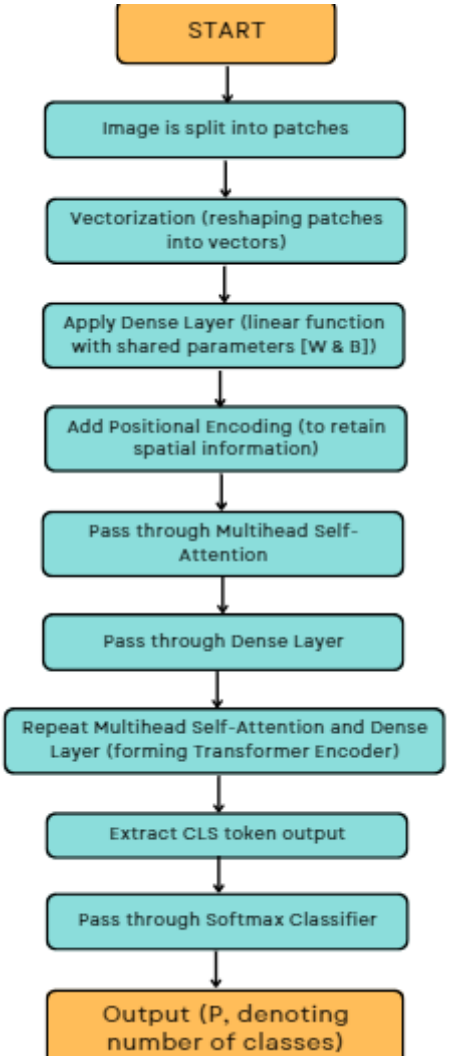
What Are Vision Transformers?

For years, Convolutional Neural Networks (CNNs) dominated computer vision. Meanwhile, Transformers revolutionized Natural Language Processing (NLP). Researchers then explored the possibility of using Transformers for image processing—thus, Vision Transformers (ViTs) were introduced.

How Do ViTs Work?



Unlike CNNs, which scan images using filters, ViTs follow a different approach:

1. Dividing the Image into Patches

- **Instead of analyzing the entire image at once, ViTs split it into smaller patches.**
- **Example: A 256×256 image can be divided into 16×16 patches.**

2. Flattening and Converting Patches into Vectors

- **Each patch is transformed into a vector representation, similar to how words in NLP are converted into embeddings.**

3. Adding Positional Encoding

- **Since Transformers do not inherently process data in a sequential order, positional embeddings are added to retain spatial information.**

4. Applying Self-Attention Mechanism

- **Self-attention allows patches to interact with each other, enabling the model to understand both local and global features of an image.**

5. Classification Layer

- **The processed image representation is passed through a fully connected (MLP) layer to make predictions.**

Advantages of ViTs

- **Better at capturing long-range dependencies – Unlike CNNs, which focus on local patterns, ViTs analyze global relationships in an image.**
- **Minimal inductive bias – ViTs learn patterns directly from data rather than relying on pre-designed convolutional filters.**
- **Highly scalable – ViTs perform exceptionally well when trained on large datasets.**

Challenges of ViTs

- **Require large datasets – ViTs perform poorly with limited data, unlike CNNs, which generalize well even on smaller datasets.**
- **Computationally expensive – Self-attention mechanisms demand significant processing power, leading to slower training times.**
- **Not ideal for small datasets – Without sufficient training data, ViTs struggle to learn meaningful representations.**

ViTs vs. CNNs: A Comparison

Feature	CNN	ViT
Handles local features	yes	no
Handles global features	no	yes
Requires a large dataset	no	yes
Faster training	yes	no
Works well on small datasets	yes	no
Scales better with more data	no	yes

Notable ViT Variants

- **DeiT (Data-efficient ViT) – Optimized for training with less data.**
- **Swin Transformer – Uses shifted windows for improved efficiency.**
- **ConvFormer – Combines CNNs and Transformers to leverage the strengths of both architectures.**

Results

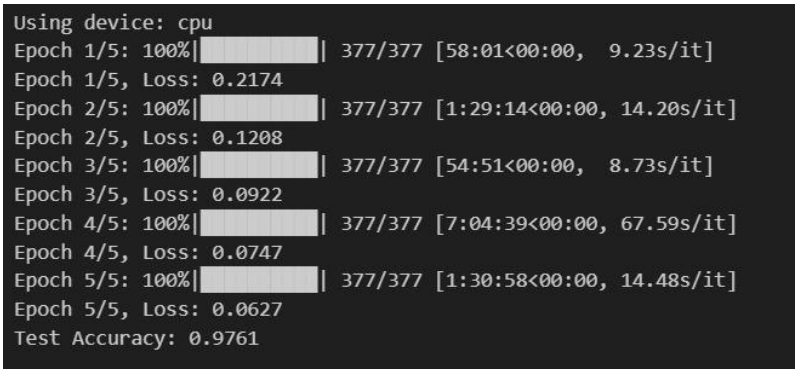


Fig 1- Vision Transformer accuracy

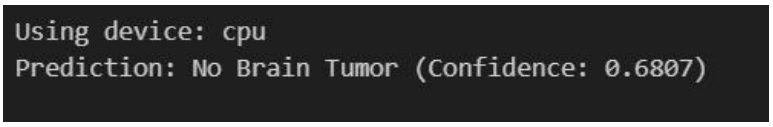


Fig 2- VIT prediction with image having no brain tumor MRI

Fig 3- VIT prediction with image having Brain Tumor MRI

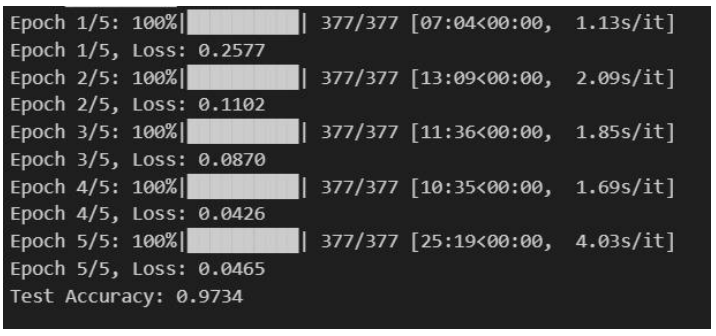


Fig 4 - CNN accuracy

Fig 5- CNN prediction with image having Brain Tumor MRI

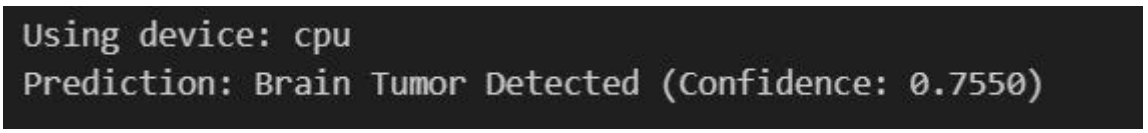


Fig 6- CNN prediction with image having no brain Tumor symptoms i.e CNN is providing false output for the same picture we gave to VIT mode

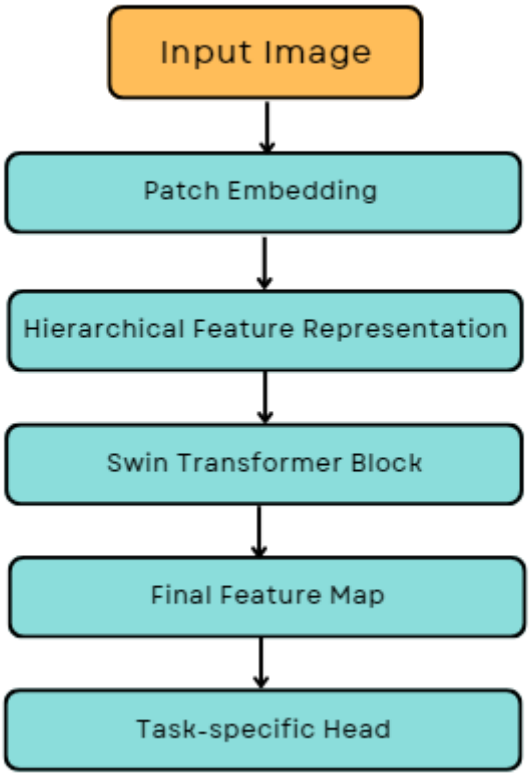
Challenges with Vision Transformers (ViTs)

ViTs are powerful, but they come with certain drawbacks:

- **They treat an image as equal-sized patches, lacking the hierarchical structure found in CNNs.**
- **Self-attention mechanisms are computationally expensive, making them slow for large images.**
- **They are inefficient for small images since they do not prioritize local details before capturing global features.**

To address these limitations, researchers combined Transformer models with CNN-inspired ideas—leading to the development of Swin Transformer.

How Swin Transformer Works



1. Splitting the Image into Small Patches

- **Similar to ViTs, Swin Transformer divides an image into small patches, treating each as a token.**

2. Local Window Attention

- **Instead of attending to all patches at once (which is computationally expensive), Swin Transformer focuses only on a small local window (e.g., 4×4 patches).**
- **This makes it significantly faster than standard ViTs.**

3. Shifted Window Mechanism

- **If a model always looks at the same window, it might miss connections between different regions of the image.**

- **Swin Transformer shifts these windows slightly at each layer, allowing better interaction between patches while keeping computations manageable.**

4. Hierarchical Feature Learning

- **Similar to CNNs, Swin Transformer builds a hierarchy:**
    - **Small patches merge into larger ones as the network deepens.**
    - **This enables it to capture both local and global details efficiently.**

Advantages of Swin Transformer

- **More computationally efficient – It limits self-attention calculations to small windows instead of the entire image.**
- **Scales well for high-resolution images – Unlike ViTs, which struggle with large images, Swin Transformer handles them effectively.**
- **Captures both fine details and global context – Its hierarchical structure ensures a balance between local and large-scale feature extraction.**
- **Versatile across vision tasks – Works well for object detection, image segmentation, and classification.**

Feature	CNN	ViT	Swin Transformer
Handles local features	Yes	No	Yes
Handles global features	No	Yes	Yes
Computational efficiency	Yes	No	Yes
Works well on small datasets	Yes	No	Yes

Scalable to large images	No	No	Yes
Suitable for object detection/segmentation	No	No	Yes

Results

```
Epoch 1/5, Loss: 0.2234
Epoch 2/5, Loss: 0.1071
Epoch 3/5, Loss: 0.1425
Epoch 4/5, Loss: 0.1766
Epoch 5/5, Loss: 0.0878
Test Accuracy: 0.8815
```

```
Test Accuracy: 0.8815
Predicted label for the image: Tumor
```

Comparison: Swin Transformer vs. ViT vs. CNNApplications of Swin Transformer

- **Image Classification – Similar to ViTs but more efficient.**
- **Object Detection – Used as a backbone in Faster R-CNN for improved detection performance.**
- **Image Segmentation – The Swin-UNet model is particularly effective for medical imaging segmentation.**

I