

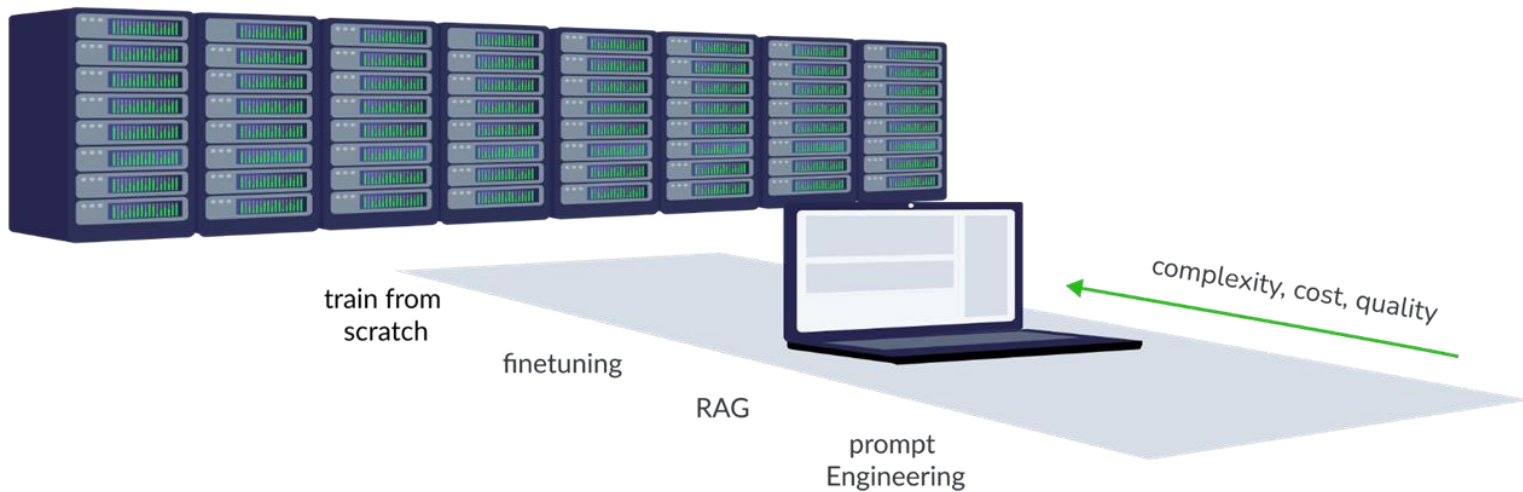


Foundations of LLM Mastery: Fine-tuning with multi GPUs

25 February 2025
ONLINE



How can you influence LLMs?



Transformer Anatomy

Attention is really all you need?

Speaker: Simeon Harrison
Trainer at EuroCC Austria

Transformer Anatomy

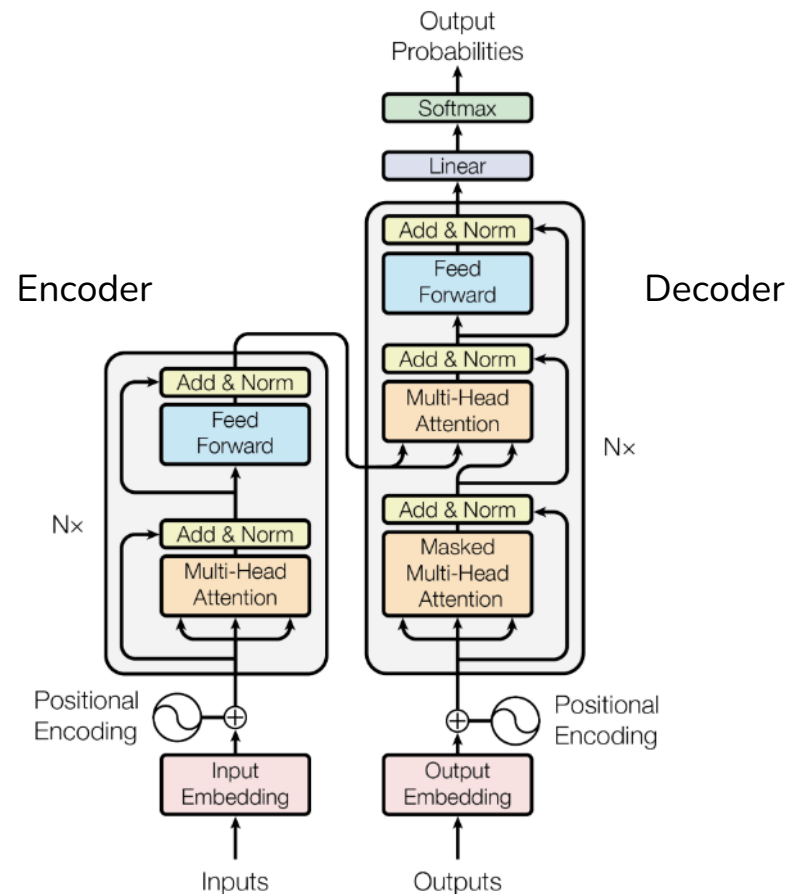
The original architecture

A transformer consists of an encoder and/or decoder block.

Words (tokens) are input as numerical representations (embeddings).

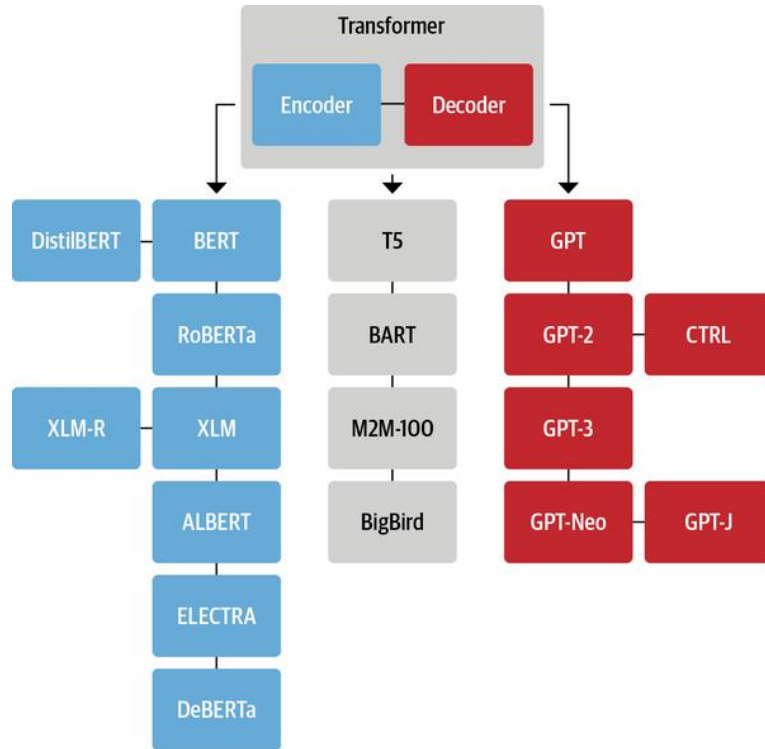
About 1/3 of all parameters are in the multi-head attention blocks

About 2/3 of all parameters are in the feed forward networks (also known as multi layer perceptron)



Source: "Attention Is All You Need", Vaswani et al.

Transformer Family



Encoder only:

These models excel at text classification, named entity recognition, and question answering

Decoder only:

Very good at predicting the next word in a sequence, therefore mostly used for text generation

Encoder-Decoder:

These models are often used for machine translation or summarization tasks.

Context Is All You Need

Embeddings

Here, we will refer to “word” instead of “token”, as it makes the content easier to explain.

A word embedding comes as a multi dimensional vector (e.g. 12.000 dim).

The initial word embedding in all of the examples of the word „mole“ is the same.



The European **mole** is a mammal



Take a biopsy of the **mole**

$$6.02 \times 10^{23}$$

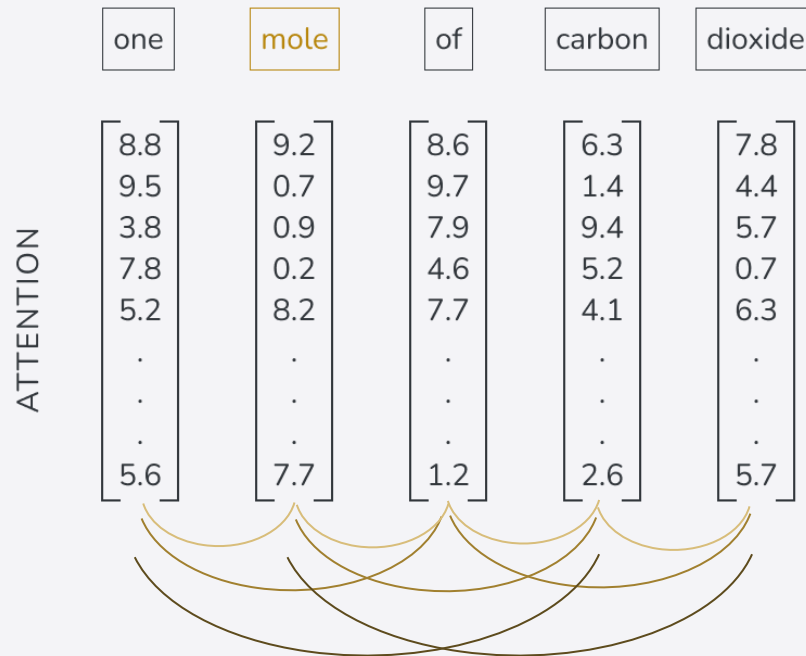
One **mole** of carbon dioxide

Context Is All You Need

Attention

The word „mole“ should be represented by a **unique vector** in the embedding space, depending on its **context**.

An **attention** block should **compute the vectors that you need to add** to the original, generic vector to get it to the correct, meaningful, rich representation, depending on the context in which the word is used.



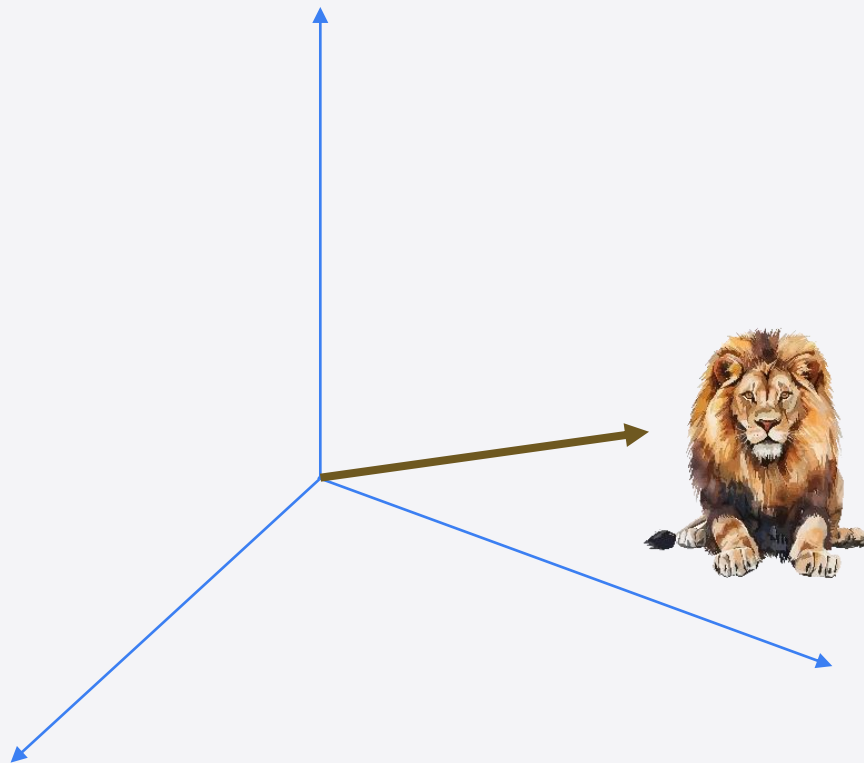
Context Is All You Need

Lion

We associate the word „lion“ with a big cat, living wild on the African continent.

We probably imagine a majestic predator with a big mane.

The embedding of the word „lion“ is a vector with a certain length and direction within the embedding space.

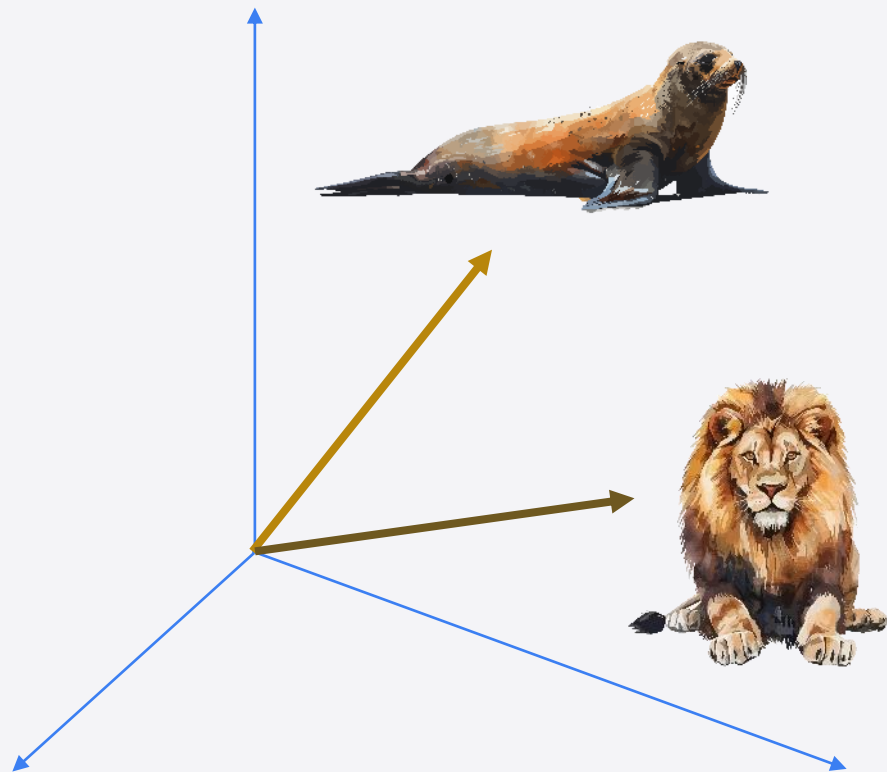


Context Is All You Need

Sea Lion

However, as soon we add the word „sea“ in front of „lion“ we imagine a totally different animal.

The same goes for the embedding. The attention mechanism needs to update the direction and length of the vector so that it represents the animal in question correctly.



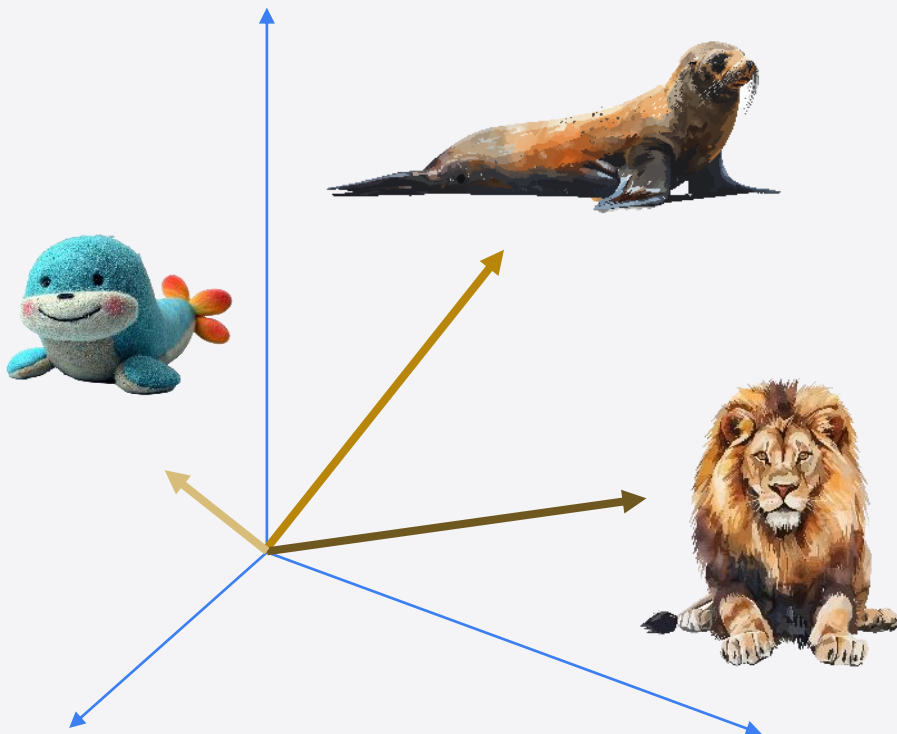
Context Is All You Need

Sea Lion Cuddly Toy

The context depends on more than just the immediate words to the left and right.

The embedding of „sea lion cuddly toy“ will certainly be very different of just „lion“.

In order to achieve that the vector for „lion“ needs to attend to all the other words in the input (context size).



THANK YOU



This project has received funding from the European High-Performance Computing Joint Undertaking (JU) under grant agreement No 101101903. The JU receives support from the Digital Europe Programme and Germany, Bulgaria, Austria, Croatia, Cyprus, Czech Republic, Denmark, Estonia, Finland, Greece, Hungary, Ireland, Italy, Lithuania, Latvia, Poland, Portugal, Romania, Slovenia, Spain, Sweden, France, Netherlands, Belgium, Luxembourg, Slovakia, Norway, Türkiye, Republic of North Macedonia, Iceland, Montenegro, Serbia