Attention Is All You Need: 2023 Architecture Updates

Abstract

This paper presents significant improvements to the Transformer architecture introduced in the original "Attention Is All You Need" work. Key enhancements include efficiency optimizations, reduced memory consumption, and improved performance on long sequences. These updates are particularly relevant for large-scale RAG systems processing extensive document collections.

Key Improvements

- **Flash Attention**: 40% reduction in GPU memory usage
- **Rotary Position Embedding**: Better handling of long sequences
- **Layer Normalization Optimization**: 15% faster training
- **Gradient Checkpointing**: Enables training of larger models
- **Mixed Precision Training**: 2x speedup with minimal accuracy loss