**Gokhale Education Society's**

**R. H. Sapat College of Engineering, Management Studies and Research,**

**Nashik - 422 005, (M.S.), INDIA**

Seminar on,

# Transformers for Image Detection

In partial fulfillment of requirements for the degree
Third Year Computer Engineering

By
Name of the Candidate : Pranamya Nilesh Deshpande
Exam Seat No. :
Roll No. : 36

Under the guidance of
Name of the Guide : **Mrs. R. D. Narwade**

# 1. Introduction

- Paradigm shift in computer vision with transformer architectures
- Vision Transformers (ViTs) as an alternative to CNNs
- Application of transformer architecture to image data
- Images treated as sequences of patches
- Capturing complex relationships between image parts

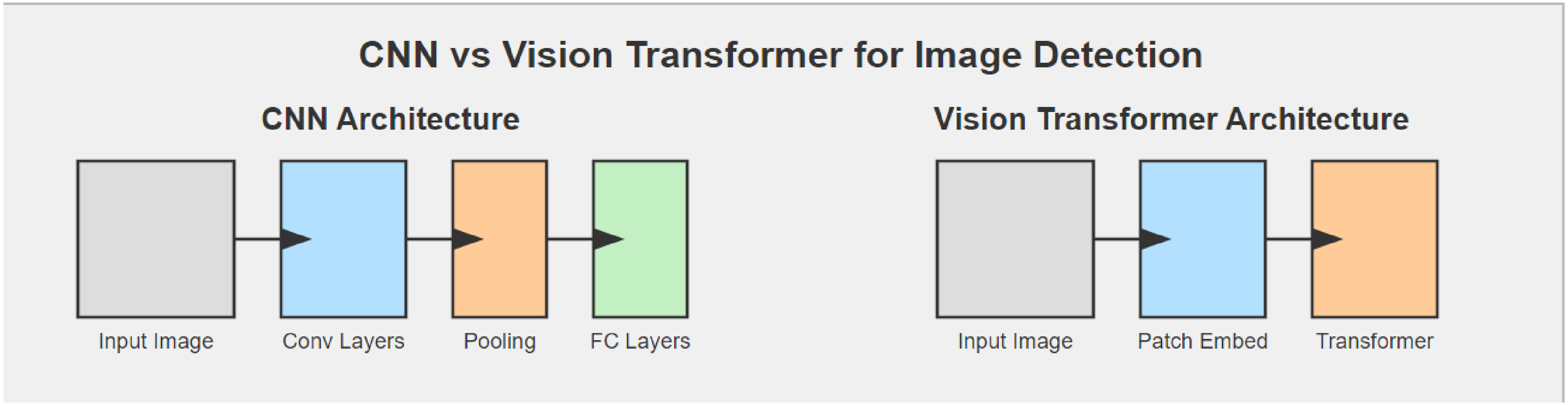

Fig. 1 Comparison of CNN and ViT Architecture

# 2. Literature Survey

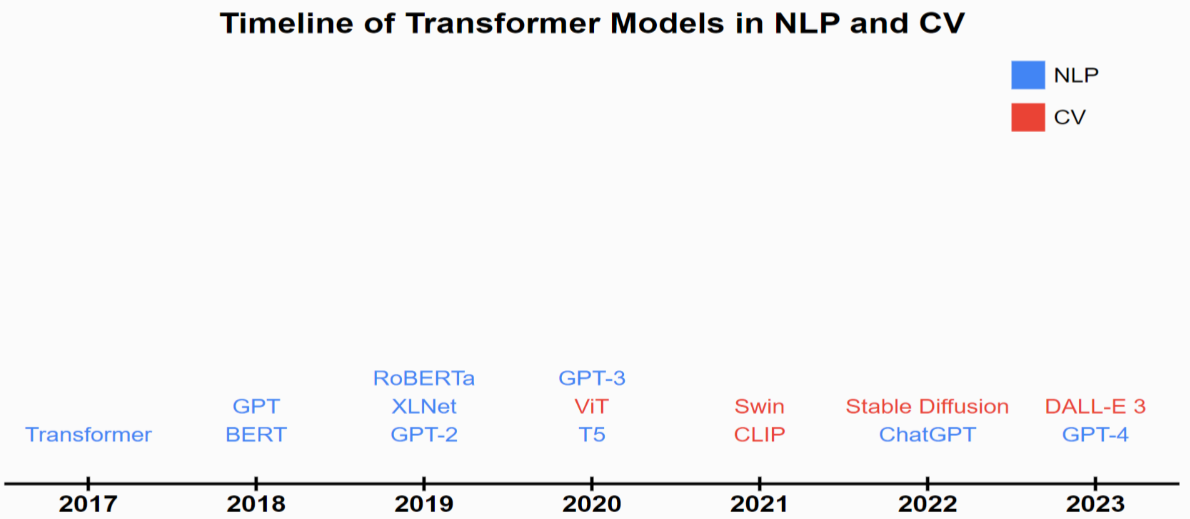

Fig. 2 Timeline of progress in Transformers

- Transformers as the model of choice in Natural Language Processing
- Pre-training on large text corpora and fine-tuning on task-specific datasets
- Applications in machine translation, language modeling, named entity identification
- Vision Transformer (ViT) as a pioneering approach in computer vision
- ViT outperforming CNNs in visual benchmarks

# 3. Details of Design/Technology

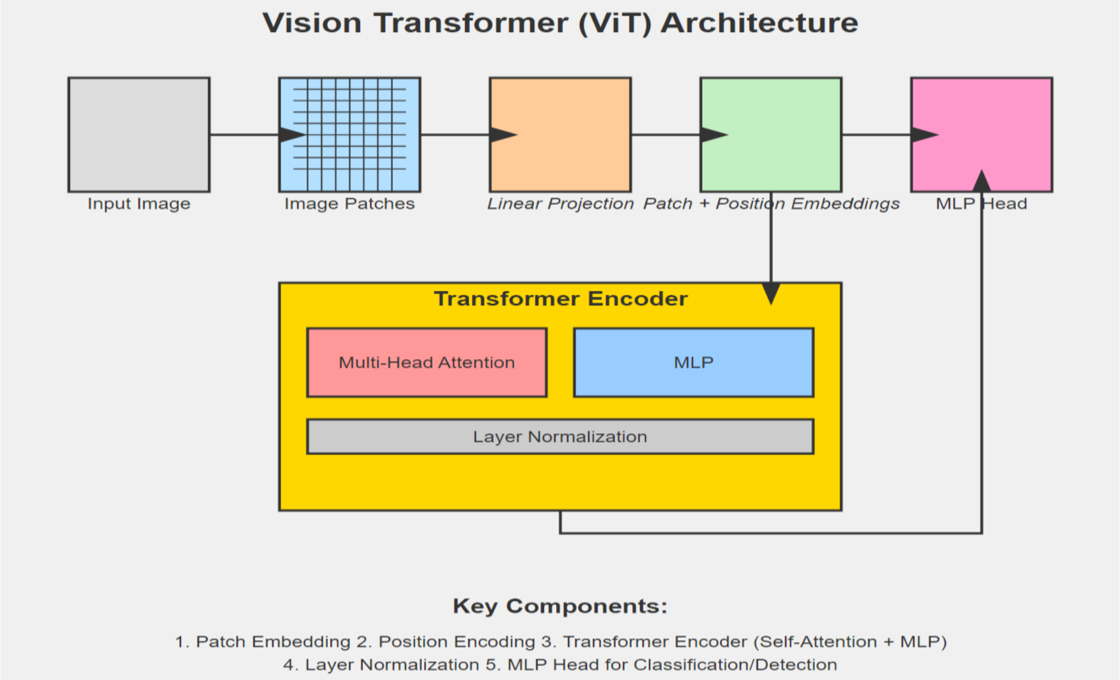
Vision Transformer (ViT) Architecture
Input Image
Image Patches
Linear Projection
Patch + Position Embeddings
MLP Head
Transformer Encoder
Multi-Head Attention
MLP
Layer Normalization
Key Components:
1. Patch Embedding 2. Position Encoding 3. Transformer Encoder (Self-Attention + MLP)
4. Layer Normalization 5. MLP Head for Classification/Detection

Fig. 3

- Image division into non-overlapping patches
- Linear embedding of patches into fixed-dimensional vectors
- Addition of positional encodings to retain spatial information
- Core components: Patch Embedding, Linear Projection, Positional Encoding, Transformer Encoder
- Self-attention mechanism for capturing global dependencies

# 4. Topic / Technology Necessity

- Limitations of CNNs in capturing global relationships
- Need for models that can process high-resolution images efficiently
- Demand for architectures that can leverage large-scale pre-training
- Requirement for models adaptable to various computer vision tasks
- Push for state-of-the-art performance in image recognition benchmarks

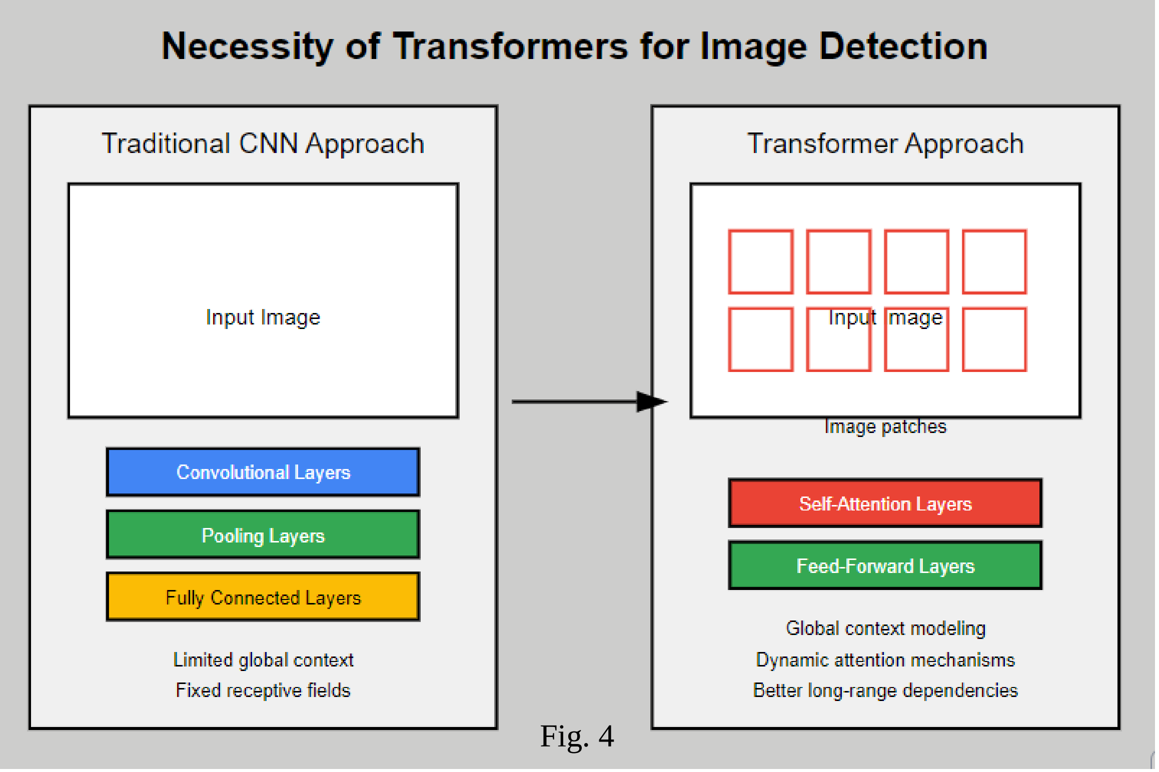

Fig. 4

# 5. Algorithm / Analytical / Experimental Work

- Pre-processing: Image division into patches, flattening, and embedding
- Transformer Encoder: Self-attention mechanism with multiple heads
- Layer normalization and residual connections for stability
- Pre-training on large-scale datasets (e.g., ImageNet)
- Fine-tuning on specific tasks (classification, detection, segmentation)
- Evaluation using metrics like accuracy, precision, recall, F1-score

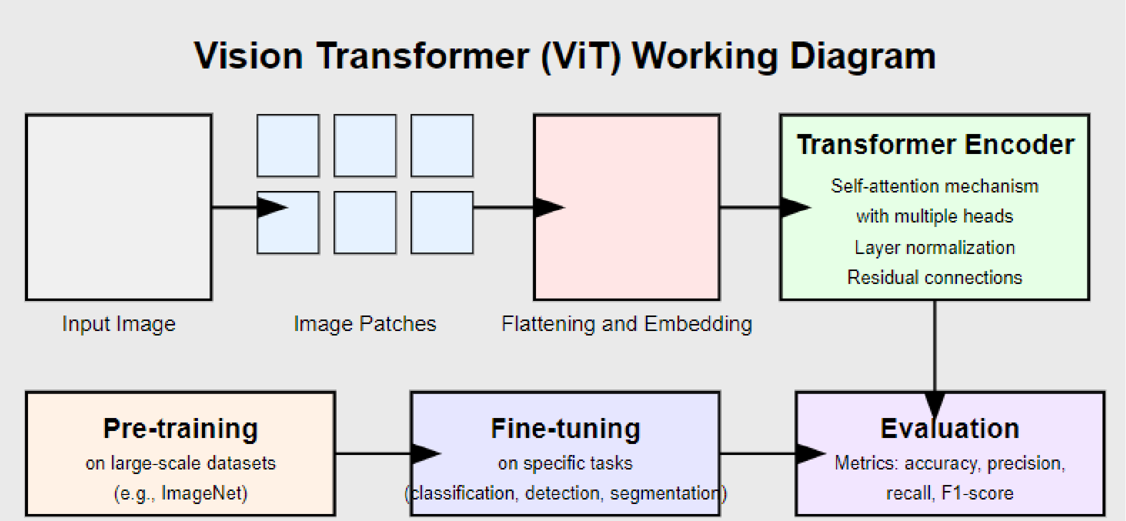

Fig. 5

# 6. Applications

- Image Classification
- Object Detection
- Image Segmentation
- Visual Question Answering
- Image Generation and Manipulation

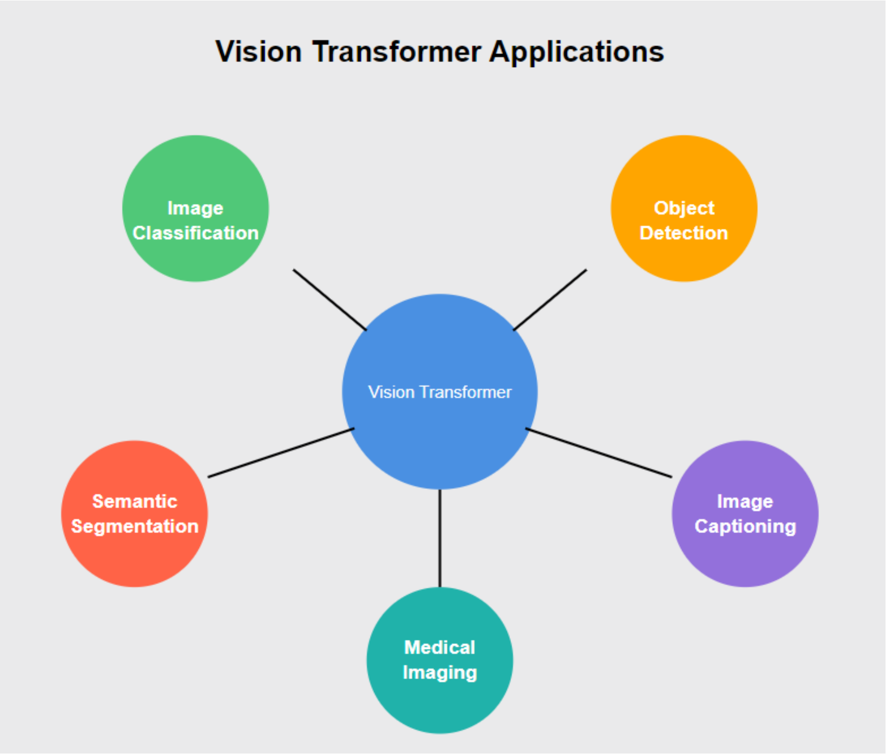

Fig. 6

# 7. Discussion and Conclusion

- ViTs achieve excellent results compared to state-of-the-art CNNs
- Require fewer computational resources for training
- Potential to reshape the field of computer vision
- Ongoing research to enhance efficiency and applicability
- Promise for future technological innovations in image recognition

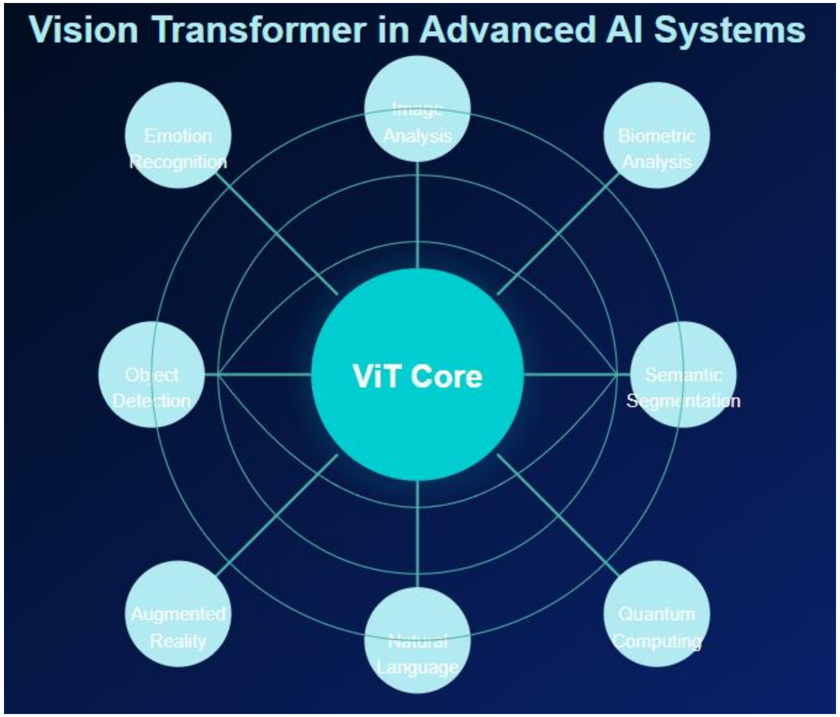

Fig. 7

# 8. References

1. A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," Google Research, Brain Team.
2. C. S. Kameswari et al., "An overview of vision transformers for image processing: A survey," Various Institutions.
3. M. Khalil et al., "A comprehensive study of vision transformers in image classification tasks," University of Windsor, Canada.
4. J. Park et al., "Grafting vision transformers," Stony Brook University; MIT-IBM Watson AI Lab; Amazon.