Transformers and Large Language Models

Poorv Kumar

December 31, 2025

1 Introduction

Recent advances in Natural Language Processing (NLP) have been largely driven by the introduction of the Transformer architecture. Transformers form the backbone of modern Large Language Models (LLMs) such as GPT, BERT, and others. This document explains the key concepts behind Transformers and LLMs in an informal and intuitive manner.

2 Limitations of RNNs and LSTMs

Before Transformers, Recurrent Neural Networks (RNNs) and Long Short-Term Memory networks (LSTMs) were widely used for sequence modeling tasks. Although effective, these models suffer from several limitations:

- Sequential processing makes training slow.
- Difficulty in learning long-range dependencies.
- Gradient-related issues such as vanishing gradients.

These drawbacks motivated the development of a new architecture that could process sequences more efficiently.

3 Attention Mechanism

The attention mechanism allows a model to focus on the most relevant parts of an input sequence while processing information. Instead of compressing an entire sequence into a single hidden state, attention enables direct interaction between different tokens.

For example, in a sentence containing pronouns, attention helps the model identify which earlier word a pronoun refers to.

4 Self-Attention

Self-attention is the core operation in Transformers. In this mechanism, each word in a sentence attends to all other words in the same sentence. Each token is projected into three vectors:

- Query (Q)
- Key (K)
- Value (V)

The attention score is computed using the formula:

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

This allows the model to assign higher importance to more relevant words.

5 Multi-Head Attention

Instead of using a single attention mechanism, Transformers use multi-head attention. Multiple attention heads allow the model to capture different types of relationships such as syntactic structure, semantic meaning, and long-distance dependencies. The outputs of all heads are concatenated and linearly transformed.

6 Positional Encoding

Transformers do not inherently understand the order of tokens. To address this, positional encodings are added to input embeddings. These encodings use sinusoidal functions to inject information about token positions, enabling the model to distinguish between different word orders.

7 Transformer Architecture

A Transformer consists of stacked layers, each containing:

- Multi-head self-attention
- Add and Layer Normalization
- Feed-forward neural network
- Add and Layer Normalization

There are two main components:

- Encoder: used for understanding tasks (e.g., BERT)
- Decoder: used for generation tasks (e.g., GPT)

8 Generative Pretrained Transformer (GPT)

GPT is a decoder-only Transformer model trained using a next-token prediction objective. Given a sequence of tokens, GPT predicts the probability of the next token. This autoregressive nature allows GPT to generate coherent text one token at a time.

9 Large Language Models

Large Language Models are characterized by:

- Very large number of parameters
- Training on massive text corpora
- Emergent abilities such as reasoning and summarization

These models act as foundation models that can be adapted to specific tasks through fine-tuning.

10 Hugging Face Ecosystem

Hugging Face provides an open-source ecosystem that includes pretrained models, datasets, and tools for training and inference. It simplifies working with state-of-the-art Transformer models and encourages reproducible research.

11 Relevance to Legal Domain

Legal documents often contain complex language and long contextual dependencies. Transformers and LLMs are well-suited for legal NLP tasks due to their ability to process long sequences and understand contextual meaning. Fine-tuning LLMs on legal data can significantly improve performance in domain-specific applications.

12 Conclusion

Transformers have revolutionized NLP by replacing recurrence with attention-based mechanisms. Large Language Models built on Transformers demonstrate strong generalization capabilities and form the foundation for many real-world AI systems today.