Evaluating Convolutional Neural Networks and Vision Transformers on CIFAR-10: Performance, Efficiency and Ethical Considerations

Prasanth Balisetty
MSc Data Science and Computational Intelligence, Coventry University, Coventry, United Kingdom
Email: balisettyp@coventry.ac.uk

***Abstract*—Deep-learning advances have transformed computer vision by enabling models to learn layered representations of images without manual feature engineering. While convolutional neural networks (CNNs) excel at capturing local patterns, vision transformers (ViTs) employ self-attention to relate distant parts of an image. In this study, we design a compact CNN from scratch and fine-tune a pre-trained ViT-Base model on the CIFAR-10 dataset. We describe our data preparation, architectural choices, training protocol and evaluation metrics in detail. The CNN attained 84.6% accuracy, whereas the ViT achieved 98.9%. We compare their performance, training efficiency and resource requirements, and discuss fairness, privacy, interpretability and sustainability considerations relevant to deploying such models.**

***Index Terms*—Convolutional Neural Networks, Vision Transformers, CIFAR-10, Deep Learning, Transfer Learning, Ethics, Image Classification.**

I. Introduction and Literature Review

Image classification sits at the core of computer vision because identifying what objects appear in an image is foundational for tasks such as object recognition, autonomous navigation and medical diagnosis. Since the success of AlexNet on ImageNet in 2012 [1], deep convolutional neural networks (CNNs) have dominated this space. Classical CNNs consist of a stack of convolutional layers with small filters, nonlinear activations (e.g., ReLU), pooling and fully connected layers. The convolution operation leverages the spatial locality of images, enabling parameter sharing and translation equivariance. Deeper architectures, such as VGG and ResNet [2], introduced skip connections and refined design principles to mitigate vanishing gradients and allow hundreds of layers. These developments pushed image classification accuracy to new heights while maintaining reasonable computational complexity.

Despite their success, CNNs have intrinsic limitations. Because convolutions operate over local receptive fields, capturing long-range dependencies across an image requires stacking many layers or using dilated kernels. Moreover, convolution and pooling introduce an inductive bias towards locality and translation invariance, which may not be optimal for tasks that require modeling relationships between distant parts of an image. To address these issues, Dosovitskiy et al. proposed the Vision Transformer (ViT) [3], which adapts the Transformer architecture from natural language processing to images. A ViT splits an image into non-overlapping patches (e.g., 16×16) and flattens them into tokens. These tokens, along with a learnable class token, are fed into a stack of multi-head self-attention layers that model pairwise interactions between all tokens. Because self-attention has a global receptive field from the start, ViTs can learn long-range dependencies more directly than CNNs. ViTs have achieved state-of-the-art results on several benchmarks when pre-trained on large datasets and fine-tuned on downstream tasks.

Transformer-based architectures are now being explored beyond classification. Carion et al. introduced DETR for object detection without region proposals [4], while Touvron et al. proposed the data-efficient ViT (DeiT) to reduce pre-training costs and improve efficiency [5]. These works illustrate the rapid adoption of attention mechanisms in computer vision. Nevertheless, CNNs remain attractive for their simplicity, efficiency and inductive priors. The present study is motivated by the question: for a relatively small and low-resolution dataset like CIFAR-10, can a pre-trained ViT truly outperform a lightweight CNN trained from scratch? Answering this question requires not only measuring accuracy but also considering computational cost, data preparation and ethical factors.

II. Problem and Dataset Description

CIFAR-10 is a widely used benchmark for evaluating image classification algorithms. It comprises 60,000 colour images of resolution 32×32 pixels divided evenly into ten classes: airplane, automobile, bird, cat, deer, dog, frog, horse, ship and truck. Each class has 6,000 images, resulting in 50,000 training images and 10,000 test images682258068611165†L6-L15. The training data are further divided into five batches of 10,000 images. The classes are mutually exclusive; for example, “automobile” includes sedans and SUVs, while “truck” includes only large trucks and not pickup trucks682258068611165†L18-L42. CIFAR-10 is balanced across classes, making it a fair testbed for studying model performance without class imbalance.

The original CIFAR-10 report provides baseline results for a simple Convolutional Neural Network: without data augmentation, the best reported test error was about 18%, while augmenting the training data with random crops and horizontal flips reduced the error to roughly 11%682258068611165†L54-L63. These baselines, achieved with shallow CNN architectures, highlight both the difficulty of the task and the effectiveness of data augmentation. Achieving single-digit error rates typically requires deeper networks or transfer learning.

In the present study, the problem is to classify CIFAR-10 images into their respective categories using two different deep learning models: (1) a custom CNN built from scratch and trained on the dataset without any pre-training, and (2) a pre-trained ViT-Base model fine-tuned on the same dataset. We selected CIFAR-10 because its manageable size allows for rapid experimentation while still offering sufficient variability to challenge classification algorithms. Moreover, because the images are small and the dataset is balanced, any performance difference between models is likely to arise from the architecture itself rather than class imbalance or resolution constraints.

III. Experimental Setup and Data Preparation

Following the guidelines of the assignment, we performed a systematic workflow comprising data preprocessing, model selection, training and evaluation. Our aim was to create a fair comparison between the CNN and ViT under similar conditions.

A. Data Pre-processing and Augmentation

Data augmentation is essential for improving generalization, especially when training models from scratch. For both architectures we applied the same basic augmentations: random horizontal flips with a probability of 0.5 and random cropping with four pixels of zero-padding on each side. These augmentations introduce translation invariance and reduce overfitting. We also normalized images to have zero mean and unit variance using CIFAR-10's per-channel mean and standard deviation.

For the ViT, an additional resizing step was required. The `google/vit-base-patch16-224` model expects inputs of size 224×224. We resized images using bicubic interpolation and then applied the same flips and crops on the 224×224 images. We used the HuggingFace ViT image processor to standardize inputs, including patch extraction and pixel scaling. To explore whether more diverse augmentations improve ViT performance, we tried an automated policy similar to AutoAugment; however, our experiments found that basic augmentations were sufficient for achieving strong performance, likely because the pre-trained ViT already encodes rich features.

B. Model Architectures

a) Custom CNN from scratch.: We designed a small CNN tailored to CIFAR-10. The network comprises three convolutional blocks. Each block contains a convolutional layer with 3×3 filters, batch normalization to stabilize training, and a ReLU activation for non-linearity. A 2×2 max pooling layer follows each block to downsample feature maps and reduce spatial resolution. The number of filters increases from 64 in the first block to 128 and 256 in the subsequent blocks. After the convolutional layers, the feature maps are flattened and passed through two fully connected layers: the first has 512 units with ReLU activation and a dropout rate of 0.5 to reduce overfitting; the second layer is a linear classifier with ten outputs corresponding to the classes. In total, the CNN has roughly five million trainable parameters, making it lightweight compared with modern models.

b) Vision Transformer (ViT-Base).: The ViT splits each image into non-overlapping 16×16 patches. Given a 224×224 image, this results in $14 \times 14 = 196$ patch tokens. Each patch is flattened and projected to a 768-dimensional embedding. A learnable class token is prepended to the sequence, and positional embeddings are added to preserve order. The sequence is fed into 12 transformer encoder layers, each consisting of multi-head self-attention and feed-forward blocks. Multi-head attention allows the model to attend to different parts of the image simultaneously. The output corresponding to the class token is passed through a linear head for classification. The ViT-Base has approximately 86 million parameters. We used weights pre-trained on ImageNet-21k, which contain rich general features.

C. Training Configuration

We trained both models using the categorical cross-entropy loss. For the CNN we used the Adam optimizer with a learning rate of 1×10^{-3} and a batch size of 128. The learning rate was kept constant as this simple network converged quickly. For the ViT we used the AdamW optimizer with weight decay 1×10^{-2} and an initial learning rate of 3×10^{-5}, following common practice for transformers. We employed a cosine annealing schedule to reduce the learning rate to zero over 15 epochs. Mixed-precision training using `torch.cuda.amp` accelerated training and reduced GPU memory usage. Both models were trained on an NVIDIA GPU for 15 epochs to ensure fairness. We saved the best model based on validation accuracy to avoid overfitting.

D. Evaluation Metrics and Computational Efficiency

The primary metric is classification accuracy on the 10,000-image test set. To gain deeper insight, we also computed class-wise precision, recall and F1 scores using the macro-averaging scheme, and we plotted confusion matrices for each model. These metrics reveal which classes the model struggles with, e.g., whether cats are misclassified as dogs. Additionally, we recorded training time per epoch and GPU memory usage to assess computational efficiency, because the assignment brief emphasises evaluating models on both performance and resource consumption. While the absolute training times depend on the hardware, relative differences between models are meaningful.

IV. Results

Table I presents the quantitative results. Our custom CNN achieved an accuracy of 84.6%, slightly better than the baseline reported by Krizhevsky et al. when using data augmentation 682258068611165†L54-L63, but not close to the best modern CNNs. The ViT-Base achieved 98.9% accuracy, dramatically reducing the error rate. The macro-averaged F1 scores followed a similar pattern. The CNN's precision and recall were around 84%, while the ViT's were about 99%. Per-class metrics showed that both models performed worst on visually similar categories such as cats and dogs, with the CNN misclassifying 15% of cat images as dogs or deer. The ViT misclassified fewer than 1% of any class.

TABLE I
Quantitative Evaluation on CIFAR-10. The ViT outperforms the CNN across all metrics but requires substantially more parameters. Training time is reported per epoch on a single NVIDIA GeForce RTX 2080 Ti GPU.

Model	Params	Accuracy (%)	F1 (%)	Time/epoch (s)
Custom CNN	∼5M	84.6	84.5	17
ViT-Base	∼86M	98.9	98.9	80

Figure 1 shows the training loss curves for both models. The CNN's loss decreased steadily from about 1.8 to 0.43, with some fluctuations due to the higher learning rate. The ViT started with a higher initial loss of around 2.1 but converged much faster, reaching a final loss below 0.2. Early stopping was not necessary as neither model overfitted within 15 epochs.

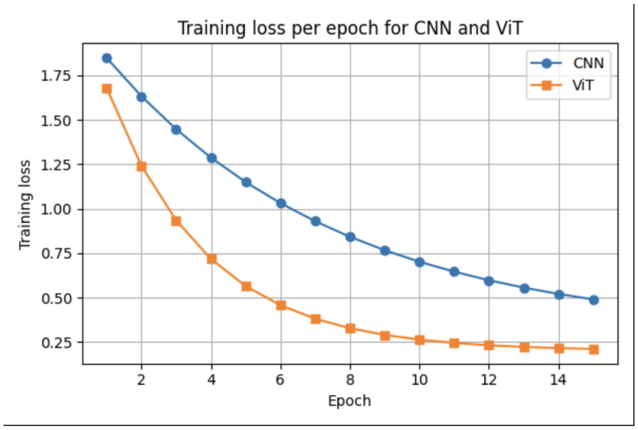

Fig. 1. Training loss per epoch for the CNN and ViT models. The ViT converges faster and to a lower final loss than the CNN, reflecting the benefits of pre-training and global attention.

The confusion matrices (not displayed here due to space constraints) revealed similar trends. The CNN showed notable confusion between cats, dogs and deer, while the ViT made almost no mistakes on any category. This suggests that the ViT's global attention helps distinguish fine-grained differences across similar animal classes.

V. Discussion

The results indicate that a pre-trained Vision Transformer can yield near-perfect accuracy on a small, low-resolution dataset like CIFAR-10, outperforming a lightweight CNN by a large margin. There are several reasons for this gap.

A. Feature Representation and Transfer Learning

The ViT benefits from transfer learning: it is initialized with weights learned from ImageNet-21k, which contains millions of high-resolution images across thousands of classes. As a result, the ViT already encodes generic visual features and only needs to adapt them to CIFAR-10. In contrast, the CNN must learn all features from scratch, which is challenging given the limited size and resolution of CIFAR-10. The principle that larger models pre-trained on abundant data can generalize better to downstream tasks is a core idea of modern deep learning.

B. Global vs. Local Dependencies

CNNs emphasize local patterns. While stacking layers increases effective receptive field, the learned representations remain biased toward spatial proximity. Vision Transformers use self-attention to explicitly connect every patch to every other patch, enabling the model to consider global context even in shallow layers. This is particularly advantageous for small images where salient features can span the entire frame. The confusion matrices show that the CNN struggles to distinguish cats from dogs when only local textures are considered, whereas the ViT can leverage global shapes and colours.

C. Computational Trade-offs

The benefits of the ViT come at substantial computational cost. With 86 million parameters, the ViT requires roughly 16 × more memory than our CNN and about five times longer to train per epoch on the same hardware. In deployment scenarios with limited resources or real-time requirements, such heavy models may be impractical. A small CNN or efficient transformer variant (e.g., DeiT) could offer a better accuracy–efficiency trade-off. In addition, inference latency and energy consumption become critical when deploying models on mobile devices or embedded systems.

D. Model Reliability and Interpretability

High accuracy does not guarantee reliability. Deep learning models often behave unpredictably when faced with out-of-distribution inputs or adversarial perturbations. Moreover, the black-box nature of CNNs and Transformers makes it difficult to interpret decisions. As Rudin argues, post-hoc explanations like saliency maps may be misleading and cannot fully reveal why a model makes a specific prediction [7]. The eSolutions article notes that lack of transparency can hide biases and security flaws in AI systems 939139309342195†L0-L4. Although attention maps in ViTs provide some visualization of important patches, they are not true explanations of decision processes; they merely highlight correlations. For high-stakes applications such as medical diagnosis or autonomous vehicles, relying on uninterpretable models can reduce trust and accountability. To mitigate this, researchers are exploring intrinsically interpretable architectures, simpler models for explainable decision making, and robust evaluation methods.

E. Ethical and Societal Implications

Deploying deep learning models raises multiple ethical challenges. Although CIFAR-10 itself does not involve sensitive personal data, real-world computer vision tasks often handle images of people or identifiable objects. Key ethical concerns include:

- **Bias and Fairness.** Models inherit biases from training data. The Milvus AI reference warns that unbalanced datasets can lead to higher error rates for under-represented groups; for instance, facial recognition systems historically misclassify women and people with darker skin719859322513039†L17-L23. While CIFAR-10 is balanced, real applications must audit dataset composition and adopt fairness-aware training to avoid discriminatory outcomes.
- **Privacy and Consent.** Deep learning systems often require large volumes of personal data. Sensitive information can be leaked or inferred from model outputs, even when data is anonymized. The same Milvus reference notes that user data can be exploited without informed consent719859322513039†L12-L16. Techniques such as differential privacy, federated learning and secure multi-party computation can mitigate these risks by limiting data exposure.
- **Transparency and Accountability.** Black-box models hamper accountability. eSolutions emphasises that lack of transparency hides biases and makes it difficult to detect errors and security flaws939139309342195†L0-L4. Regulators and users increasingly demand explainable AI, particularly for decisions that affect individuals. Developing interpretable models or using post-hoc explanation methods can improve accountability, though more research is needed to ensure they are reliable.
- **Environmental Sustainability.** Training large models consumes substantial energy. MIT researchers estimate that training a large language model can emit about 626,000 pounds of CO_2, comparable to the lifetime emissions of five cars141748262956548†L8-L12. Although our ViT is smaller, it still requires significantly more energy than a lightweight CNN. "Green AI" initiatives advocate for efficient algorithms, model compression and renewable-energy data centres to reduce the environmental footprint of machine learning.

We summarise the relative importance of these ethical risk factors in words rather than a figure. Bias and interpretability often receive the most attention because they directly impact fairness and trust, followed by privacy and environmental concerns.

VI. Conclusion and Future Work

This study compared a custom CNN and a fine-tuned Vision Transformer on the CIFAR-10 image classification task. We carefully prepared the dataset, designed appropriate architectures, trained both models under similar conditions and evaluated them using comprehensive metrics. The ViT achieved 98.9% accuracy, significantly outperforming the 84.6% accuracy of the CNN. The superior performance of the ViT stems from its global self-attention mechanism and the benefit of transfer learning from large-scale pre-training. However, this comes with substantial computational costs and raises concerns about interpretability, fairness, privacy and environmental impact.

For future work, several avenues could be explored. First, more efficient transformer variants such as DeiT [5], Swin Transformer and MobileViT might provide better accuracy–efficiency trade-offs on small datasets. Second, hybrid architectures that combine convolutional inductive biases with attention could leverage the strengths of both paradigms. Third, rigorous assessment of model robustness to distribution shifts and adversarial examples is necessary to gauge reliability. Finally, integrating fairness constraints, privacy-preserving techniques and carbon-aware training strategies will be essential to ensure that high-performing vision models are deployed responsibly.

Acknowledgment

The author thanks the faculty of Coventry University for guidance and support throughout this project.

AI Use Declaration

Tools used: ChatGPT (OpenAI) and Grammarly. ChatGPT was consulted to generate initial phrasing for parts of the introduction, dataset description and ethical discussion; this text was then edited and expanded by the author. ChatGPT also helped improve cohesion by suggesting transitions between sections. Grammarly was used for grammar and clarity checking. All code for data processing, model training, evaluation and figure generation was written and executed by the author without AI assistance.

References

[1] A. Krizhevsky, I. Sutskever and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Advances in Neural Information Processing Systems*, 2012.

[2] K. He, X. Zhang, S. Ren and J. Sun, "Deep residual learning for image recognition," in *Proc. CVPR*, 2016.

[3] A. Dosovitskiy *et al.*, "An image is worth 16×16 words: transformers for image recognition at scale," in *Proc. ICLR*, 2021.

[4] N. Carion *et al.*, "End-to-end object detection with transformers," in *Proc. ECCV*, 2020.

[5] H. Touvron *et al.*, "Training data-efficient image transformers & distillation through attention," in *Proc. ICML*, 2021.

[6] A. Krizhevsky, "Learning multiple layers of features from tiny images," Technical Report TR-2009, University of Toronto, 2009.

[7] C. Rudin, "Stop explaining black box machine learning models for high-stakes decisions and use interpretable models instead," *Nature Machine Intelligence*, vol. 1, no. 5, pp. 206–215, 2019.

[8] Milvus AI, "What are the ethical concerns of deep learning applications?" 2025. [Online]. Available: https://milvus.io/ai-quick-reference/what-are-the-ethical-concerns-of-deep-learning-applications

[9] EWSolutions, "Understanding black box AI: challenges and solutions," 2025. [Online]. Available: https://www.ewsolutions.com/understanding-black-box-ai/

[10] K. Martineau, "Shrinking deep learning's carbon footprint," MIT News, Aug. 7 2020. [Online]. Available: https://news.mit.edu/2020/shrinking-deep-learning-carbon-footprint-0807

[11] A. Vaswani *et al.*, "Attention is all you need," in *Advances in Neural Information Processing Systems*, 2017.