Implementation of Transformer Architecture Through Visual Learning

Harshika Dehariya
Department of Computer Science & Engineering, Technocrats Institute of Technology (Excellence), Bhopal, Madhya Pradesh, India

Dr. Vikas Gupta
Department of Computer Science & Engineering, Technocrats Institute of Technology (Excellence), Bhopal, Madhya Pradesh, India

Arjun Rajput
Department of Computer Science & Engineering, Technocrats Institute of Technology (Excellence), Bhopal, Madhya Pradesh, India

Saurabh Karsoliya
Department of Computer Science & Engineering, Technocrats Institute of Technology (Excellence), Bhopal, Madhya Pradesh, India

Abstract

Transformer infrastructures have emerged as a basis for artificial intelligence research in recent years, moving from natural language processing to visual literacy. The Transformer's tone-attention medium makes it possible to model long-range dependencies in visual data, which improves point representation and contextual comprehension. The use of Transformer infrastructures in visual literacy operations is thoroughly examined in this review of the literature. It examines how early Vision Mills (ViT) evolved into sophisticated hierarchical, cold-blooded, multimodal infrastructures like Swin Transformer, DETR, and CLIP. The review also covers widely used datasets, training approaches, assessment standards, and difficulties with computation and data efficacy. Similarly, it looks at real-world applications in various computer vision fields, such as independent systems, multimodal understanding, and medical imaging. According to the results, visual Mills have demonstrated a paradigm shift in computer vision research by outperforming conventional convolutional models in terms of rigidity and scalability.

Keywords

Computer Vision, Deep Learning, Image Recognition, Multimodal Systems, Self-Attention, Vision Transformer, Visual Learning.