Inside ChatGPT

What’s Transformer architecture?

A little bit about me..

Emilia Zawadzka-Gosk
zawadzka.emilia@gmail.com

Data Engineer
@PKO Faktoring

ML/AI enthusiast

since 2011 I am a data person

2018-2024 lecturer and researcher
@PJAIT

Agenda

This presentation is for you if:

- you know the basics of neural network
- and you are curious what Transformer is

ked ChatGPT to Write Me a *HIT* Song - YouTube

I COOKED what ChatGPT told me to - YouT

Today I am going to use GPT3 to cr

those **recipes**. GPT3 is a natural la

YouTube · Flavor Lab · 17 gru 2022

YouTube · A.I. Awareness · 22 maj 2023

How it started

Attention Is All You Need

Ashish Vaswani* Google Brain avaswani@google.com

Noam Shazeer* Google Brain noam@google.com

Niki Parmar* Google Research nikip@google.com

Jakob Uszkoreit* Google Research usz@google.com

Llion Jones* Google Research llion@google.com

Aidan N. Gomez* † University of Toronto aidan@cs.toronto.edu

Łukasz Kaiser* Google Brain lukaszkaiser@google.com

Illia Polosukhin* ‡ illia.polosukhin@gmail.com

The headline of the article „Attention Is All You Need”

Transformer architecture

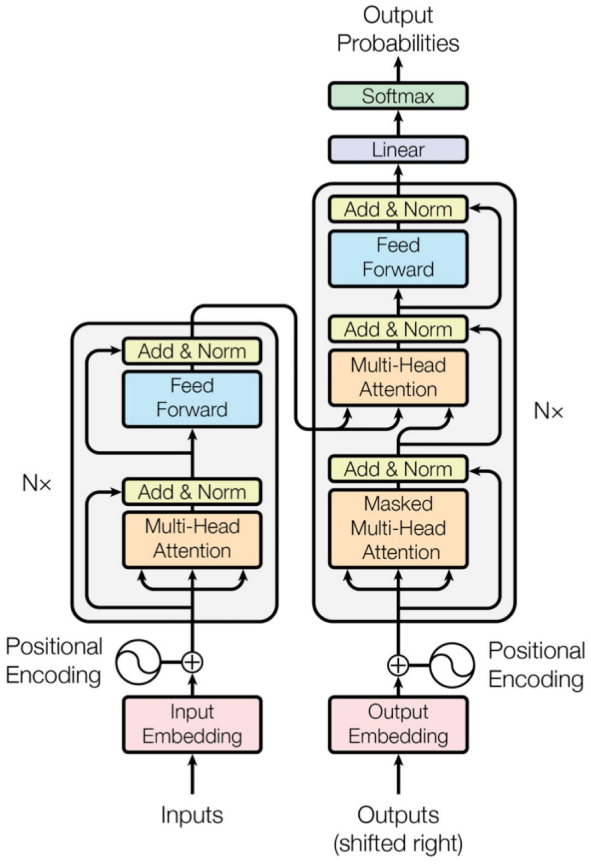


Figure 1: The Transformer - model architecture.

Encoder

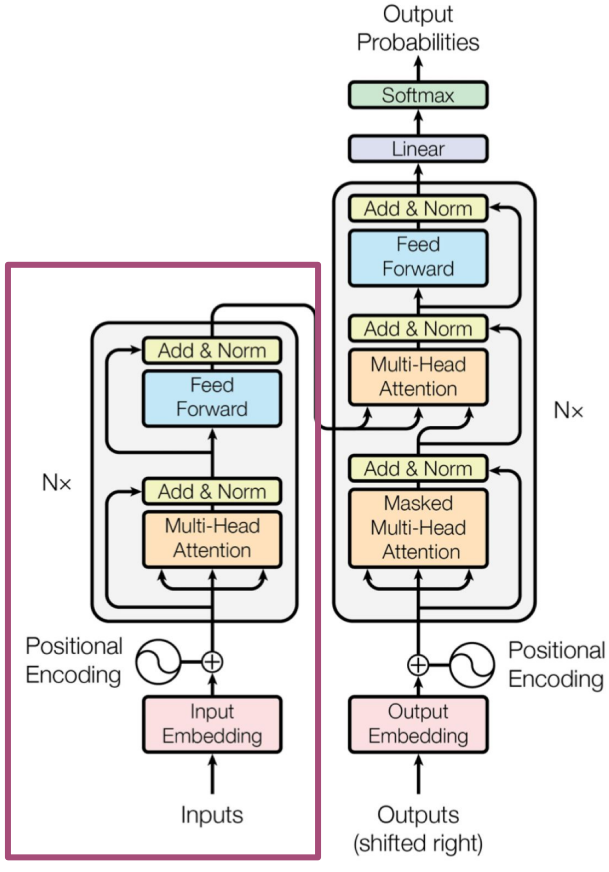
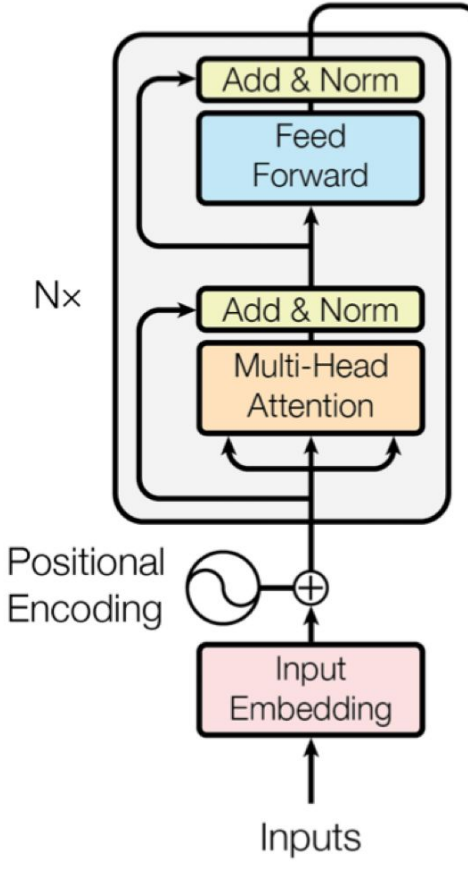


Figure 1: The Transformer - model architecture.

Encoder


Add & Norm
Feed Forward
Nx
Add & Norm
Multi-Head Attention
Positional Encoding
Input Embedding
Inputs

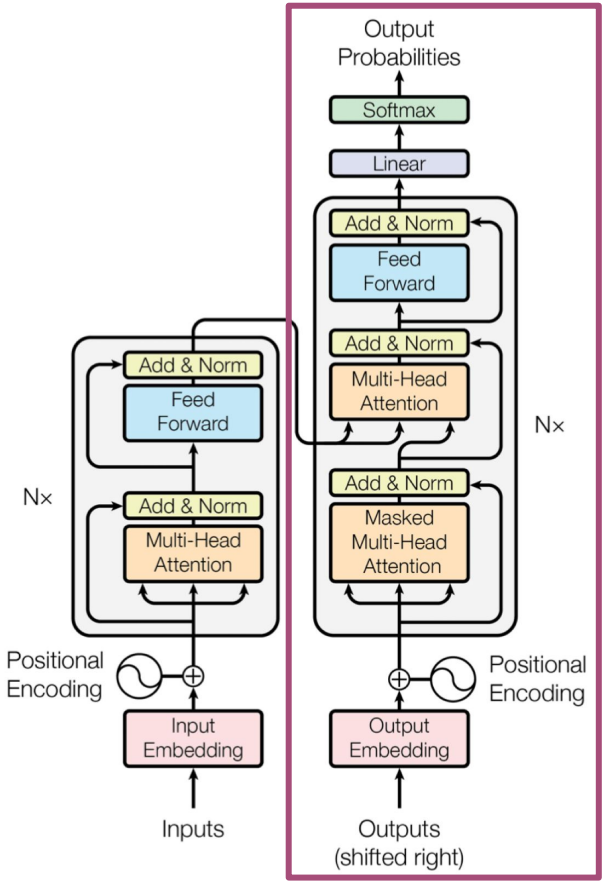
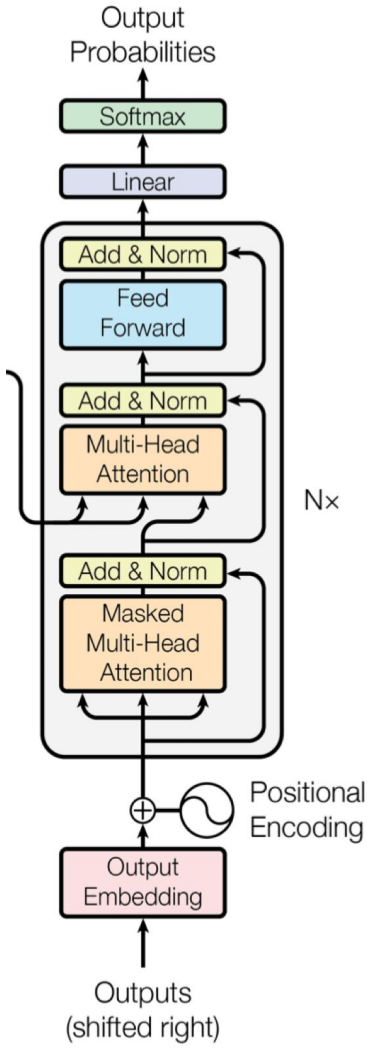


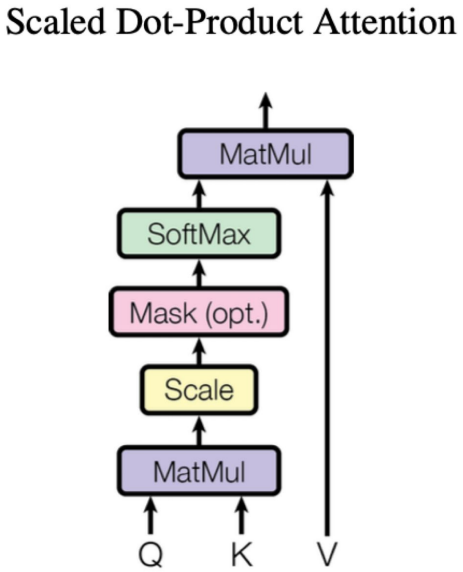
Figure 1: The Transformer - model architecture.

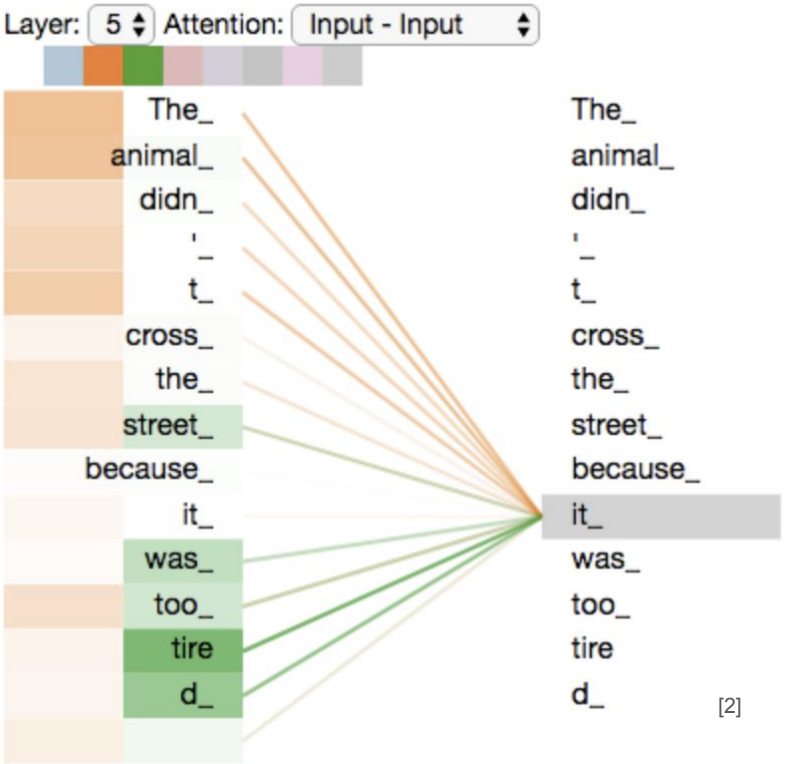
Decoder


Output Probabilities
Softmax
Linear
Add & Norm
Feed Forward
Add & Norm
Multi-Head Attention
Add & Norm
Masked Multi-Head Attention
Nx
Positional Encoding
Output Embedding
Outputs (shifted right)

Decoder

Attention


Scaled Dot-Product Attention
MatMul
SoftMax
Mask (opt.)
Scale
MatMul
Q
K
V


Layer: 5
Attention: Input - Input
The_
animal_
didn_
'_
t_
cross_
the_
street_
because_
it_
was_
too_
tire
d_
The_
animal_
didn_
'_
t_
cross_
the_
street_
because_
it_
was_
too_
tire
d_

[2]

Training objective

1. Mask some percentage of the input tokens at random and then predict those masked tokens.
2. Next sentence prediction: In every sample there are two sentences: A nad B. Model needs to predict if B follows A.

Transformer Architecture: My Project Experience

Visual Question Answering Project (2023)

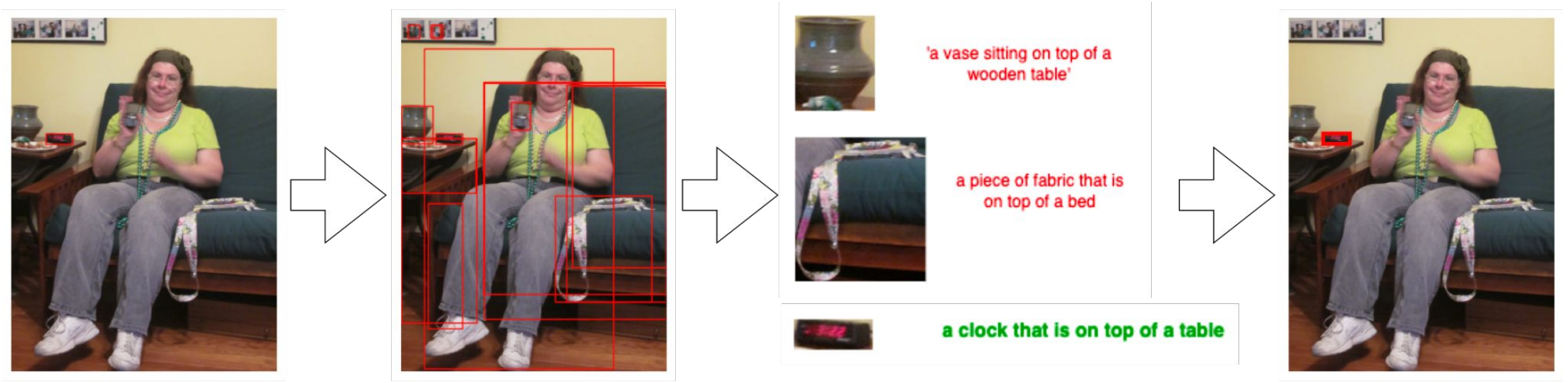
Pytanie: Which is different from the group?

Koordynaty: 409, 172, 432, 206

Pytanie: What do we drive for personal use?

Koordynaty: 161, 181, 569, 367

Visual Question Answering Project (2023)



Pytanie: „Where do we look to see time”

Experiment

Example 1

A: What animal is fluffy and furry?

B: the cat sleeps on a windowsill;

Answer: yes

Example 2

A: What can be used to cut bread?

B: a person on a bike;

Answer:no

The ancestor: GPT-2

Context (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.
GPT-2: 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.

Fine-tuning

- Train the entire model
- Train only the part of it (e.g. last layer)

Prompt-tuning

1. Freeze the model.
2. Concatenate trainable vector (soft prompt) at the beginning of your prompt.
3. Train the vector.
4. Use it at inference time.

Results of my experiment

	Accuracy
“Manual” prompt without examples	0%
“Manual” prompt with 4 examples	18.7%
Prompt tuning	77%
Fine-tuning	81.9%

https://github.com/zawemi/PW

Valuable resources

1. Attention is all you need. https://arxiv.org/pdf/1706.03762.pdf

2. Illustrated Transformer: http://jalammar.github.io/illustrated-transformer/

3. Language Models are Unsupervised Multitask Learners.

https://d4mucfpksywv.cloudfront.net/better-language-models/language_models_are_unsupervised_multitask_learners.pdf

4. Prompt tuning: https://huggingface.co/docs/peft/conceptual_guides/prompting

5. Fine-tuning: https://huggingface.co/docs/transformers/training

Thank you :)