WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Vision Transformer: A New Frontier in Computer Vision

Yashas S[1], Tejas N K[2], S Mokshith Reddy[3], Ms Manusha A[4], Shamun S[5]

[1] Dayananda Sagar Academy of Technology and Management yashas1598@gmail.com
[2] Dayananda Sagar Academy of Technology and Management tejas7892250557@gmail.com
[3] Dayananda Sagar Academy of Technology and Management saimokshith1221@gmail.com
[4] Assistant prof Dayanada sagar Academy of technology and management manusha-cse@dsatm.edu.in
[5] Dayananda Sagar Academy of Technology and Management shamunshaik@gmail.com

ABSTRACT -

Vision Transformer (ViT) is a neural network architecture that applies the Transformer model, previously used in natural language processing, to image recognition tasks. ViT divides an image into patches, which are then linearly embedded and fed into a Transformer encoder. This approach eliminates the need for convolutional layers, which are commonly used in traditional computer vision models. ViTs have shown promising results in various image recognition tasks, including image classification, object detection, and image segmentation

INDEX TERMS -

- Vision Transformer
- Image Recognition
- Computer Vision
- Deep Learning
- Neural Networks
- Transformer Model

INTRODUCTION :

Gesture recognition is essential in Transformer, first applied to the field of natural language processing, is a type of deep neural network mainly based on the self-attention mechanism. Thanks to its powerful representation capabilities, researchers are exploring at methods to use transformer to computer vision applications. In a range of visual benchmarks, transformer-based models perform similar to or better than other types of networks such as convolutional and recurrent neural networks. Given its strong performance and minimal requirement for vision-specific inductive bias, transformer is garnering more and more interest from the computer vision community. We evaluate these vision transformer models in this study by classifying them in various tasks and weighing their benefits and drawbacks. We primarily investigate the following areas: video processing, low-level vision, high/mid-level vision, and the backbone network. In order to push transformer into actual device-based applications, we also incorporate effective transformer approaches. Additionally, since the self-attention process is the foundation of transformer, we also briefly examine it in computer vision. We address the difficulties and offer a number of recommendations for future research on vision transformers towards the end of this paper.

Initiative: A Revolutionary Approach to Vision Transformer (ViT) is a groundbreaking neural network architecture that leverages the power of Transformer models, initially developed for natural language processing, to revolutionize image recognition tasks. By dividing images into patches and feeding them into a Transformer encoder, ViT eliminates the need for convolutional layers, a cornerstone of traditional computer vision models. This innovative approach has demonstrated remarkable results in various image recognition tasks, including image classification, object detection, and image segmentation.

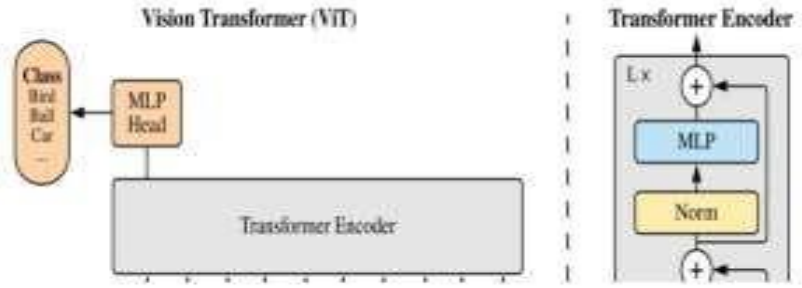
OpenCV (Open-Source Computer Vision Library) is a powerful open-source library designed for real-time computer vision and image processing. It provides a wide range of tools for tasks such as object detection, tracking, image filtering, and more, making it an essential framework for developing innovative computer vision applications. In our project, OpenCV plays a critical role in enabling the detection and tracking of hand gestures through the live webcam feed. By leveraging its robust image processing capabilities, we can accurately detect the user's fingertip and track its motion in real time, facilitating the creation of frames from the Video.

Within our initiative: Hybrid architectures: Combine ViTs with CNNs for improved performance and efficiency.

- Inductive biases: Incorporate prior knowledge about image structures into the model.
- Interpretability: Develop methods to understand the model's decision-making process.

- Computational efficiency: Optimize training and inference for resource-constrained devices.

This comprehensive guide provides a basic framework for implementing Vision Transformers. By understanding the core concepts and experimenting with different architectures and techniques, you can unlock the full potential of this powerful approach to image recognition.



Source: transformer Blog[1]

FIGURE 1: visual representation of image extraction using transformer decoder

LITERATURE SURVEY :

Vision Transformers (ViTs) have emerged as a powerful paradigm in computer vision, demonstrating remarkable performance in various tasks like image classification, object detection, and segmentation. This survey aims to provide a comprehensive overview of the current state-of-the-art in ViT research Transformer Model: The Transformer model, originally introduced in the context of natural language processing (NLP), employs self-attention mechanisms to capture longrange dependencies within a sequence.

* ViT (Vision Transformer): The seminal work by Dosovitskiy et al. (2020) adapted the Transformer model to image recognition by dividing images into patches, treating them as a sequence of tokens, and feeding them into a Transformer encoder.Combining the strengths of Convolutional Neural Networks (CNNs) and ViTs has led to significant improvements.. Swin Transformer (Liu et al., 2021): Introduces hierarchical feature maps using shifted windows, enabling efficient long-range dependencies and multi-scale representations.. Local Window Attention (Wang et al., 2021): Divides the feature map into local windows, reducing computational complexity while preserving spatial locality.

Efficient ViTs: leverages computer vision to track finger paths and generate corresponding text output. However, your methodology distinguishes itself by emphasizing simplicity, resource efficiency, and enhanced accessibility. Your implementation avoids complex object identification and behavior analysis steps, instead focusing on gesture detection for drawing paths and directly converting them into meaningful outputs like messages or drawings. While the Virtual Air Canvas expands its scope to applications like autonomous surveillance and video indexing, your approach prioritizes practical usability, especially in scenarios requiring rapid and intuitive human-computer interaction. Both projects align in addressing the needs of the hearing-impaired community but differ in their technical depth and target use cases.

METHODOLOGY :

The Vision Transformer methodology is structured The Vision Transformer (ViT) is a groundbreaking neural network architecture that applies the Transformer model, initially developed for natural language processing (NLP), to image recognition tasks. Here's a breakdown of its key methodology:

1. Image Patching:

- The input image is divided into a sequence of smaller, fixed-size patches.
- Each patch is essentially treated as a "token" or "word" in the image.

2. Linear Embedding:

- Each image patch is flattened into a one-dimensional vector.
- These vectors are then linearly projected into a higherdimensional space, creating a set of patch embeddings.

3. Positional Encoding:

- Positional information is crucial for the Transformer to understand the spatial relationships between the patches.
- Positional embeddings are added to the patch embeddings to encode their relative positions within the image.

CONCLUSION :

Vision Transformer demonstrates ViTs leverage the power of Transformer models, initially developed for NLP, to process images. They treat images as a sequence of patches and utilize self-attention mechanisms to capture long-range dependencies .They have shown strong performance across a range of computer vision tasks, often surpassing traditional CNN-based models.. Global Context: They can effectively capture global dependencies within images, crucial for tasks like object detection and scene understanding. Flexibility: The Transformer architecture is highly adaptable and can be applied to various computer vision tasks. ViTs represent a significant advancement in computer vision, opening up new possibilities for image recognition and analysis.

Ongoing research focuses on improving their efficiency, exploring new architectures, and expanding their applications. As research progresses, ViTs are likely to play an increasingly important role in shaping the future of computer vision.

REFRENCES :

1. [1][https://journals.bilpubgroup.com/index.php/jcsr/article/view/5610]
2. Dosovitskiy, A., Beyer, L., Kolesnikov, A., et al. (2020). An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale..
3. Liu, Z., Lin, Y., Cao, Y., et al. (2021). Swin Transformer: Hierarchical Vision Transformer using Shifted Windows.
4. Wang, X., Xie, H., Li, C., et al. (2021). Towards a Unified Architecture of Convolution and Self-Attention for Vision..
5. Chen, H., Fan, C., Xie, R., et al. (2021). MobileViT: Multi-Scale Mobile Vision Transformers.
6. D.
7. P. Rai, R. Gupta, V. Dsouza and D. Jadhav, "Virtual Canvas for Interactive Learning using OpenCV," 2022 IEEE 3rd Global Conference for Advancement in Technology (GCAT), Bangalore, India, 2022, pp. 1-5.
8. D. Lee, H. Yoon and J. Kim, "Continuous gesture recognition by using gesture spotting," 2016 16th International Conference on Control, Automation and Systems (ICCAS), Gyeongju, Korea (South), 2016, pp. 1496-1498.