
**Savitribai Phule Pune University**
**Gokhale Education Society's**

**R. H. Sapat College of Engineering, Management Studies and Research,**
**Nashik - 422 005, (M.S.), INDIA**

**DEPARTMENT OF COMPUTER ENGINEERING**
**Third Year Computer Engineering**
**Year 2024– 2025**

**Roll No: 36**

**Name of Student: Pranamya Nilesh Deshpande**

**Mobile No.: (+91) 8180812144**

**Official-Mail ID: pranamyadeshpande14@gmail.com**

**Seminar Title: Transformers for Image Recognition**

**Seminar Guide: Prof. Mrs. R. D. Narwade**

**Area of the Seminar: Computer Vision**

## Abstract

While the Transformer architecture has become the de-facto standard for natural language processing tasks, its applications to computer vision remain limited. Vision Transformers (ViTs) have recently become the state-of-the-art across many computer vision tasks. In contrast to convolutional networks (CNNs), ViTs enable global information sharing even within shallow layers of a network, i.e., among high-resolution features. Vision transformer-based models have been highly successful in various domains of artificial intelligence, including natural language processing and computer vision, which have generated significant interest from academic and industrial researchers. These models have outperformed other networks like convolutional and recurrent networks in visual benchmarks, making them a promising candidate for image processing applications.

We show that this reliance on CNNs is not necessary and a pure transformer applied directly to sequences of image patches can perform very well on image classification tasks. When pre-trained on large amounts of data and transferred to multiple mid-sized or small image recognition benchmarks (ImageNet, CIFAR-100, VTAB, etc.), Vision Transformer (ViT) attains excellent results compared to state-of-the-art convolutional networks while requiring substantially fewer computational resources to train.

# Introduction

In recent years, the field of computer vision has witnessed a paradigm shift with the advent of transformer architectures, which have revolutionized natural language processing (NLP) and are now being effectively applied to image recognition tasks. Transformers, initially designed for sequence modelling in NLP, have shown great promise in handling visual data, offering an alternative to traditional convolutional neural networks (CNNs).

Self-attention-based architectures, particularly Transformers have become the model of choice in Natural Language Programming. The dominant approach is to pre-train on a large text corpus and then fine-tune on a smaller task-specific dataset. Transformers are models that focus entirely on the self-attention process to establish global dependencies between input and output, and they have dominated natural language modelling in recent years. Transformers and their variations have been thoroughly explored and used in natural language processing tasks such as machine translation, light-weight transformers, dynamic mask attention networks, language modelling, routing transformers, positional encoding schemes, and named entity identification.

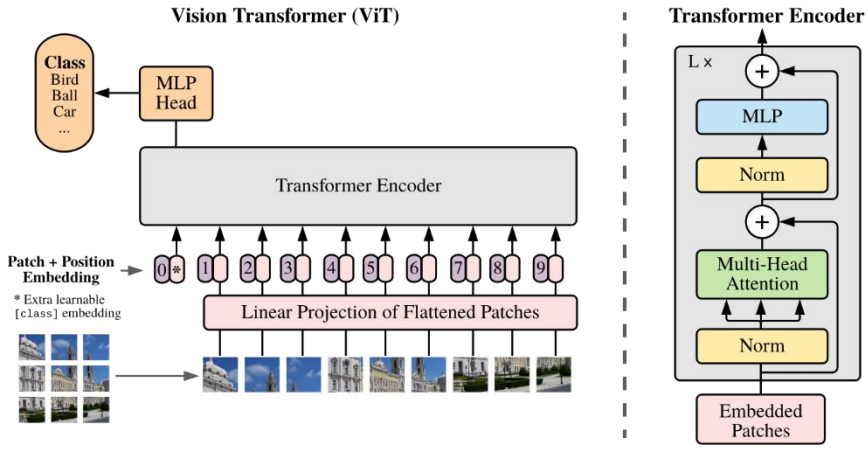
The Vision Transformer (ViT) is a pioneering approach that applies a pure transformer architecture to image data by treating images as sequences of patches. Each image is divided into fixed-size patches, linearly embedded, and fed into a transformer model. This method allows the transformer to process images similarly to how it handles sequences of words in NLP, capturing complex relationships between different parts of the image.

Transformers have established themselves as a groundbreaking technology in the realm of image recognition, providing a formidable alternative to convolutional neural networks (CNNs). By utilizing self-attention mechanisms and the capacity to capture global dependencies, vision transformers (ViTs) have achieved new performance benchmarks in visual data processing. Current research efforts are focused on enhancing these models, increasing their applicability and efficiency across a broad spectrum of image recognition tasks. As transformer models continue to advance, they possess the potential to significantly reshape the field of computer vision, facilitating further technological innovations and breakthroughs.

## Method/Algorithm:

### Data Preprocessing

The input images are first pre-processed by dividing each image into non-overlapping patches. Each patch is then linearly embedded into a fixed-dimensional vector. These embedded patches are then augmented with positional encodings to retain spatial information, which is crucial for effective image recognition.




### Vision Transformer Architecture

The Vision Transformer (ViT) architecture deviates from traditional convolutional neural networks by treating images as sequences of patches. The core components of the ViT model include:

### 1. Patch Embedding

**Image Division**:

- **Input Image**: Consider an input image $I$ of dimensions $H \times W \times C$, where $H$ is the height, $W$ is the width, and $C$ represents the number of colour channels.
- **Patch Creation**: The image is divided into $N$ patches, each of size $P \times P$. This means the image is broken down into smaller, non-overlapping square regions.

**Flattening Patches**:

- **Vector Representation**: Each patch, initially a $P \times P$ grid of pixels, is flattened into a one-dimensional vector. Since each pixel has C color channels, the size of this vector becomes $P^2 \cdot C$.

**2. Linear Projection**

**Projection to Embeddings**:

- **Embedding Space**: Each flattened patch vector is linearly projected into a higher-dimensional space to create the input embeddings. This process transforms the vector into another vector of dimension $D$.
- **Learnable Weights**: This transformation is achieved using a learnable weight matrix, enabling the model to adjust the projections during training.

**3. Positional Encoding**

**Adding Positional Information**:

- **Spatial Structure**: To maintain the spatial structure of the image patches, positional encodings are added to each patch embedding. This helps the model understand the position of each patch within the original image.
- **Combined Embedding**: The positional encodings are combined with the patch embeddings to form a new set of embeddings that include positional information. Additionally, an extra learnable embedding is included, which is used for classification purposes.

**4. Transformer Encoder**

**Self-Attention Mechanism**:

- **Input Sequence**: The sequence of combined embeddings, including positional information, is fed into a transformer encoder.

- **Attention Scores**: The transformer encoder uses a self-attention mechanism to calculate relationships between different patches, capturing global dependencies across the image. This mechanism involves computing attention scores using queries, keys, and values derived from the embeddings.

**Multi-Head Self-Attention**:

- **Multiple Attention Heads**: The self-attention mechanism operates with multiple heads, allowing the model to focus on different parts of the input sequence simultaneously. This provides a more comprehensive understanding of the image.

**Layer Norm and Residual Connections**:

- **Normalization**: Each attention layer is followed by a layer normalization step and a residual connection, which helps stabilize and accelerate the training process.
- **Feed-Forward Network**: After the self-attention layer, a feed-forward neural network is applied. This is followed by another layer normalization and residual connection, ensuring that the model can learn complex representations.

**Training Procedure**

The ViT model is pre-trained on large-scale image datasets to learn a robust representation of visual features. The pre-training involves optimizing a cross-entropy loss function using the Adam optimizer. After pre-training, the model is fine-tuned on specific downstream tasks such as image classification, object detection, and segmentation.

- **Pre-Training**: The model is trained on a diverse set of large-scale image datasets such as ImageNet, with data augmentation techniques applied to improve generalization.
- **Fine-Tuning**: For specific tasks, the pre-trained model is fine-tuned on smaller, task-specific datasets. During fine-tuning, only a subset of the parameters may be updated to adapt the model to the new data.

## Evaluation Metrics

The performance of the ViT model is evaluated using standard metrics such as accuracy, precision, recall, and F1-score. For image classification tasks, top-1 and top-5 accuracy are commonly reported. For object detection and segmentation, mean Average Precision (mAP) is used as the evaluation metric.

## Computational Considerations

The ViT model leverages the parallelization capabilities of transformers, enabling efficient training on high-resolution images. Techniques such as mixed-precision training and distributed computing are employed to manage the computational complexity and memory usage during training.

## Implementation Details

The implementation is done using popular deep learning frameworks such as TensorFlow and PyTorch. The code is optimized for GPU acceleration, and hyperparameters such as learning rate, batch size, and number of epochs are tuned through extensive experimentation.

## References:

[1] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," Google Research, Brain Team.

[2] C. S. Kameswari, J. Kavitha, T. S. Reddy, B. Chinthaguntla, S. Jagatheesaperumal, S. Gaftandzhieva, and R. Doneva, "An overview of vision transformers for image processing: A survey," in Dept. of Computer Science and Engineering (AI&ML), Keshav Memorial Institute of Technology, Hyderabad, India; Dept. of Information Technology, BVRIT HYDERABAD College of Engineering for Women, Hyderabad, India; Dept. of Electronics and Communication Engineering, Malla Reddy Engineering College, Secunderabad, India; Dept. of Electronics and Communication Engineering, Sheshadri Rao Gudlavalleru Engineering College, Gudlavalleru, India; Dept. of Electronics and Communication Engineering, Mepco Schlenk Engineering College, Sivakasi, India; University of Plovdiv ―Paisii Hilendarskiǁ, Plovdiv, Bulgaria.

[3] M. Khalil, A. Khalil, and A. Ngom, "A comprehensive study of vision transformers in image classification tasks," in Dept. of Computer Science, University of Windsor, Windsor, Ontario, Canada.

[4] J. Park, K. Kahatapitiya, D. Kim, S. Sudalairaj, Q. Fan, and M. S. Ryoo, "Grafting vision transformers," in Stony Brook University; MIT-IBM Watson AI Lab; Amazon.