KIIT UNIVERSITY

Department of Computer Science and Engineering

Assignment Title:
Exploring Transformers and Vision Transformers (ViT): Theory, Implementation, and Real-Time Evaluation

Submitted By: Shubhi Tiwari
Roll No: 22052412
B.Tech (CSE – AI & ML)
Section: CSE–24

Submitted To:
Mr. Himanshu Ranjan
Subject: Deep Learning / Advanced Neural Networks

Index

1. Introduction

Transformers have revolutionized deep learning by introducing a novel architecture based entirely on self-attention. Unlike recurrent neural networks, Transformers process data in parallel, enabling faster training and improved handling of long-range dependencies. Originally proposed for Natural Language Processing tasks, the Transformer has since evolved into a powerful tool for computer vision through the Vision Transformer (ViT). This report aims to explain the core architecture of Transformers, highlight how ViT differs from CNNs, and explore its advancements and limitations.

2. Transformer Architecture

The Transformer consists of two primary components: the encoder and the decoder. Each layer is built using multi-head self-attention and feed-forward networks, supported by residual connections and normalization layers.

2.1 Encoder: The encoder processes input data and captures contextual information using self-attention. Each encoder block transforms token embeddings through parallel attention heads.

2.2 Decoder: The decoder uses masked attention to generate outputs step by step while referencing encoder outputs for contextual alignment.

2.3 Attention Mechanism: Attention allows the model to weigh relationships between tokens. The key formula is Attention(Q, K, V) = softmax(QK^{T} / $\sqrt{d_k}$)V, where Q, K, and V are query, key, and value matrices respectively.

3. Vision Transformer (ViT)

The Vision Transformer adapts the Transformer architecture for image processing by dividing an image into patches (e.g., 16×16 pixels), flattening them, and treating each patch as a token. These token embeddings are then processed by the Transformer encoder layers to extract relationships across the entire image.

3.1 Working Principle: ViT embeds image patches using a linear projection layer and adds positional encodings to preserve spatial information. The embeddings are processed through multiple Transformer layers, followed by a classification head for final prediction.

3.2 Differences Between CNN and ViT: CNNs use local filters to capture spatial hierarchies, while ViTs apply global attention, enabling them to learn relationships between distant regions. However, ViTs typically require large datasets and more computational resources to perform effectively.

4. Limitations of ViT:

Despite their success, Vision Transformers have some drawbacks. They require large-scale datasets and extensive computational power, making them less efficient for small-scale applications. Additionally, ViTs lack the inductive biases of CNNs, which naturally capture local spatial features. Without sufficient data, ViTs may struggle to generalize well or converge efficiently during training.

5. Recent Improvements (Swin Transformer, DeiT)

Recent architectures have been proposed to improve the efficiency of ViTs. The Swin Transformer introduces a hierarchical feature representation using shifted windows, reducing computational cost while maintaining strong performance. DeiT (Data-efficient Image Transformer) enhances training efficiency through knowledge distillation and strong data augmentation techniques. These improvements enable ViTs to perform competitively even with limited data, bridging the gap between CNNs and pure attention-based architectures.

6. Research Paper References

1. Vaswani, A., et al. (2017). 'Attention Is All You Need.' *Advances in Neural Information Processing Systems (NeurIPS).*
2. Dosovitskiy, A., et al. (2020). 'An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale.' *International Conference on Learning Representations (ICLR).*
3. Liu, Z., et al. (2021). 'Swin Transformer: Hierarchical Vision Transformer using Shifted Windows.' *IEEE International Conference on Computer Vision (ICCV).*