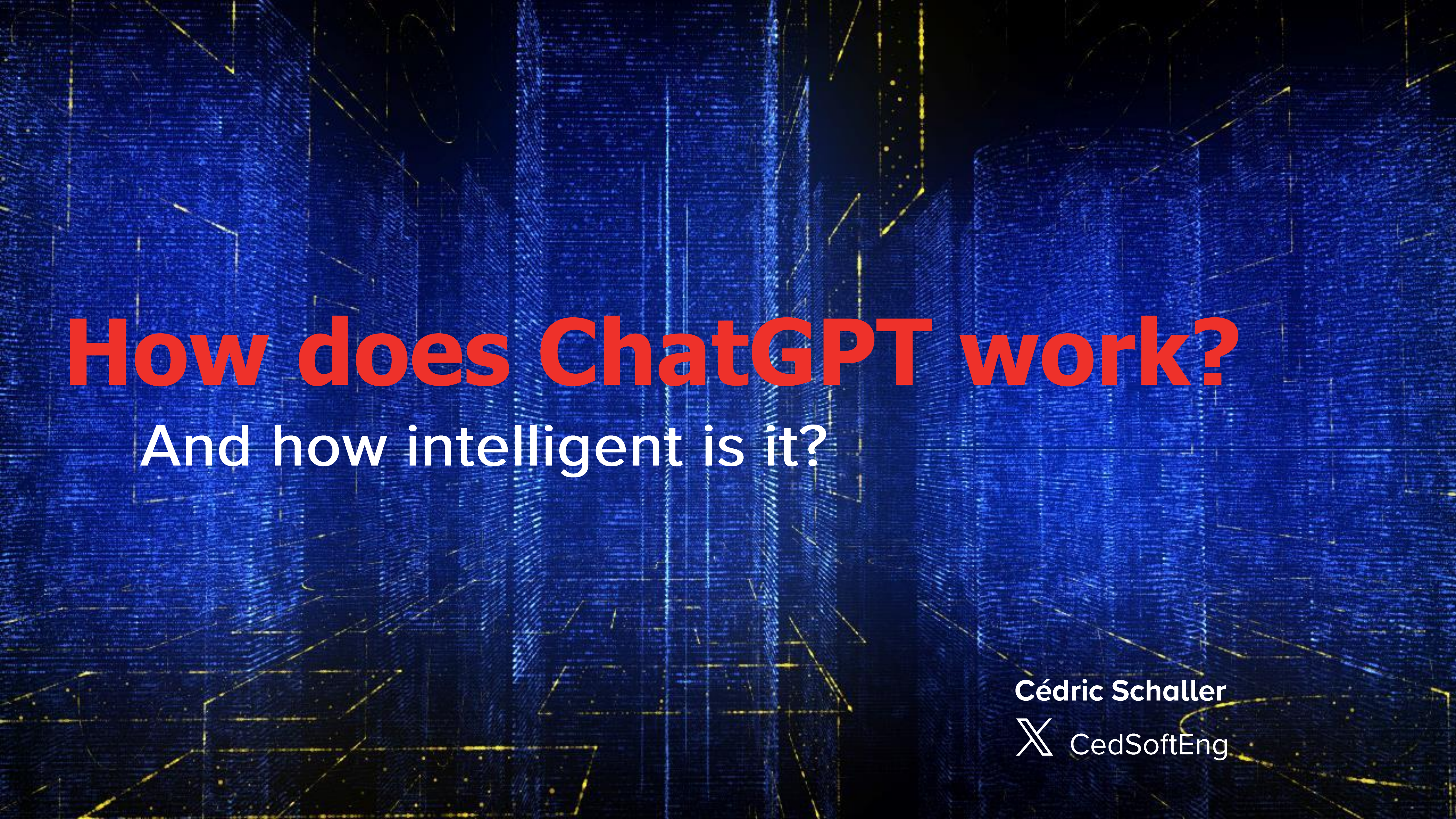


How does ChatGPT work?

And how intelligent is it?

Cédric Schaller

𝕏 CedSoftEng

Here we have a book, 9 eggs, a laptop, a bottle and a nail. Please tell me how to stack them onto each other in a stable manner.

GPT-4

- Place the book flat on a level surface, such as a table or a floor. The book will serve as the base of the stack and provide a large and sturdy support.
- Arrange the 9 eggs in a 3 by 3 square on top of the book [...] Make sure the eggs are not cracked or broken, and handle them gently to avoid dropping them.
- Place the laptop on top of the eggs [...]
- Place the bottle on top of the laptop, with the cap facing up and the bottom facing down. [...]
- Place the nail on top of the bottle cap, with the pointy end facing up and the flat end facing down. The nail will be the final and smallest object in the stack, and its sharp tip will prevent it from rolling or sliding off the bottle. [...]

ELCA Tech Days

Source: https://doi.org/10.48550/arXiv.2303.12712

How much is 173 * 863?

GPT-4

173 multiplied by 863 equals 149,309. // it is in fact 149,299

```
      173
x     863
---------
     1519    (173 x 3)
     5190    (173 x 60)
   138600    (173 x 800)
---------
  149,309
```

Agenda

HOW DOES CHATGPT WORK?

LARGE LANGUAGE MODELS (LLM)

- It's all about architecture!
- The transformer architecture
- Tokenization and embeddings
- Attention
- Reinforcement Learning From Human Feedback (RLHF)

DISCUSSION ON TRANSFORMERS

- Potential
- Shortcomings and dangers

HOW INTELLIGENT IS CHATGPT?

DISCUSSION ON ARTIFICIAL GENERAL INTELLIGENCE (AGI)

Just guessing the next token

Large Language Models (LLM) are neural networks trained to guess the next token in a sequence.

Eighty percent of success is showing

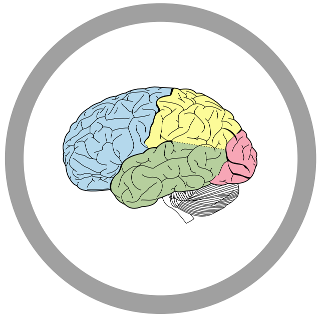
up.	12%
off.	5%
through.	2%
respect.	0.1%

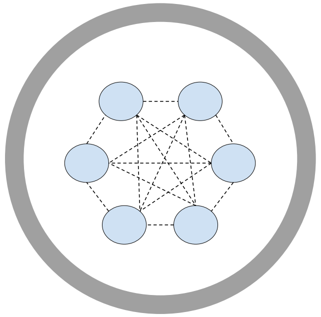
It’s all about architecture!

HUMAN BRAIN

Each region specializes in a specific domain.

Example: visual cortex (red)





FULLY CONNECTED NEURAL NETWORKS ARE NOT MANAGEABLE AT SCALE

`O(n^2)` connections for n neurons

i.e. 10^12 connections for a million neurons

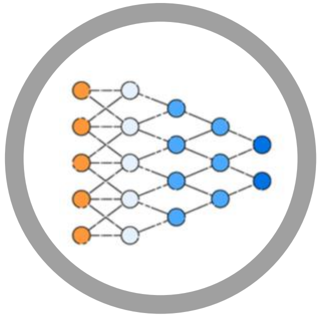
CONVOLUTIONAL NEURAL NETWORKS (CNN)

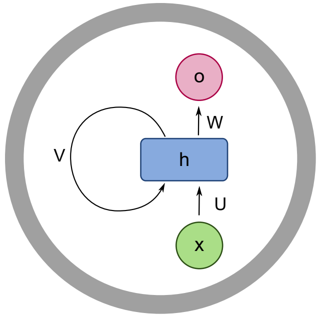
Widely used for image classification.

(1) CNN extract features (e.g. edges, spots).

(2) These features are combined into higher-level features (e.g. eyes).

(3) These higher-level features are then aggregated into faces, cars, etc.



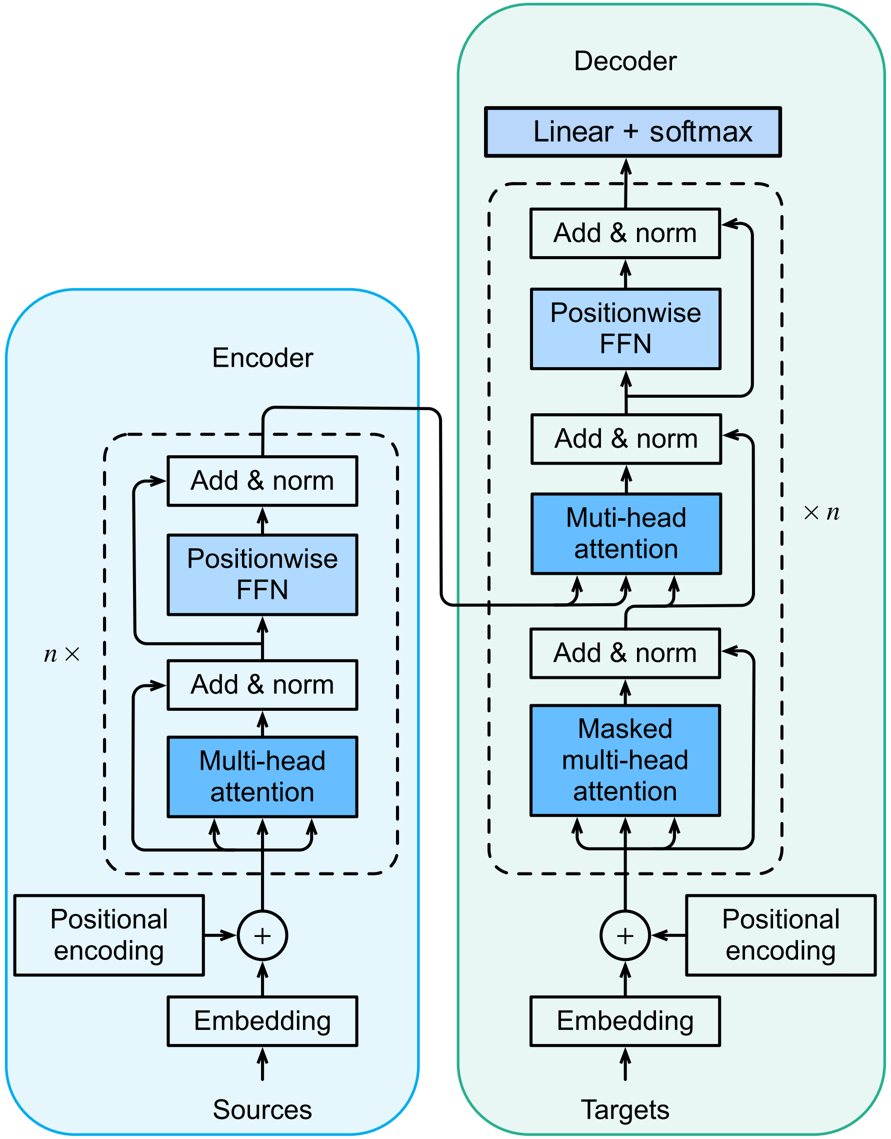


RECURRENT NEURAL NETWORKS (RNN)

Were widely used in Natural Language Processing (NLP).

Cannot be trained in parallel due to their sequential nature.

The Transformer architecture



THE PAPER

Attention is all you need (2017)

Authors mainly from Google

Was focused on translation.

A LANDMARK IN AI HISTORY

Many innovations (self-attention, multi-head attention, positional encoding)

Allows a high degree of parallelization

Cited ~ 90’000 times

Architecture has remained nearly unchanged since then.

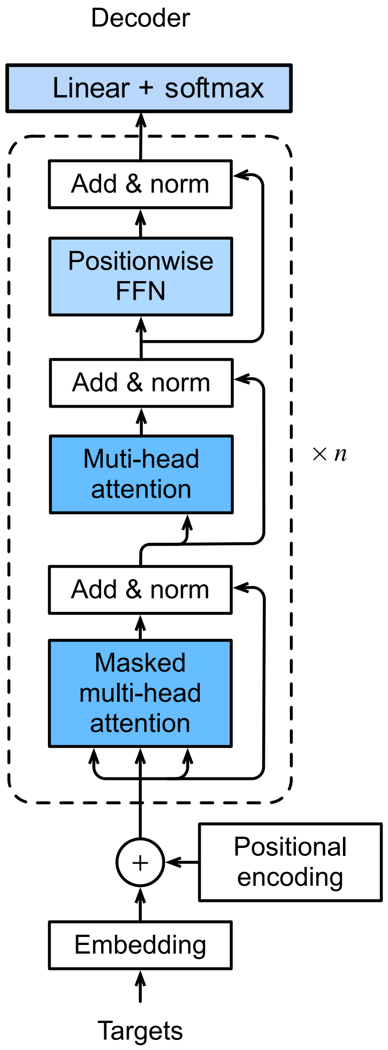
ENCODERS AND DECODERS

Can also use the encoder only (e.g. sentiment analysis).

Conversational agents (like ChatGPT) use only the decoder part.

Source: https://arxiv.org/abs/1706.03762

The decoder



Decoder

Linear + softmax

Add & norm

Positionwise FFN

Add & norm

Muti-head attention

Add & norm

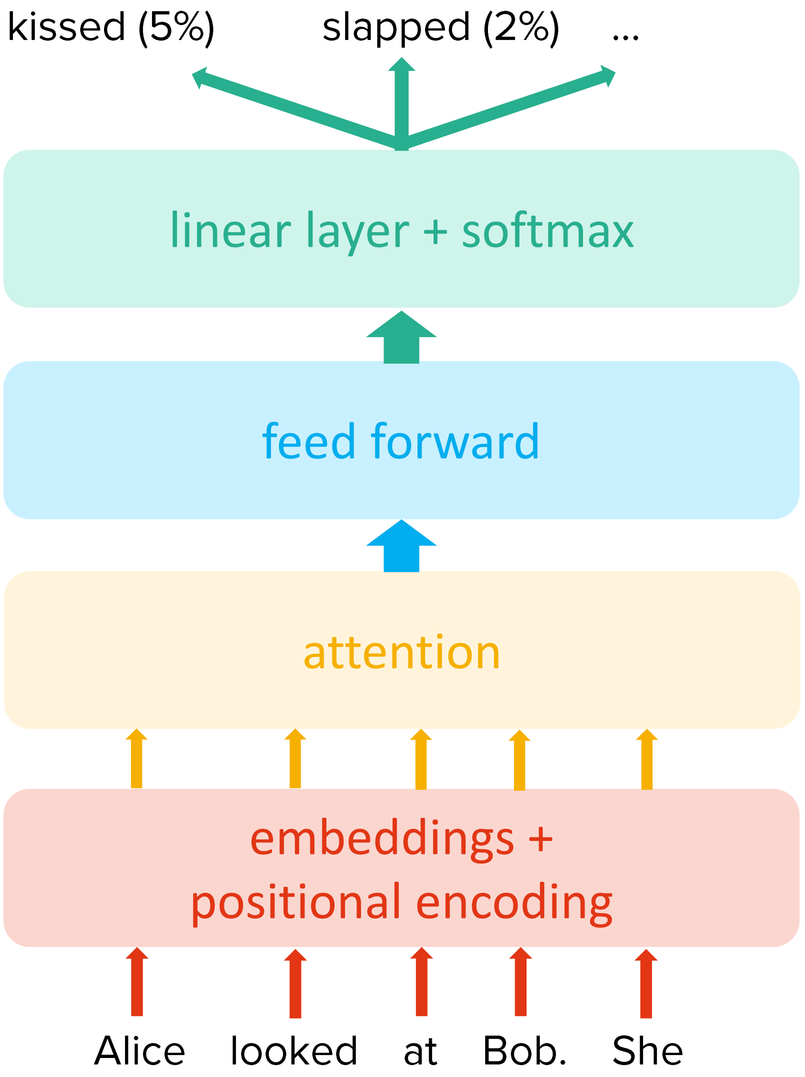
Masked multi-head attention

× *n*

Positional encoding

Embedding

Targets



kissed (5%)

slapped (2%)

...

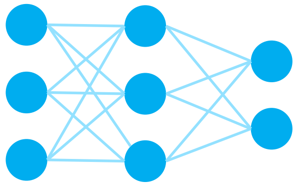
linear layer + softmax

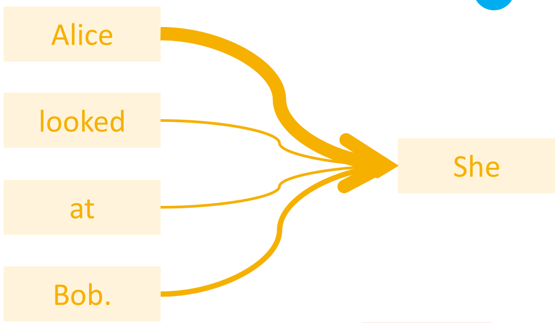
feed forward

attention

embeddings + positional encoding

Alice looked at Bob. She





Alice

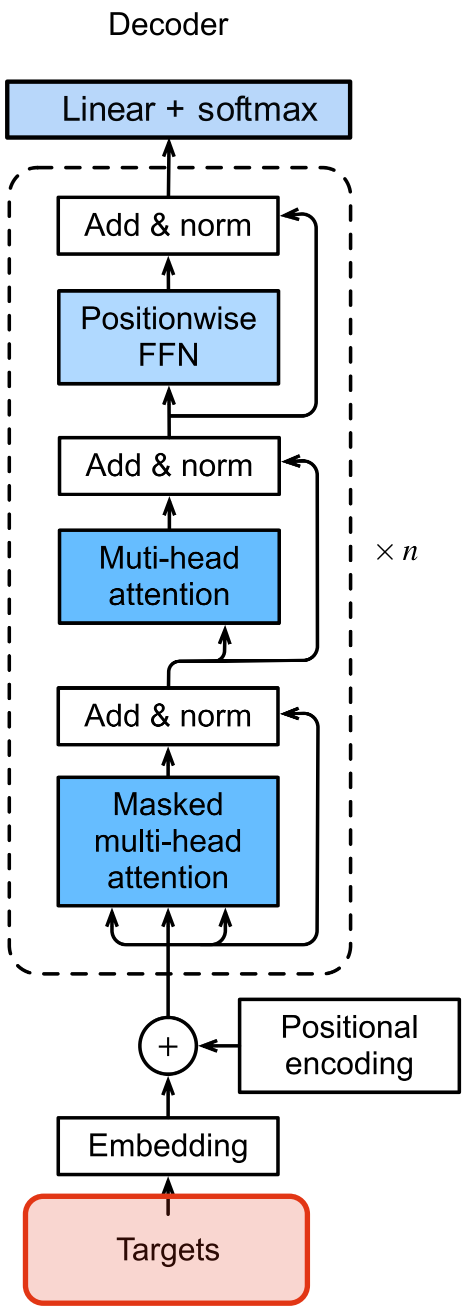
looked

at

Bob.

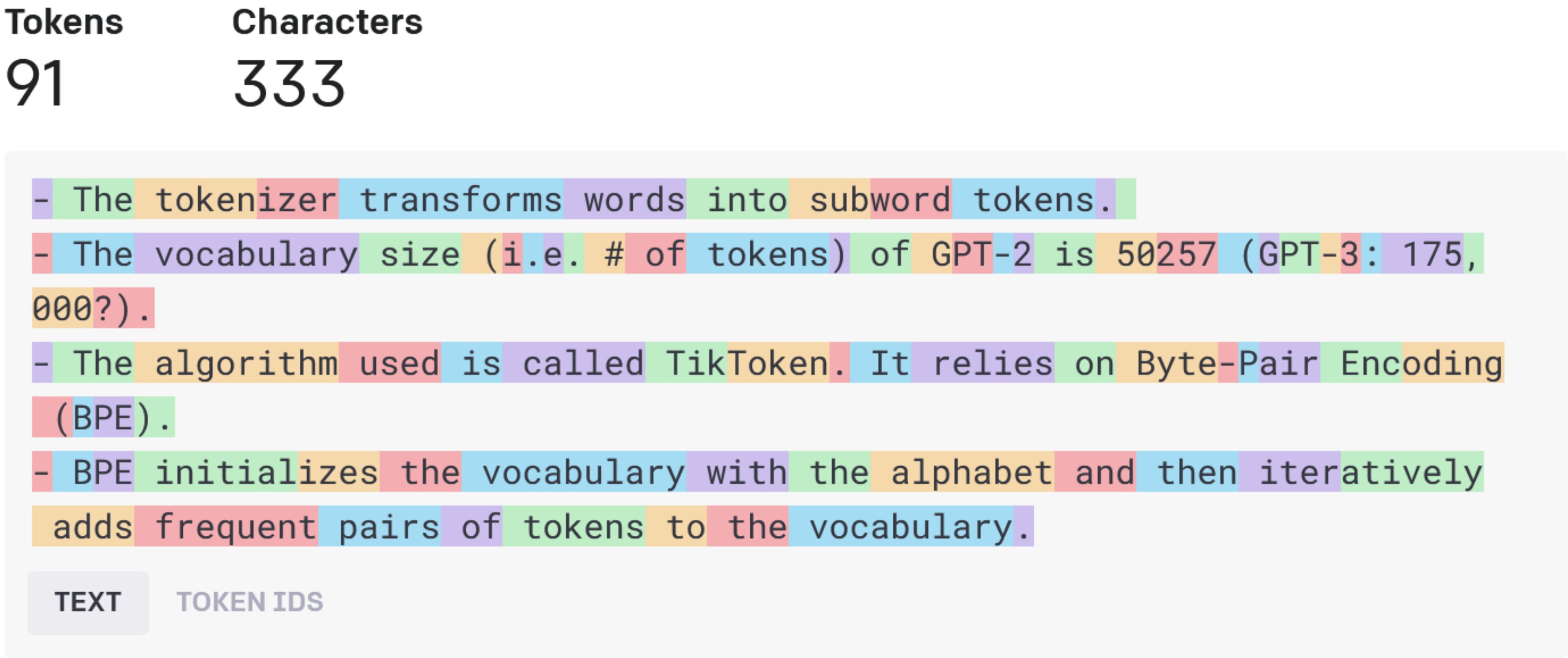
She

Tokenization



ELCA Tech Days

Tokenization



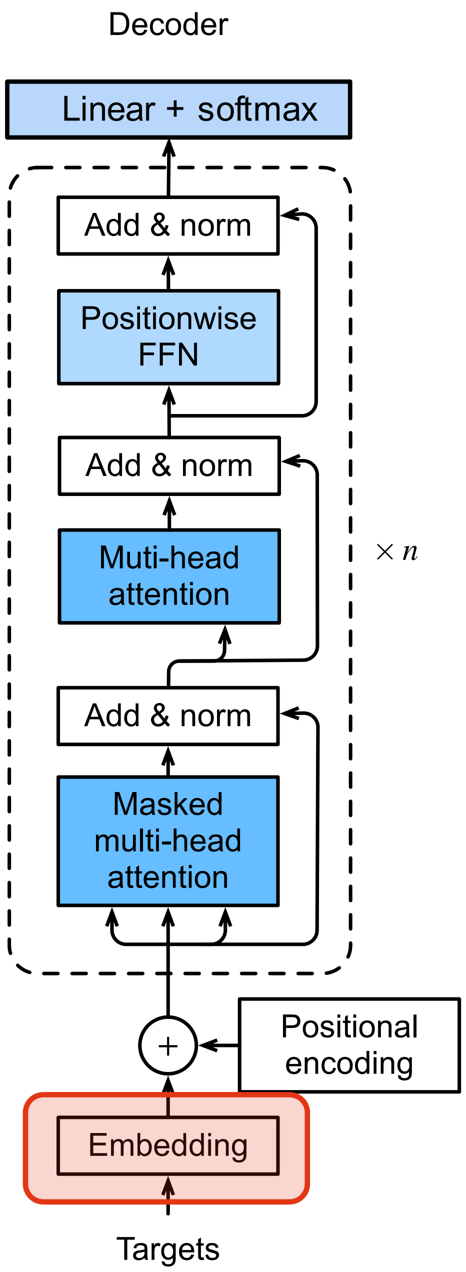
Tokens
91

Characters
333

- The tokenizer transforms words into subword tokens.
- The vocabulary size (i.e. # of tokens) of GPT-2 is 50257 (GPT-3: 175,000?).
- The algorithm used is called TikToken. It relies on Byte-Pair Encoding (BPE).
- BPE initializes the vocabulary with the alphabet and then iteratively adds frequent pairs of tokens to the vocabulary.

Source: https://platform.openai.com/tokenizer

Embeddings


Decoder
Linear + softmax
Add & norm
Positionwise FFN
Add & norm
Muti-head attention
Add & norm
Masked multi-head attention
× n
+
Positional encoding
Embedding
Targets

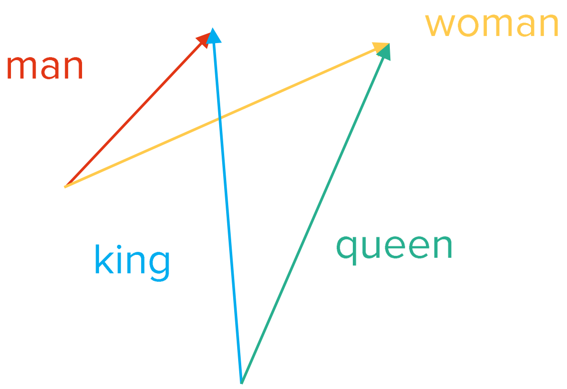
WHY EMBEDDINGS

Neural networks work with vectors.

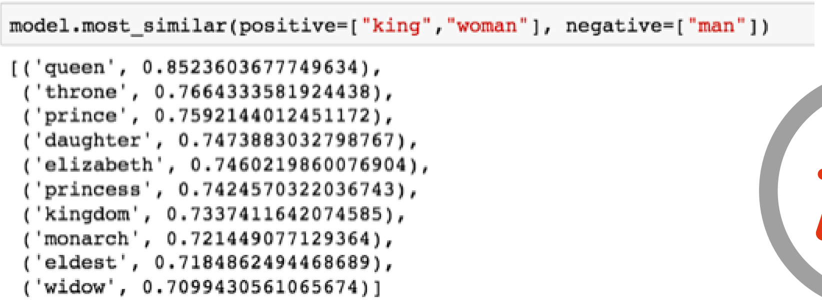
Embeddings are vectors that capture the semantics of the words/tokens in an n-dimensional space.

THE INTUITION

king - man + woman = queen



THE IMPLEMENTATION



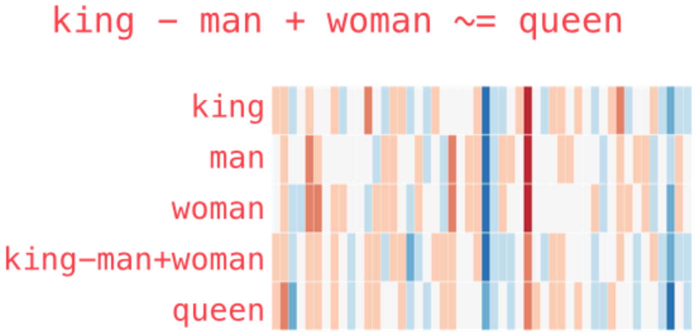
```
model.most_similar(positive=["king","woman"], negative=["man"])
```

```
[('queen', 0.8523603677749634),
 ('throne', 0.7664333581924438),
 ('prince', 0.7592144012451172),
 ('daughter', 0.7473883032798767),
 ('elizabeth', 0.7460219860076904),
 ('princess', 0.7424570322036743),
 ('kingdom', 0.7337411642074585),
 ('monarch', 0.721449077129364),
 ('eldest', 0.7184862494468689),
 ('widow', 0.7099430561065674)]
```

Similarity is measured by the angle between the vectors (cosine similarity).

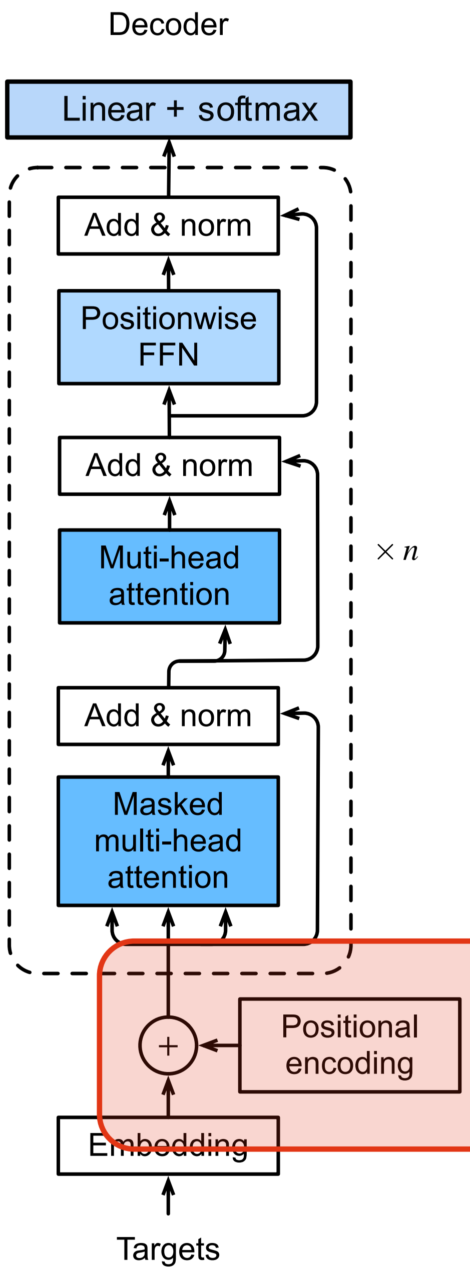
OpenAI are very secretive about their embedding algorithm as it gives them a huge competitive advantage.

THE VISUALIZATION



Source: https://kazemnejad.com/blog/transformer_architecture_positional_encoding/

Positional encoding



The position is encoded as a vector which is added to the embedding vector.

WHY POSITIONAL ENCODING?

The transformer processes the tokens in parallel and it has therefore no sense of positioning.

This information must be included somehow in the vector representing the token.

THE IMPLEMENTATION

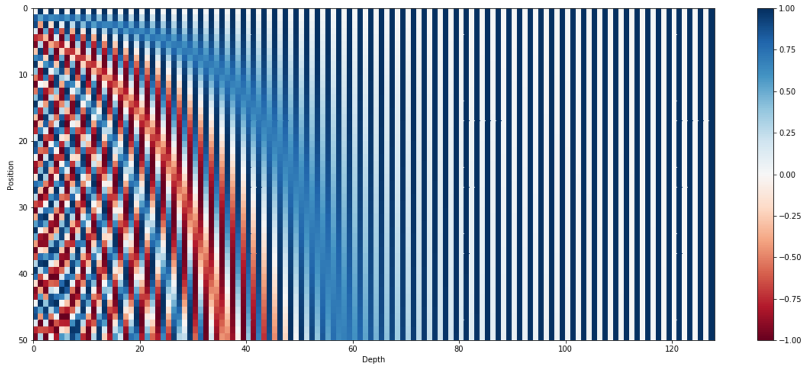
$$\vec{p_t} = \begin{bmatrix} \sin(\omega_1 . t) \\ \cos(\omega_1 . t) \\ \\ \sin(\omega_2 . t) \\ \cos(\omega_2 . t) \\ \\ \vdots \\ \\ \sin(\omega_{d/2} . t) \\ \cos(\omega_{d/2} . t) \end{bmatrix}_{d \times 1}$$

THE INTUITION

0 : 0 0 0 0
1 : 0 0 0 1
2 : 0 0 1 0
3 : 0 0 1 1
4 : 0 1 0 0
5 : 0 1 0 1
6 : 0 1 1 0
7 : 0 1 1 1

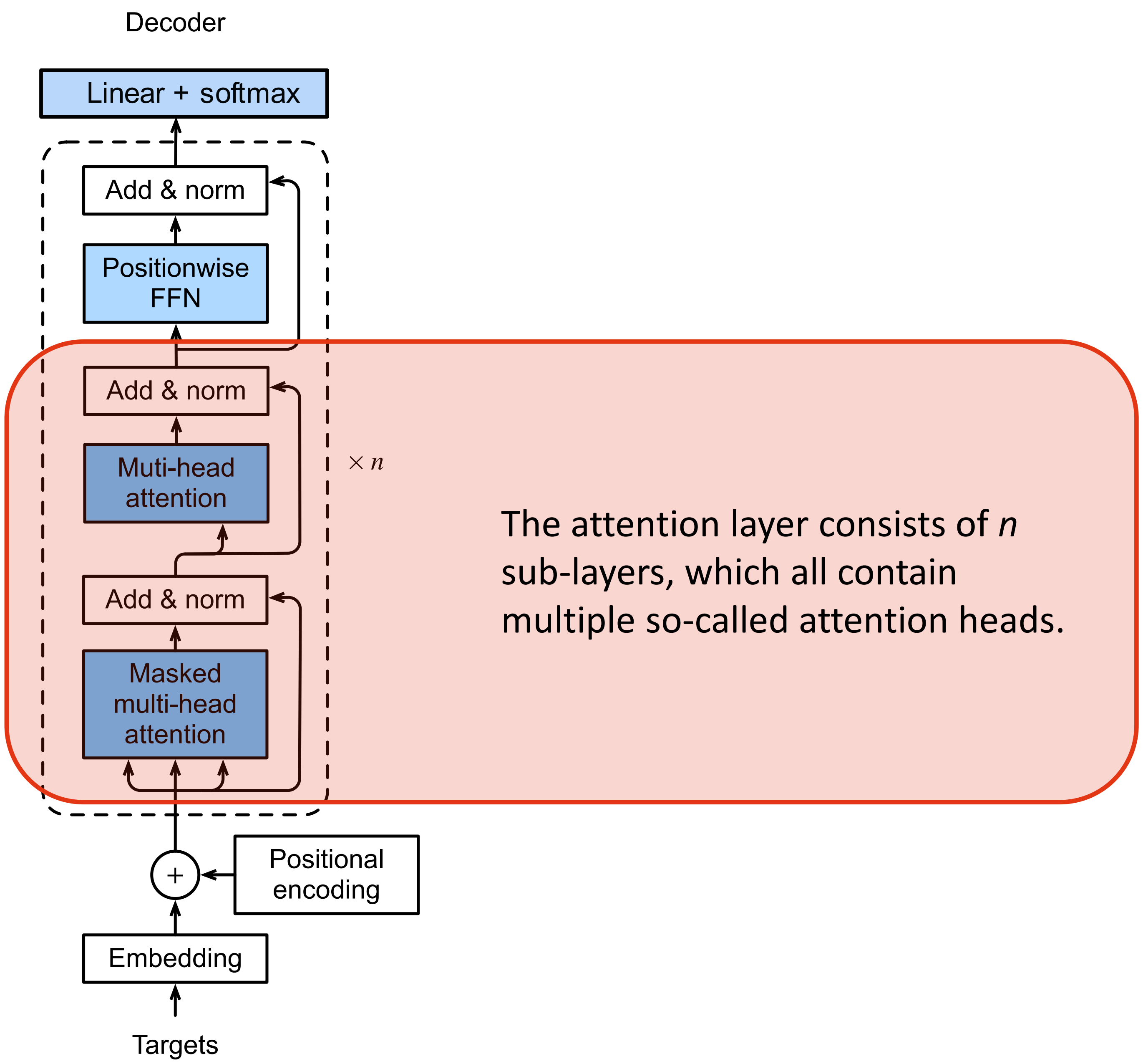
8 : 1 0 0 0
9 : 1 0 0 1
10 : 1 0 1 0
11 : 1 0 1 1
12 : 1 1 0 0
13 : 1 1 0 1
14 : 1 1 1 0
15 : 1 1 1 1

THE VISUALIZATION



Source: https://kazemnejad.com/blog/transformer_architecture_positional_encoding/

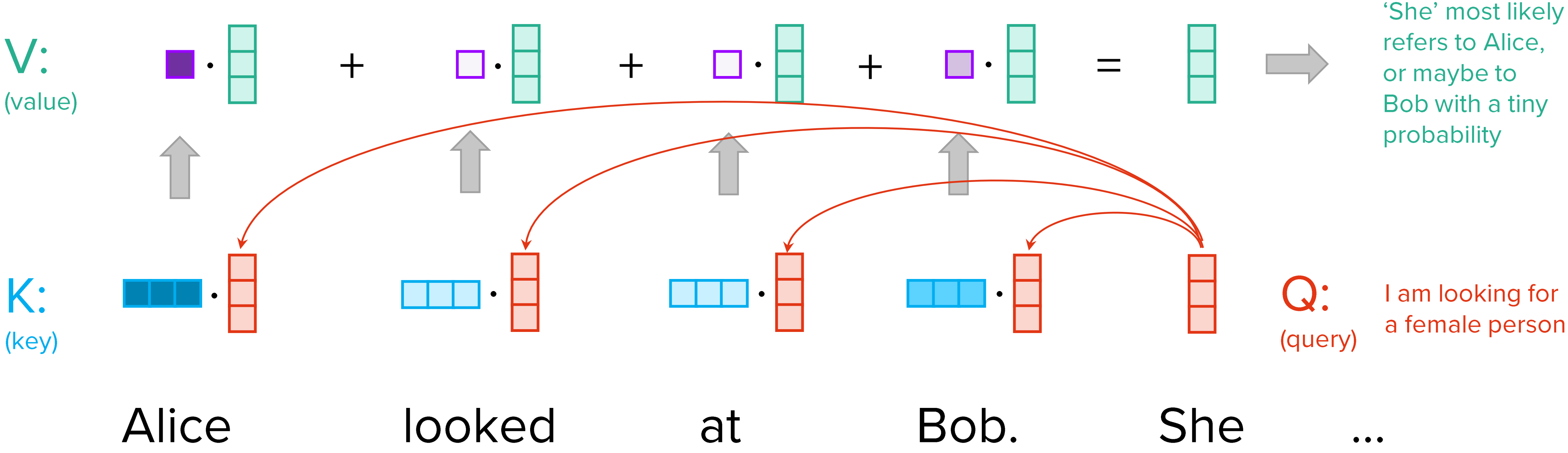
Multi-head attention



The attention layer consists of *n* sub-layers, which all contain multiple so-called attention heads.

ELCA Tech Days

(Self-)Attention


V:
(value)
+
+
+
=
'She' most likely refers to Alice, or maybe to Bob with a tiny probability
K:
(key)
Q:
(query)
I am looking for a female person
Alice
looked
at
Bob.
She
...

Q: (query)	What the query head is interested in (female?, happy?, ...).		K: (key)	How strongly the given token matches the query		V: (value)	The information we want to pass on to the next layer

ELCA Tech Days

But how does it answer questions?

Write a poem about bread and cheese.

Write a poem about angel food cake.
Write a poem about someone who choked on a ham sandwich.

BASE MODELS VS. ASSISTANTS

- Base models do not answer questions. They only want to find the next token for a given sequence.
- A base model can be trained to become an assistant with Reinforcement Learning from Human Feedback (RLHF).

RLHF

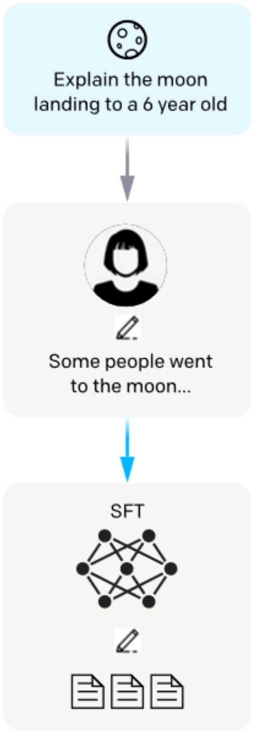
Step 1

Collect demonstration data, and train a supervised policy.

A prompt is sampled from our prompt dataset.

A labeler demonstrates the desired output behavior.

This data is used to fine-tune GPT-3 with supervised learning.


Explain the moon landing to a 6 year old
Some people went to the moon...
SFT

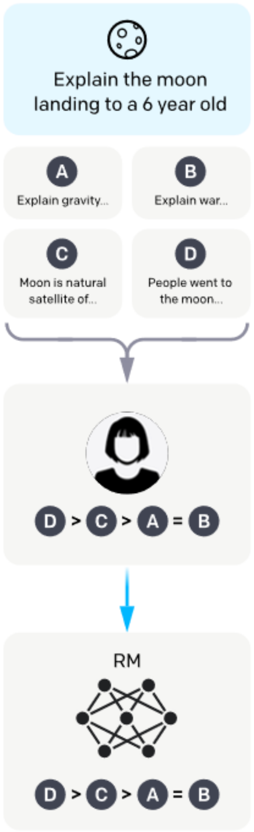
Step 2

Collect comparison data, and train a reward model.

A prompt and several model outputs are sampled.

A labeler ranks the outputs from best to worst.

This data is used to train our reward model.


Explain the moon landing to a 6 year old
A
Explain gravity...
B
Explain war...
C
Moon is natural satellite of...
D
People went to the moon...
D > C > A = B
RM
D > C > A = B

Step 3

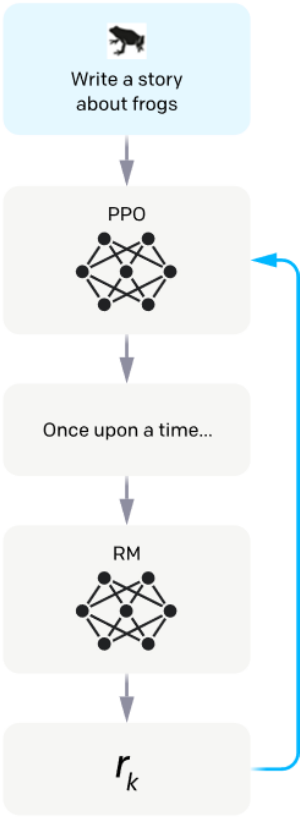
Optimize a policy against the reward model using reinforcement learning.

A new prompt is sampled from the dataset.

The policy generates an output.

The reward model calculates a reward for the output.

The reward is used to update the policy using PPO.



Source: https://openai.com/research/instruction-following

Agenda

HOW DOES CHATGPT WORK?

LARGE LANGUAGE MODELS (LLM)

- It’s all about architecture!
- The transformer architecture
- Tokenization and embeddings
- Attention
- Reinforcement Learning From Human Feedback (RLHF)

DISCUSSION ON TRANSFORMERS

- Potential
- Shortcomings and dangers

HOW INTELLIGENT IS CHATGPT?

DISCUSSION ON ARTIFICIAL GENERAL INTELLIGENCE (AGI)

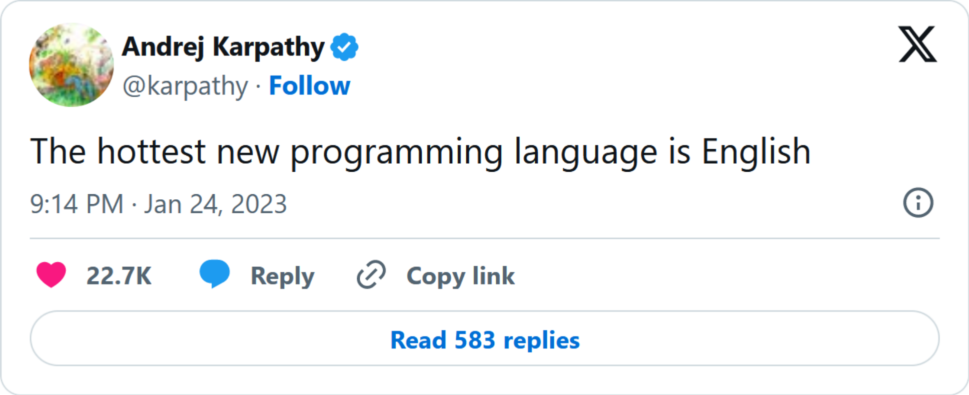
From training and fine-tuning to prompting

Before

An AI project typically looked like this:

1. Collect training data
2. Choose the proper model/architecture
3. Train the model
4. Fine-tune the model
5. Operationalize
6. Restart from scratch for the next project

Now



Andrej Karpathy
@karpathy · Follow

The hottest new programming language is English

9:14 PM · Jan 24, 2023

Hallucinations

Example

Who was the sole survivor of the Titanic?

ChatGPT

The sole survivor of the Titanic disaster among the ship's passengers was a baby named Elizabeth Gladys Dean, who was later known as "Millvina" Dean. […]

Why?

- Transformers have **no notion of truth**. They just want to **complete a sequence of tokens**.
- A few “wrong” tokens can make the process completely **derail**.
- This typically happens when the probability distribution is not sharp enough (due to **lack of training data**) or when the **provided context** “tricks” the transformer (as illustrated on the left).

?

Mitigation

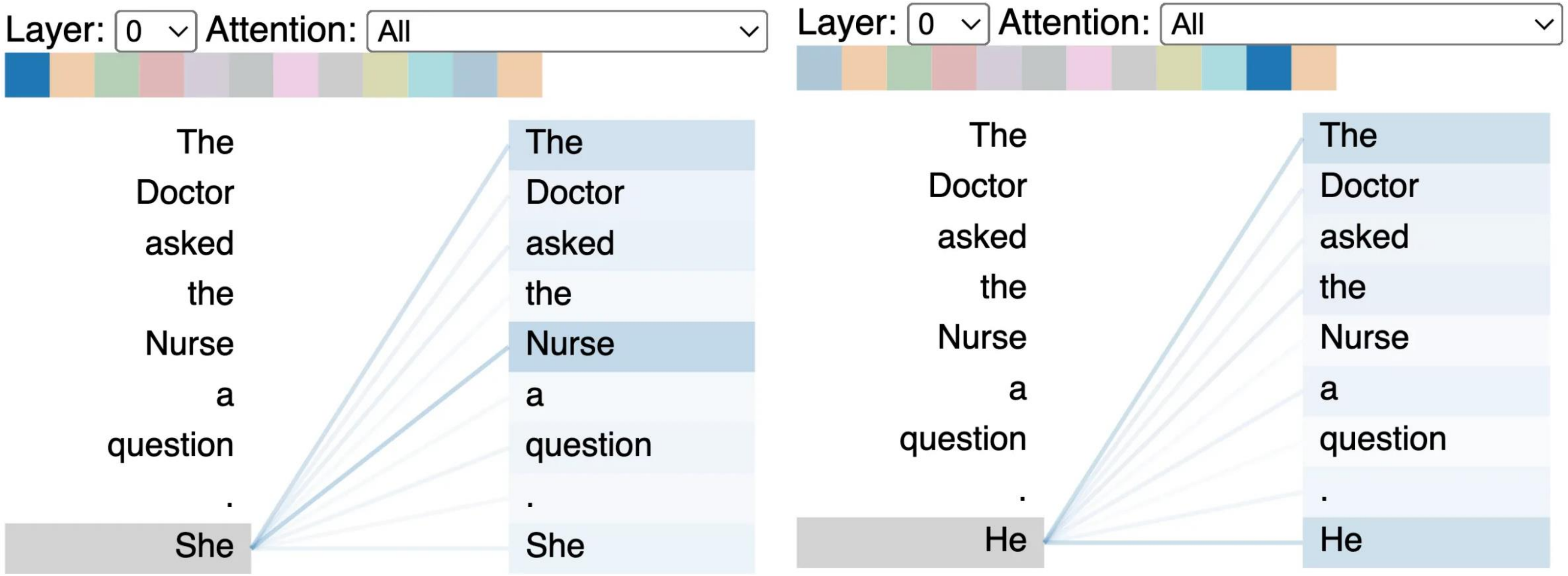
- Domain-specific context can be retrieved from a vector database (see upcoming talks).
- With this approach, we can even cite sources.

Bias

This picture illustrates gender bias in the attention mechanism.

The model (here GPT-2) associates “she” with “nurse” and “he” with “The Doctor”

This bias is induced by training data and perpetuated by the transformer.



Layer: 0 Attention: All

Layer: 0 Attention: All

Source: https://generativeai.pub/explainable-ai-visualizing-attention-in-transformers-4eb931a2c0f8

Agenda

HOW DOES CHATGPT WORK?

LARGE LANGUAGE MODELS (LLM)

- It’s all about architecture!
- The transformer architecture
- Tokenization and embeddings
- Attention
- Reinforcement Learning From Human Feedback (RLHF)

DISCUSSION ON TRANSFORMERS

- Potential
- Shortcomings and dangers

HOW INTELLIGENT IS CHATGPT?

DISCUSSION ON ARTIFICIAL GENERAL INTELLIGENCE (AGI)

So, is ChatGPT intelligent or not?

This analysis relies on a definition of intelligence published in 1997 which consists of 6 areas, and which now draws a consensus among psychologists.

REASON

✓

Can e.g. reason about existing code and provide suggestions for improvements.

LEARN QUICKLY AND LEARN FROM EXPERIENCE

~

Can learn from the context provided (within a session) but the model is frozen is time.

PLAN

✗

Due to its sequential generation, a transformer thinks linearly and is not able to backtrack. Example: fails at solving Towers of Hanoi.

COMPREHEND COMPLEX IDEAS

✓

Can reason about complex relationships (e.g. impact of interest rates on inflation).

SOLVE PROBLEMS

✓

Can solve complex real-world problems (see example at beginning of presentation).

THINK ABSTRACTLY

✓

Can prove mathematical theorems but also understand abstract concepts like civilization, culture and art.

Source: https://arxiv.org/pdf/2303.12712.pdf

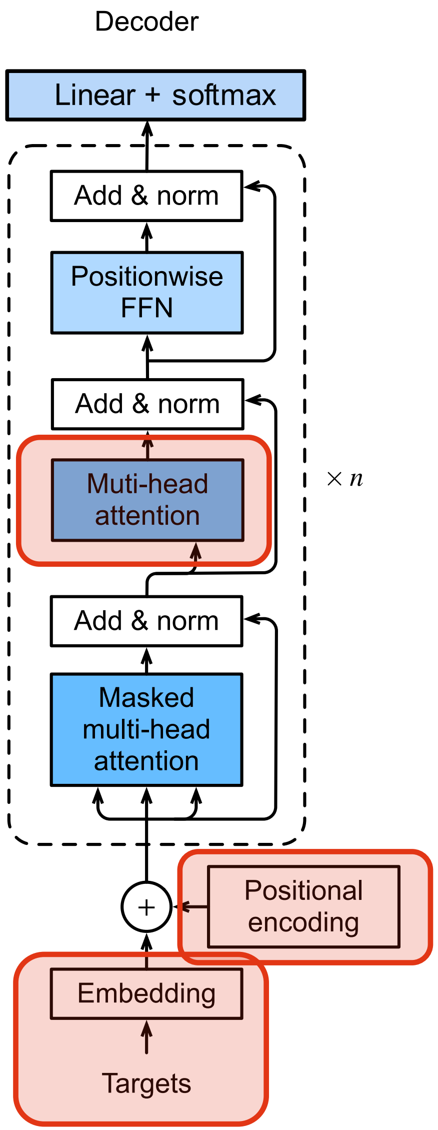
Let’s wrap-up!

Large Language Models

Transformer

RLHF

Decoder vs. Encoder



training vs. prompting

hallucinations

(gender) bias

Definition of intelligence

ELCA Tech Days

Bonus – And so why is it so bad at maths?

TOKENIZING

Tokenizing has not been optimized for maths. Big numbers are split “randomly” into tokens.

TRANSFORMER ARCHITECTURE

The transformer is optimized for *natural* language processing. It is probabilistic by nature, while maths obeys strict rules.

TRAINING DATA

Training data more focused on maths might help, but the problems inherent to the architecture remain.

PLUGINS

A more promising approach is the use of plugins (e.g. calculator, Wolfram)