Image Classification on CIFAR-10 Using Feature-Level Fusion of ResNet50 and MobileNetV2

Isaiah Marc R. Postrero, Hurley W. Banaag, and Keely Shay Englatiera

Department of Computer Science, University of Science and Technology of Southern Philippines
cs.cdo@ustp.edu.ph

Abstract. This mini case study explores a Graphics and Visual Computing approach to image classification by combining features from two pretrained convolutional neural networks: ResNet50 and MobileNetV2. The goal is to check if merging the strengths of two different architectures can improve performance on a small but diverse dataset like CIFAR-10. We briefly describe the dataset, preprocessing, fusion strategy, and training pipeline, and we present visual diagnostics such as Grad-CAM heatmaps, accuracy curves, and sample predictions. Results show that simply fusing frozen features performs poorly due to feature mismatch, while lightly finetuning the last blocks of both networks helps them align better and increases accuracy.

1 Introduction

Graphics and Visual Computing often deal with extracting meaningful information from images. Image classification is one of its most common tasks. Modern convolutional neural networks (CNNs) learn layered features that capture shapes, textures, and patterns. Using pretrained networks is helpful because they already understand general visual structures. Combining multiple models, or *model fusion*, attempts to mix different viewpoints or feature styles to get a stronger representation. In this study, we explore feature-level fusion between two well-known architectures: ResNet50, a high-capacity residual network, and MobileNetV2, a lighter model designed for efficiency.

2 Problem and Relevance

The task is to classify CIFAR-10 images into ten everyday object categories. Although CIFAR-10 is small, it contains many variations in color, angle, and shape, which makes it a good benchmark for testing feature representations.

This topic is relevant to visual computing because it:

- demonstrates transfer learning using pretrained deep models,
- uses Grad-CAM to visually inspect what parts of the image the model focuses on,
- explores feature fusion, which is widely used in multimodal systems (vision + text, or multi-network ensembles).

3 Dataset

CIFAR-10 contains 50,000 training images and 10,000 test images, all in RGB format with a resolution of 32×32. The classes are airplane, automobile, bird, cat, deer, dog, frog, horse, ship, and truck. For compatibility with ResNet50 and MobileNetV2, all images are resized to 224×224.

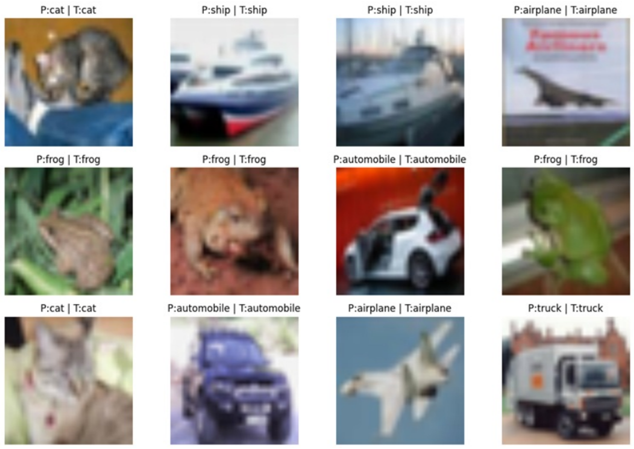


Fig. 1: CIFAR-10 sample images.

4 Methodology

4.1 Fusion Architecture

We evaluate four setups:

1. **ResNet50 (finetuned)** – baseline large model.
2. **MobileNetV2 (finetuned)** – baseline lightweight model.
3. **Fusion (Frozen)** – both networks kept frozen and their features simply concatenated.
4. **Fusion (Finetune Last Blocks)** – selected final layers are unfrozen so the two models can adjust and produce more compatible features.

Both CNNs produce their feature vectors, which are combined as:

$$F_{\text{fusion}} = \text{Concat}(F_{\text{ResNet50}}, F_{\text{MobileNetV2}})$$

The classifier head is:

$$\text{Linear(3328}\rightarrow\text{512)} \rightarrow \text{ReLU} \rightarrow \text{Dropout(0.2)} \rightarrow \text{Linear(512}\rightarrow\text{10)}$$

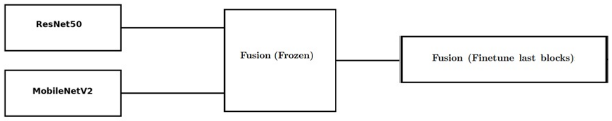


Fig. 2: Fusion pipeline overview.

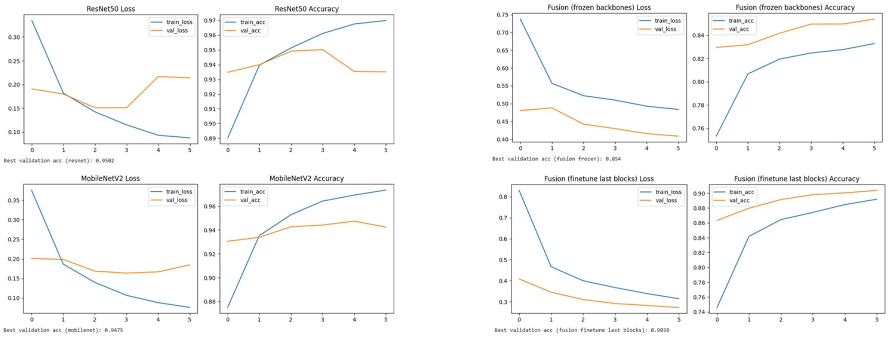
4.2 Training Setup

- Basic augmentations: resize, random crop, horizontal flip
- Optimizer: AdamW (1×10^{-4} learning rate)
- Training duration: 6 epochs
- Mixed precision for faster training

5 Results

Table 1: Validation accuracy of all models.

Model	Val Acc (%)
ResNet50 baseline	95.02
MobileNetV2 baseline	94.75
Fusion (frozen)	85.40
Fusion (finetuned)	90.38

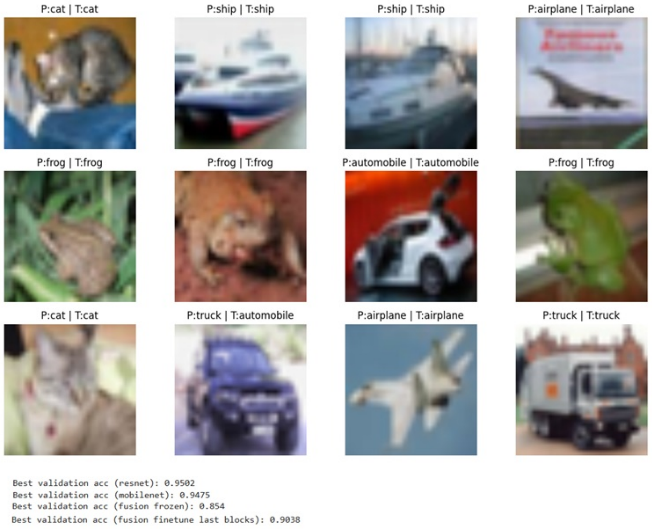


```
Best validation acc (resnet): 0.9502
Best validation acc (mobilenet): 0.9475
Best validation acc (fusion frozen): 0.854
Best validation acc (fusion finetune last blocks): 0.9038
```

(a) Baseline models

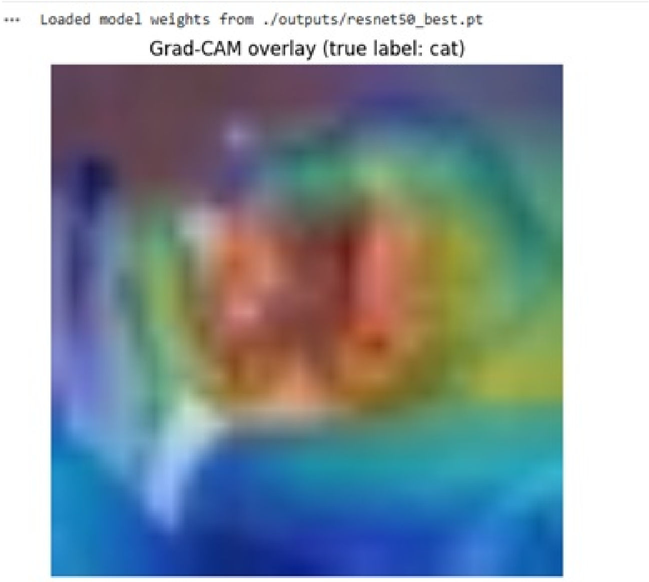
(b) Fusion models

Fig. 3: Accuracy curves.



```
Best validation acc (resnet): 0.9502
Best validation acc (mobilenet): 0.9475
Best validation acc (fusion frozen): 0.854
Best validation acc (fusion finetune last blocks): 0.9038
```

Fig. 4: Predictions of the finetuned fusion model.



```
··· Loaded model weights from ./outputs/resnet50_best.pt
```

Fig. 5: Grad-CAM heatmaps showing model attention.

6 Discussion

The frozen fusion model performed the worst. Since both networks were pretrained on ImageNet and left untouched, their feature styles did not match. One network may focus more on texture patterns while the other may focus on object outlines. When these incompatible features are merged directly, the classifier struggles to interpret them.

Allowing the last blocks to be finetuned improves accuracy because the networks can adjust their features to align better. However, even with finetuning, the fusion still lags behind single backbones. True fusion often requires deeper integration, such as attention layers, projection networks, or transformer-based fusion that better blends high-level features.

7 Conclusion

This mini case study shows how feature-level fusion behaves in a Graphics and Visual Computing setting. Simple frozen concatenation is not effective, while selective fine-tuning helps stabilize the combined representation. Tools like Grad-CAM and learning curves provide helpful insights into model behavior. Future work may explore attention-based fusion, better feature alignment layers, or hybrid CNN Transformer models for improved performance.