Hybrid CNN-Transformer Architecture for Fine-Grained Classification of Philippine Currency

Imroz Khan[1] and Febron Sedoriosa[1]

University of Science and Technology of Southern Philippines,
Cagayan de Oro City, Misamis Oriental 9000, Philippines

Abstract. This case study explores the efficacy of "Architecture Fusion" in computer vision by combining a Convolutional Neural Network (CNN) with a Vision Transformer (ViT) for the fine-grained classification of Philippine currency. Leveraging a dataset of 16 distinct classes, including both coins and banknotes, we implemented a hybrid ResNetViTHybrid model. The architecture fuses the local feature extraction capabilities of a pretrained ResNet18 with the global context modeling of a Transformer Encoder. To address class imbalance, we employed Stratified Random Sampling. The model demonstrated rapid convergence, achieving $> 98\%$ validation accuracy within the first epoch, suggesting that hybrid architectures effectively capture the subtle visual distinctions in currency denominations.

Keywords: Architecture Fusion · ResNet · Vision Transformer · Image Classification · Philippine Currency

1 Introduction

The automation of currency recognition is a critical task for financial systems, automated vending machines, and assistive technologies designed for the visually impaired. However, distinguishing between currency denominations presents significant challenges due to subtle inter-class similarities, such as the visual resemblance between the old and new series of 10 Peso coins, and intra-class variations caused by wear and tear. While standard Convolutional Neural Networks (CNNs) excel at detecting local patterns like edges and textures [1], they often struggle to model long-range spatial dependencies required to understand the full context of an image. Conversely, Vision Transformers (ViTs) are adept at modeling global relationships through self-attention mechanisms but typically require massive datasets to learn low-level features effectively [2]. This study proposes a hybrid architecture that fuses a CNN backbone with a Transformer Encoder. The primary objective is to combine the inductive bias of CNNs with the attention mechanisms of Transformers to achieve robust, fine-grained classification of Philippine currency.

2 Dataset Description

The study utilized a custom dataset of Philippine Currency obtained from Roboflow Universe [3]. The dataset includes 16 distinct classes covering the full spectrum of circulating legal tender. These classes encompass banknotes in denominations of 20, 50, 100, 200, 500, and 1000 Pesos (including the distinct Polymer variant), as well as coins ranging from 25 Centavos to the 20 Peso coin (including both "Old" and "New" series variants).

To address the significant class imbalance present in the raw data, such as the disparity between thousands of coin images and fewer high-value banknote samples, we implemented a Stratified Random Sampling strategy. This involved randomly sampling exactly 100 images per class for the training set, resulting in approximately 1,600 total training images, and exactly 20 images per class for the validation set, totaling approximately 320 images. Regarding preprocessing, all images were cropped using provided bounding box annotations to remove background noise and subsequently resized to standard 224 × 224 pixel dimensions to match the input requirements of the backbone architecture.

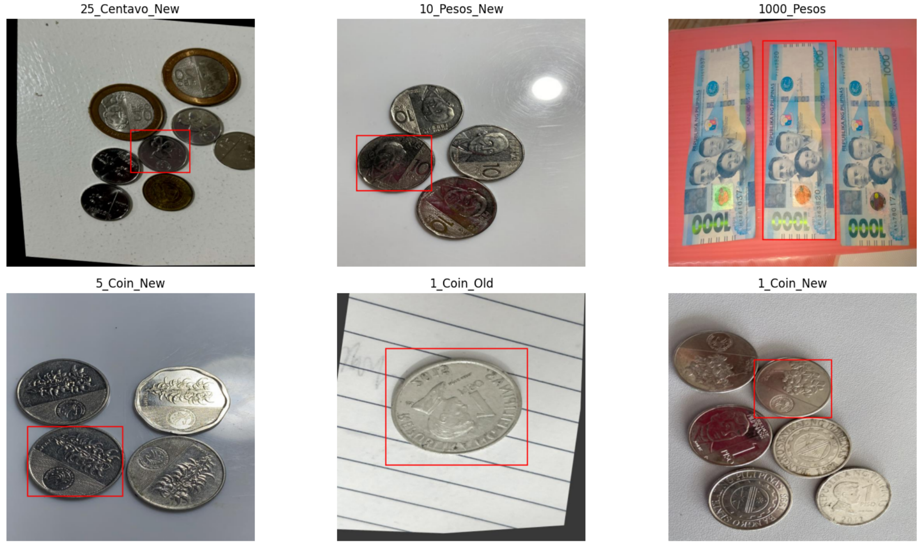

Fig. 1. Visualization of the preprocessing step. The red bounding box indicates the region of interest defined by the COCO annotations. The model crops this region to focus solely on the currency object, removing background clutter before resizing to 224 × 224.

3 Methodology

The core contribution of this study is the implementation of an architecture fusion strategy, defined in PyTorch as the `ResNetViTHybrid` class. We constructed a dual-stage pipeline that processes images sequentially through a CNN and a Transformer.

The first stage involves Local Feature Extraction using a ResNet18 backbone [1]. We removed the final fully connected classification layer, allowing the ResNet18 to process the input image ($224 \times 224 \times 3$) and output a dense feature map of shape ($512 \times 7 \times 7$). This step captures high-level "local" features such as the texture of the paper money or the ridges on coins. In the second stage, these spatial features are flattened into a sequence of 49 tokens, effectively treating the image as a sequence of patches similar to Natural Language Processing techniques. These tokens are then passed into the Global Context Modeling stage using a Transformer Encoder Layer [2]. The Multi-Head Self-Attention mechanism allows the model to compare every patch of the image against every other patch, learning global structural relationships. Finally, the output of the Transformer is pooled and passed through a final Linear Layer to map the features to the 16 class logits.

The model was trained in a Google Colab environment accelerated by a T4 GPU. We trained for 5 epochs using a Batch Size of 16, the Adam Optimizer with a learning rate of 1×10^{-4}, and the CrossEntropyLoss function. Regularization was achieved primarily through Transfer Learning, relying on the pretrained weights of ResNet18 to prevent overfitting on the stratified subset.

4 Results and Visualizations

The hybrid model showed exceptional performance, converging significantly faster than anticipated. Quantitatively, the training loss decreased rapidly from 0.23 in Epoch 1 to 0.04 by Epoch 5. Simultaneously, the validation accuracy stabilized at approximately 99% starting from the very first epoch. Qualitatively, the model successfully distinguished between visually similar classes. For instance, it correctly identified the *1000_Pesos_Polymer* versus the standard *1000_Pesos*, likely due to the Transformer's ability to attend to the specific texture patterns, such as the shiny surface of polymer notes, globally across the image.

To further analyze the model's classification performance, we visualized the Normalized Confusion Matrix (Figure 4). The diagonal dominance confirms high precision and recall across all 16 classes.

5 Discussion

The "Architecture Fusion" proved highly effective for this task. The ResNet backbone provided a strong visual foundation, detecting edges and shapes immediately, while the transformer component fine-tuned the decision boundaries

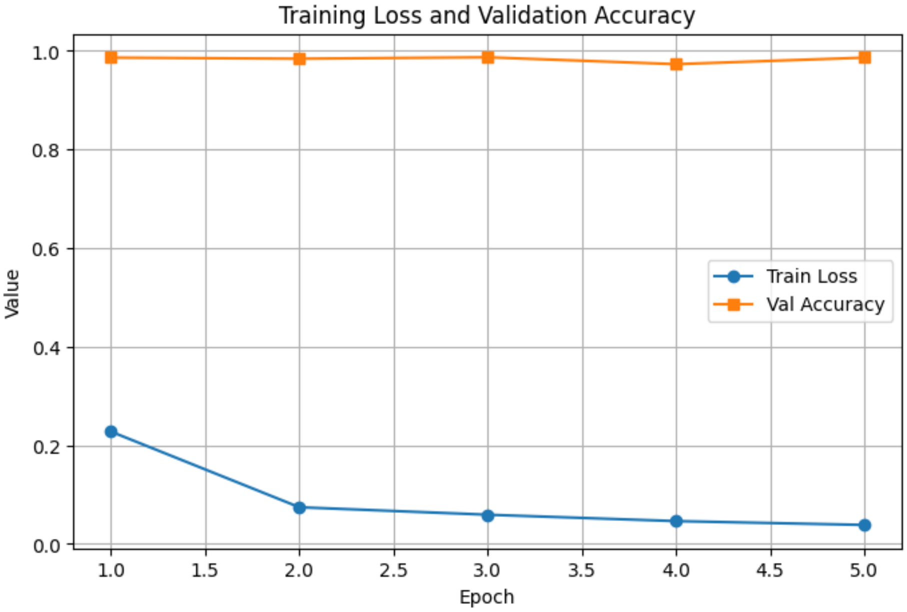

Fig. 2. Training Loss and Validation Accuracy Curves

between complex classes. The high accuracy can be attributed largely to transfer Learning. Since ResNet18 was pretrained on ImageNet, it already possessed robust feature extractors. The Stratified Sampling strategy was also crucial; by limiting the data to 100 balanced samples per class, we prevented the model from optimizing solely for the majority class, forcing it to learn the features of rare bills equally well. A limitation of this approach is the reliance on pre-cropped images; in a real-world scenario, this classifier would need to be paired with an object detector.

6 Conclusion

This mini case study successfully demonstrated the power of fusing CNNs and Transformers. The `ResNetViTHybrid` model achieved near-perfect accuracy on the Philippine Currency dataset with minimal training time. The fusion strategy effectively combined the efficiency of CNNs with the contextual power of Transformers, providing a robust solution for fine-grained image classification.

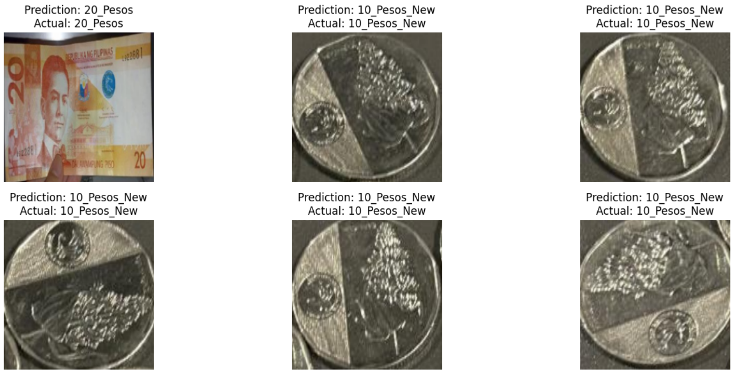

Fig. 3. Sample Predictions on Test Data

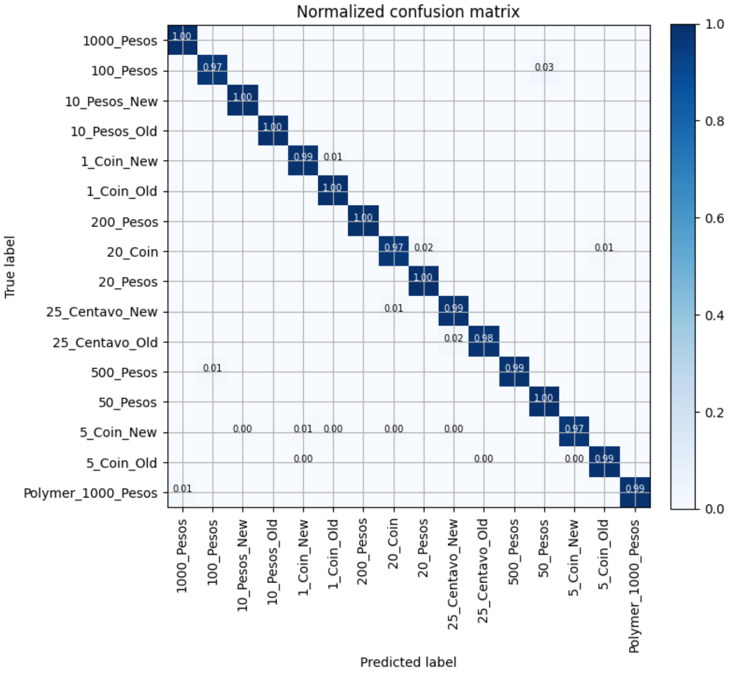

Fig. 4. Normalized Confusion Matrix. The strong diagonal axis indicates that the model rarely misclassifies denominations, effectively distinguishing even between visually similar classes like the old and new 10 Peso coins.

References

1. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770–778 (2016)
2. Dosovitskiy, A., et al.: An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. arXiv preprint arXiv:2010.11929 (2020)
3. Philippine Banknotes Dataset. Roboflow Universe.: `https://universe.roboflow.com/philippine-banknotes/philippine-banknotes/dataset` (Accessed: December 2025)