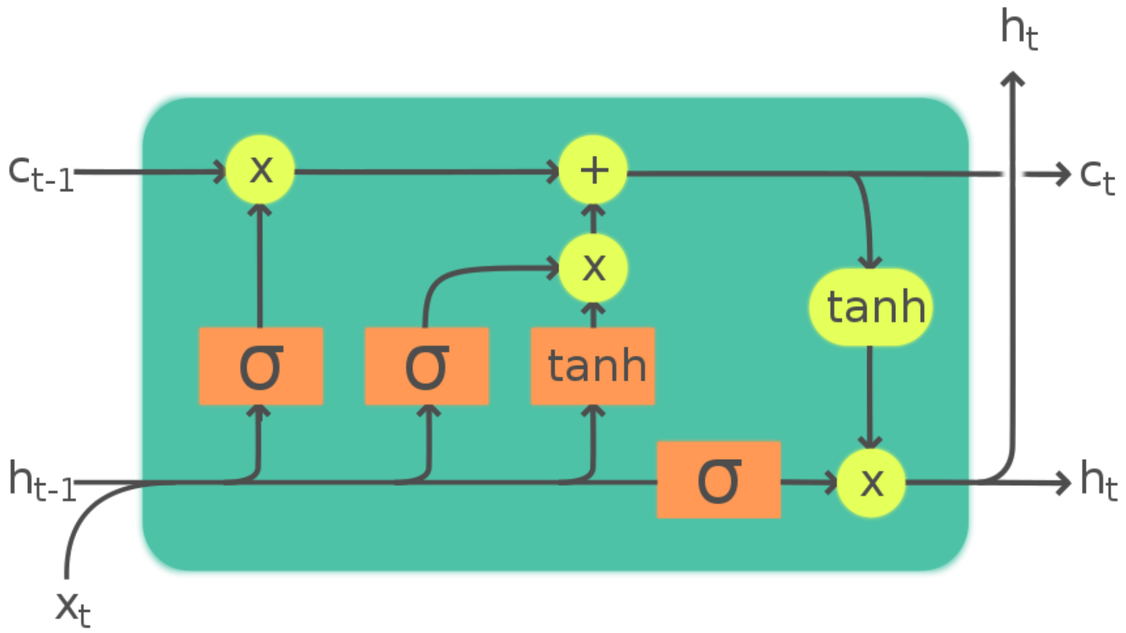
Transformers

AI CAMP

school of ai Algiers

The problem of LSTMs

Translate: “Don’t use LSTMs for long and complex translations because they don’t capture very very very very very large sequences, use attention instead”



- Processing is done sequentially.
- There will be a moment where we need to “forget” some data (forget gates)
- Sometimes we forget essential information.
- We can’t differentiate between apple as a fruit or brand

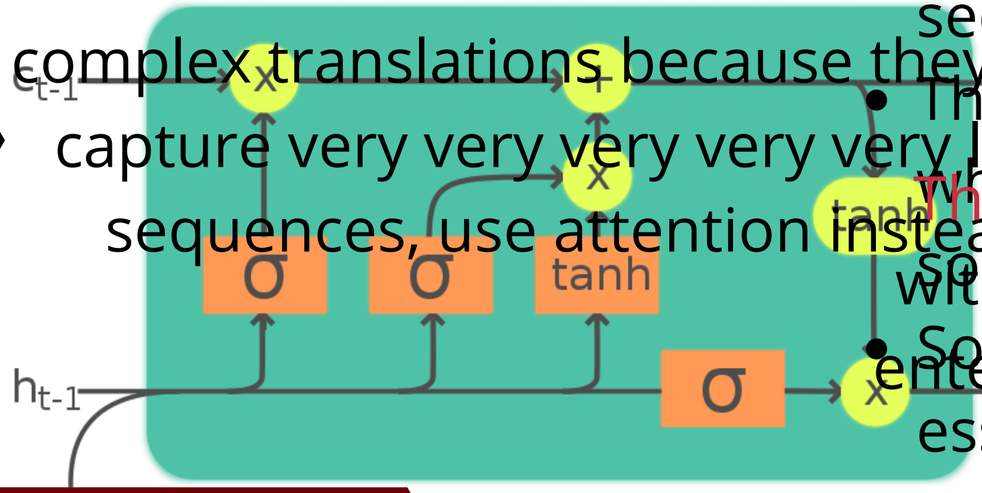
The boy was going outside with his friends ..... He entered the shop and bought oranges.

AI CAMP

school of ai Algiers

The problem of LSTMs

Translate: “Don’t use LSTMs for long and complex translations because they don’t capture very very very very very very large sequences, use attention instead”



- Processing is done sequentially.
- There will be a moment where we need to “forget” some data (forget gates)
- Sometimes we forget essential information

The boy was going out with his friends……..…… entered the shop and bou oranges.

AI CAMP

school of ai Algiers

The problem of LSTMs

We need a solution that remembers details as long as possible and that processes in parallel !

The solution step by step

Embeddings

Encoder-decoder

Attention

Transformers

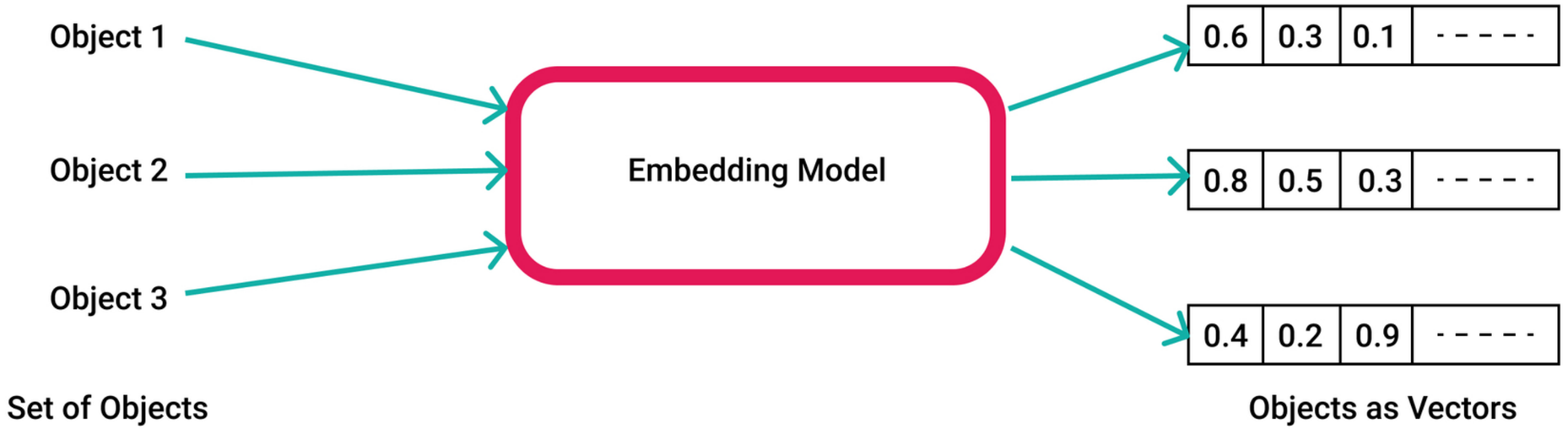
AI CAMP

school of ai Algiers

The solution step by step

Embeddings

1. We want to turn words into vectors for computers to process



2. We want close words in meaning to have a close representation as vectors !

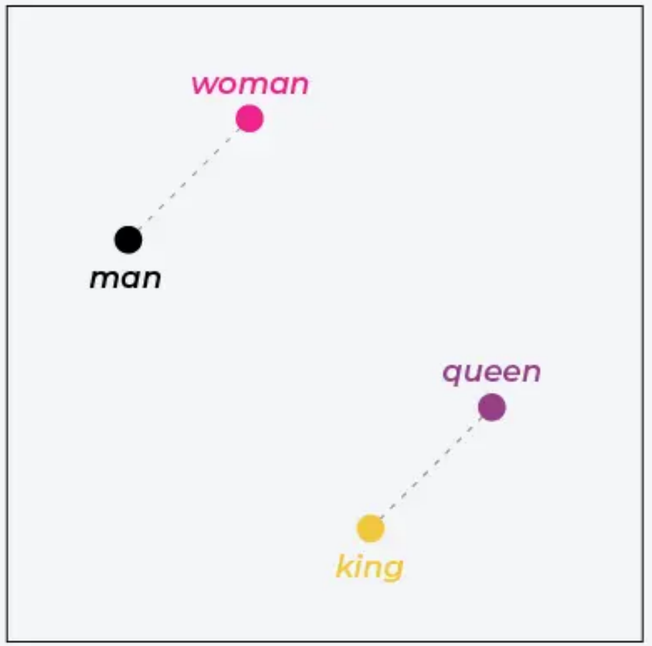
AI CAMP

school of ai Algiers

The solution step by step

word	living being	feline	human	gender	royalty	verb	plural
man	0.6	-0.2	0.8	0.9	-0.1	-0.9	-0.7
woman	0.7	0.3	0.8	-0.7	0.1	-0.5	-0.4
king	0.5	-0.4	0.7	0.8	0.9	-0.7	-0.6
queen	0.8	-0.1	0.8	-0.9	0.8	-0.5	-0.9

Word embedding



Visualization of word embedding

queen = king + woman

AI CAMP

school of ai Algiers

The solution step by step

Where to classify the word “windows” ? Is it closer to words that are related to house or that are related to technology ?

The solution step by step

solution = depends on the context

- Translate: I closed the windows
- Translate: I bought a new windows for my computer

AI CAMP

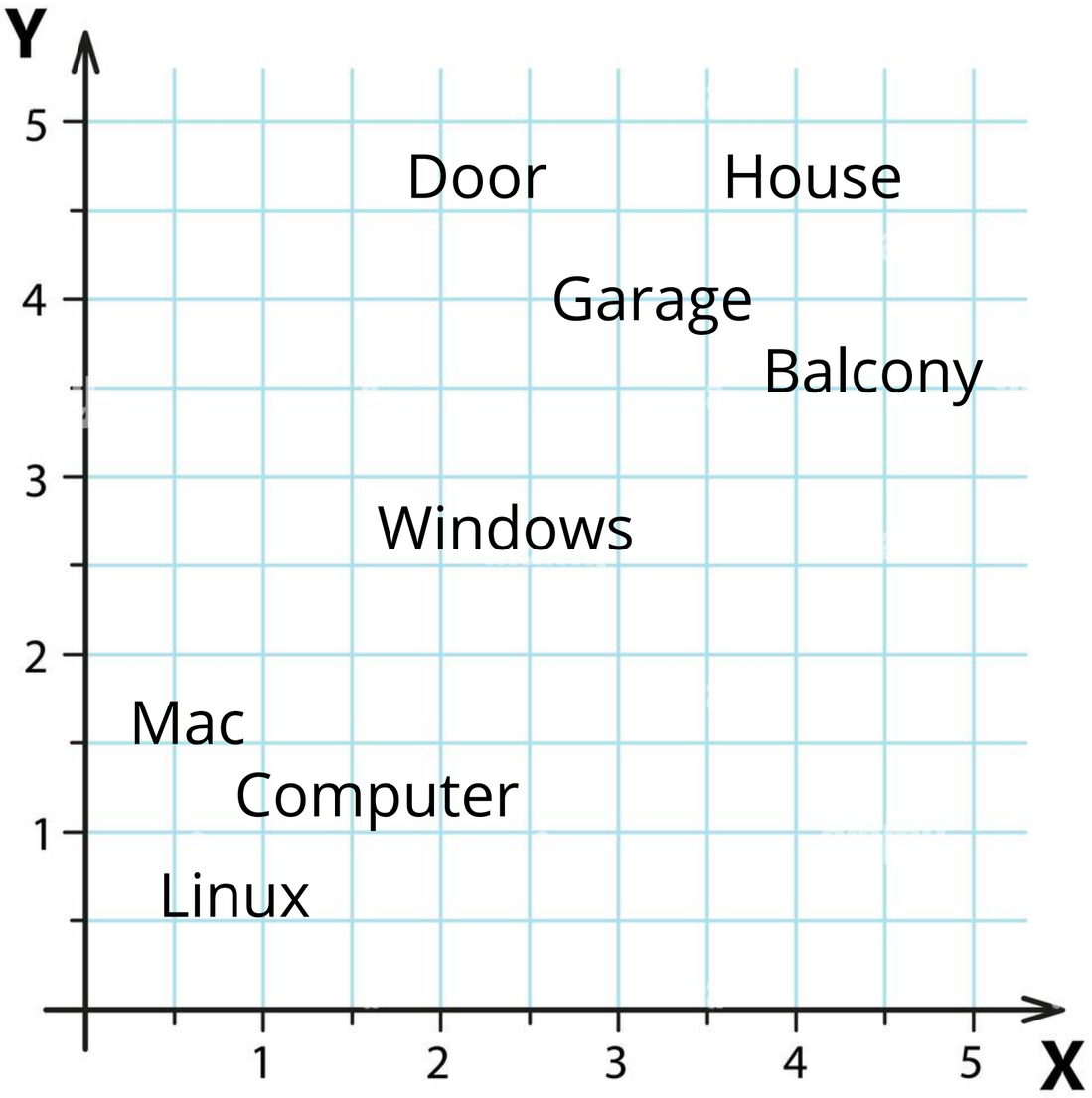
school of ai Algiers

The solution step by step

solution = depends on the context

How can we inject the context to a model ?

The solution step by step

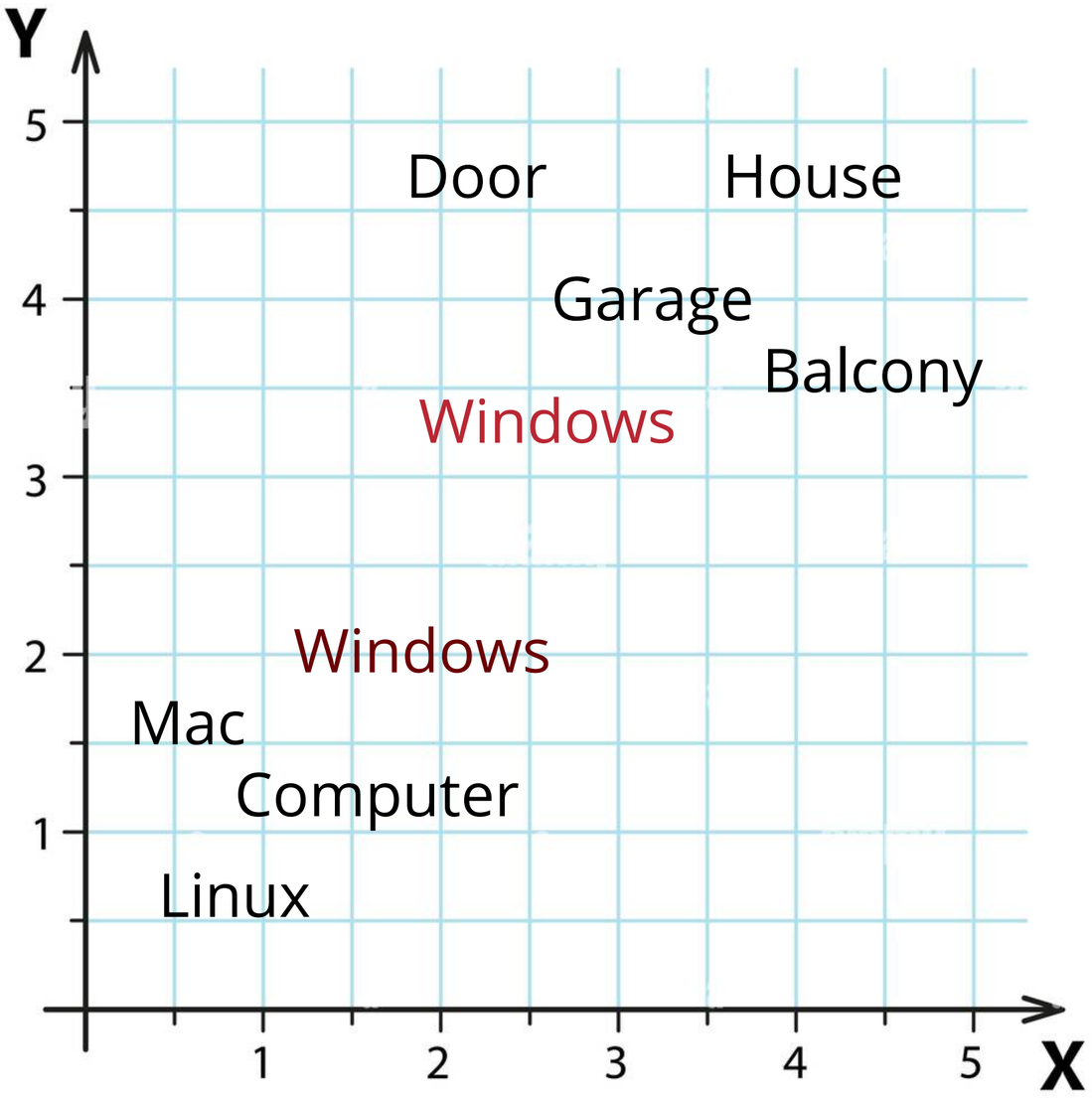


1. Translate: I closed the windows
2. Translate: I bought a new windows for my computer

AI CAMP

school of ai Algiers

The solution step by step



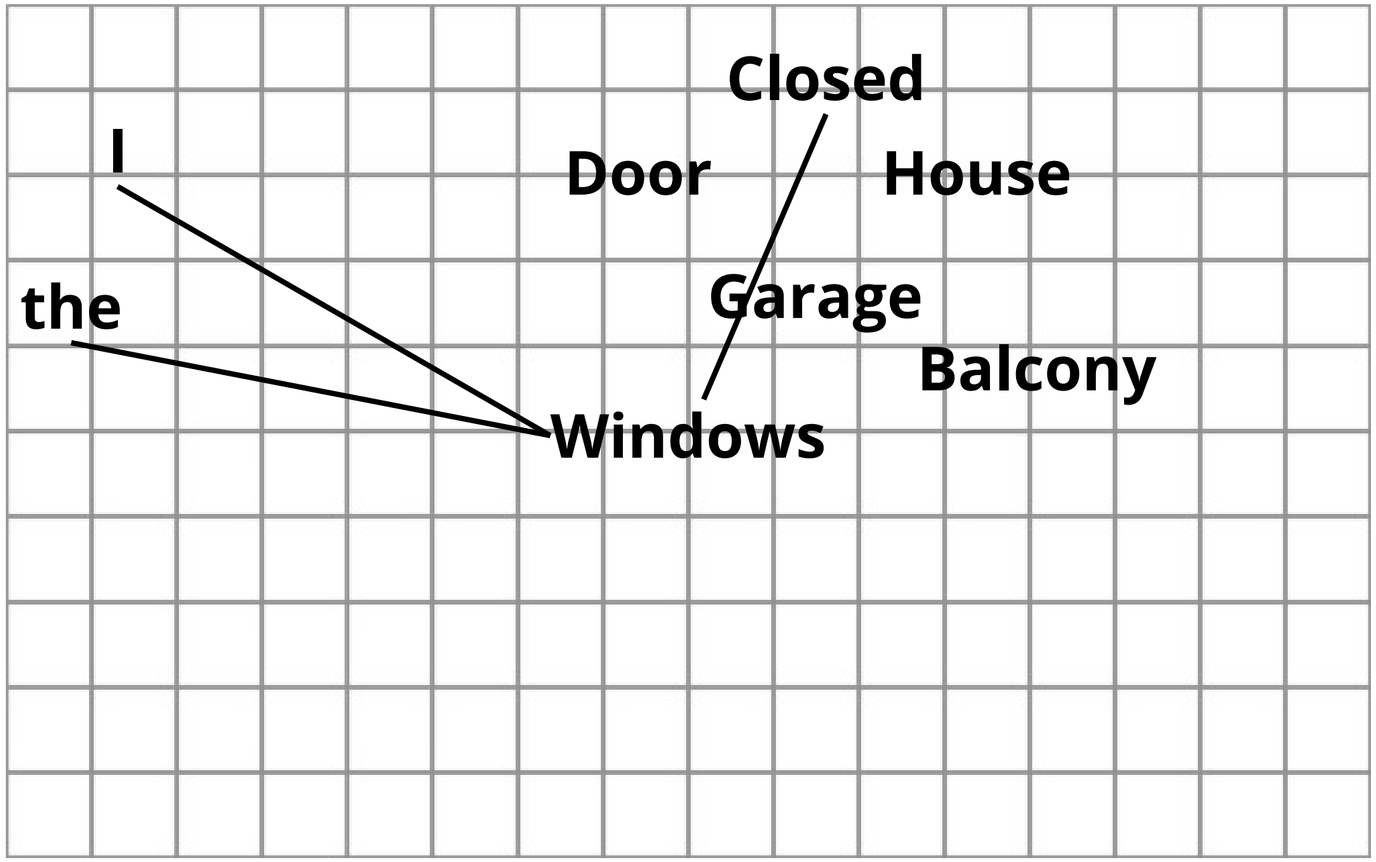
1. Translate: I closed the windows
2. Translate: I bought a new windows for my computer

AI CAMP

school of ai Algiers

The solution step by step

Because we are using a computer we need to do processing in all the words of the context !



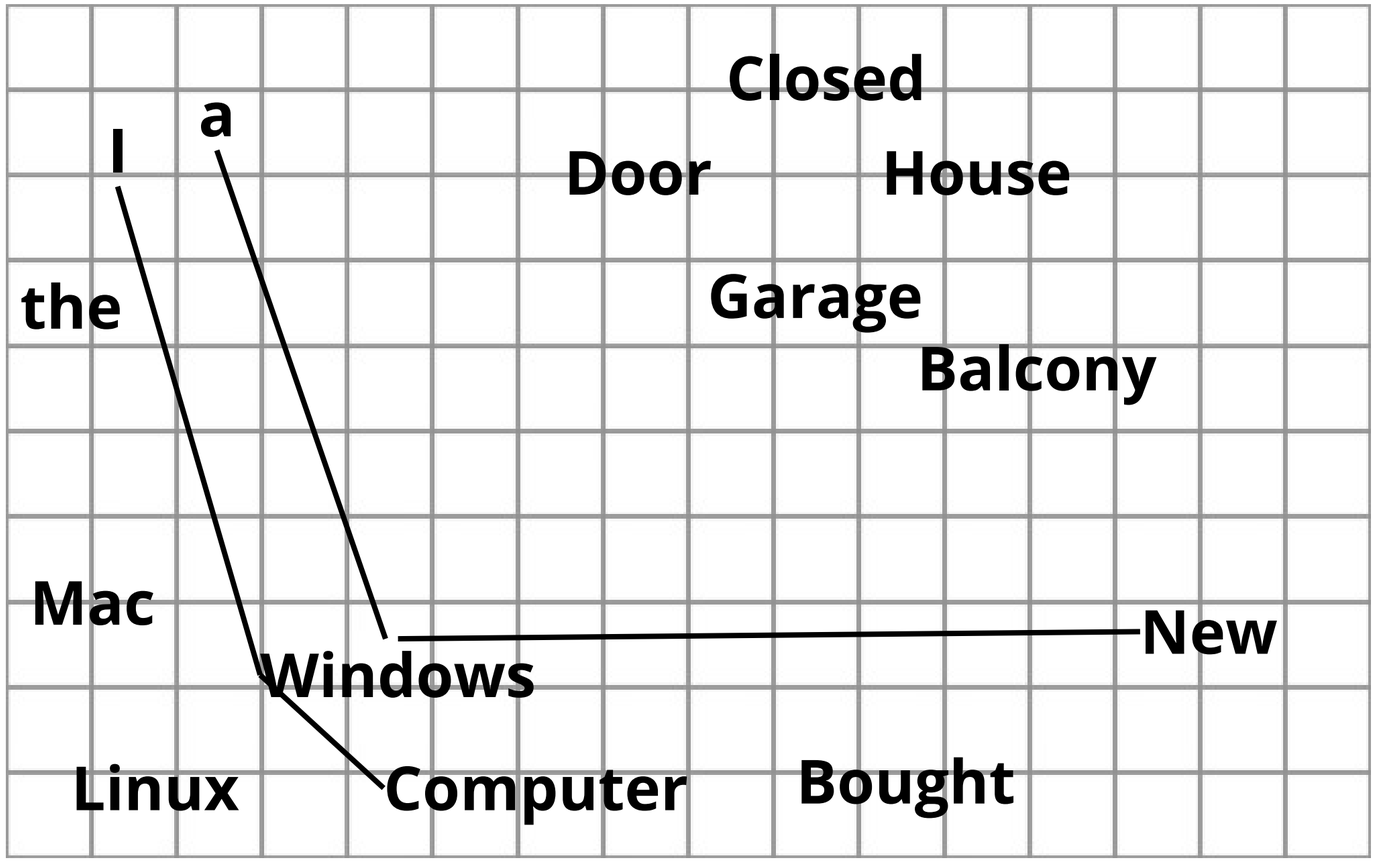
You can see it as gravitational forces

AI CAMP

school of ai Algiers

The solution step by step

Because we are using a computer we need to do processing in all the words of the context !



Not all words in the context have the same impact on “Windows”

AI CAMP

school of ai Algiers

The solution step by step

How can we measure the similarity between words embedding ?

(1,4).(0,3)=1*0 + 4*3 = 12

(1,4).(3,0)=1*3 + 0*3 = 3

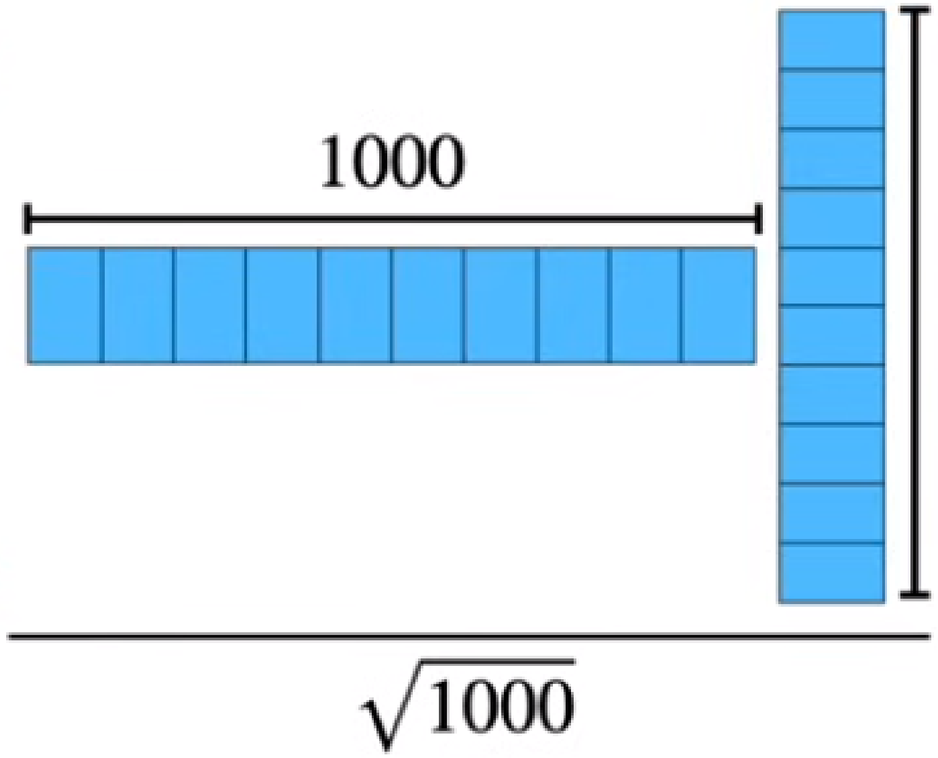
The dot product !

AI CAMP

school of ai Algiers

The solution step by step

What to do when we have embeddings that has more than 1000 values ?



We keep the definition of similarity and just divide by the square of the length of the embedding to make the value smaller -> more manageable

The scaled dot product !

AI CAMP

school of ai Algiers

The solution step by step

1. Translate: I closed the windows
2. Translate: I bought a new windows for my computer

I: [1, 0] closed: [1, 1] the: [0, 1] windows: [2, 0]

	I	closed	the	windows
I	1.000	0.000	0.707	0.000
closed	1.414	1.000	0.707	1.414
the	0.707	0.707	1.000	1.414
windows	0.000	1.414	1.414	1.000

windows = 1.414 closed + 1.414 the

AI CAMP

school of ai Algiers

The solution step by step

Normalization

Want coefficients to add to 1

$$\text{Orange} \longrightarrow \frac{1\ \text{Orange} + 0.71\ \text{Apple}}{1 + 0.71} = 0.58\ \text{Orange} + 0.42\ \text{Apple}$$

Need coefficients to be positive

$$\text{Orange} \longrightarrow \frac{1\ \text{Orange} - 1\ \text{Motorcycle}}{1 - 1} =$$

Solution? $x \longrightarrow e^x$

AI CAMP

school of ai Algiers

The solution step by step

Softmax

$$x \longrightarrow e^{x}$$

$$\text{Orange} \longrightarrow \frac{e^{1}\,\text{Orange} + e^{0.71}\,\text{Apple}}{e^{1} + e^{0.71}} = 0.57\,\text{Orange} + 0.43\,\text{Apple}$$

$$\text{Orange} \longrightarrow \frac{e^{1}\,\text{Orange} + e^{-1}\,\text{Motorcycle}}{e^{1} + e^{-1}} = 0.88\,\text{Orange} + 0.12\,\text{Motorcycle}$$

AI CAMP

school of ai Algiers

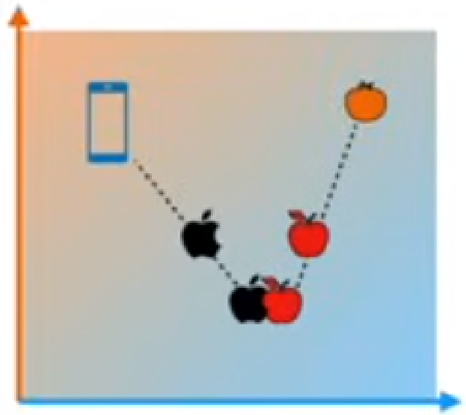
The solution step by step

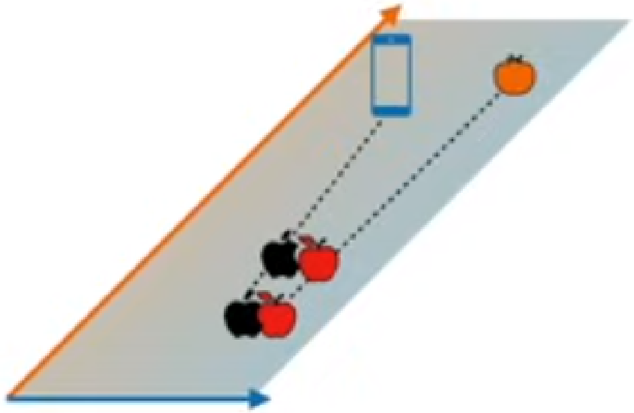
1. Translate: I closed the windows
2. Translate: I bought a new windows for my computer

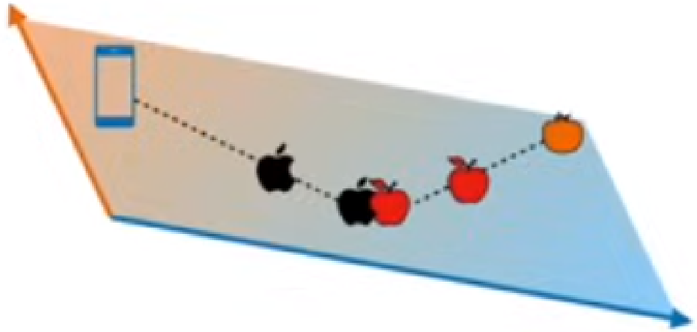
So in a nutshell we have found two embeddings for the word windows that can be used in two different contexts !!!! Our objective is to find good transformations

The solution step by step

- We have 3 different embeddings (shown as planes)
- Which embedding (plane) is the best ?







But how do we mathematically get new embeddings ?

AI CAMP

school of ai Algiers

The solution step by step

Embeddings

Encoder-decoder

Attention

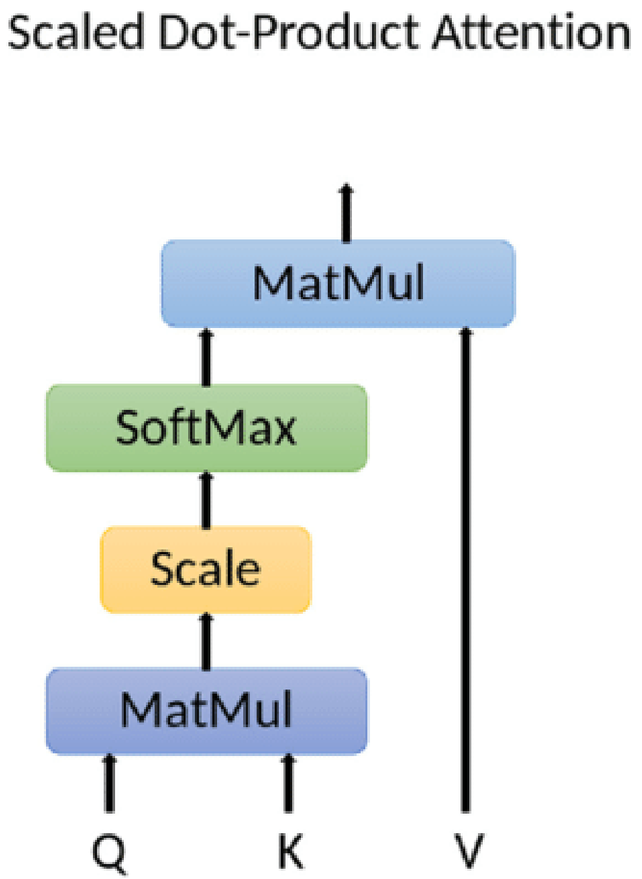
Transformers

AI CAMP

school of ai Algiers

The solution step by step

Attention



AI CAMP

school of ai Algiers

The solution step by step

Self-Attention allows the model to relate words to each other.

In this simple case we consider the sequence length **seq** = 6 and $\mathbf{d_{model}} = \mathbf{d_k} = 512$.

The matrices **Q**, **K** and **V** are just the input sentence.

$$Attention(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

$$\text{softmax}\left(\frac{\mathbf{Q}\ (6, 512) \times \mathbf{K^T}\ (512, 6)}{\sqrt{512}}\right) =$$

	YOUR	CAT	IS	A	LOVELY	CAT	Σ
YOUR	0.268	0.119	0.134	0.148	0.179	0.152	1
CAT	0.124	0.278	0.201	0.128	0.154	0.115	1
IS	0.147	0.132	0.262	0.097	0.218	0.145	1
A	0.210	0.128	0.206	0.212	0.119	0.125	1
LOVELY	0.146	0.158	0.152	0.143	0.227	0.174	1
CAT	0.195	0.114	0.203	0.103	0.157	0.229	1

(6, 6)

* all values are random.

* for simplicity I considered only one head, which makes $d_{model} = d_k$.

Umar Jamil – https://github.com/hkproj/transformer-from-scratch-notes

AI CAMP

school of ai Algiers

The solution step by step

$$Attention(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

	YOUR	CAT	IS	A	LOVELY	CAT
YOUR	0.268	0.119	0.134	0.148	0.179	0.152
CAT	0.124	0.278	0.201	0.128	0.154	0.115
IS	0.147	0.132	0.262	0.097	0.218	0.145
A	0.210	0.128	0.206	0.212	0.119	0.125
LOVELY	0.146	0.158	0.152	0.143	0.227	0.174
CAT	0.195	0.114	0.203	0.103	0.157	0.229

(6, 6)

X

V

(6, 512)

=

Attention

(6, 512)

Each row in this matrix captures not only the meaning (given by the embedding) or the position in the sentence (represented by the positional encodings) but also each word's interaction with other words.

This results on a context aware representation of the initial sentence !

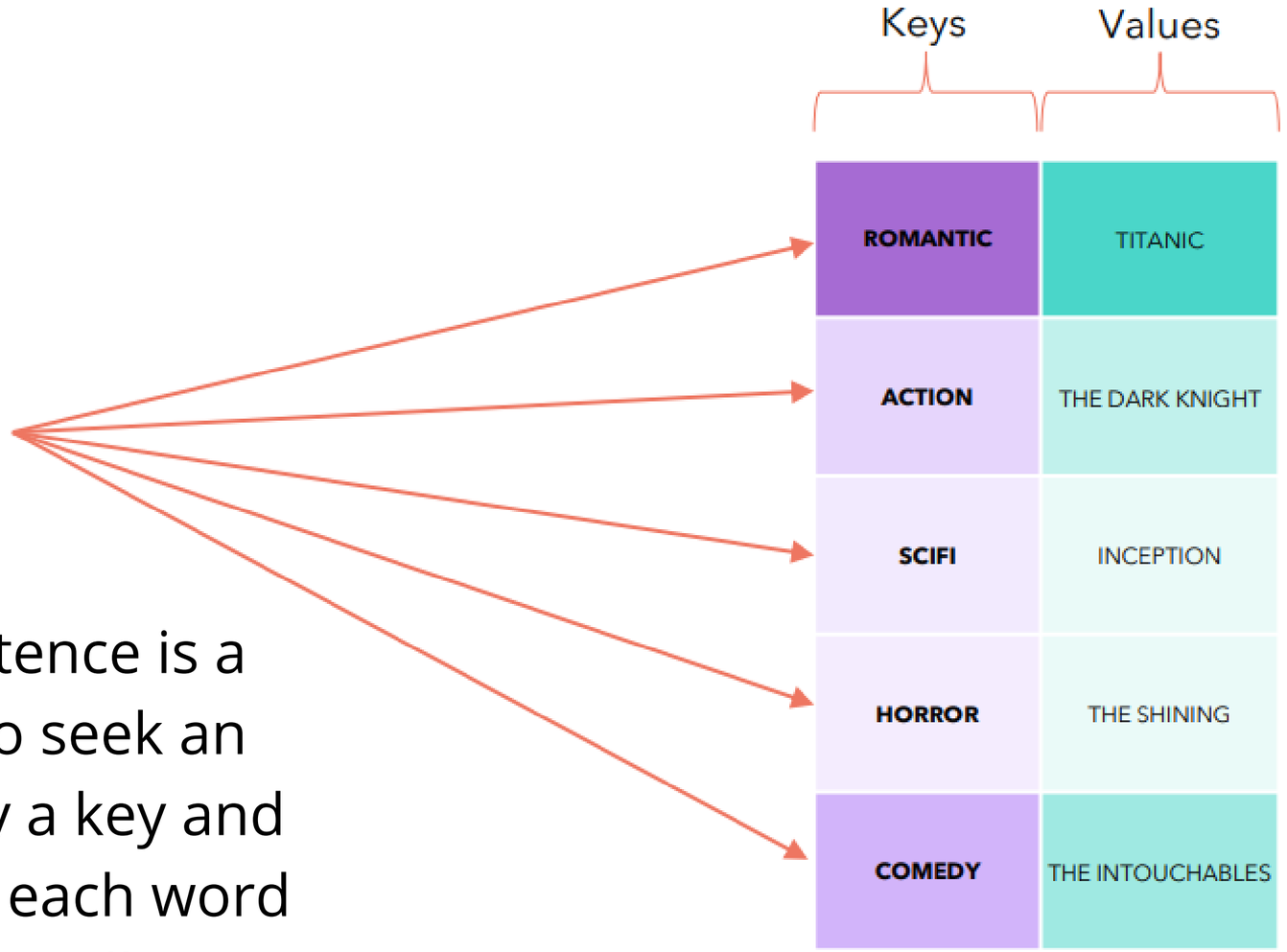
Umar Jamil – https://github.com/hkproj/transformer-from-scratch-notes

AI CAMP

school of ai Algiers

The solution step by step

Query = "**love**"



Each word of the sentence is a query and we use it to seek an entry that is defined by a key and a value (which are also each word in the sentence as well)

* this could be a Python dictionary or a database table.

AI CAMP

school of ai Algiers

The solution step by step

Ok this is all good but where is training ? where is machine learning ? where are the parameters ?

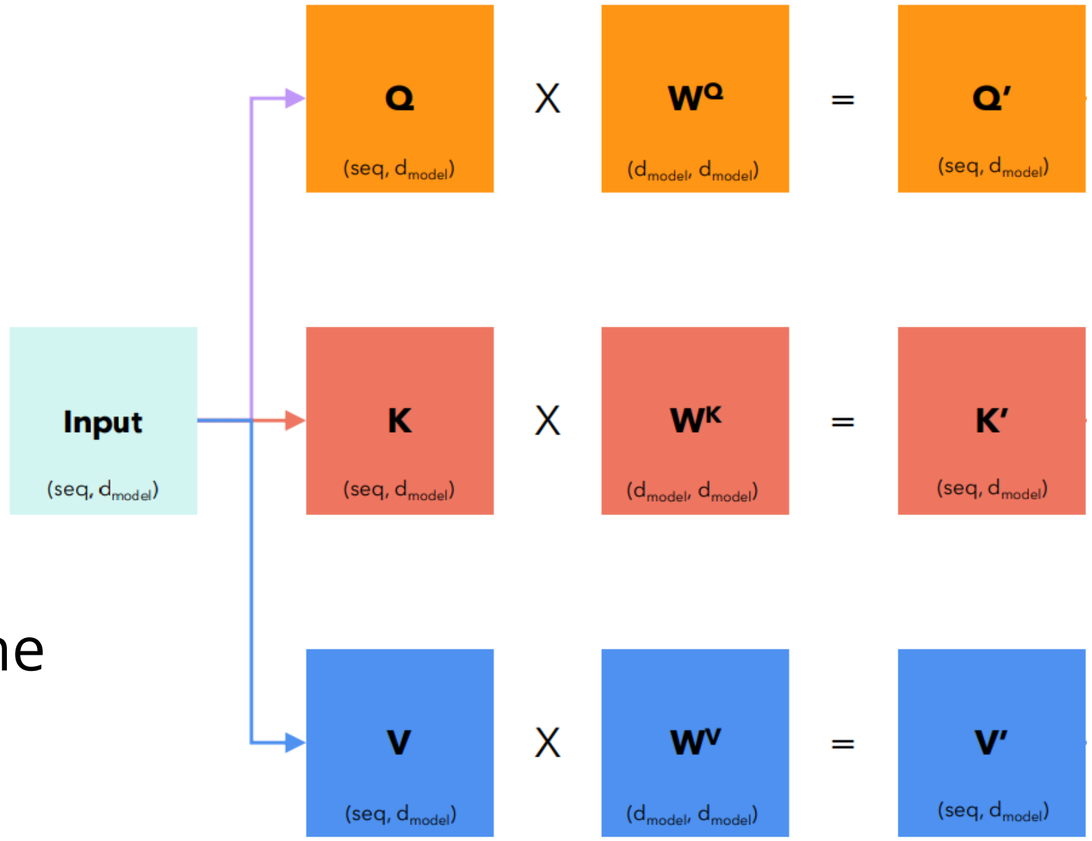
AI CAMP

school of ai Algiers

The solution step by step

- If we apply attention each time a sequence is fed in, we will have as much models as inputs.
- Our objective is to have one model that learns to detect context for each input sequence.

We multiply by weights !!!



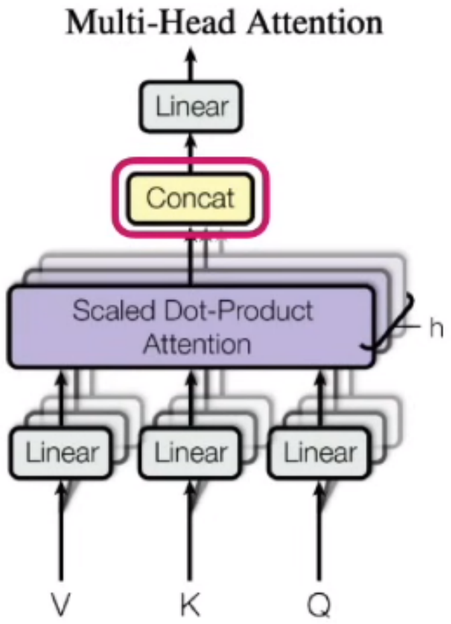
The training is done in the Transformers !

AI CAMP

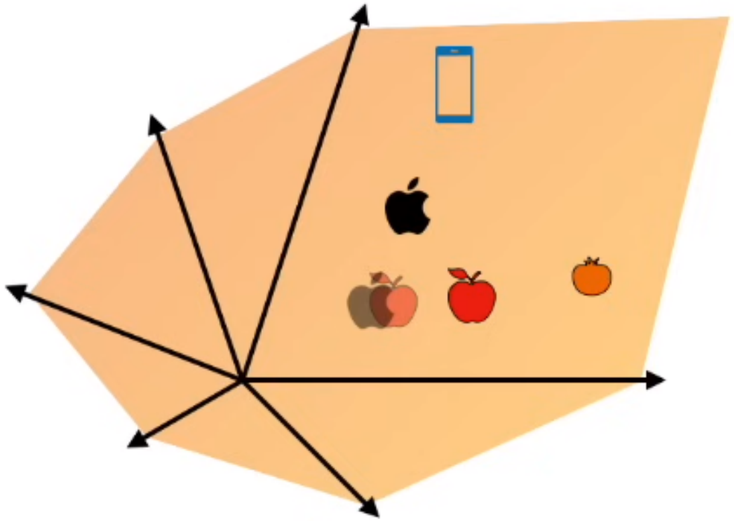
school of ai Algiers

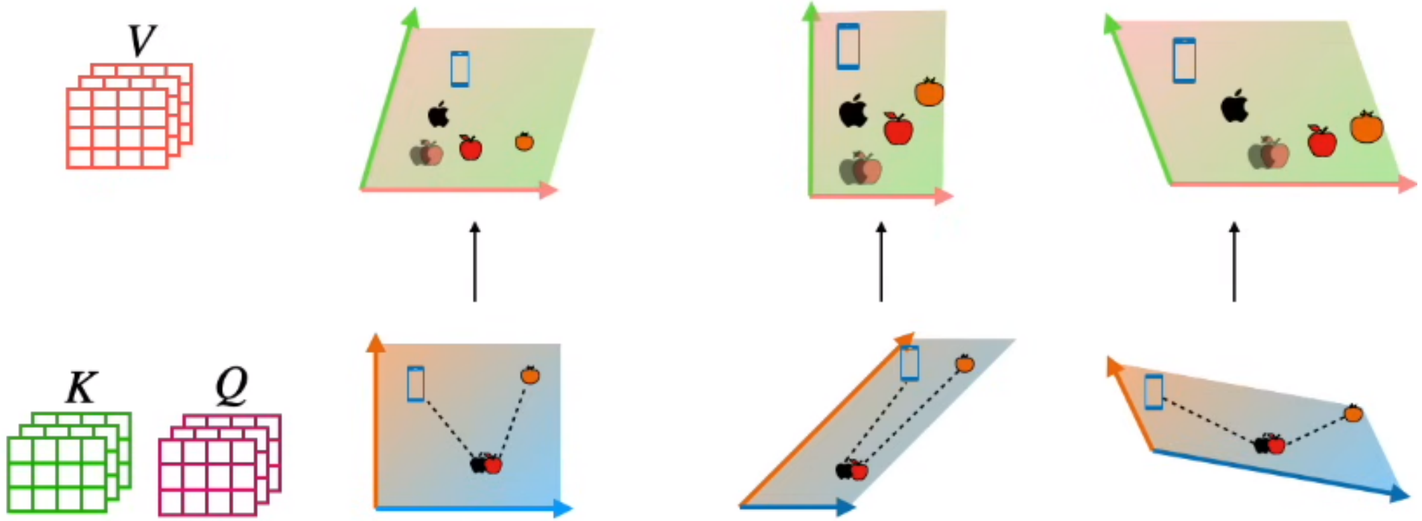
The solution step by step

Multi-head attention



$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, ..., \text{head}_\text{h})W^O$$
$$\text{where head}_\text{i} = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V)$$



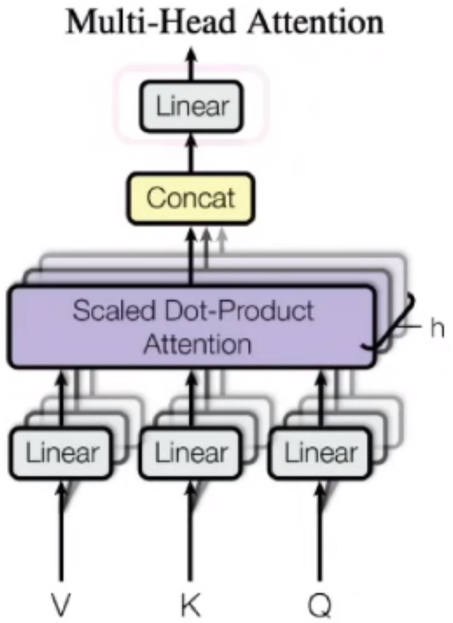


AI CAMP

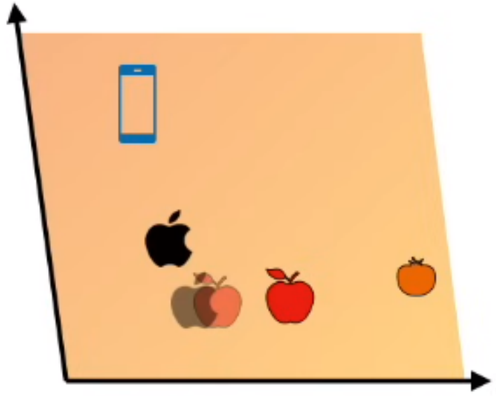
school of ai Algiers

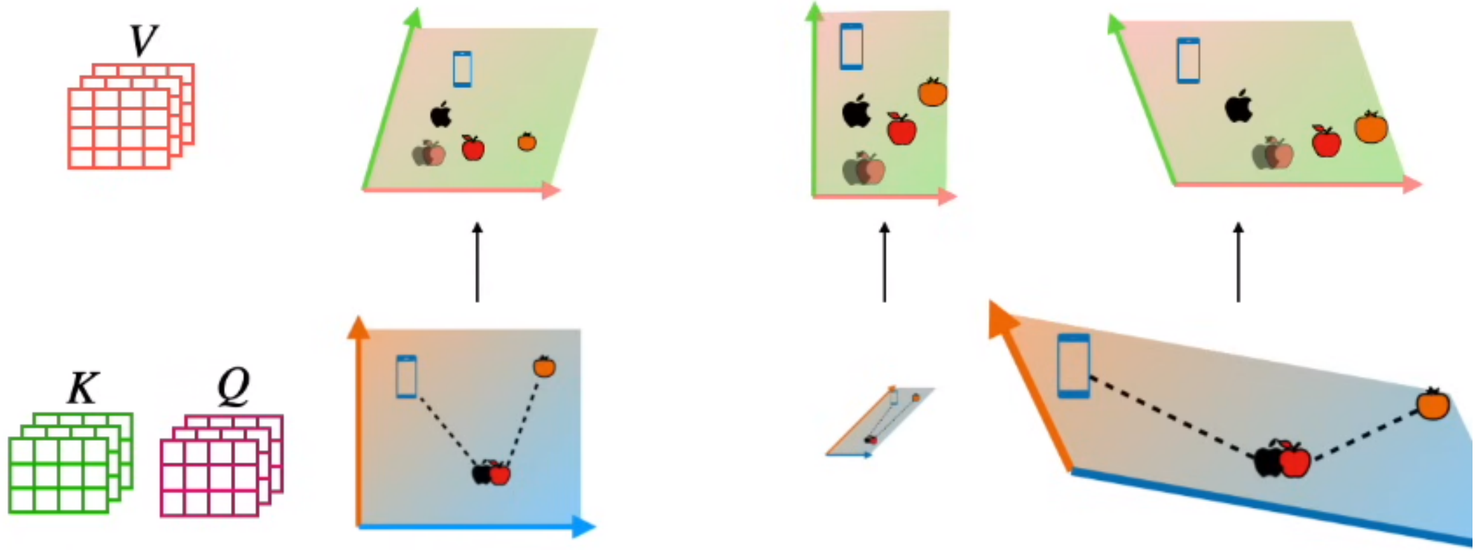
The solution step by step

Multi-head attention



$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, ..., \text{head}_\text{h})W^O$$
$$\text{where head}_\text{i} = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V)$$

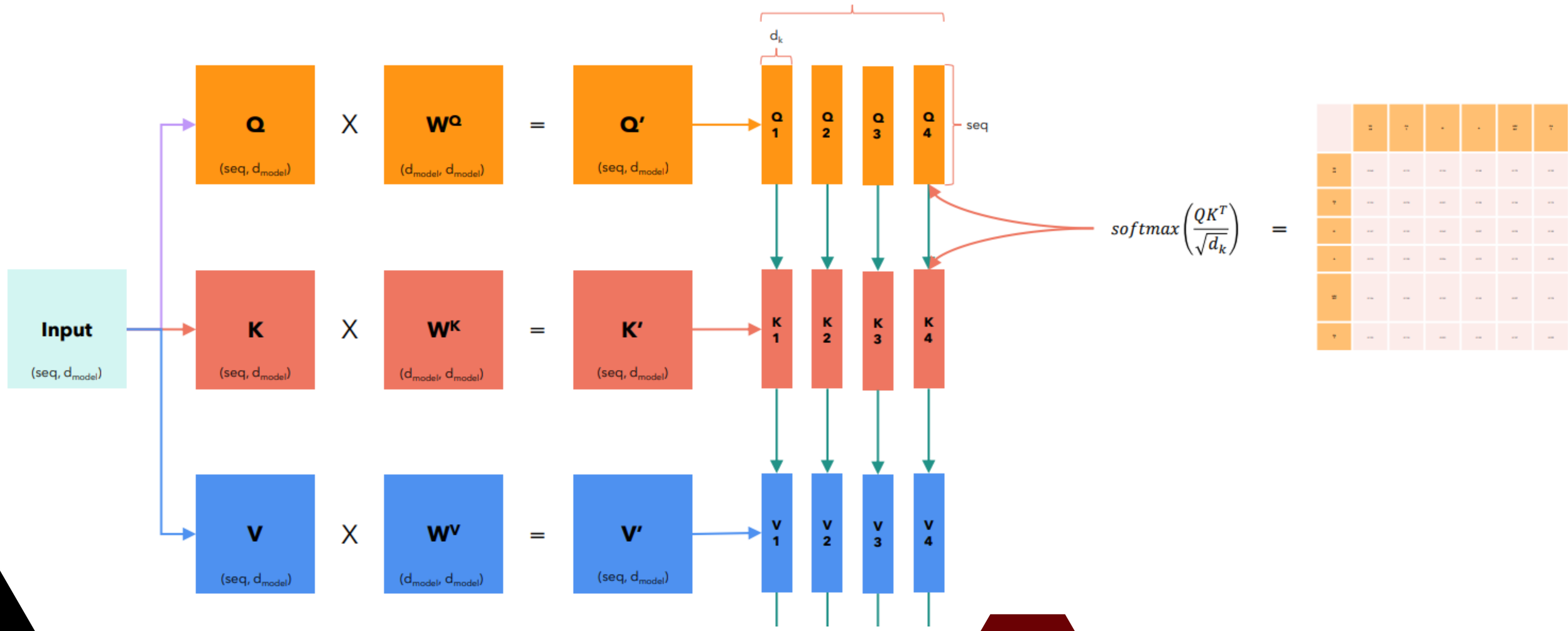




AI CAMP

school of ai Algiers

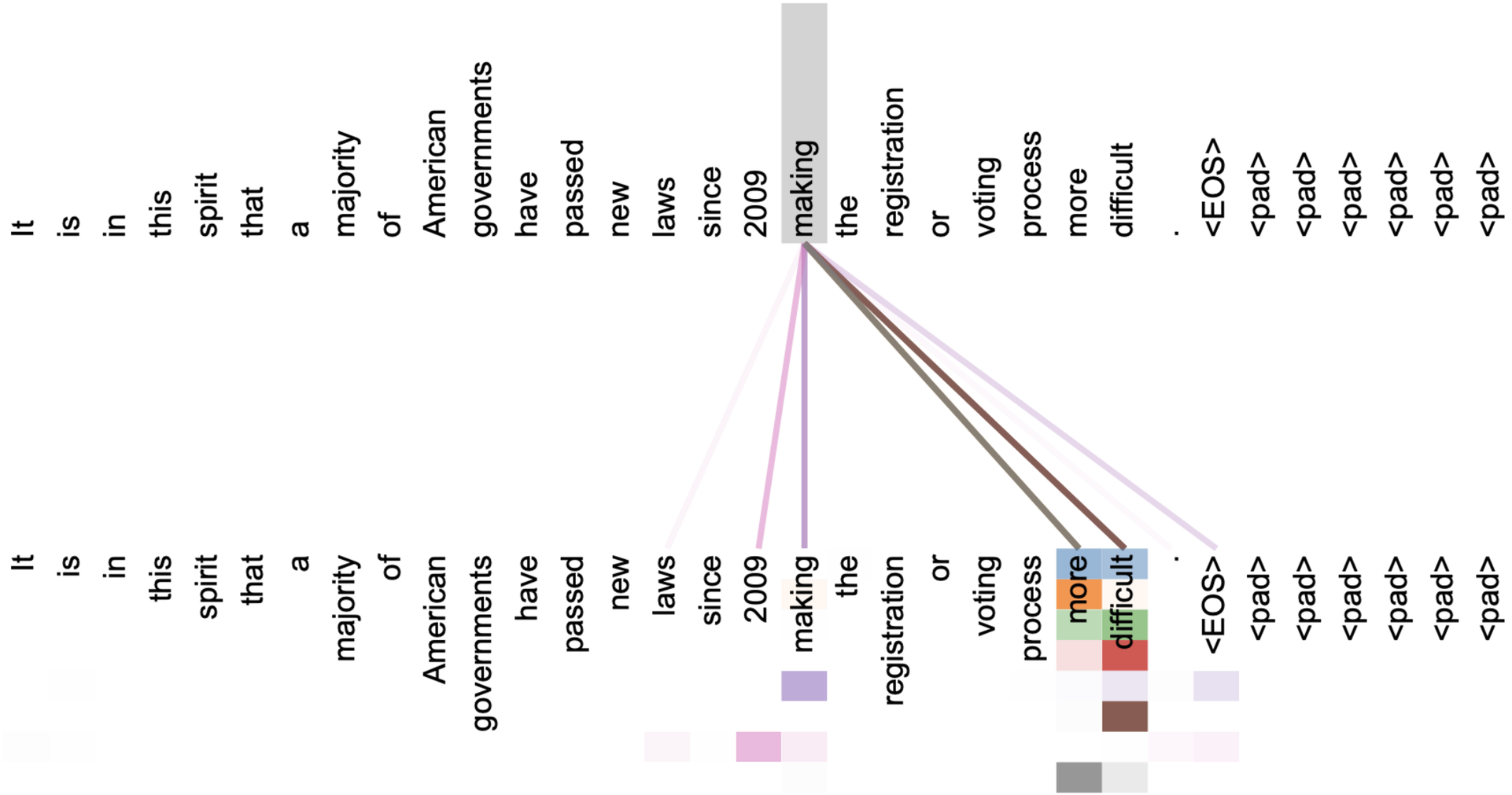
The solution step by step



AI CAMP

school of ai Algiers

The solution step by step



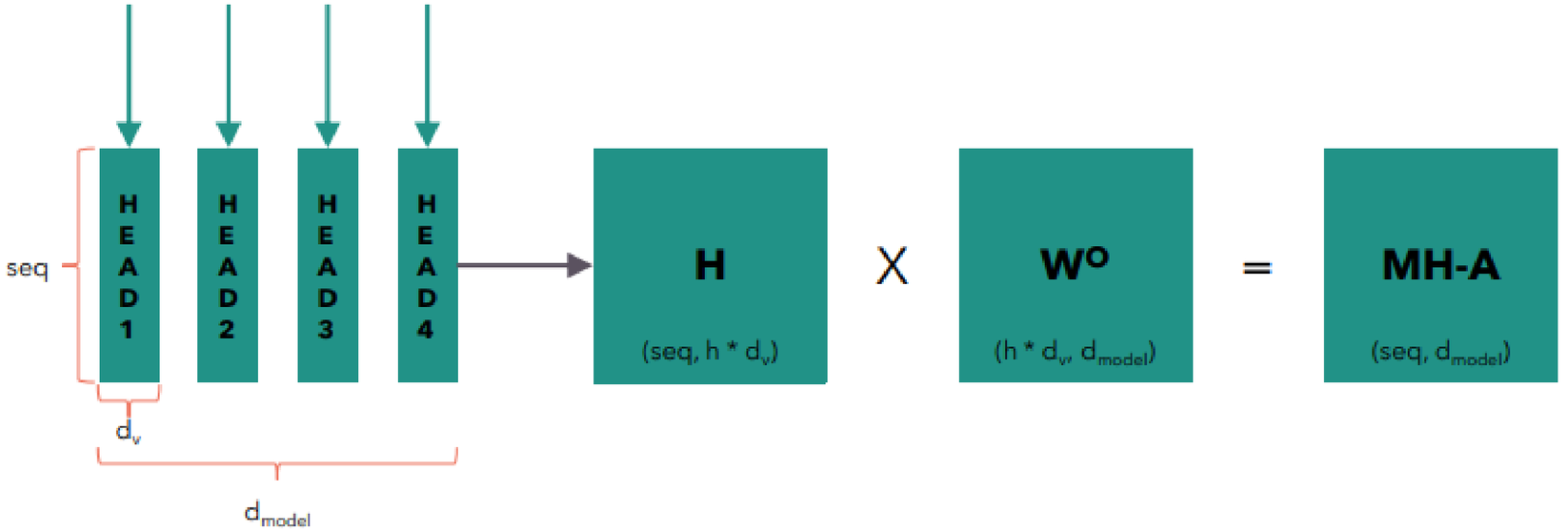
AI CAMP

school of ai Algiers

The solution step by step

$$Attention(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

$$head_i = Attention(QW_i^Q, KW_i^K, VW_i^V)$$



$$MultiHead(Q, K, V) = Concat(head_1 \ \ldots head_h)W^O$$

AI CAMP

school of ai Algiers

The solution step by step

Embeddings

Encoder-decoder

Attention

Transformers

AI CAMP

school of ai Algiers

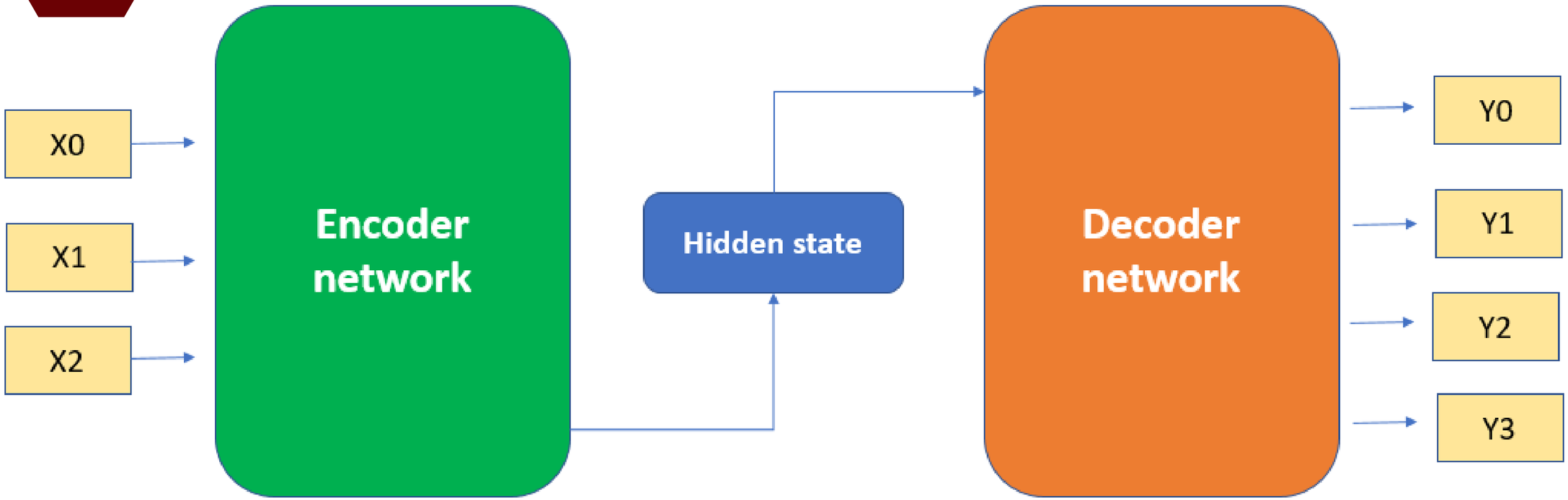
The solution step by step

Encoder-decoder

AI CAMP

school of ai Algiers

The solution step by step



The encoder network processes the input sequence

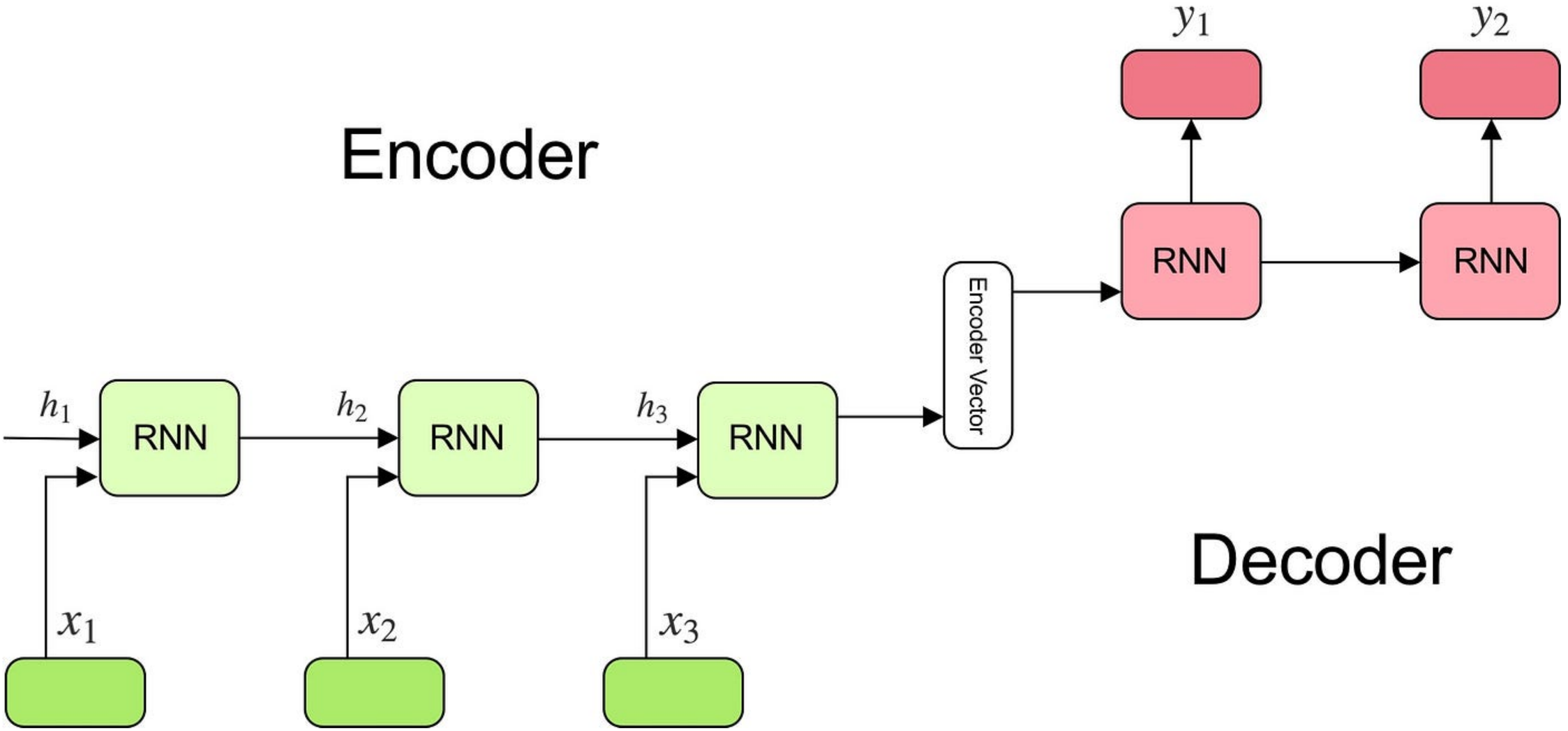
The hidden state (sometimes also called the context vector) captures the essence of the input sequence.

The decoder network takes the hidden state from the encoder and begins generating the output sequence

AI CAMP

school of ai Algiers

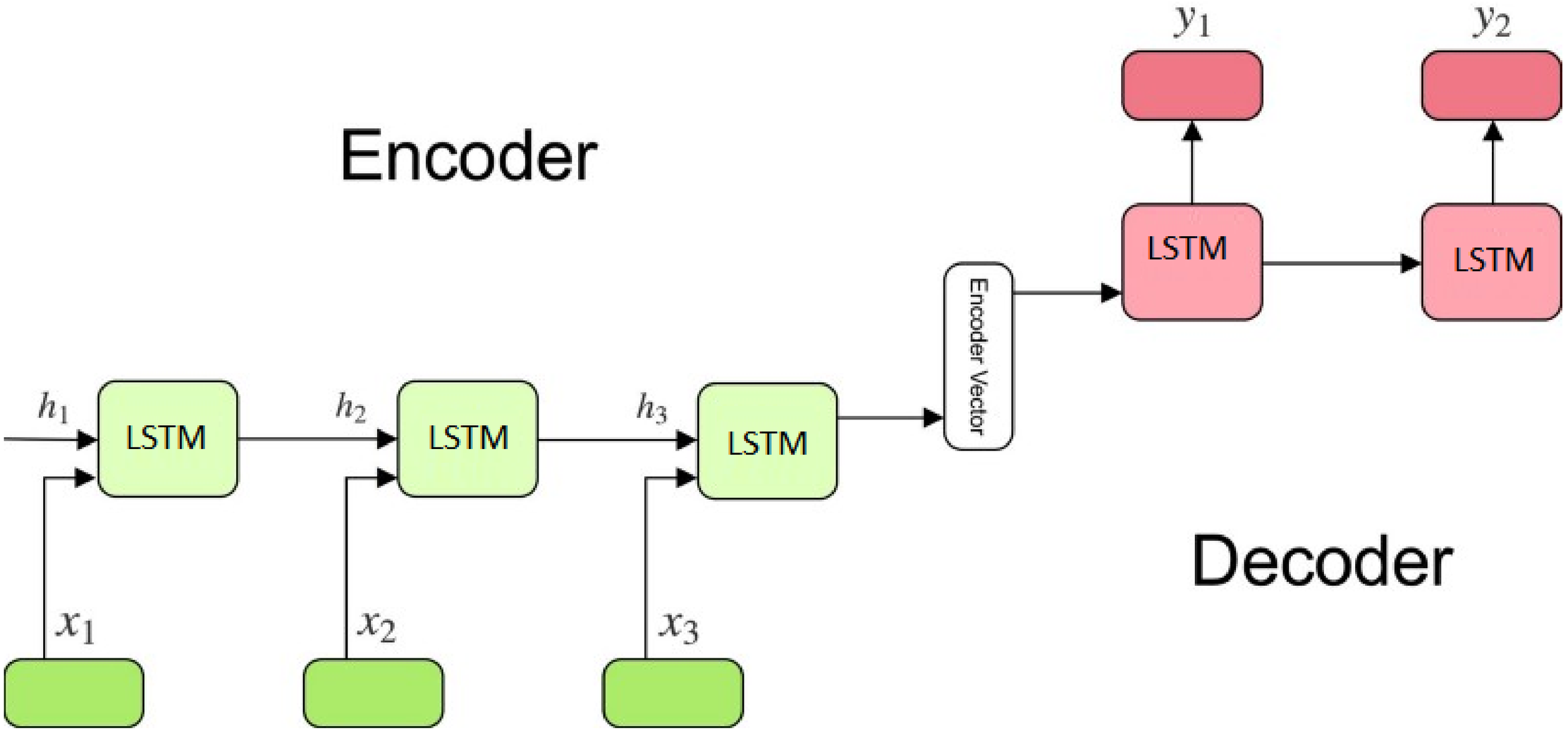
The solution step by step



AI CAMP

school of ai Algiers

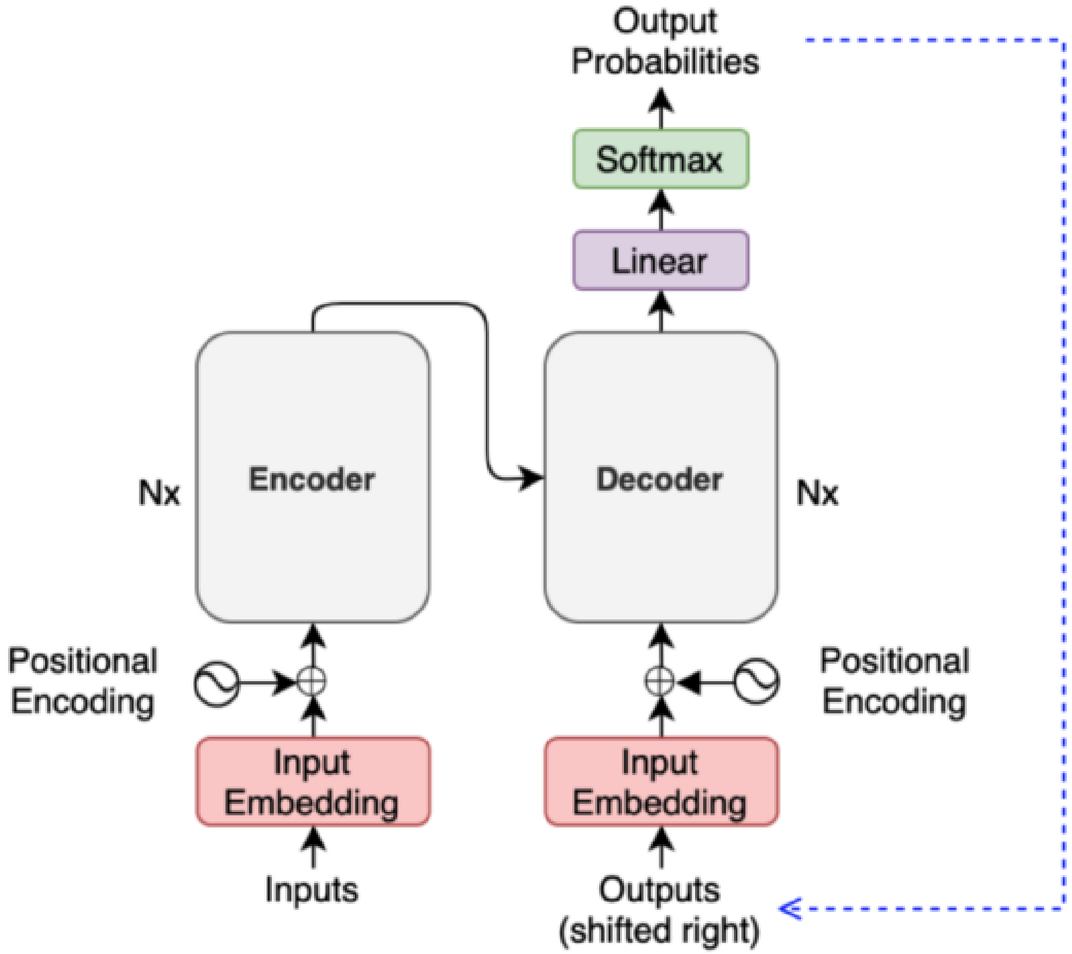
The solution step by step



AI CAMP

school of ai Algiers

The solution step by step



Transformers architecture

AI CAMP

school of ai Algiers

The solution step by step

Embeddings

Encoder-decoder

Attention

Transformers

AI CAMP

school of ai Algiers

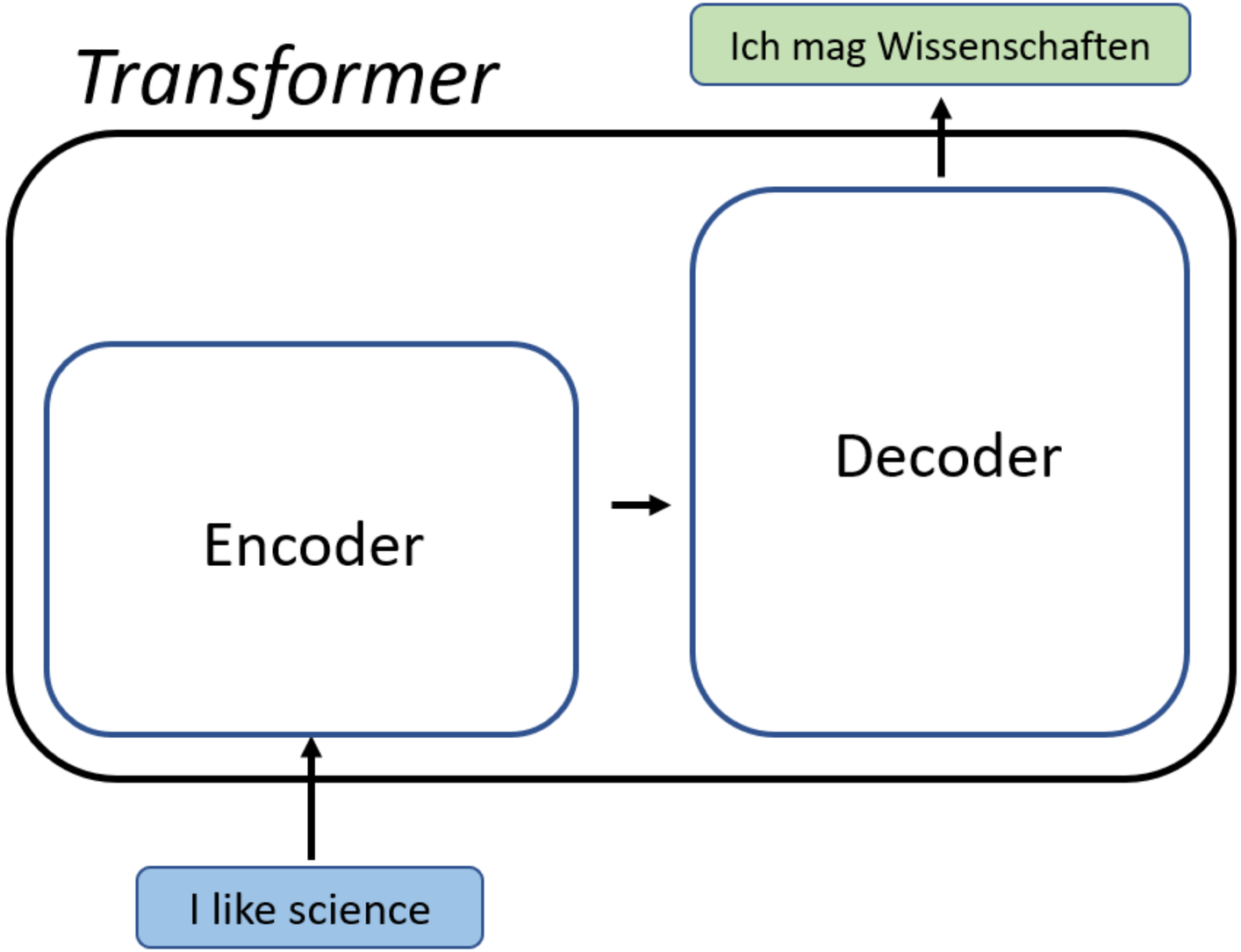
The solution step by step

Transformers

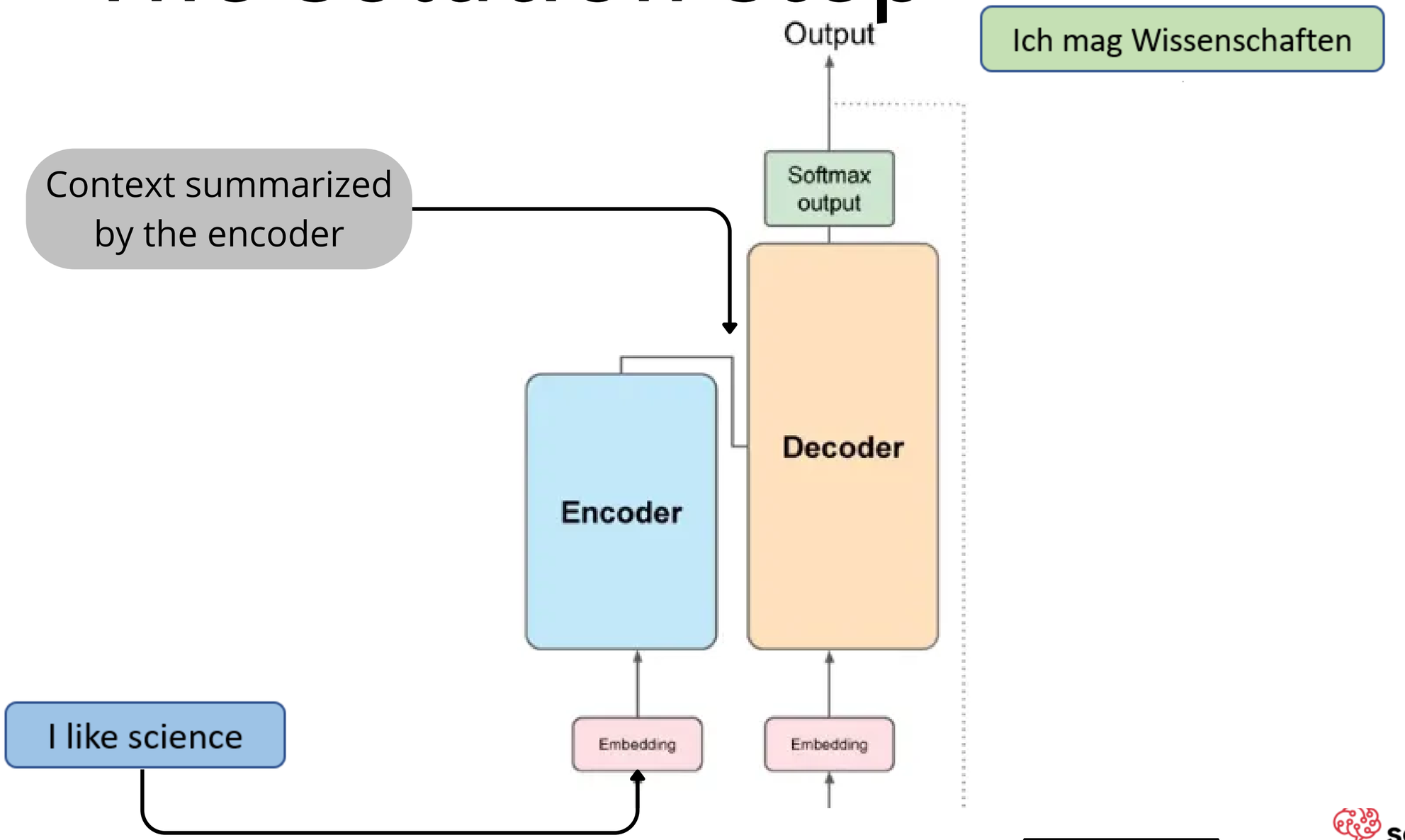
AI CAMP

school of ai Algiers

The solution step by step



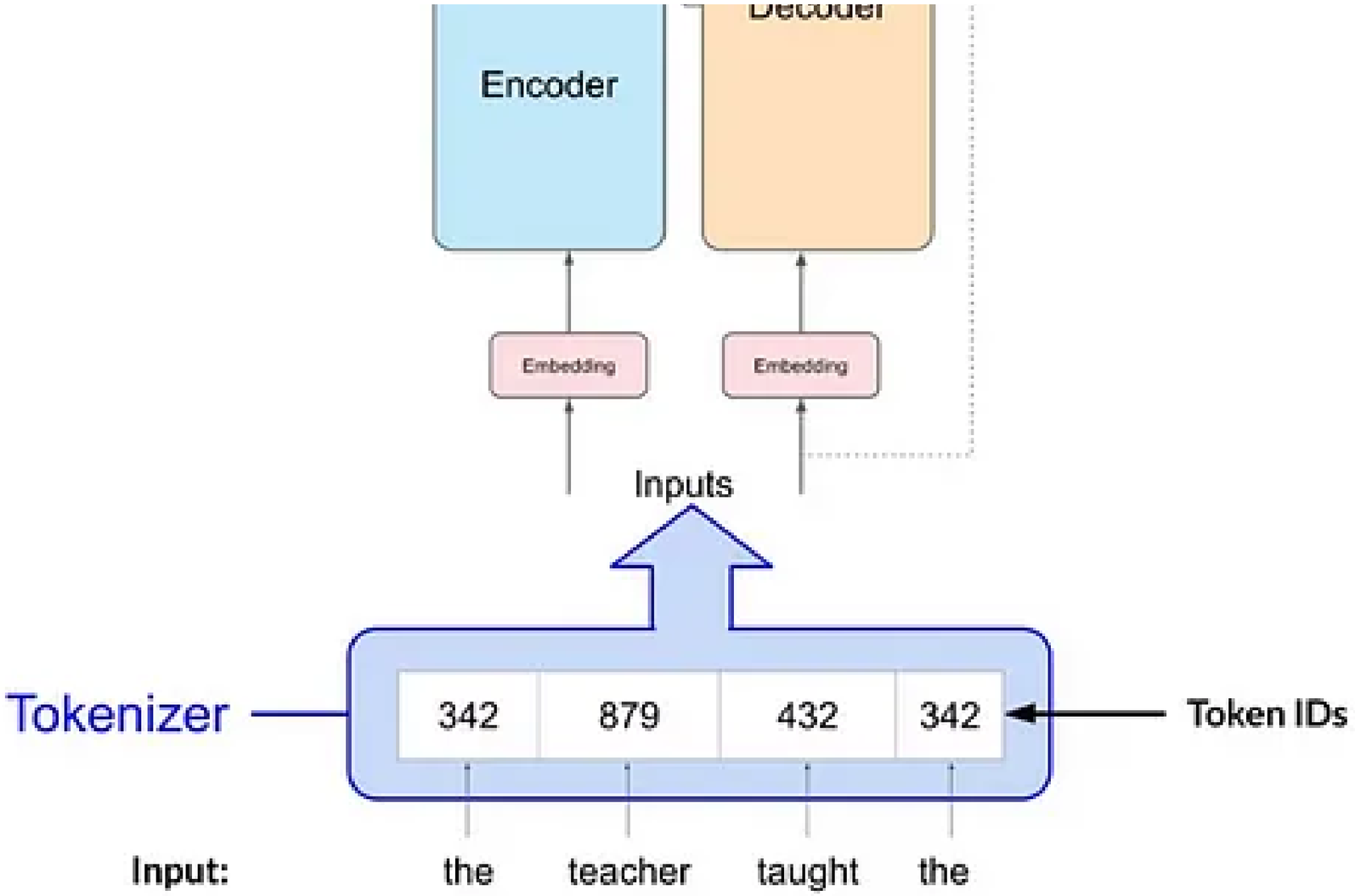
The solution step by step



AI CAMP

school of ai Algiers

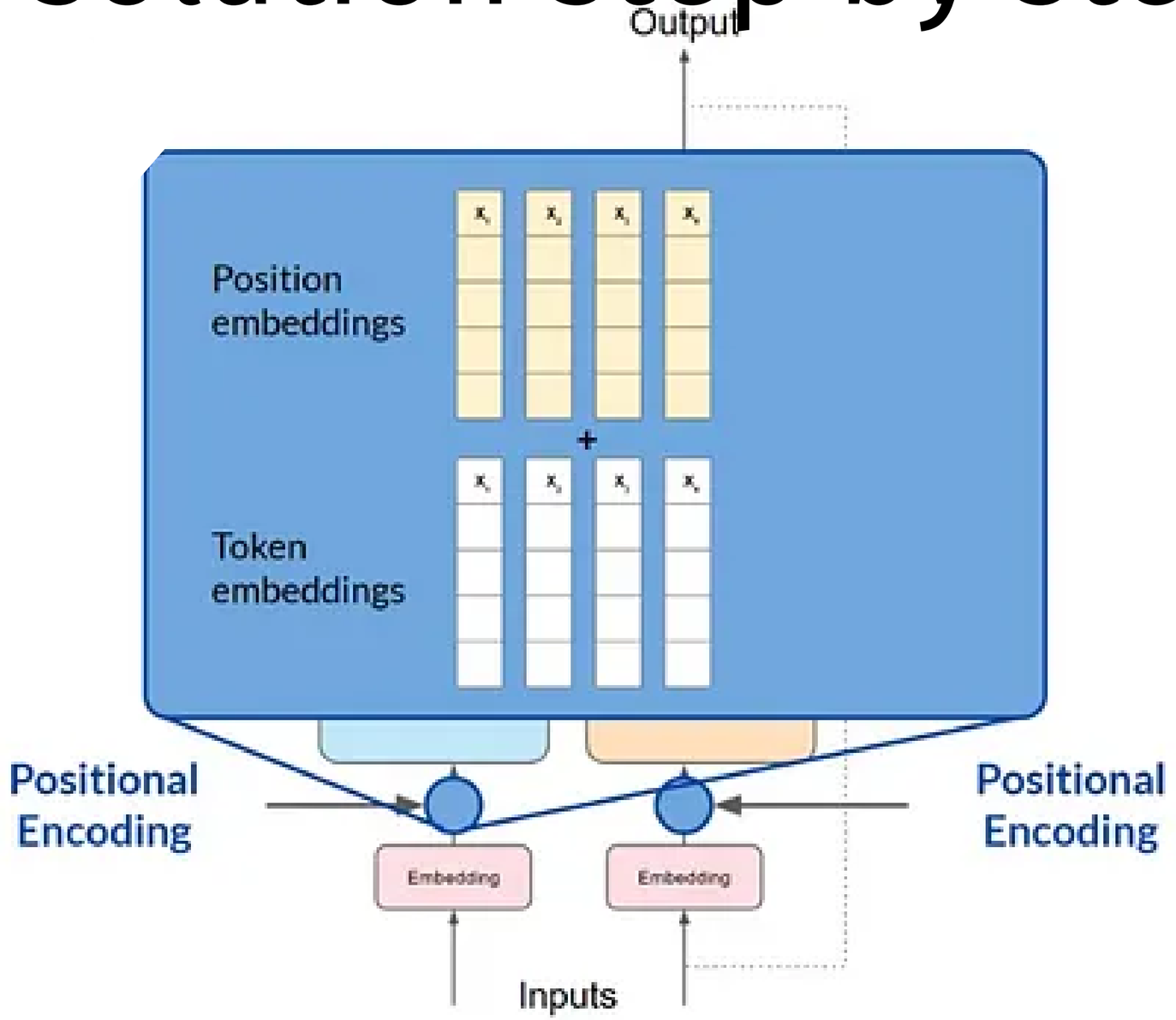
The solution step by step



AI CAMP

school of ai Algiers

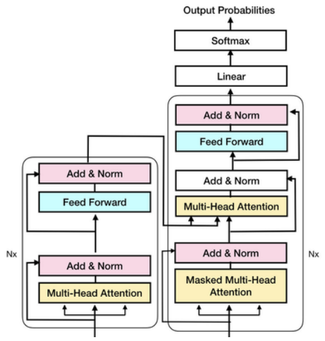
The solution step by step

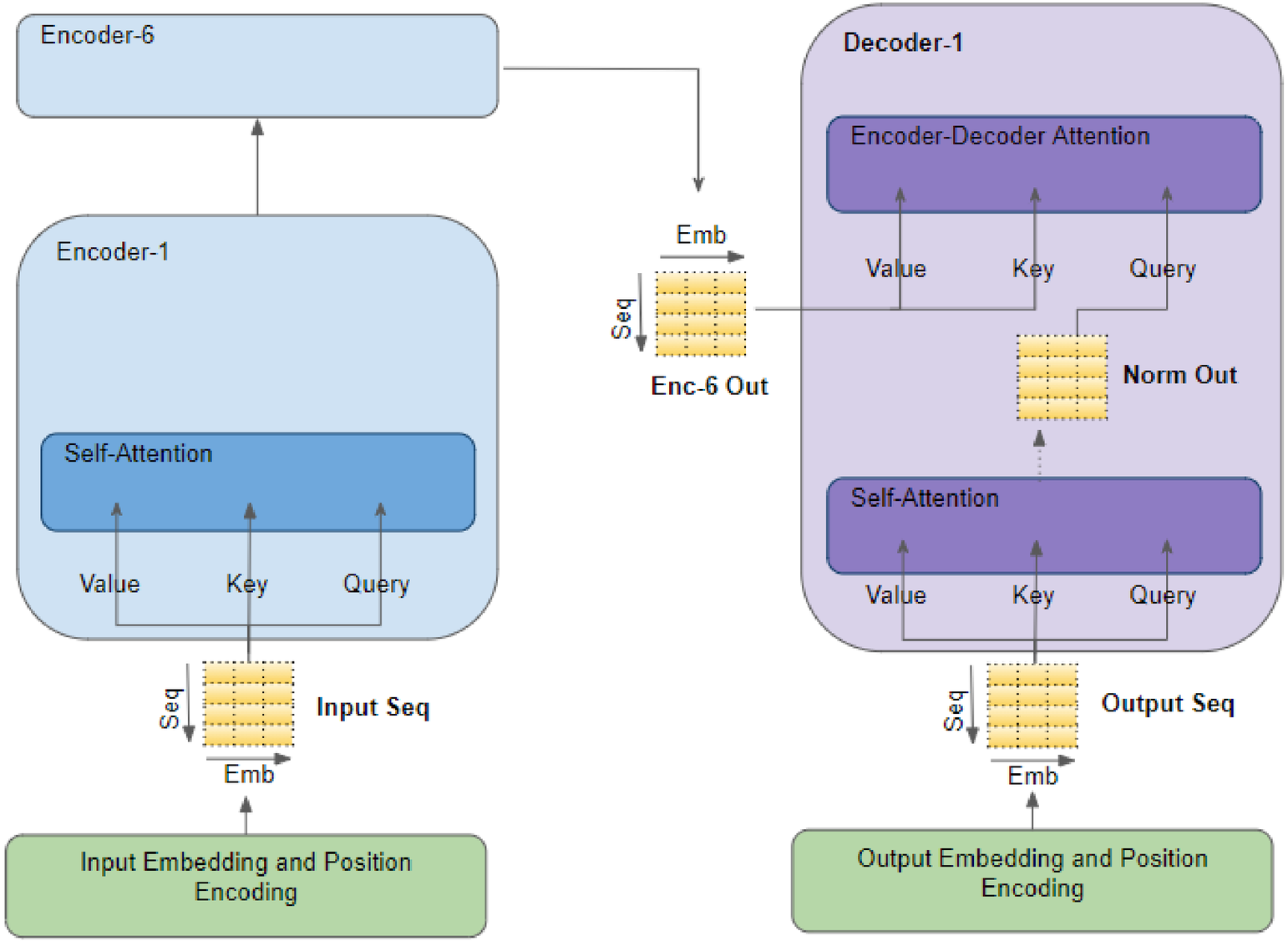


AI CAMP

school of ai Algiers

The solution step by step





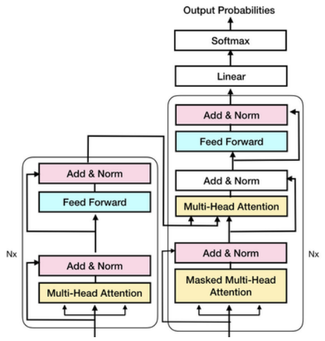
The most important part

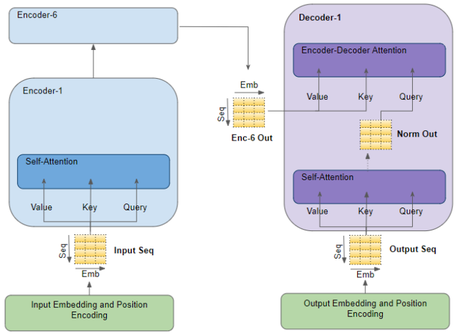
We will query out entities that are of the same context as the input !!!

AI CAMP

school of ai Algiers

The solution step by step





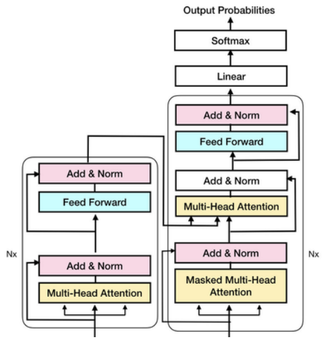
	YOUR	CAT	IS	A	LOVELY	CAT
YOUR	0.268	~~0.148~~ $-\infty$	~~0.134~~ $-\infty$	~~0.148~~	~~0.179~~	~~0.152~~
CAT	0.124	0.278	~~0.201~~	~~0.128~~	~~0.154~~	~~0.115~~
IS	0.147	0.132	0.262	~~0.097~~	~~0.218~~	~~0.145~~
A	0.210	0.128	0.206	0.212	~~0.119~~	~~0.125~~
LOVELY	0.146	0.158	0.152	0.143	0.227	~~0.17~~ $-\infty$
CAT	0.195	0.114	0.203	0.103	0.157	0.229

The tokens at a certain position must depend on previous tokens only. We don't want our model to see future tokens

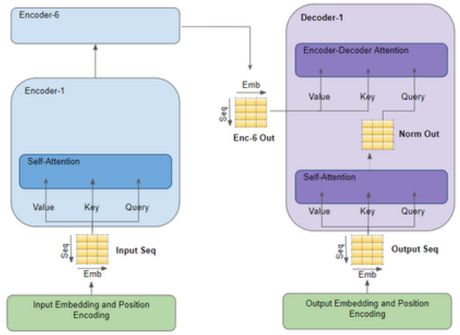
AI CAMP

school of ai Algiers

The solution step by step



	YOUR	CAT	IS	A	LOVELY	CAT
YOUR	0.268	~~0.119~~	~~0.134~~	~~0.148~~	~~0.179~~	~~0.152~~
CAT	0.124	0.278	~~0.201~~	~~0.128~~	~~0.154~~	~~0.115~~
IS	0.147	0.132	0.262	~~0.097~~	~~0.218~~	~~0.145~~
A	0.210	0.128	0.206	0.212	~~0.119~~	~~0.125~~
LOVELY	0.146	0.158	0.152	0.143	0.227	~~0.174~~
CAT	0.195	0.114	0.203	0.103	0.157	0.229



We have all the building blocks to train and infer with a transformers model !

AI CAMP

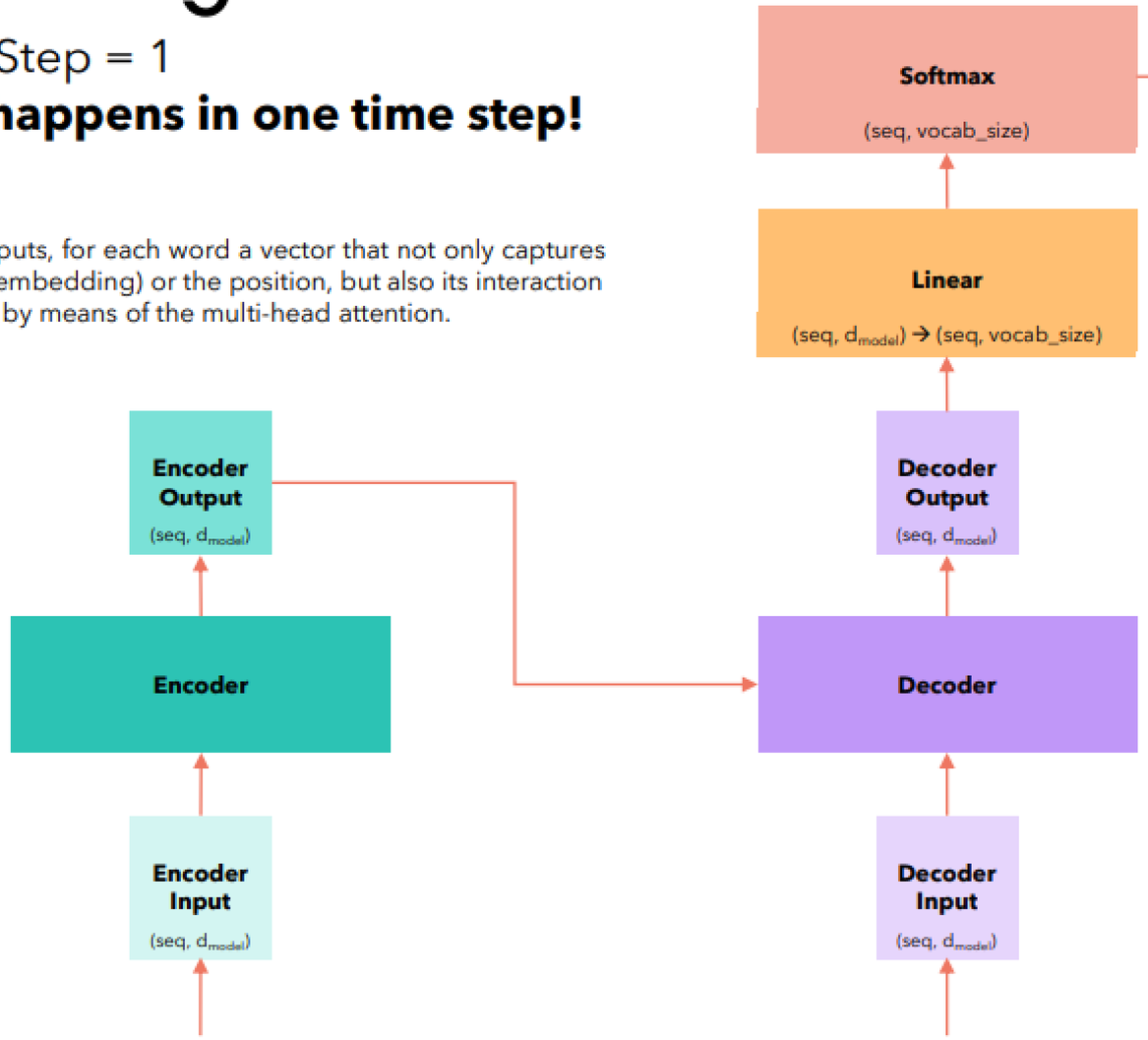
school of ai Algiers

Training

Time Step = 1

It all happens in one time step!

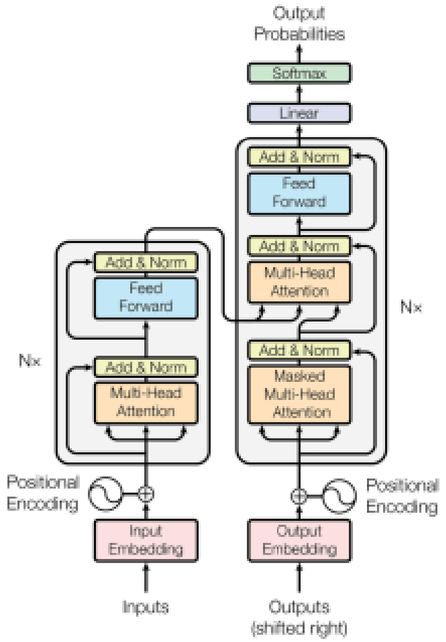
The encoder outputs, for each word a vector that not only captures its meaning (the embedding) or the position, but also its interaction with other words by means of the multi-head attention.



Ti amo molto <EOS>

* This is called the "label" or the "target"

Cross Entropy Loss



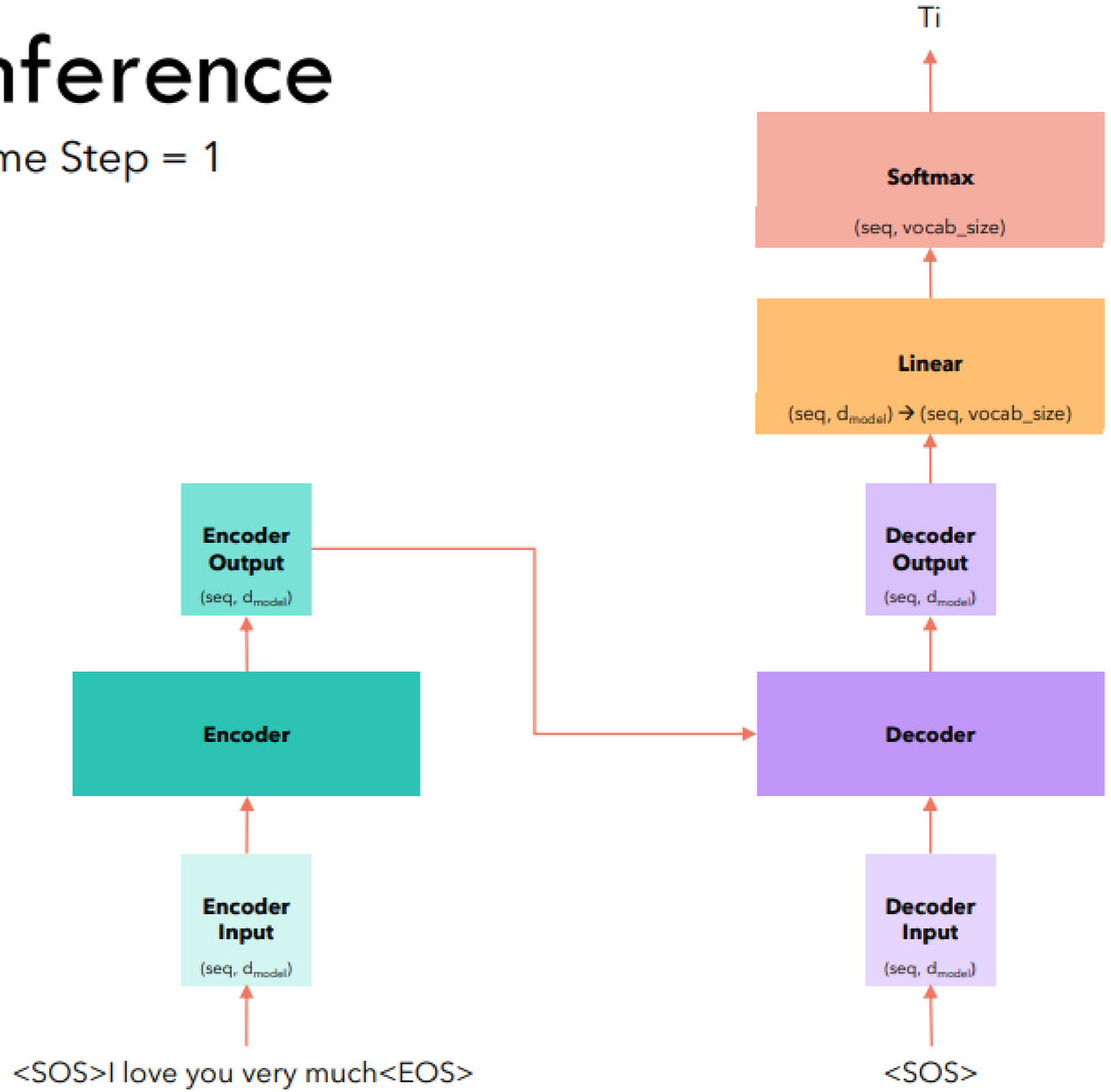
We prepend the <SOS> token at the beginning. That's why the paper says that the decoder input is shifted right.

AI CAMP

school of ai Algiers

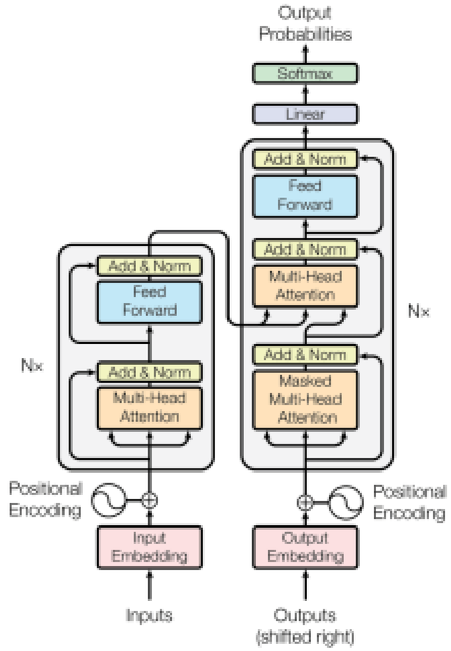
Inference

Time Step = 1



We select a token from the vocabulary corresponding to the position of the token with the maximum value.

The output of the last layer is commonly known as **logits**



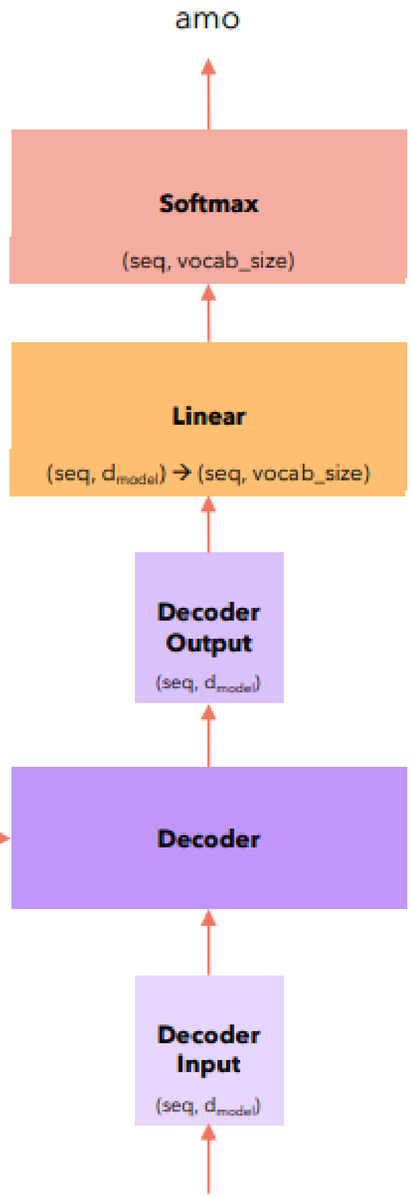
* Both sequences will have same length thanks to padding

AI CAMP

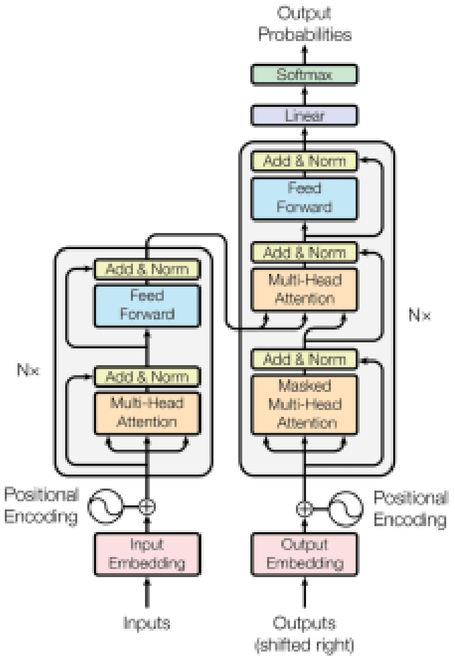
school of ai Algiers

Inference

Time Step = 2



Since decoder input now contains **two** tokens, we select the softmax corresponding to the second token.



Append the previously output word to the decoder input

AI CAMP

school of ai Algiers

The solution step by step

Embeddings

Encoder-decoder

Attention

Transformers

AI CAMP

school of ai Algiers

See you tomorrow for LLMs

AI CAMP

school of ai Algiers