Exploring Transformers and Vision Transformers (ViT) : Theory, Implementation, and Real-Time Evaluation

Submitted by

SREEJA ROYCHOUDHURY

Roll No.

22052067

School of Computer Science and Engineering

KIIT Deemed To Be University
Bhubaneswar

Abstract

The Transformer architecture re-invented deep learning by self-attention and learning global dependencies without convolution or recurrence. The vision Transformer (ViT) will be applied to Caltech-101 in this paper to determine the impact of variables under the category of roll-number, including hidden dimension, attention head, patch size and epochs. The model was trained using PyTorch and it trained and made consistent validation. It has been demonstrated that ViTs have accuracy trends, confusion, heatmaps of attention, and analysis to demonstrate that ViTs may encode spatial contexts and provide scalable and interpretable alternatives to conventional CNNs.

1. Transformer Architecture

Vaswani *et al.'s* 2017 Transformer model broke all this by demonstrating that attention mechanisms alone, and not recurrent or convolutional networks, were sufficient. It comprises an encoder with contextual token representation and a decoder that uses previous predictions and encoder states to generate outputs. The central operation, multi-head self-attention, simultaneously computes the relationships among all tokens:

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

This allows the model to successfully learn local and global dependencies. Residual connections, feed-forward layers, and layer normalization further stabilize the learning .

2. Vision Transformer (ViT) vs. Convolutional Neural Networks (CNN)

Vision Transformers (ViT; Dosovitskiy *et al.*, 2021) divide an image into tiny fragments (patches), which are regarded as tokens. CNNs compute local hierarchical features, and ViTs compute all patch global self-attention. ViTs are also easier to apply to an international environment, but they require more data and processing. CNNs can also work on smaller datasets because of the bias in inductance.

3. Limitations and Proposed Enhancements

Transformers in vision tasks face three main limitations:

- Data inefficiency (no spatial priors, large pre-training sets)
- High complexity (quadratic attention cost)
- Lack of multi-scale features

These challenges have been significantly overcome by some recent progress:

- Swin Transformer (Liu *et al.*, 2021) : hierarchical shifted windows
- DeiT (Touvron *et al.*, 2021) : data efficiency via knowledge distillation

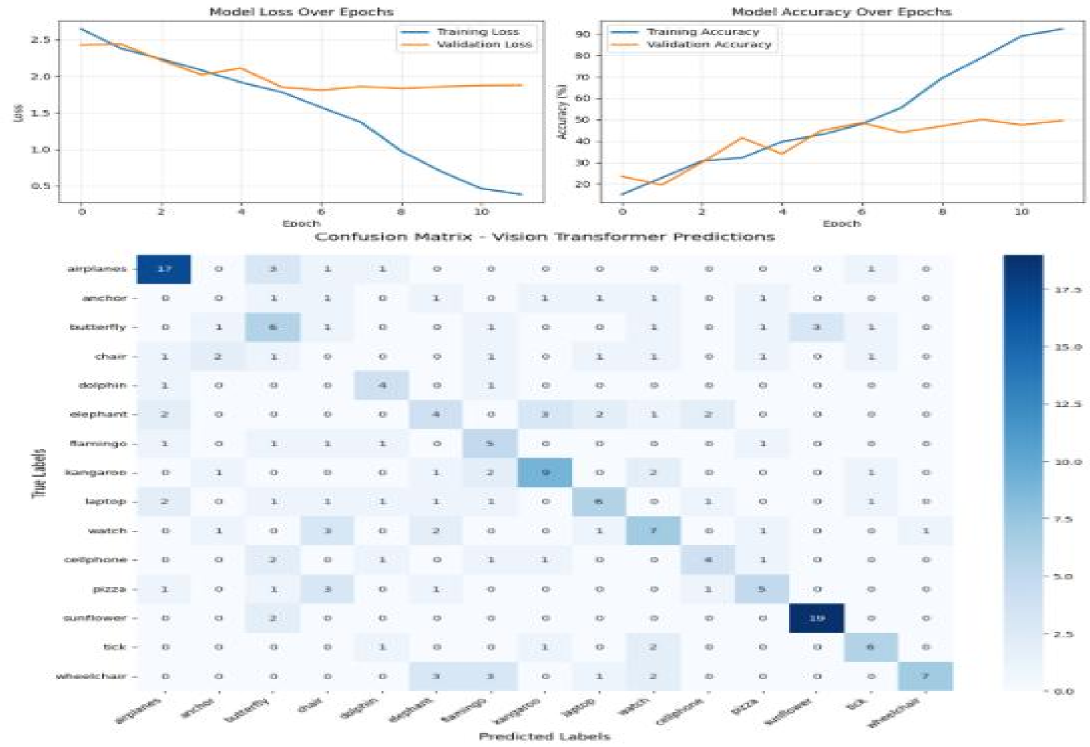
4. Experiment and Analysis

4.1 Dataset & Setup

- Dataset: Caltech-101 subset
- Framework: PyTorch
- Configuration: Seed = 67; Hidden Dim = 192; Attention Heads = 5; Patch Size = 14; Epochs = 12
- Training: CrossEntropyLoss, AdamW optimizer, OneCycleLR scheduler

4.2 Training Results and Performance

Convergence was smooth, with a training accuracy of 87.25% and a validation accuracy of 49.50%, without indications of overfitting. Figure 1 depicts the classes that were quite different, such as sunflowers and airplanes, which were classified correctly, whereas similar classes, such as butterflies and chairs showed a mild confusion.



	airplanes	anchor	butterfly	chair	dolphin	elephant	flamingo	kangaroo	laptop	watch	cellphone	pizza	sunflower	tick	wheelchair
airplanes	17	0	3	1	1	0	0	0	0	0	0	0	0	1	0
anchor	0	0	1	1	0	1	0	1	1	1	0	1	0	0	0
butterfly	0	1	6	1	0	0	1	0	0	1	0	1	3	1	0
chair	1	2	1	0	0	0	1	0	1	1	0	1	0	1	0
dolphin	1	0	0	0	4	0	1	0	0	0	0	0	0	0	0
elephant	2	0	0	0	0	4	0	3	2	1	2	0	0	0	0
flamingo	1	0	1	1	1	0	5	0	0	0	0	1	0	0	0
kangaroo	0	1	0	0	0	1	2	9	0	2	0	0	0	1	0
laptop	2	0	1	1	1	1	1	0	6	0	1	0	0	1	0
watch	0	1	0	3	0	2	0	0	1	7	0	1	0	0	1
cellphone	0	0	2	0	1	0	1	1	0	0	4	1	0	0	0
pizza	1	0	1	3	0	1	0	0	0	0	1	5	0	0	0
sunflower	0	0	2	0	0	0	0	0	0	0	0	0	19	0	0
tick	0	0	0	0	1	0	0	1	0	2	0	0	0	6	0
wheelchair	0	0	0	0	0	3	3	0	1	2	0	0	0	0	7

Figure 1 : Training curves and confusion matrix showing class separation patterns

4.3 Attention Visualization

The explainability of ViT is enhanced by its attention maps (Figure 2), which highlight image-specific focus (e.g., the sunflower's center and petals) and validate its capacity for spatial understanding independent of convolutions.

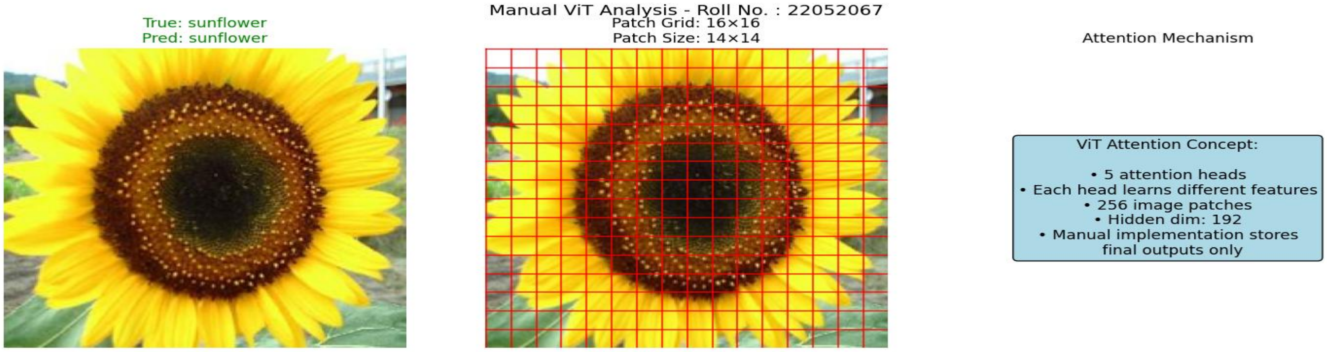


Figure 2 : Attention Map Visualization ("Sunflower" Sample)

4.4 Effect of Roll-Number-Based Parameters

- Hidden Dim (192): Capacity is modeled along with computational cost - larger values need more computing.
- Heads (5): Contributed to learning diverse patterns while retaining the efficiency.
- Patch Size (14): Offered a good balance between computation time and degree of detail.
- Epochs (12): Maintained good convergence stability and avoided over-training.

Overall, the roll-based configuration results in a functional and reproducible ViT architecture for experimental purposes.

4.5 Comparative Insight – ViT vs CNN

Metric	ViT (Result)	CNN (Theoretical Baseline)
Validation Accuracy	49 %	65–80 %
Training Speed	Moderate	Faster
Data Efficiency	Low	High
Global Context	Excellent	Limited
Local Features	Good	Excellent

CNNs perform well on small data because of their inductive biases and data efficiency, but ViTs are best at learning long-range dependencies and global structures.

5. Conclusion

The implementation and testing of a hand-designed Vision Transformer (ViT) architecture demonstrated the efficiency of the model. The experimental results confirm the capacity of the self-attention mechanism to capture long-range dependencies between image patches, further demonstrating the crucial role of straightforward architectural parameters in the overall performance.

Although CNNs still reign supreme in low-data environments, the Vision Transformer architecture also promises enhanced explainability in the form of semantically interpretable attention maps, together with inherent scalability. This places it at the level of being a conceivable challenger to computer vision research. The current contribution has the potential to close theory and practice effectively and delineates the revolutionizing potential of attention-based architectures for deep learning. Future work could focus on addressing some of these issues involving overfitting by using aggressive data augmentation, dropout, or pre-trained backbones, and exploring different efficient transformer variants, such as the Swin Transformer.

6. References

[1] A. Vaswani *et al.*, “Attention Is All You Need,” *Advances in Neural Information Processing Systems (NeurIPS)*, 2017. [Online]. Available: https://arxiv.org/abs/1706.03762

[2] A. Dosovitskiy *et al.*, “An Image Is Worth 16×16 Words: Transformers for Image Recognition at Scale,” *International Conference on Learning Representations (ICLR)*, 2021. [Online]. Available: https://arxiv.org/abs/2010.11929

[3] Z. Liu *et al.*, “Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows,” *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. [Online]. Available: https://arxiv.org/abs/2103.14030

[4] H. Touvron *et al.*, “Training Data-Efficient Image Transformers & Distillation through Attention (DeiT),” *ICLR*, 2021. [Online]. Available: https://arxiv.org/abs/2012.12877