Introduction to transformer

Plan of attack

The plan of this lecture notes is laid out as follows:

1. A bit background on why people think about transformers.
2. The transformer architecture.
3. Some questions about the transformer architecture.
4. More architecture detail.
5. Multimodal transformers
6. Applications
7. Tricks people use to turn a transformer into an LLM.

Background

In the first part of this lecture, we are going to explore the basic idea behind a transformer with language modeling. Once you have the foundation down, generalizing the idea to other modality is conceptually straightforward.

Here are some useful references for interested students: [1]

Transformer Architecture

Tokenization

Given a sentence, the first step we need is to represent the sentence in a format that a computer can understand, i.e. numbers. To do that, we need to tokenize the sentence. There are multiple ways to tokenize a sentence, including character-level tokenization, word-level tokenization, and subword-level tokenization.

Position encoding

If we just feed the vector representing our sentence into a transformer, it will not do what one may expect it to do since it is lacking some understanding of the order of token. Imagine processing the sentence: “Tom Marvolo Riddle” with a character level tokenization, since the transformer does not know the order of the token, the representations “I am Lord Voldemort” and “IaLVoldmorte or dm “ are indistinguishable from the original sentence.

[2]

Attention Mechanism

The core of a transformer is the attention mechanism, as suggested by the name of the original paper which popularized the transformer architecture: “Attention is all you need” [3].

Difference between MLP and Transformer

Common architectures

Multimodal Transformers

Applications

Introduction to transformer

Bibliography

[1] M. Phuong and M. Hutter, “Formal Algorithms for Transformers,” *ArXiv*, 2022, [Online]. Available: https://api.semanticscholar.org/CorpusID:250644473

[2] J. Su, Y. Lu, S. Pan, B. Wen, and Y. Liu, “RoFormer: Enhanced Transformer with Rotary Position Embedding,” *ArXiv*, 2021, [Online]. Available: https://api.semanticscholar.org/CorpusID:233307138

[3] A. Vaswani *et al.*, “Attention is All you Need,” in *Neural Information Processing Systems*, 2017. [Online]. Available: https://api.semanticscholar.org/CorpusID:13756489