Introduction to Attention Mechanism in LLMs

”Attention is all you need”

SkyRain

Yanshan University

May 13, 2025

Outline

Introduction about Attention Mechanism

Learning Attention Mechanism

Conclusion

Introduction about Attention Mechanism

- Attention mechanism is a key component of LLMs
- It allows the model to focus on different parts of the input
- Helps in understanding context and relationships

Why Attention?

- Traditional models struggled with long-range dependencies
- Attention mechanism overcomes this limitation
- Enables parallel processing of input data

Compare with traditional models

Traditional Models

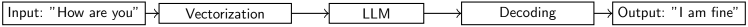
- RNNs and LSTMs
- Sequential processing
- Difficulty in capturing long-range dependencies

Attention Mechanism

- Processes all tokens simultaneously
- Captures relationships between all tokens
- More efficient and effective for long sequences

Key Concept

LLM process numbers



Input: "How are you"

Vectorization

LLM

Decoding

Output: "I am fine"

- Input text is tokenized into matrixes
- Each vector in matrix represents a token(a word)
- Output is decoded back into text
- LLM processes the matrixes
- Attention mechanism is used to understand the relationships between tokens

Attention Mechanism

Attention Mechanism

- Key component of LLMs
- Allows the model to focus on different parts of the input
- Helps in understanding context and relationships

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right) V \tag{1}$$

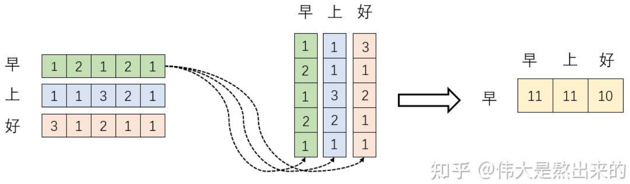
$$Q = X \times W_Q, K = X \times W_K, V = X \times W_V \tag{2}$$

Given that Q, K, V are the linear transformation of the input X, we can simplify the attention mechanism as:

$$\text{Attention}(X) = \text{softmax}\left(XX^T\right) X \tag{3}$$

Understanding Attention Mechanism

A word (token) is represented as a vector, and a sentence is represented as a matrix X.



早	1	2	1	2	1
上	1	1	3	2	1
好	3	1	2	1	1

早	上	好
1	1	3
2	1	1
1	3	2
2	2	1
1	1	1

	早	上	好
早	11	11	10

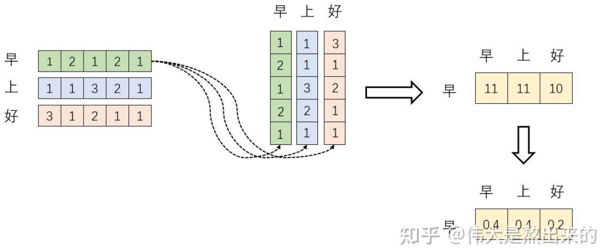
- Vector $A \times B$ means how much relation it have between A and B.
- $X \times X^T$ means how much relation it have between each token in the sentence.

Understanding Attention Mechanism

The Softmax function

- Converts raw scores into probabilities
- Ensures that the sum of probabilities equals 1

$$\text{softmax}(x_i) = \frac{e^{x_i}}{\sum_j e^{x_j}} \tag{4}$$



早	1	2	1	2	1
上	1	1	3	2	1
好	3	1	2	1	1

早	上	好
1	1	3
2	1	1
1	3	2
2	2	1
1	1	1

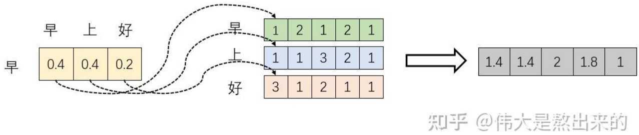
	早	上	好
早	11	11	10

	早	上	好
早	0.4	0.4	0.2

Understanding Attention Mechanism

The last X

- The last X is the output of the attention mechanism
- It is a weighted sum of the input vectors
- Helps in generating the final output



Conclusion

- Attention mechanism is a key component of LLMs
- It can focus different part of input with different weights
- It helps in understanding context and relationships

Future Work

- Can we write a C++ inference engine of ChatGLM like llama.cpp?
- Can we train a LLM from scratch?
- Can we use the attention mechanism in other fields?

Thank You

Thank you for your **Attention**!

Find this slide on Github: KamijoToma/slides