Advances in Large Language Models

A Comprehensive Review of Transformer Architectures

Abstract

This paper presents a comprehensive review of recent advances in large language models (LLMs), focusing on transformer architectures and their applications. We examine the evolution from early attention mechanisms to modern architectures like GPT, BERT, and T5, analyzing their strengths, limitations, and real-world applications.

1. Introduction

Large language models have revolutionized natural language processing by demonstrating remarkable capabilities in understanding and generating human-like text. The transformer architecture, introduced by Vaswani et al. in 2017, has become the foundation for most modern LLMs. These models use self-attention mechanisms to process input sequences in parallel, enabling them to capture long-range dependencies more effectively than previous recurrent architectures.

2. Key Concepts

2.1 Self-Attention Mechanism

The self-attention mechanism allows models to weigh the importance of different words in a sequence when processing each word. This is computed using queries (Q), keys (K), and values (V) matrices. The attention scores are calculated as: Attention(Q,K,V) = softmax(QK^T/√d_k)V

2.2 Multi-Head Attention

Multi-head attention allows the model to attend to different positions simultaneously, learning various types of relationships. Each head performs attention independently, and their outputs are concatenated and linearly transformed.

3. Applications

LLMs have found applications in numerous domains including: • Text Generation: Creating human-like text for various purposes • Translation: High-quality machine translation between languages • Question Answering: Understanding and responding to queries • Code Generation: Writing functional code from natural language descriptions • Summarization: Condensing long documents into key points

Table 1: Comparison of Popular LLMs

Model	Parameters	Architecture	Key Innovation
GPT-3	175B	Decoder-only	Scaling laws, few-shot learning
BERT	340M	Encoder-only	Bidirectional pre-training
T5	11B	Encoder-Decoder	Text-to-text framework
LLaMA	65B	Decoder-only	Efficient training

4. Conclusion

Large language models represent a significant breakthrough in AI, demonstrating capabilities that were thought impossible just a few years ago. As we continue to scale these models and improve their architectures, we can expect even more impressive applications. However, challenges remain in terms of computational requirements, bias mitigation, and ensuring safe deployment.