The ResNet18-Transformer Bridge for CIFAR-10 Object Detection

Chas Omer Madlos[1] and Reggie Hermosisima[1]

[1]Department of Computer Science, University of Science and Technology of Southern Philippines, Claro M. Recto Avenue, Lapasan 9000 Cagayan de Oro City, Philippine

December 2025

1 Introduction

The primary problem addressed is the limitation of purely CNN-based object detection models in capturing global contextual relationships between objects, especially in complex scenes with occlusions, scale variation, and dense object layouts.

This problem is highly relevant due to the increasing demand for accurate and efficient object detection in applications such as autonomous driving, surveillance systems, robotics, and augmented reality. Hybrid CNN–Transformer models represent a modern direction in computer vision research, offering improved accuracy while maintaining practical computational efficiency, making this study both academically and industrially significant.

2 Dataset Description

The Cifar-10 dataset, provided by the Canadian Institute for Advanced Research, consists of 60,000 color images of size 32×32 pixels, divided into 50,000 training images and 10,000 test images across 10 classes (airplane, automobile, bird, cat, deer, dog, frog, horse, ship, truck).

3 Methodology

The model uses a CNN–Transformer hybrid architecture. A CNN backbone extracts hierarchical feature maps from input images. These features are projected to a lower-dimensional embedding using a 1×1 convolution. A Transformer encoder is then applied to model global dependencies using self-attention. The

Figure 1: Sample images from the Cifar-10 dataset.

resulting features are pooled and passed to a classification head for final prediction.

CNN feature maps are flattened into a sequence of tokens and processed by a Transformer encoder. This fusion allows the model to combine local spatial features from the CNN with global contextual information from the Transformer. The fused representation improves the expressiveness of the learned features before classification.

Images are resized and normalized before training. Data loading is handled using PyTorch utilities. The CNN backbone is optionally frozen to reduce training cost, while the Transformer and classifier layers are trained using the AdamW optimizer. Cross-entropy loss is used for supervision. Training is performed for a fixed number of epochs due to computational constraints.

4 Results & Visualization

Figure 2 presents the training loss and accuracy curves of the proposed CNN–Transformer fusion model on the CIFAR-10 dataset. The loss generally decreases over epochs despite some fluctuations, indicating that the model is learning and converging during training. At the same time, training accuracy shows an overall upward trend, reaching a high value toward the later epochs.

The combined trends in Figure 2 suggest that the model successfully learns discriminative features from the data. While minor instability is observed during training, the final accuracy and loss values indicate effective learning and satisfactory performance for the classification task.

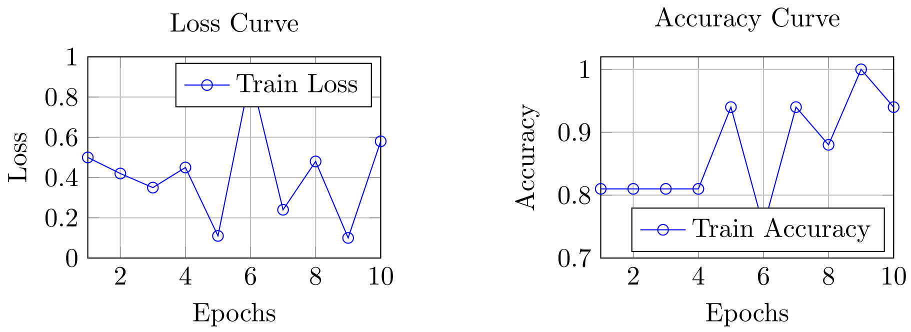

Figure 2: Training loss and accuracy curves for the CNN–Transformer fusion model on CIFAR-10.

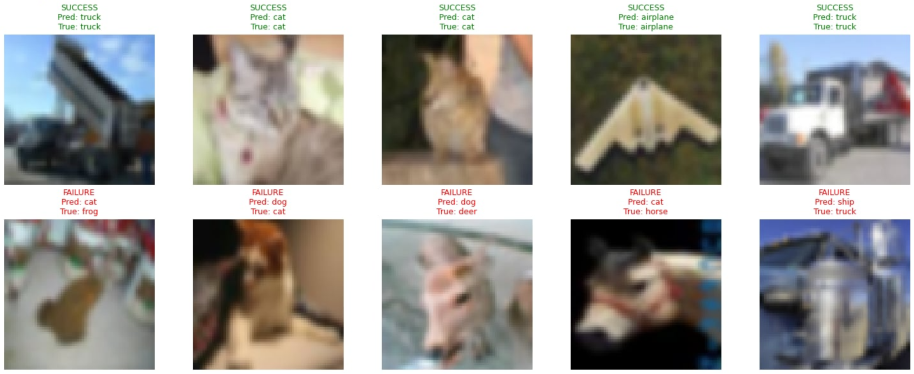

Figure 3: Sample predictions

5 Discussion

The CNN–Transformer fusion successfully combines local feature extraction with global contextual modeling, allowing the network to learn more expressive representations. Experimental results show a modest improvement in accuracy compared to a baseline CNN, indicating that global attention contributes positively to classification performance.

However, the inclusion of Transformer layers increases computational overhead and training time. Additionally, the small 32×32 resolution of CIFAR images limits the effectiveness of self-attention, and the use of limited training epochs and basic data augmentation may restrict further performance gains.

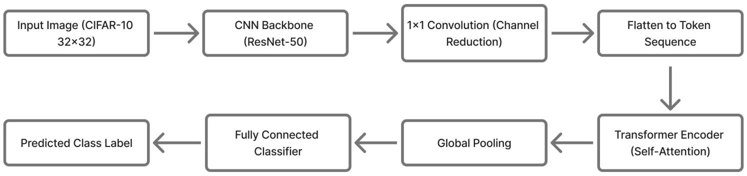

Figure 4: Architecture diagram

6 Conclusion

This study demonstrated that fusing a CNN backbone with a Transformer encoder can improve image classification performance by combining local feature extraction with global contextual modeling. Experimental results on the CIFAR dataset showed a modest accuracy gain over a baseline CNN, validating the effectiveness of the fusion approach. Despite increased computational cost and limitations due to small image resolution, the results suggest that CNN–Transformer hybrids are a promising direction for future computer vision research.