A Concise Report on Attention, Transformers, and the Road to Artificial General Intelligence

Dipanjan and Yash

1 ABSTRACT

This report presents a concise overview of the evolution of neural network architectures, focusing on Multilayer Perceptrons (MLPs), attention mechanisms, and the Transformer architecture that underpins modern Large Language Models (LLMs). It explains how Transformers combine self-attention and MLP blocks to achieve scalable, parallel computation and powerful representation learning. The report also discusses the significance of Transformers in the pursuit of Artificial General Intelligence (AGI), highlighting recent progress and remaining challenges.

2 REPORT

2.1 Multilayer Perceptrons (MLPs)

Modern artificial intelligence has evolved through several key architectural innovations, beginning with Multilayer Perceptrons (MLPs). MLPs are feedforward neural networks composed of multiple layers of neurons that apply weighted sums followed by nonlinear activation functions. Using backpropagation, MLPs can approximate complex functions and learn hierarchical representations of data. Although early MLPs were constrained by limited computational resources and training challenges such as vanishing gradients, they introduced the core principles of representation learning and layered abstraction that form the foundation of modern deep learning systems.

2.2 Attention and Transformer Architecture

A major breakthrough occurred with the introduction of attention mechanisms, culminating in the Transformer architecture proposed in *Attention Is All You Need.* A Transformer is built as a stack of identical layers, each consisting of two fundamental components: a self-attention block and a position-wise Multilayer Perceptron (MLP). The self-attention mechanism enables each token in a sequence to directly relate to every other token, allowing efficient modeling of long-range dependencies and rich contextual understanding.

The MLP layers then transform these attended representations independently at each position, increasing the expressive capacity of the model. By eliminating recurrence and convolution, Transformers process all tokens simultaneously, achieving high parallelism and efficient GPU-based training. This architecture powers modern LLMs such as GPT, Gemini, and Claude, as well as applications in vision, speech, and multimodal artificial intelligence.

2.3 Road to Artificial General Intelligence (AGI)

Transformer-based models have renewed interest in Artificial General Intelligence (AGI), defined as AI capable of human-level performance across diverse tasks. Although current systems remain narrow, large Transformers exhibit emergent abilities such as limited reasoning and cross-domain transfer. However, true AGI will require advances beyond scaling, including improved memory, grounding, and continual learning. These directions remain central to ongoing research efforts worldwide today globally increasingly.