Feature Fusion of ResNet18 and MobileNetV2 for CIFAR-10 Image Classification

Via Nicole A. Preciado[1] and Ismael Sulpicio G. Zarate[1]

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

Abstract. Image classification relies on effective feature representation and model architecture. While single convolutional neural network (CNN) backbones perform well, they may be limited by architectural bias. This mini case study investigates feature-level fusion of two pretrained CNN backbones, ResNet18 and MobileNetV2, for CIFAR-10 image classification. Classifier heads are removed, backbone features are concatenated, and a lightweight multilayer perceptron performs final prediction. On the CIFAR-10 test set, the fusion model achieves 96.24% accuracy, slightly outperforming ResNet18 (96.12%) and improving over MobileNetV2 (94.51%). The results demonstrate that simple feature concatenation can yield competitive and stable performance while satisfying the architecture fusion requirement.

Keywords: Image Classification · Feature Fusion · CNN · CIFAR-10

1 Introduction

1.1 Problem Definition

Image classification aims to automatically assign semantic labels to images based on visual content. Although deep CNNs achieve strong performance, a single backbone may not capture all relevant visual patterns due to architectural constraints and representational bias.

1.2 Motivation and Relevance

Feature representation and visualization are central topics in graphics and visual computing. Feature fusion enables combining complementary representations learned by different architectures, improving robustness and generalization while remaining simple and interpretable.

2 Dataset Description

The CIFAR-10 dataset consists of 60,000 RGB images belonging to 10 object classes: airplane, automobile, bird, cat, deer, dog, frog, horse, ship, and truck.

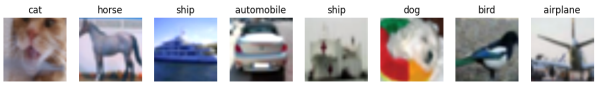

Fig. 1. Representative CIFAR-10 sample images after resizing for pretrained CNN backbones.

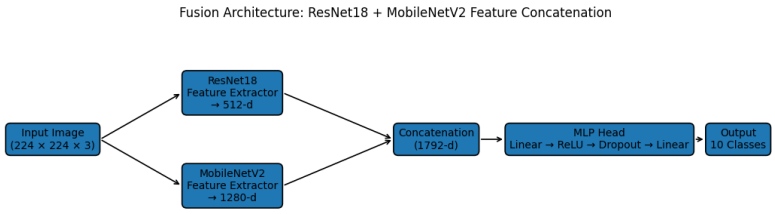

Fig. 2. Feature-level fusion architecture combining ResNet18 and MobileNetV2 through feature concatenation and an MLP classifier.

The dataset is divided into 50,000 training images and 10,000 test images. From the training set, 45,000 images are used for training and 5,000 for validation.

Images are resized from 32×32 to 224×224 to match ImageNet-pretrained model inputs. Figure 1 shows representative samples from the dataset.

3 Methodology

3.1 Architectures Used

Two pretrained CNN backbones are employed as feature extractors:

- **ResNet18**, which utilizes residual connections to learn deep hierarchical features.
- **MobileNetV2**, a lightweight architecture optimized for efficiency using depthwise separable convolutions.

The original classification layers are removed to obtain fixed-length feature embeddings.

3.2 Fusion Strategy

ResNet18 produces a 512-dimensional feature vector, while MobileNetV2 outputs a 1280-dimensional feature vector. These embeddings are concatenated into a 1792-dimensional representation and passed to a lightweight MLP classifier consisting of linear, ReLU, dropout, and linear layers. Figure 2 illustrates the fusion pipeline.

Table 1. Test accuracy comparison of baseline and fusion models.

Model	Test Accuracy
ResNet18 Baseline	0.9612
MobileNetV2 Baseline	0.9451
ResNet18 + MobileNetV2 Fusion	0.9624

3.3 Preprocessing and Training Details

All experiments are implemented using PyTorch and executed in Google Colab with GPU acceleration. Images are normalized using ImageNet statistics. Data augmentation includes random resized cropping, horizontal flipping, and color jittering. Mixed-precision training is employed to improve efficiency.

Training configuration:

- Batch size: 64
- Data loader workers: 2
- Optimizer: AdamW
- Learning rate: 3×10^{-4}
- Weight decay: 1×10^{-4}
- Loss: Cross-entropy with label smoothing (0.1)
- Epochs: 8
- Backbone freezing: first 3 epochs before fine-tuning

4 Results and Visualizations

4.1 Accuracy and Loss Curves

Figure 3 stacks the training and validation accuracy and loss curves for the ResNet18 baseline, MobileNetV2 baseline, and the fusion model. All models converge smoothly after backbone unfreezing, with stable validation behavior.

4.2 Quantitative Results

The fusion model achieves the highest CIFAR-10 test accuracy (96.24%), slightly exceeding ResNet18 (96.12%) and improving over MobileNetV2 (94.51%). This indicates that feature concatenation can provide modest gains by combining complementary representations.

4.3 Sample Predictions

Figures 4 and 5 show representative correct and incorrect predictions from the fusion model. Misclassifications typically involve visually similar categories or low-resolution ambiguity.

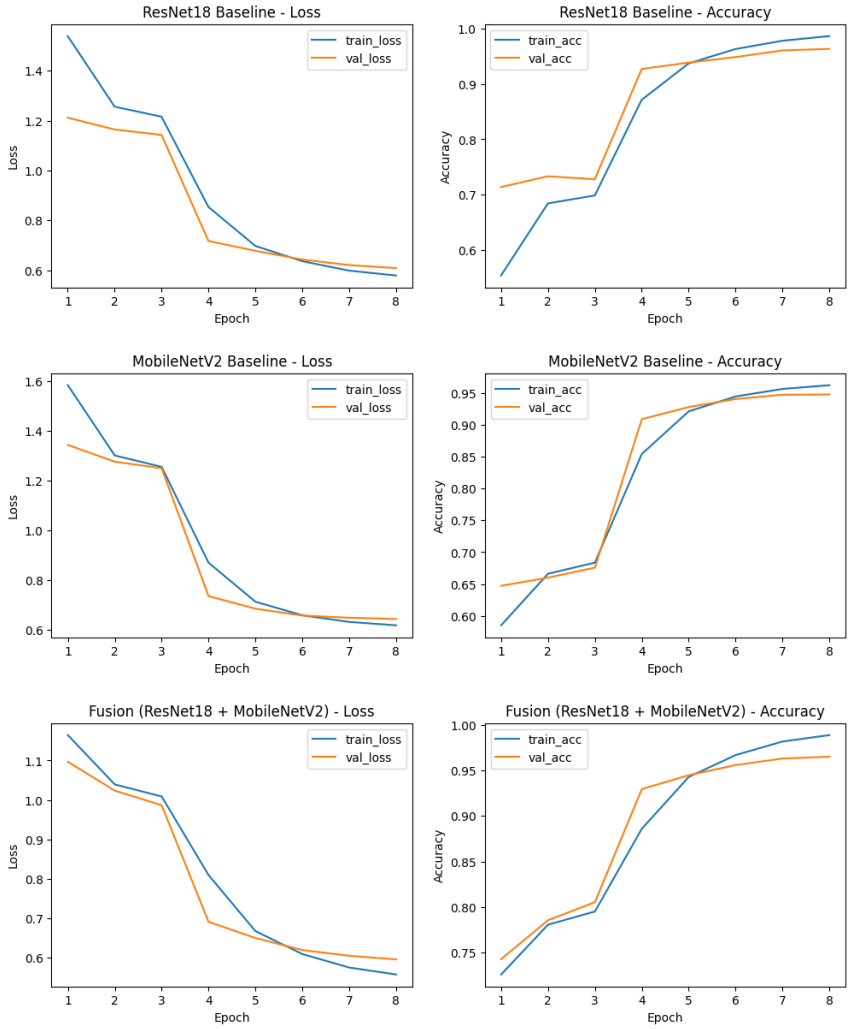

Fig. 3. Training and validation accuracy and loss curves for ResNet18, MobileNetV2, and the proposed fusion model on CIFAR-10.

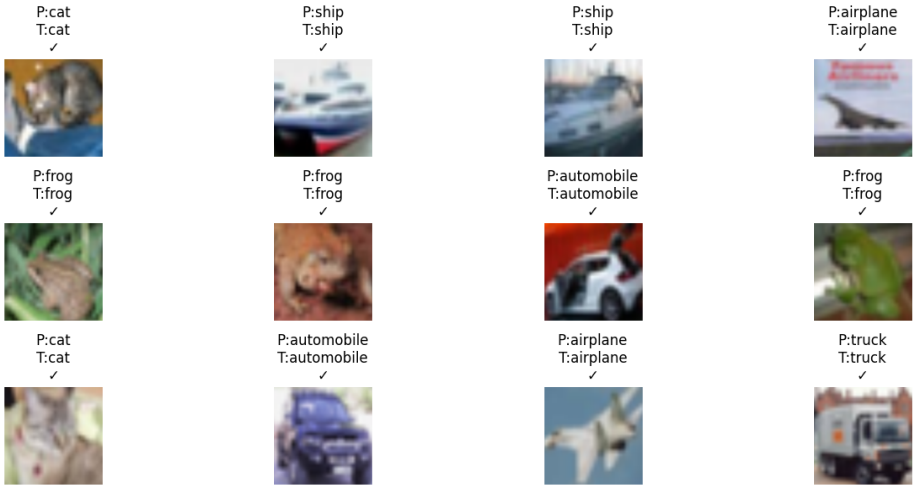

Fig. 4. Sample correct predictions produced by the fusion model.

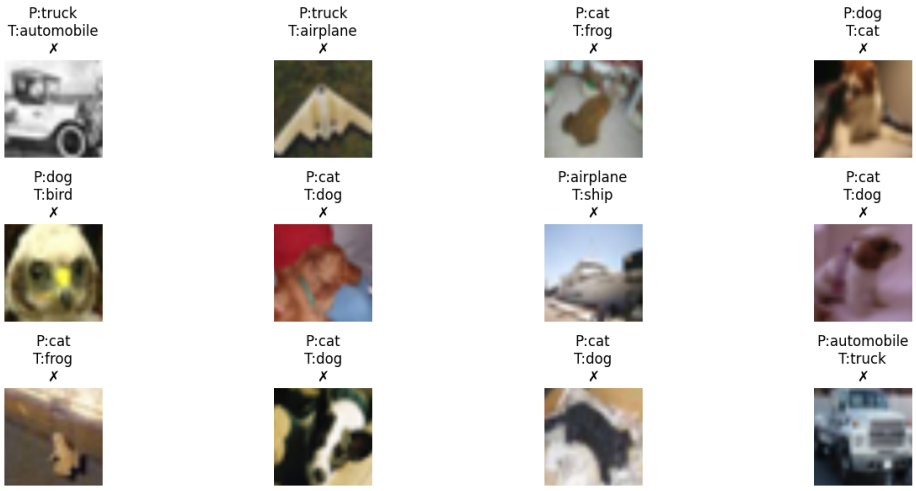

Fig. 5. Sample incorrect predictions produced by the fusion model. Errors commonly occur between visually similar classes.

5 Discussion

5.1 Contribution of Fusion

The fusion approach combines complementary representations learned by ResNet18 and MobileNetV2. Although the improvement over ResNet18 is modest, the fusion model achieves the highest test accuracy and substantially improves over MobileNetV2, demonstrating the benefit of combining deep residual and lightweight mobile features.

5.2 Limitations and Observations

Feature concatenation proved effective, easy to implement, and straightforward to explain. However, confusion between visually similar classes (e.g., cat and dog) remains, suggesting potential benefits from attention-based fusion mechanisms or higher-resolution datasets.

6 Conclusion

This mini case study demonstrates that feature-level fusion of pretrained CNN backbones is a practical and effective strategy for CIFAR-10 image classification. By concatenating ResNet18 and MobileNetV2 features and applying a lightweight MLP classifier, the fusion model achieves 96.24% test accuracy and slightly outperforms single-backbone baselines.

Acknowledgments. The authors thank the course instructor for guidance and feedback throughout the Graphics and Visual Computing course.

Disclosure of Interests. The authors declare that they have no competing interests.

References

1. Krizhevsky, A., Hinton, G.: Learning multiple layers of features from tiny images. Technical report, University of Toronto (2009). https://www.cs.toronto.edu/ kriz/cifar.html
2. 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 (CVPR), pp. 770–778 (2016). https://arxiv.org/abs/1512.03385
3. Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., Chen, L.-C.: MobileNetV2: Inverted residuals and linear bottlenecks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4510–4520 (2018). https://arxiv.org/abs/1801.04381
4. Zhang, Y., Yang, Q., Wang, H.: Feature-level fusion of convolutional neural networks for image classification. IEEE Access **6**, 59439–59447 (2018). https://ieeexplore.ieee.org/document/8462313