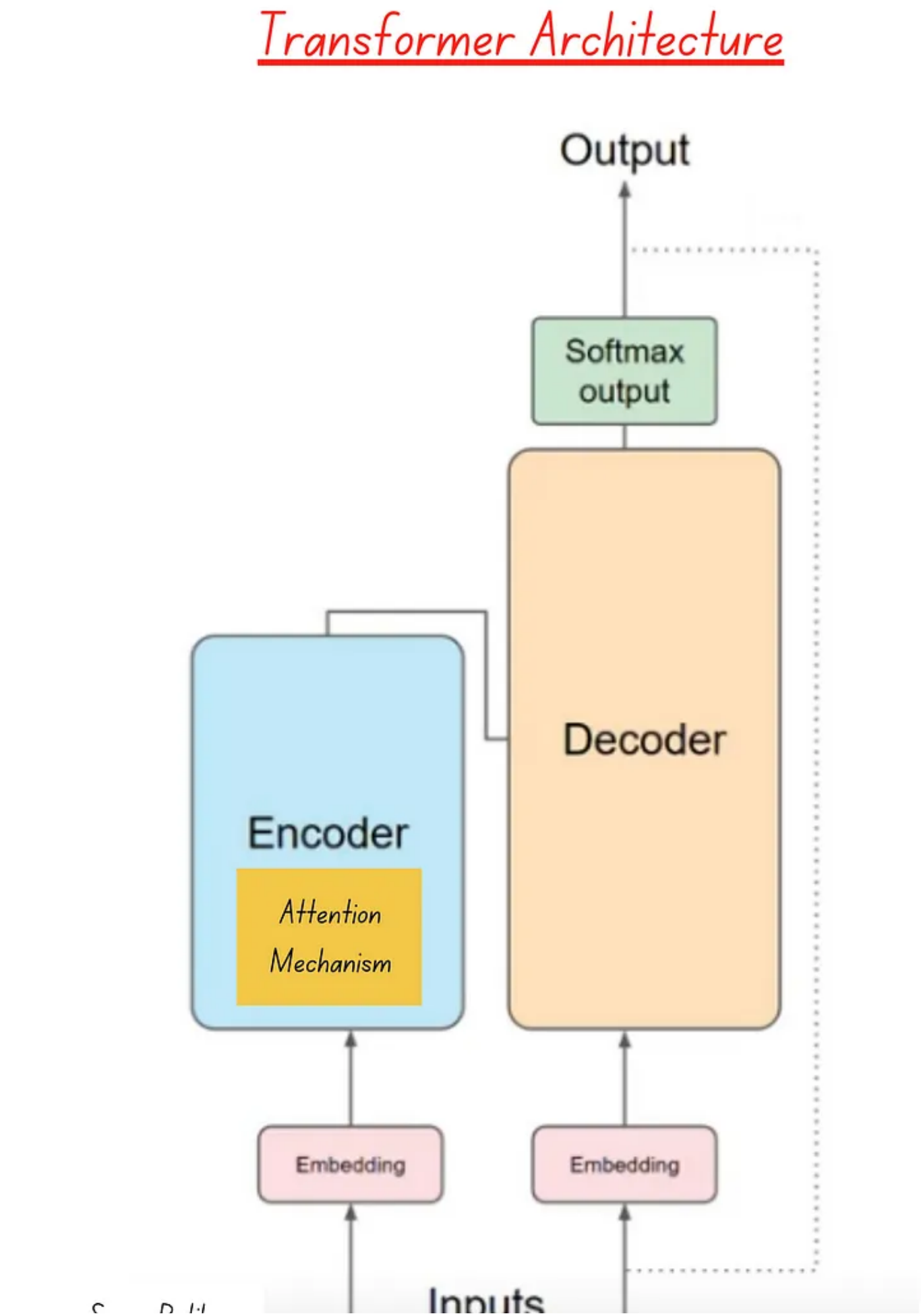
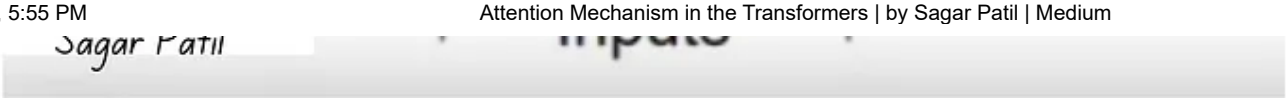
Attention Mechanism in the Transformers

Sagar Patil · Follow
4 min read · Sep 25, 2023

> *In the world of natural language processing and machine learning, few innovations have been as transformative as the attention mechanism. With its inception in the realm of neural networks, the attention mechanism has evolved and found its most remarkable application in the form of Transformers.*

If you've ever wondered how language models like BERT, GPT, and T5 manage to understand and generate human-like text, the answer lies in attention mechanisms.


Transformer Architecture
Output
Softmax output
Decoder
Encoder
Attention Mechanism
Embedding
Embedding
Inputs

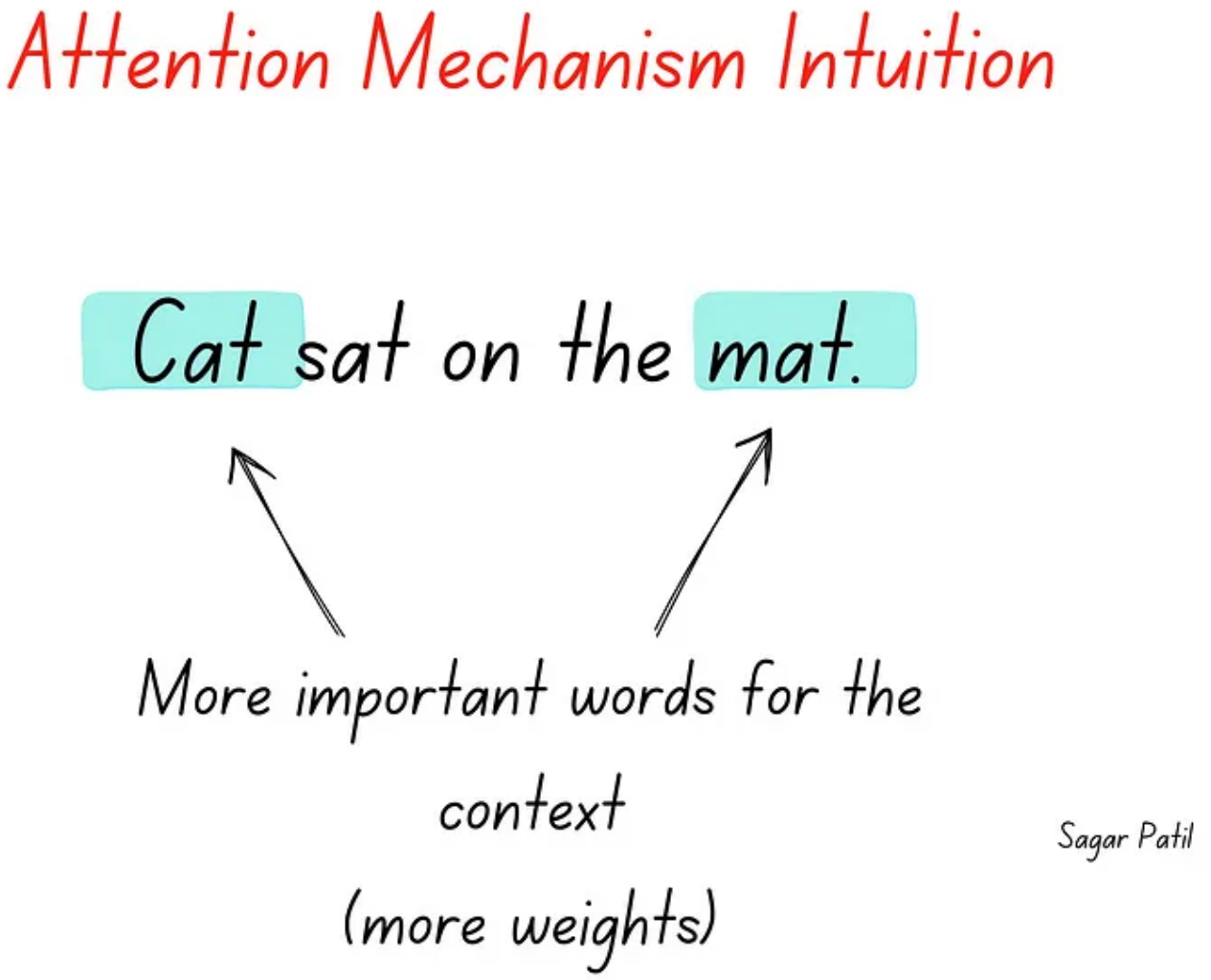


Transformer Architecture and attention mechanism

In this article, I will try to simplify the attention mechanisms in Transformers, exploring what they are, why they are essential, and how they work their magic.

What is the Attention Mechanism?

- At its core, an attention mechanism is a **computational method** inspired by human cognition that helps models in artificial intelligence focus on specific parts of input data while processing it.
- Attention is the ability of a model to **focus on important part** of sentence. It is helps in understanding the context in the text data or any sequential data.
- Think of it as a way for the model to decide which parts of the data are more relevant or important at any given moment.

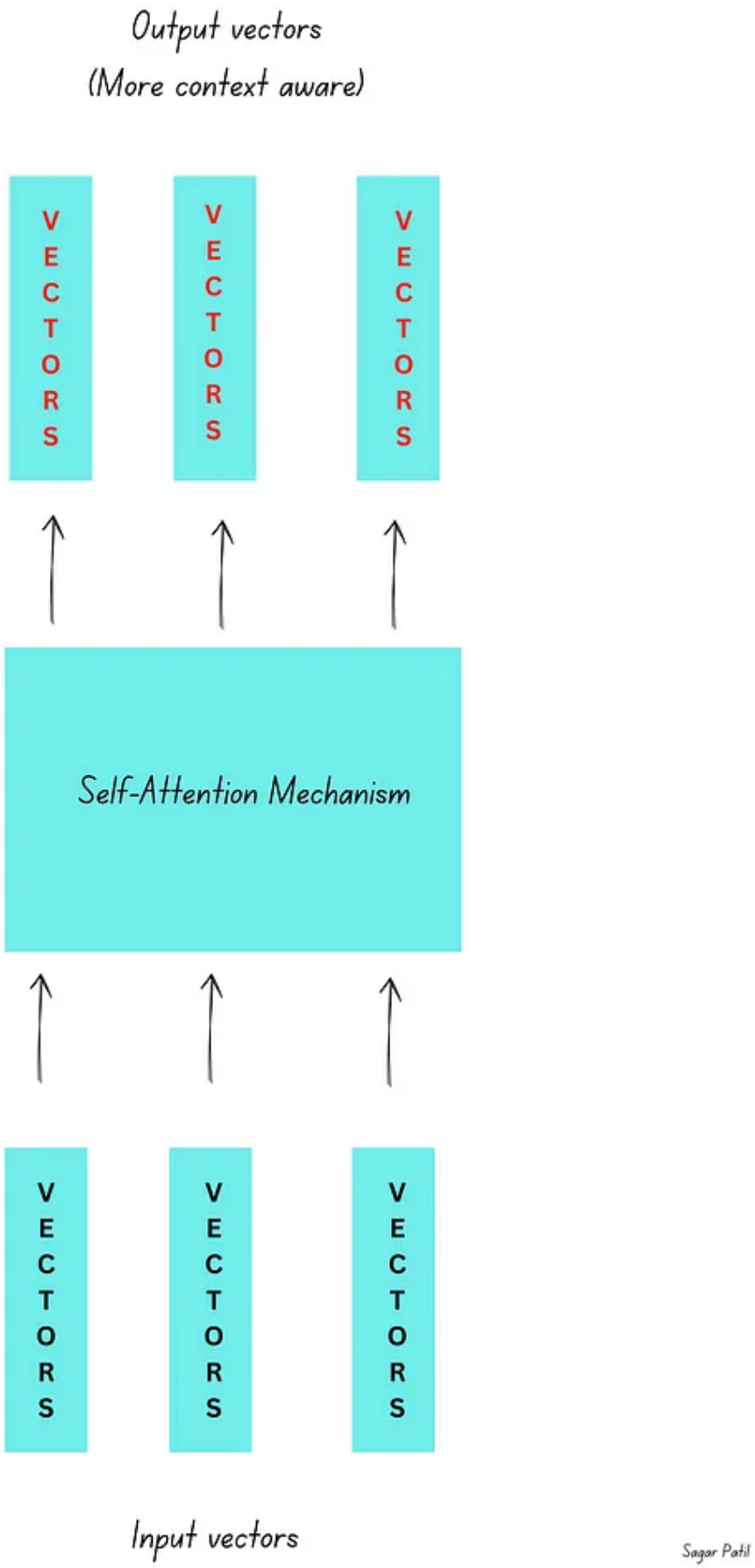


attention

- Imagine you're reading a book: Your attention naturally shifts from word to word, emphasizing certain words or phrases based on context and relevance to understand the text better. Attention mechanisms in AI work somewhat similarly.
- Thus, Attention is a mechanism that enables models to focus on specific parts of input data, **assigning varying degrees of importance** to different elements (words). It's like telling the model where to pay attention when processing information.

> *In essence, with the help of attention mechanism, we* ***try to understand context of the sentence*** *by understanding the* ***importance of specific words*** *by assigning*

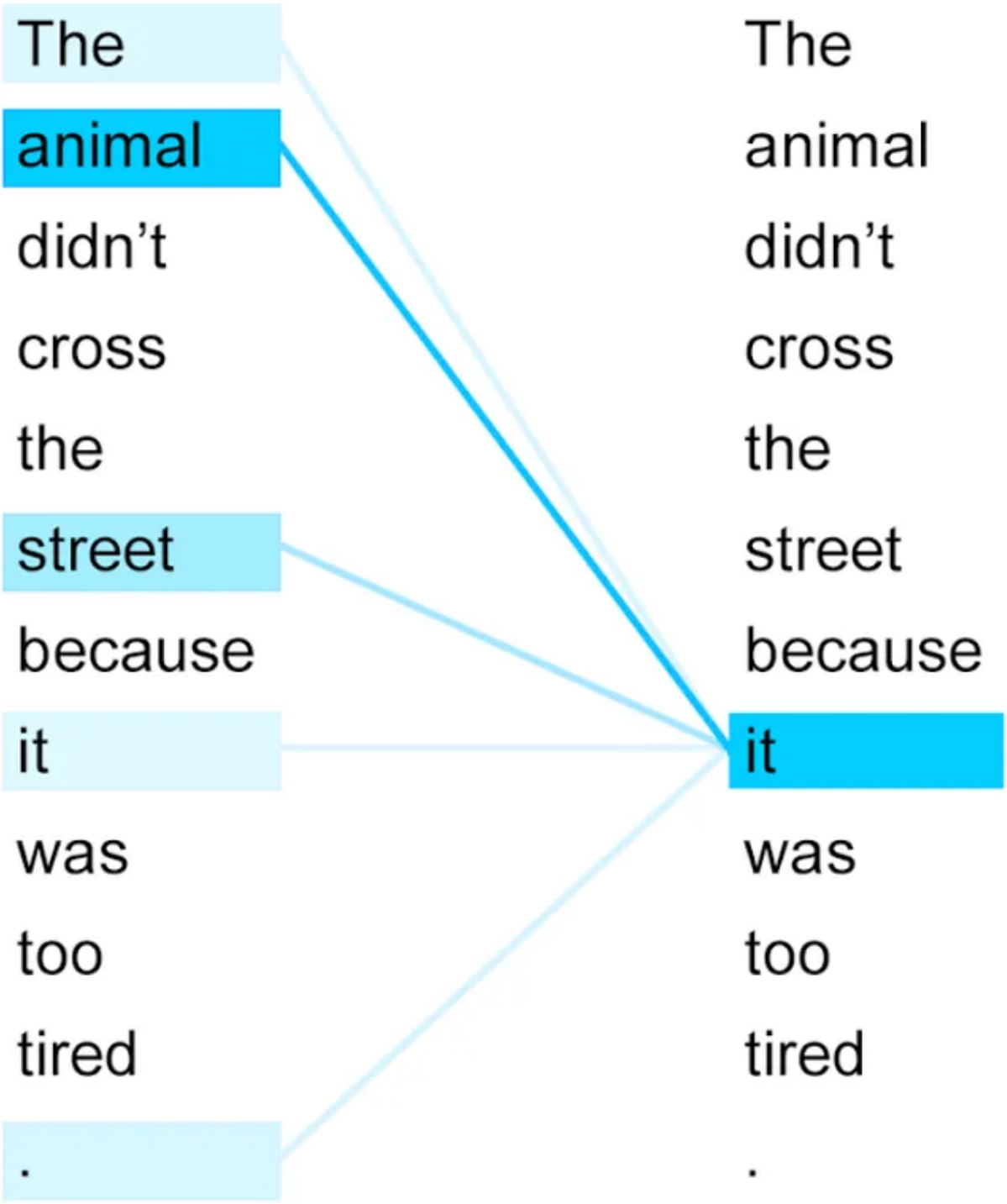
> ***more or less weights*** *to the words.*



Self attention mechanism

What is Self-attention then?

- Self-attention is like looking at different words within the **same sentence** and deciding how much importance to give to each word when understanding the meaning of the sentence. It helps the model consider the **relationships between words within the text.**
- We compare the words of sentence to the words of same sentence during computation.



self-attention (source -internet)

for example, in our sentence, self-attention helps the model understand that “cat” and “mat” are related because they have high attention weights, indicating they are the subject and object of the action “sat.”

How Do Attention Mechanisms Work?

let’s dive deeper into attention weights and how they work within the context of self-attention.

Attention Weights:

- In self-attention, attention weights are numerical values that indicate how much focus or importance each word in a sentence should receive

from other words in the same sentence.

- The goal is to calculate these weights so that the model can assign higher values to words that are more relevant or have stronger relationships with the current word being processed.

Calculation of Attention Weights:

Query, Key, and Value: In self-attention, each word in the sentence (or token) is associated with three vectors:

- Query: A vector representing the current word that we want to calculate attention weights for.
- Key: Vectors associated with all other words in the sentence, serving as a comparison.
- Value: Vectors associated with all other words, which will be used for the weighted sum.

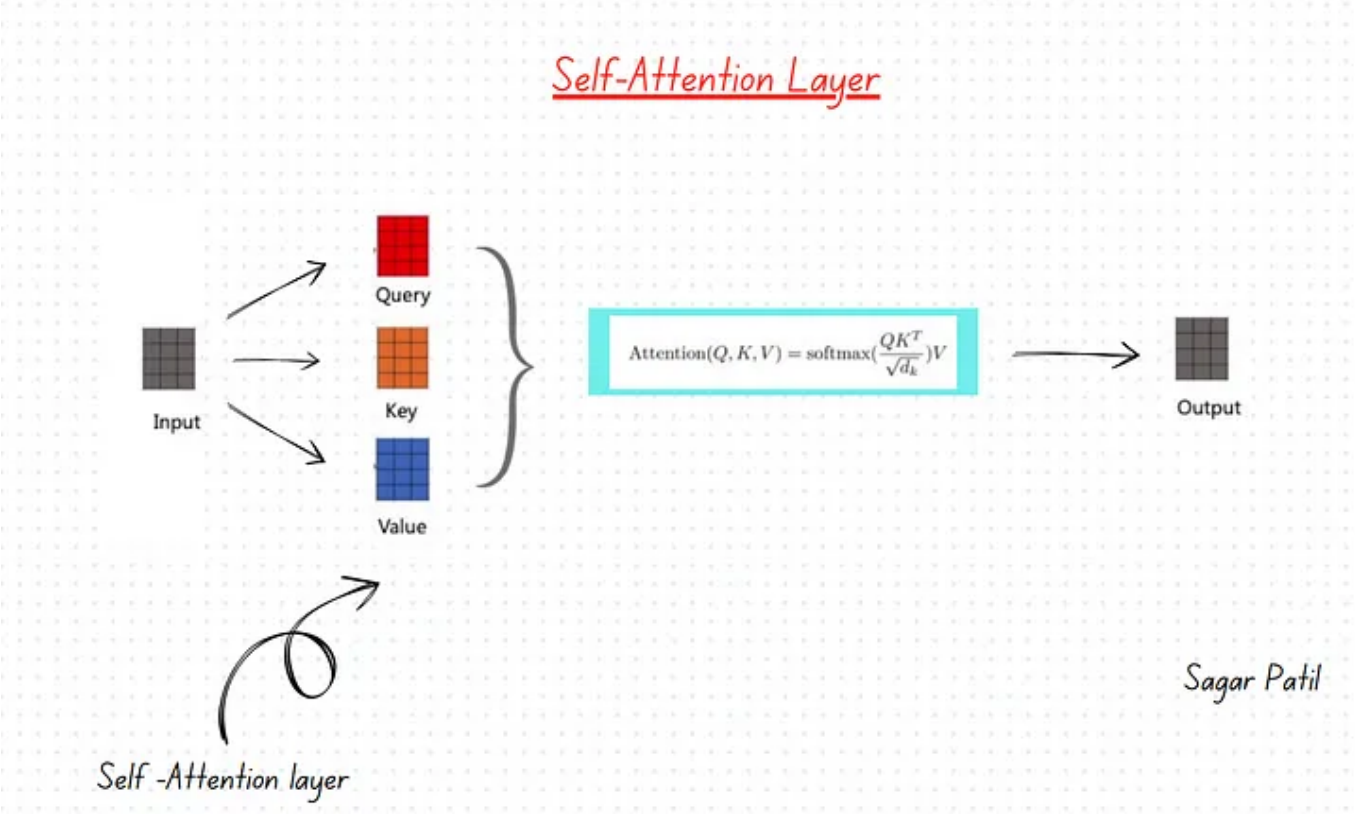
Scoring by dot product: To calculate the attention weights, the model performs a dot product between the query vector of the current word and the key vectors of all other words. This dot product measures how similar or relevant each word is to the current word.

Scaling: The dot products are scaled down (**divided by the square root of the dimension of the key vectors**) to prevent the gradients from becoming too large.

SoftMax: The scaled dot products are then passed through a softmax function. This function **normalizes the scores**, turning them into a

probability distribution where higher scores get higher probabilities. This ensures that the **attention weights sum up to 1**.

Weighted Sum: Finally, the attention weights obtained from the softmax are used to calculate a weighted sum of the value vectors of all words. This weighted sum represents the new representation of the current word, incorporating information from all other words based on their importance scores.



Self-Attention layer

Interpretation:

- If the attention weight for a specific word is high, it means that the model considers that word highly relevant to understanding the current word's context.

- Lower attention weights indicate that the word is less important or less related to the current word.

We delved into what, why and how of the attention mechanism. In conclusion, attention mechanisms have transformed the way machines understand and generate human-like text, and they continue to be at the heart of innovation in the field.

Generative Ai Tools | Deep Learning | Artificial Intelligence | Machine Learning | NLP

Written by Sagar Patil

40 Followers

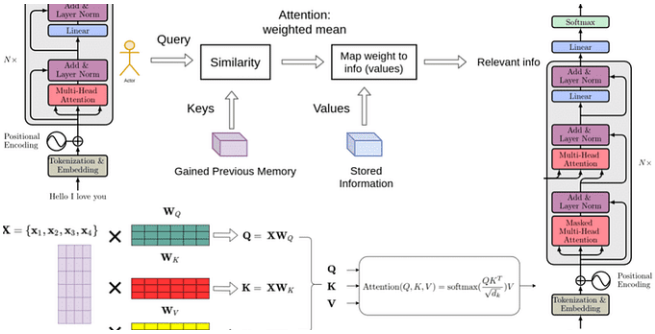
Data Scientist

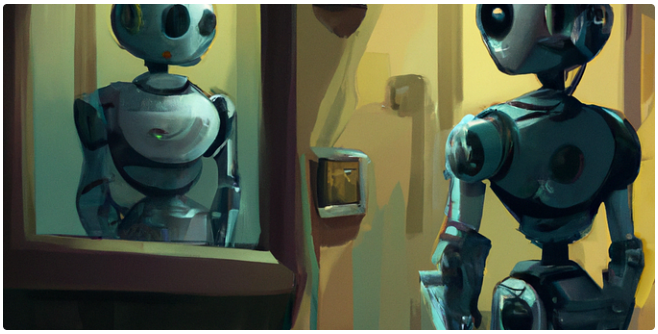
Follow

More from Sagar Patil

See all from Sagar Patil

Recommended from Medium





Amanatullah

Transformer Architecture explained

Transformers are a new development in machine learning that have been making a lo...

10 min read · Sep 1, 2023

308 2

Thomas van Dongen in Towards Data Science

Demystifying efficient self-attention

A practical overview

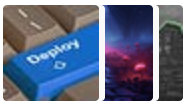
20 min read · Nov 7, 2022

623 2

Lists

Natural Language Processing

1196 stories · 667 saves



Predictive Modeling w/ Python

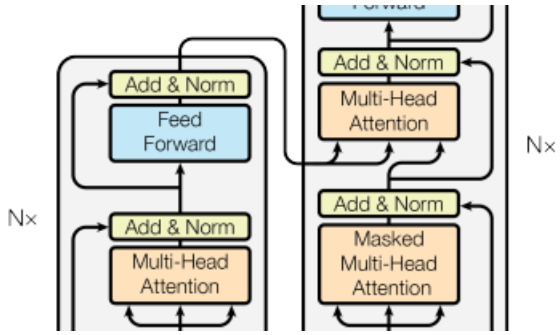
20 stories · 902 saves

AI Regulation

6 stories · 317 saves

Practical Guides to Machine Learning

10 stories · 1057 saves



Dr. Ernesto Lee

FAILED

An Intuitive Explanation of 'Attention Is All You Need': The...

This is the technology that makes ChatGPT works!

9 min read · Oct 14, 2023

28

FAILED

FAILED

See more recommendations