Transformers detailed report

Hariish Jayakumar and Moksh

December 2025

1 Introduction

Transformer models form the foundation of most modern Large Language Models (LLMs) that are widely used today. The architecture was introduced in the landmark paper "Attention Is All You Need", and it fundamentally changed how sequence-based data such as natural language is processed.Unlike earlier models that handled words sequentially, transformers use attention mechanisms that allow them to process an entire sequence in parallel.This shift is important because language is inherently contextual, and being able to relate all words in a sentence simultaneously leads to richer and more stable representations.

2 Limitations of Earlier Sequence Models

Before transformers, models such as Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) networks were commonly used.Although effective for short sequences, these models suffered from slow training and difficulty in capturing long-range dependencies. Information from earlier parts of a sentence often weakened as the sequence grew longer. Transformers addressed these issues by completely removing recurrence and instead relying on attention.This allows every token in a sequence to directly interact with every other token, regardless of their position.

3 Self-Attention and Intuition

The core idea behind transformers is the self-attention mechanism.Self-attention enables the model to assign varying importance to different words depending on context.Each word is represented using query, key, and value vectors, and attention scores are computed based on similarity between these representations.Visualizing attention as a weighted flow of information between tokens makes it easier to understand how meaning is distributed across a sentence.

4 Multi-Head Attention and Model Structure

Instead of using a single attention operation, transformers employ multi-head attention.Each attention head focuses on different types of relationships, such as grammatical structure or semantic similarity.This allows the model to learn multiple perspectives of the same input simultaneously.A standard transformer layer consists of multi-head self-attention followed by a feed-forward neural network, along with residual connections and normalization.Stacking multiple such layers enables the model to learn increasingly abstract language representations.

5 Transformers and GPT

Generative Pre-trained Transformers (GPT) are models built entirely on the transformer architecture. GPT models are trained in two stages: pre-training on large-scale unlabeled text data, followed by task-specific fine-tuning.During generation, GPT predicts the next token based on previously seen tokens, which explains its ability to generate coherent and context-aware text.

6 Hugging Face and Practical Implementation

Hugging Face provides an ecosystem that makes transformer models accessible for practical use. It offers pre-trained models, datasets, and development tools that allow users to experiment with and deploy LLMs without training them from scratch. This significantly lowers the barrier to entry for students and researchers.

7 Transformers and Artificial General Intelligence

Artificial General Intelligence (AGI) refers to an AI system capable of performing a wide range of intellectual tasks at a human level, without being restricted to specific domains. While current transformer-based LLMs show impressive generalization abilities, they still lack true understanding, reasoning autonomy, and consciousness. Transformers represent an important step toward AGI, but achieving true general intelligence will likely require architectural and conceptual advances beyond scaling existing models.

8 Conclusion

Transformers represent a fundamental shift in natural language processing by enabling parallel, attention-based learning. Their success has led to powerful models such as GPT and has shaped the modern AI landscape. Although they do not yet achieve true intelligence, transformers remain central to ongoing research in artificial intelligence.

References

1. A comprehensive guide to large language model applications with hugging face. Accessed 2025.

2. Transformer model. Accessed 2025.

3. What is gpt? Accessed 2025.

4. 3Blue1Brown. But what is a language model?, 2023. YouTube Video.

5. Unknown. Self-attention and transformers explained, 2024. YouTube Video.