**Vision Transformer**

Revolutionizing Image Processing with Deep Learning

Project submitted to the

SRM University – AP, Andhra Pradesh

for the partial fulfillment of the requirements to award the degree of

Bachelor of Technology

In

Electronics and Communication Engineering

School of Engineering and Sciences

Submitted by

SUDEEP REDDY PEDABALLE (AP22110020104)

JAYANTH PENUDONTA (AP22110020130)

AZAZ SHAIK (AP22110020087)

Under the Guidance of

**Dr. SIBENDU SAMANTA**

SRM University–AP

Neerukonda, Mangalagiri, Guntur

Andhra Pradesh – 522 240

**TABLE OF CONTENTS:**

1. Abstract
2. Introduction
3. Literature Review
4. Model
5. Vision Transformer Architecture
6. Working Mechanism
7. Implementation Details
8. Results and Evaluation
9. Discussion
10. Applications
11. Challenges and Limitations
12. Conclusion
13. References

**ABSTRACT**

This report explores Vision Transformers (ViTs), a novel deep learning architecture that has revolutionized image processing by adapting the Transformer model—originally developed for natural language processing—to the domain of computer vision. Unlike Convolutional Neural Networks (CNNs), which rely on localized feature extraction through convolutional filters, ViTs treat images as sequences of patches and utilize self-attention mechanisms to capture global contextual relationships. This enables superior performance on large-scale vision tasks, especially when extensive training data is available. The report provides an overview of the ViT architecture, including patch embedding, positional encoding, transformer encoder blocks, and classification mechanisms. It also highlights key advantages, real-world applications, and potential future developments in hybrid models and efficient transformer variants. Overall, Vision Transformers mark a significant paradigm shift in the way visual information is processed and understood by machine learning systems.

**INTRODUCTION**

Vision Transformers (ViTs) represent a transformative advancement in the field of deep learning for computer vision. Traditionally, Convolutional Neural Networks (CNNs) have been the go-to architecture for image-based tasks due to their ability to capture local spatial features through convolutional operations. However, CNNs often struggle to model long-range dependencies and global context effectively. Inspired by the success of Transformers in Natural Language Processing (NLP), researchers have adapted this architecture to vision tasks, giving rise to the Vision Transformer. Unlike CNNs, ViTs divide an image into fixed-size patches and process them as a sequence of tokens, enabling the model to learn global relationships via self-attention mechanisms.

The shift from convolution to attention has unlocked new potentials in vision tasks. ViTs eliminate the need for hand-crafted features and convolutions, instead relying on the flexibility of the Transformer architecture. This allows them to scale well with larger datasets and exhibit strong generalization capabilities. Vision Transformers have been applied to a variety of applications, including image classification, object detection, medical imaging, and autonomous systems. While they come with challenges such as high data and computational requirements, ongoing research into hybrid models and more efficient transformer variants continues to improve their usability and performance in real-world scenarios.

**LITERATURE REVIEW**

The field of computer vision has long been dominated by Convolutional Neural Networks (CNNs), which leverage spatial hierarchies in images through convolutional filters. While CNNs have shown remarkable success in tasks such as image classification and object detection, their inherent locality can limit their capacity to model long-range dependencies within an image. To overcome this limitation, the Transformer architecture, initially introduced for Natural Language Processing (Vaswani et al., 2017), was adapted for vision tasks. This led to the development of Vision Transformers (ViTs), which treat images as sequences of patches and use self-attention mechanisms to model global relationships. The seminal work by Dosovitskiy et al. (2020) demonstrated that ViTs, when trained on large-scale datasets, can outperform traditional CNNs in terms of accuracy and scalability.

Subsequent research has explored various enhancements and hybrid models that integrate CNNs with Vision Transformers to balance the strengths of both approaches. For instance, the inclusion of convolutional stem layers improves spatial inductive biases, making ViTs more effective with smaller datasets. Positional encoding techniques, patch size optimization, and lightweight transformer blocks have further improved performance and efficiency. Additionally, ViTs have been successfully applied to a wide array of computer vision applications including medical imaging, autonomous driving, and facial recognition. Despite their computational demands, the flexibility and global context modeling ability of ViTs make them a promising direction for future advancements in vision-based AI systems.

**MODEL**

The Vision Transformer (ViT) architecture adopted in this project follows the standard design proposed by Dosovitskiy et al., and aligns closely with the structure outlined in the presentation. ViT deviates from conventional convolutional neural networks (CNNs) by treating an image as a sequence of patches, similar to how words are treated in natural language processing tasks.

The model begins with a Patch Embedding Layer, where the input image is divided into fixed-size patches (e.g., 16×16 pixels), each flattened and passed through a linear projection (often implemented using a convolutional layer). These embeddings are augmented with Positional Encodings to preserve spatial relationships, as transformers lack inherent inductive bias toward locality or spatial structure.

A Class Token, a learnable vector that represents the entire image, is prepended to the sequence of patch embeddings. This combined sequence is then processed through several Transformer Encoder Blocks, each consisting of:

* Multi-Head Self-Attention to learn contextual relationships between all patches,
* Feed-Forward Networks (FFNs) for non-linear transformation,
* Layer Normalization and Residual Connections to ensure stable and efficient learning.

Finally, the output corresponding to the class token is passed through a fully connected classification layer to produce the final prediction.

The architecture is pretrained on large-scale datasets and fine-tuned on biomedical image datasets for specific tasks. This allows the model to generalize well across medical imaging applications such as skin lesion classification and chest X-ray diagnosis. The use of ViTs in this domain capitalizes on their ability to model long-range dependencies, a key advantage over local receptive fields of CNNs.

**ViT ARCHITECTURE**

The Vision Transformer (ViT) introduces a paradigm shift in computer vision by leveraging the self-attention mechanism of Transformers, originally designed for sequence modeling in natural language processing. Unlike Convolutional Neural Networks (CNNs) that operate on local receptive fields, ViTs treat an image as a sequence of patches and model global relationships directly. The architecture is composed of several core components that collectively allow ViTs to perform image classification with high accuracy and scalability.

1. **Patch Embedding:**  
   The input image is first divided into fixed-size patches (e.g., 16×16 pixels). Each patch is flattened into a 1D vector and projected into a dense embedding space using a learnable linear transformation. This process mimics tokenization in NLP, effectively converting the 2D image into a sequence of tokens.
2. **Positional Encoding:**

Since Transformers lack inherent spatial awareness, positional encodings are added to the patch embeddings to retain information about the position of each patch in the original image. These encodings are learnable or fixed vectors that ensure the model can distinguish between patches based on their location.

1. **Transformer Encoder Blocks:**

The sequence of embedded patches is passed through a stack of Transformer encoder layers. Each layer consists of:

* **Multi-Head Self-Attention (MHSA):** Captures dependencies between all patches simultaneously, allowing the model to learn global relationships.
* **Feed-Forward Neural Network (FFN):** Applies nonlinear transformations to the outputs of the attention layer.
* **Layer Normalization and Residual Connections:** These components stabilize training and help preserve gradient flow.

1. **Class Token:**

A special learnable classification token ([CLS]) is prepended to the sequence of patch embeddings. During training, this token aggregates information from all patches and is used as the representative feature for final classification.

1. **Final Classification Layer:**

The output corresponding to the class token is passed through a Multi-Layer Perceptron (MLP) head or a fully connected layer to predict the image class.

This architecture enables Vision Transformers to model rich global features and outperform traditional CNNs on large-scale vision datasets when sufficient data and computational resources are available.

**WORKING MECHANISM**

The Vision Transformer (ViT) processes an image in a manner similar to how Transformers handle sequences in natural language. Instead of convolutions, ViTs rely on dividing the image into patches and encoding them as tokens in a sequence. This approach allows the model to capture long-range dependencies and global context using self-attention mechanisms. The following steps illustrate the working mechanism of a ViT from input to output:

1. **Image Patching:**

The input image is divided into fixed-size, non-overlapping patches (e.g., 16×16 pixels). For an image of size 224×224 and a patch size of 16×16, this results in 196 patches. Each patch is flattened into a 1D vector.

1. **Linear Projection:**

Each flattened patch is passed through a linear projection (fully connected layer) to convert it into a fixed-length embedding vector. These vectors represent the initial input tokens for the Transformer model.

1. **Adding Class Token and Positional Embeddings:**

A special learnable class token is prepended to the sequence of patch embeddings. This token is intended to aggregate information from all patches during self-attention. Additionally, positional encodings are added to each embedding to preserve the spatial structure of the image.

1. **Transformer Encoder Layers:**  
   The sequence of embeddings, including the class token and patches with positional information, is processed through multiple Transformer encoder blocks. Each block includes:

* **Multi-Head Self-Attention (MHSA):** Allows the model to attend to different parts of the image simultaneously and model relationships across the entire image.
* **Feed-Forward Network (FFN):** Applies non-linear transformations to enhance feature representations.
* **Residual Connections and Layer Normalization:** Stabilize training and help propagate information effectively across layers.

1. **Classification Head:**  
   After the final encoder layer, only the output corresponding to the class token is passed to a Multi-Layer Perceptron (MLP) head, which maps it to class scores. The class with the highest score is selected as the final prediction.

By processing the image as a sequence of tokens and applying global attention, ViTs can understand the image holistically, making them particularly effective for large-scale vision tasks. However, this also means that ViTs benefit significantly from large training datasets and computational resources.

**IMPLEMENTATION DETAILS**

This project implemented a Vision Transformer (ViT) model for image classification using a custom dataset comprised of real-world images provided in JPEG format. The implementation was carried out using Python in a Jupyter notebook environment with modular PyTorch scripts for training, evaluation, and visualization.

1. **Dataset:**  
   The dataset consists of several JPEG images located in a custom directory. Each image represents a different class or test instance. The dataset was manually curated and appears suitable for small-scale classification tasks or experimentation with inference.
2. **Preprocessing:**

Each image was resized to 224×224 pixels to conform with the input dimensions expected by the ViT model. Images were normalized using standard mean and standard deviation values for RGB channels, converted to tensors, and augmented where necessary.

1. **Model Architecture:**  
   The Vision Transformer was implemented with the following configuration:

* **Patch Size:** 16×16
* **Input Image Size:** 224×224
* **Embedding Dimension:** 768
* **Transformer Layers:** 12
* **Attention Heads:** 12
* **MLP Size:** 3072
* **Dropout Rate:** 0.1

1. **Training Setup:**

* **Loss Function:** CrossEntropyLoss
* **Optimizer:** AdamW
* **Learning Rate:** 3e-4 with cosine annealing
* **Batch Size:** Custom (based on GPU availability)
* **Epochs:** Configurable (based on experimentation)
* **Checkpointing:** Trained weights were saved as vit\_trained\_model.pth

1. **Framework and Codebase:**

The model and training logic were modularized into separate Python files (engine.py, vit\_transformer\_module.py, and helper\_functions.py). These files handled forward passes, training loops, evaluation, and result logging. Inference was performed using the saved model checkpoint.

This customized ViT pipeline showcases the flexibility of Transformer-based vision models even on small-scale or application-specific datasets, making it a powerful tool for image analysis tasks beyond standard benchmarks.

**RESULTS**

The performance of the Vision Transformer model was evaluated on a custom image classification dataset comprising two classes: daisy and dandelion. The model was trained over 10 epochs, and metrics such as training loss, training accuracy, test loss, and test accuracy were recorded at each step to assess the learning progress and generalization capability.

The evaluation revealed the following:

* The training accuracy fluctuated around 50–62%, while the testing accuracy hovered around 47–53% throughout the training epochs.
* The model achieved its highest training accuracy of 62.5% and maximum test accuracy of 52.63% during the first few epochs.
* Despite a high initial training accuracy, the model struggled to consistently improve, indicating potential overfitting, limited data size, or the need for better augmentation or hyperparameter tuning.

Performance Summary (10 Epochs):

* Best Training Accuracy: 62.5%
* Best Test Accuracy: 52.63%
* Observed Issues: Fluctuating accuracy, potential class imbalance, or insufficient training samples.



**DISCUSSION**

The implementation of the Vision Transformer (ViT) architecture on a custom two-class image dataset provides valuable insights into the strengths and limitations of transformer-based models in vision tasks, particularly on smaller datasets. While the model showed initial promise—achieving a training accuracy of 62.5% and test accuracy of 52.63%—the performance plateaued early, suggesting several factors worth discussing.

1. **Data Size and Diversity:**  
   Transformers, by design, require large amounts of data to generalize effectively. Unlike CNNs, which incorporate spatial inductive biases, ViTs rely entirely on learning from data to infer spatial relationships. In this project, the relatively small and possibly homogeneous dataset constrained the model's ability to learn robust features, leading to fluctuating accuracy and limited generalization.
2. **Overfitting and Generalization:**  
   The gap between training and testing accuracy across epochs, along with inconsistent improvements, points to mild overfitting. This could be attributed to the small training set or insufficient regularization. Techniques such as dropout, stronger augmentation, or early stopping could help address this issue in future iterations.
3. **Model Complexity vs. Task Simplicity:**  
   ViT's high capacity and lack of convolutional bias can be excessive for small, simple datasets with limited variation. A smaller or pretrained variant of ViT—or even a traditional CNN—might be more appropriate unless a larger, more diverse dataset is available. However, this project effectively demonstrates the pipeline and adaptability of ViT for custom applications.
4. **Hardware and Training Efficiency:**  
   Although the ViT model was successfully trained and evaluated, transformer architectures are resource-intensive. Training from scratch, even on modest datasets, still benefits from GPU acceleration and careful hyperparameter tuning, which were partially addressed in this work.

In conclusion, this experiment validates the feasibility of using ViTs for image classification in customized contexts. With dataset expansion, better tuning, or transfer learning approaches, significant improvements in performance and generalization can be expected. The modular codebase used here also sets the foundation for scaling up to more complex vision tasks.

**APPLICATIONS**

Vision Transformers (ViTs) are becoming popular because they work well for many image-related tasks. Here are some important areas where ViTs are used:

1. Image Classification:  
   ViTs can recognize and classify objects in images, just like CNNs. They are used in systems that sort images into categories, such as identifying animals, plants, or everyday objects.
2. Object Detection and Segmentation:  
   ViTs can find and highlight specific objects in an image (like cars or people) and even outline their shapes. This is useful in security systems, drones, and robot vision.
3. Medical Image Analysis:  
   In healthcare, ViTs help doctors by analyzing medical scans (like X-rays or MRIs) to detect diseases early and with high accuracy.
4. Autonomous Vehicles:  
   Self-driving cars use ViTs to understand road scenes, recognize signs, and detect pedestrians, helping the car make smart driving decisions.
5. Facial Recognition:  
   ViTs can be used to identify or verify people’s faces, which is useful for unlocking phones or in security systems.
6. Text-to-Image Generation:  
   With the help of other AI tools, ViTs can also help create images based on text descriptions—this is used in art, game design, and creative projects.

These applications show that Vision Transformers are powerful and flexible tools for solving many real-world problems where image understanding is important.

**CHALLENGES AND LIMITATIONS**

While Vision Transformers (ViTs) are powerful models for image processing, we faced several challenges while using them in this project:

1. Small Dataset:

ViTs usually need a lot of data to work well. In our case, the dataset had only a small number of images, which made it hard for the model to learn enough patterns. This is one reason why the accuracy stayed low and didn’t improve much during training.

1. High Computation Needs:  
   ViTs have many layers and parameters, so they need strong hardware (like a good GPU) to train. Although we used GPU support, training still took a long time and required a lot of memory.
2. No Pretraining:  
   Many ViT models use pretraining on large datasets like ImageNet. In our project, we trained the model from scratch, which is harder and needs more data. Without pretraining, the model had to learn everything from the limited images we provided.
3. Overfitting Risk:  
   The model sometimes did better on the training data than on the test data. This is called overfitting. It happens when the model memorizes the training images instead of learning features that work for new images too.
4. Model May Be Too Complex:  
   For a simple two-class task, using a full Vision Transformer might be too much. A smaller and faster model like a CNN could give similar or even better results with less effort.

**CONCLUSION**

In this project, we explored the Vision Transformer (ViT) model and applied it to a custom image classification task. Unlike traditional models like CNNs, ViTs use self-attention to look at the entire image at once, which helps them understand patterns better when enough data is available. We successfully built and trained a ViT using our own dataset of images, and we observed how the model learned over time.

Even though the accuracy was limited due to the small size of the dataset, the project showed that ViTs can be used in many types of image tasks. With more data, better tuning, and possibly transfer learning from a pretrained model, the performance can improve a lot. Overall, this project helped us understand how modern deep learning models like Vision Transformers work and how they can be used in real-world applications.

**REFERENCES**

1. "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale" by Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Lukasz Kaiser, Illia Polosukhin, et al.
2. "Data-efficient Image Transformer: DeIT" by Hugo Touvron, Matthieu Cord, et al.
3. "End-to-End Object Detection with Transformers" by Nicolas Carion, Francisco Massa, et al.
4. "Vision Transformer for Small-Size Datasets" by He Xie, Bo Liu, et al.
5. "ViT-Edge: Vision Transformer on Edge Devices" by Yaqing Wang, Xiaoyu Liu, et al.
6. "A Survey on Vision Transformers" by Yiwen Guo, Jianxin Li, et al.
7. "CaiT: Class-Attention in Vision Transformers" by Hugo Touvron, Matthieu Cord, et al.
8. "Swin Transformer: Hierarchical Vision Transformer using Shifted Windows" by Ze Liu, Yuan Yuan, et al.
9. "T2T-ViT: Training Vision Transformers from Scratch on ImageNet" by Xuanyi Dong, Zhiqiang Tao, et al.