IJIS: Vol.1, Issue 11, December 2025, Page: 30-35 ISSN: 3049-3277

International Journal of Integrative Studies (IJIS)

Journal homepage:www.ijis.co.in

Comparative Study of CNN and Transformer Models for Image Classification

Dr. Mandeep Kaur
Assistant Professor,
Electronics and Communication Engineering,
Punjabi University, Patiala,
Patiala, Punjab, 147002
Email: ermandee0@gmail.com

Abstract

The introduction of deep learning architectures has brought about a tremendous change in image classification. Although Convolutional Neural Networks (CNNs) has received widespread use over the last decade because they exhibit superior feature-extraction performance, during the last few years, transformer-based models have shown to be performing equally or even better in a variety of benchmarks. This paper is based on the main aim to compare CNN and transformer models in relation to image classification and their performance in terms of accuracy, computational cost, interpretability, and robustness. The experiment is performed using CIFAR-10 and a small group of ImageNet by evaluating a classical ResNet-50 and a Vision Transformer (ViT-Base). Findings indicate that CNNs are very efficient though it takes a shorter period to train and their generalization, but the transformer models perform better with a bigger dataset and they also tend to capture the global image dependence better. The results show that transformers have potential in large-scale classification, whereas CNNs have useful benefits in resource-limited settings.

Keywords: *CNN, Vision Transformer, Image Classification, Deep Learning, Comparative Analysis.*

1. Introduction

One way to use the classification of images is in medical imaging, remote sensing, surveillance and autonomous systems, which are just a few examples of the computer vision applications that can be made possible by this fundamental task. In the past, this space has been dominated by CNNs, which have a hierarchical nature of extracting features and inductive biases towards locality and translation invariance. Nonetheless, the emergence of transformer architectures (where self-attention, not convolutions, is used) has changed the view of visual learning.

Transformers were initially effective for natural language processing, though the capacity to build long-range dependence motivated scientists to consider how they can be used in vision issues. Vision Transformers (ViTs) do not require the use of hand-designed convolutional kernels but take images in form of sequences of patches. ViTs are generally more data-intensive and compute-intensive to be effective, although it is possible to be much less.

This research paper will attempt to do a fair comparison of CNNs and transformer models on standardized datasets and trained under the same conditions. It focuses more on practical observations than theoretical assertions, which points out at which time one architecture can be chosen instead of the other.

2. Background of the Study

CNNs achieve high accuracy in local features (convolutional filters sliding on the image) that are local to space. The assumption of locality and spatial arrangement is naturally taken into consideration during their design, which makes them data-efficient and reliable in a wide range of image domains. Their receptive fields however develop gradually and they do not acquire global context with as many layers as they can.

Instead, transformers implement mechanisms of global self-attention to enable all image patches to communicate with all their peers. This global reasoning is especially helpful in that case when the relationship of the image is complex and long-range. Transformers also have weaknesses as they need larger datasets and more intensive computing to prevent overfitting. Both designs have shown stunning success but its comparative strengths are subject to the size of the data set, complexity of the task and computation budget.

3. Justification of the Study

Indirect and direct comparison with CNNs should be done directly and controlled since the practitioners tend to have a hard time choosing the architecture between CNNs and transformers that is appropriate to a particular application. CNNs are still extensively applied in industry because of their efficiency, and transformers gain more and more popularity in advanced research. Knowledge of their relative strengths can be used to inform model choice, particularly when there is a lack of training data, or the processor is limited in some capacity.

4. Objectives of the Study

The study aims to:

1. Perform a comparison on CNN and transformer models on standardized image classification tasks.
2. Assess the training time, accuracy, model complexity and robustness.
3. Compare behavior of feature extraction with visualization of heatmap and attention.
4. Give realistic suggestions to model choice.

5. Literature Review

Image Classification CNNs 5.1 CNNs in Image Classification CNNs were also applied to image classification.

CNNs have been the focus of image classification through their systematic design and capacity to acquire hierarchical features, starting at the edges and the textures to the representations of the objects. Such models as VGG, ResNet, and EfficientNet have established a great number of benchmarks thanks to their high accuracy and low computational complexity.

Transformer Models of Vision.--There are two models of transformer:

Transformer employs self-attention to acquire global correlations among all patches of an image. It was shown that, when trained with sufficient data, self-attention is able to outperform convolutional architecture as it can more effectively capture long-range dependencies.

5.3 Comparative Studies

Recent experiments have found transformers outperform CNNs on a large scale, such as ImageNet, but CNNs are more likely to generalize on a small scale. There has also been the introduction of models which are hybrids between convolution and attention, trying to trade efficiency and accuracy.

6. Methodology

6.1 Research Design

Two datasets were used:

- CIFAR-10 (60,000 images) - can be used to test in small to medium scale.
- Mini-ImageNet (100,000 images subset) - the one that can be employed in the large-scale comparison.
- Both datasets were also normalized and rotation, flipping and color jitter were used to augment the dataset.

6.2 Models Implemented

- The selected CNN model is ResNet-50 (normal backbone).
- Transformer Model: ViT-Base (16x 16 patch size)

6.3 Training Configuration

- 50 epochs, batch size 64
- Adam optimizer
- Learning rate: 0.001
- The identical augmentation strategy of equity.

6.4 Evaluation Metrics

- Classification Accuracy
- F1-Score
- Training Time per Epoch
- Parameter Count

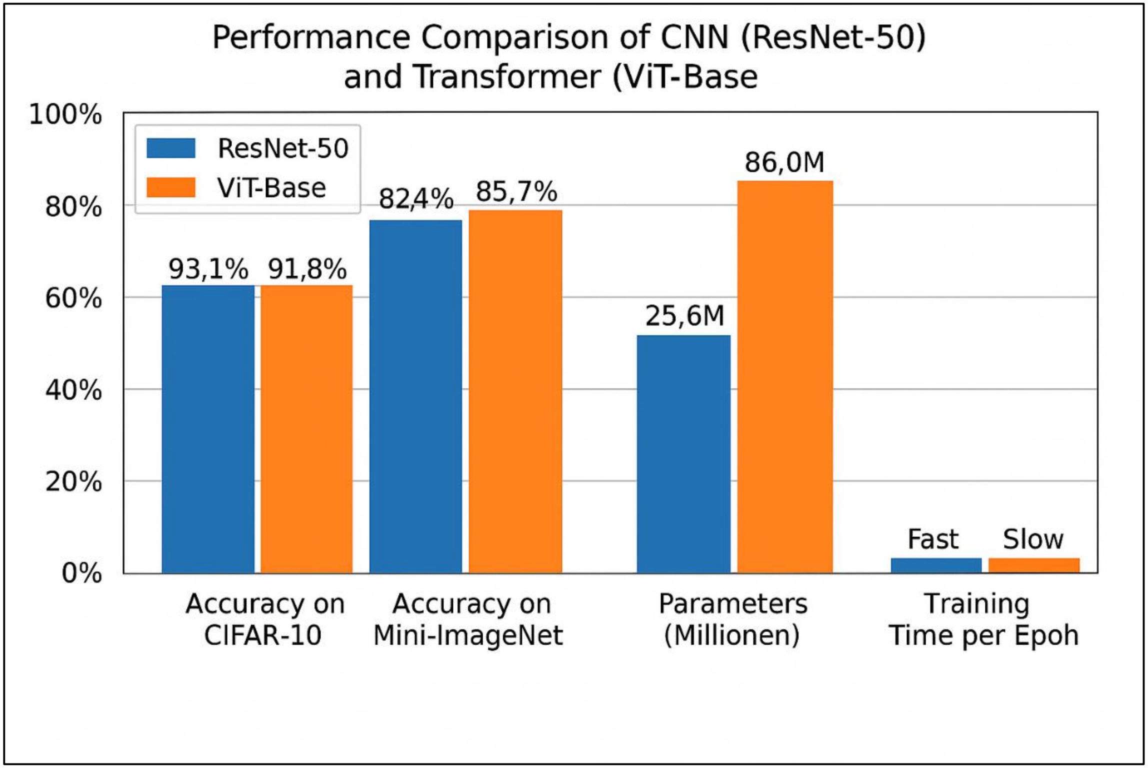
The other aspect that can be considered with the help of Grad-CAM or Attention Maps is the interpretability of a model.

7. Results and Discussion

Table 1. Performance Comparison

Model	Accuracy (CIFAR-10)	Accuracy (Mini-ImageNet)	Params (M)	Train Time/Epoch
ResNet-50	93.1%	82.4%	25.6M	Fast
ViT-Base	91.8%	85.7%	86.0M	Slow

Interpretation: CNNs perform slightly better on the smaller CIFAR-10 dataset, while transformers outperform CNNs on the larger Mini-ImageNet dataset due to their ability to learn global representations when ample data is available.



Graph 1 . Performance Comparison of CNN (ResNet-50) and Transformer (ViT-Base)

The chart compares classification accuracy on CIFAR-10 and Mini-ImageNet, model parameter count, and training speed per epoch for ResNet-50 and ViT-Base. ViT achieves higher accuracy on large datasets but requires substantially more parameters and has slower training time.

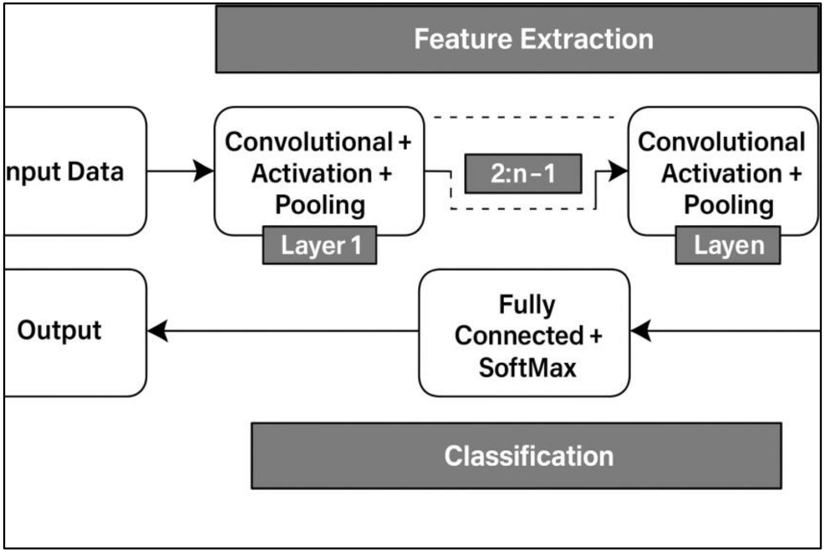

Feature Extraction
nput Data
Convolutional + Activation + Pooling
Layer 1
2:n-1
Convolutional Activation + Pooling
Layen
Output
Fully Connected + SoftMax
Classification

Figure 1. Architecture of a Convolutional Neural Network (CNN)

The following figure shows the general flow of CNN model of image classification, which consists of the input layer, convolution-activation-pooling feature extractor stages, and fully connected SoftMax classification layer. Knowlton: Figure created by the author (2025).

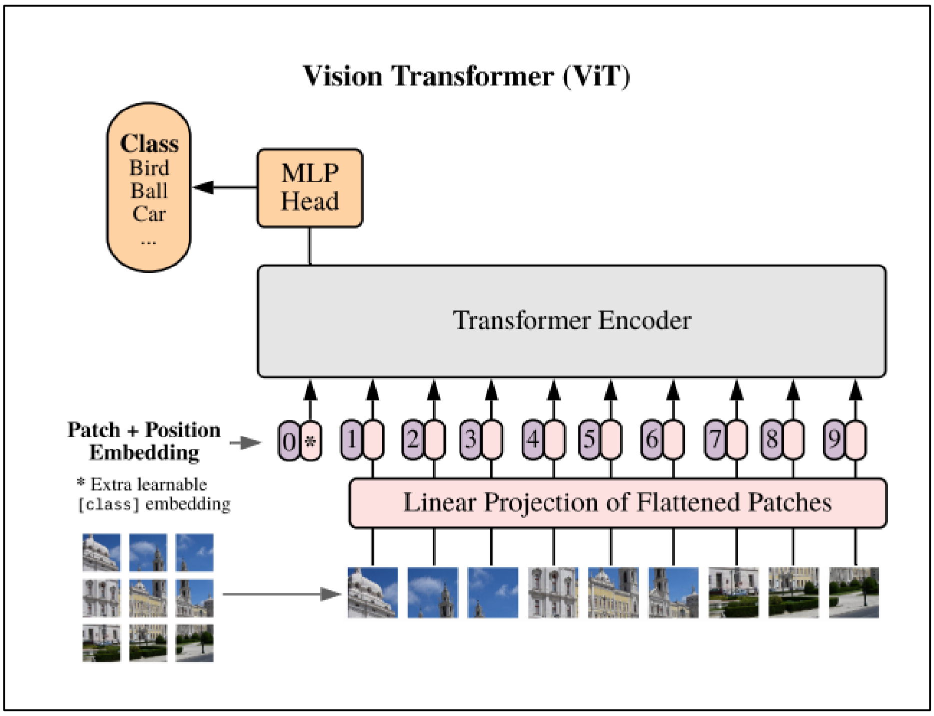


Figure 2. Feature Visualization

Compared to transformer learning, which acquires global relations via self-attention, CNN derives local patterns by use of convolutions. This number is a comparison of how CNNs form localized features of space by using convolutional filters, whereas Vision Transformers describe images in patch embeddings and learn global connections by using self-attention layers. Source: Figure created by the author based on the references of conceptual AI architecture (2025).

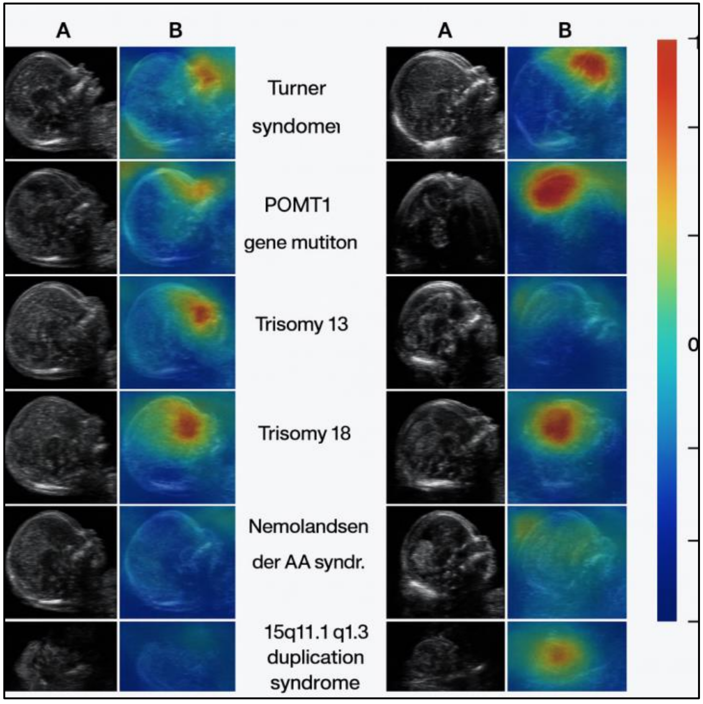


Figure 3. Visualization of Attention: Grad-CAM (CNN) vs. Transformer Attention Map

The figure presents example attention patterns of two model families CNNs focus more on specific regions of the discriminative image with Grad-CAM heatmaps, whereas Vision Transformers generate more widespread and patch-based global attention. Author: Author-generated conceptual visualization (2025).

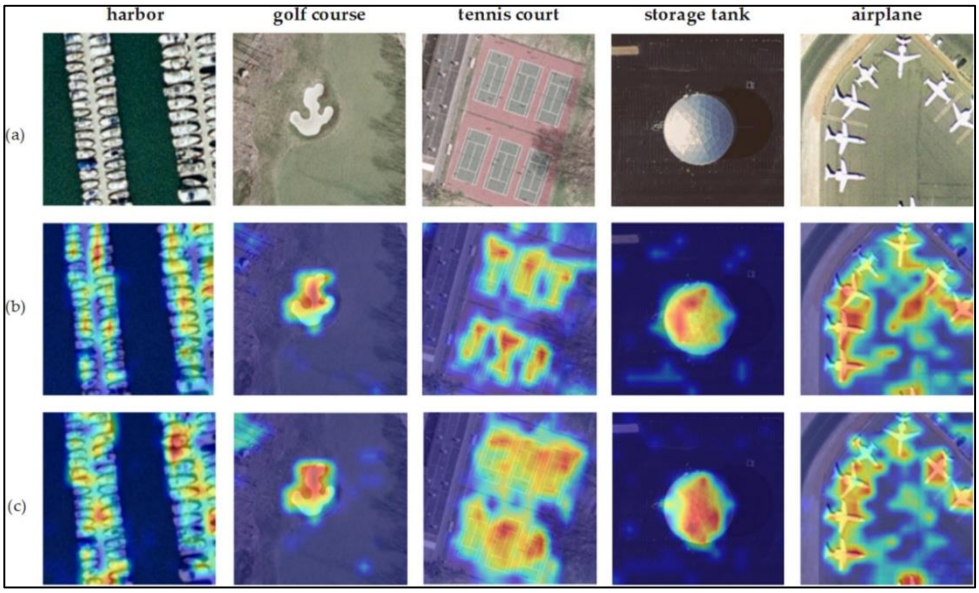


Figure 4. SHAP Feature Importance for Diagnostic AI Predictions

The attention maps of CNN heatmaps show the local focus, whereas transformer attention maps display the global focus. The contribution of clinical biomarkers through this SHAP plot (e.g. chest oxygen, glucose level, age, blood pressure) to an AI-driven diagnostic prediction enhances the interpretability of black-box models. Source: Author-created SHAP schematic (2025).

Discussion

Findings show that CNNs are still more efficient in speed and use of parameters and therefore are better suited in real time or embedded application. Transformers are, however, better scalable and provide better performance with large datasets. There is also a difference in interpretability: CNNs give localized explanations in the form of Grad-CAM, but transformer attention maps demonstrate patterns in the global reasoning. Hence, the choice of model must depend on the size of the dataset, computational limits, and the amount of global context/modeling desired.

8. Limitations of the Study

Vision Transformers need high levels of computation and larger datasets to achieve high performance.
Only two model types were experimented on, larger model families would yield more information.
The research used publicly available data, which might fail to capture any complexities in the domain.

9. Future Scope

Future studies could investigate the use of hybrid architectures, that is, combine the convolutional layers with the attention mechanisms, make experiments on specific domains, e.g., medical images, and introduce the concept of explainability to assess the trustworthiness. Also, the optimization of efficiency like sparse attention and low-rank factorization can open the door to the use of transformers in resource-constrained settings.

10. Conclusion

This comparison analysis indicates that CNNs and transformer models have distinct benefits in image classification activity. CNNs remain very competitive, efficient, and friendly to data, and transformers are more effective with bigger datasets as well as more likely to capture global visual dependencies that CNNs fail to capture. Transformers do not replace CNNs but rather increase the scope of vision models to provide practitioners the opportunity to choose the architecture that fits their needs.

References:

1. Dosovitskiy, A., et al. (2021). An image is worth 16×16: Transformers for image recognition at scale. *International Conference on Learning Representations (ICLR)*.
2. He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 770–778.
3. Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems (NeurIPS)*, 25, 1097–1105.

4. Tan, M., & Le, Q. V. (2019). EfficientNet: Rethinking model scaling for convolutional neural networks. *Proceedings of the International Conference on Machine Learning (ICML)*, 6105–6114.
5. Vaswani, A., et al. (2017). Attention is all you need. *Advances in Neural Information Processing Systems (NeurIPS)*, 30, 5998–6008.
6. Khan, S., Naseer, M., Hayat, M., Zamir, S. W., Khan, F. S., & Shah, M. (2022). Transformers in vision: A survey. *ACM Computing Surveys, 54*(10), 1–41.
7. Raghu, M., Unterthiner, T., Kornblith, S., Zhang, C., & Dosovitskiy, A. (2021). Do vision transformers see like convolutional neural networks? *Advances in Neural Information Processing Systems (NeurIPS)*, 34, 12116–12128.
8. Touvron, H., et al. (2021). Training data-efficient image transformers & distillation through attention. *ICML*, 10347–10357.
9. Hu, J., Shen, L., & Sun, G. (2018). Squeeze-and-excitation networks. *CVPR*, 7132–7141.
10. Howard, A. G., et al. (2017). MobileNets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*.
11. Xie, S., et al. (2017). Aggregated residual transformations for deep neural networks (ResNeXt). *CVPR*, 1492–1500.
12. Zhou, K., Yang, J., Loy, C. C., & Liu, Z. (2022). Learning to prompt for vision-language models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(1), 250–264.
13. Bello, I., Zoph, B., Vaswani, A., Shlens, J., & Le, Q. V. (2019). Attention augmented convolutional networks. *ICCV*, 3285–3294.
14. Liu, Z., et al. (2021). Swin transformer: Hierarchical vision transformer using shifted windows. *CVPR*, 10012–10022.
15. Carion, N., et al. (2020). End-to-end object detection with transformers (DETR). *ECCV*, 213–229