

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

Transformer Architecture

Tokenization

Attention Mechanism

Position encoding

Difference between MLP and Transformer

Common architectures

Multimodal Transformers

Applications