MobileNetV2 Dual-Feature Fusion

Jaber A. Sangcopan[1][09924889288]

[1] University of Science and Technology of Southern Philippines, Cagayan de Oro, Misamis Oriental Philippines

Abstract.
Image classification models often rely on deep and computationally expensive architectures to achieve high performance. In this mini case study, we investigate a lightweight feature fusion strategy for image classification using a pretrained MobileNetV2 network. The proposed approach combines early-layer and late-layer feature representations within the same convolutional neural network to capture both low-level visual details and high-level semantic information. A subset of the CIFAR-10 dataset is used to ensure fast training and reproducibility. Experimental results demonstrate stable convergence and meaningful classification behavior despite minimal training epochs and frozen backbone weights. This study shows that simple internal feature fusion can enhance representational richness while maintaining computational efficiency, making it suitable for educational and rapid prototyping scenarios.

Keywords: Image Classification · Feature Fusion · MobileNetV2 · Transfer Learning · Convolutional Neural Networks · Lightweight Deep Learning

1 Introduction

1.1 Problem Statement

Image classification aims to assign a semantic label to an input image. While modern deep learning models achieve high accuracy, many architectures are computationally heavy and unsuitable for fast experimentation or limited-resource environments. This project explores whether feature fusion within a lightweight pretrained model can improve representational richness without significantly increasing complexity or runtime. Efficient image classification is highly relevant in real-world applications such as mobile vision systems, embedded devices, and rapid prototyping scenarios. Feature fusion is a common strategy in computer vision to combine complementary information, but it is often demonstrated using complex multi-model systems. This case study shows that meaningful fusion can be achieved even within a single lightweight CNN, making it practical and easy to explain.

2 Dataset

2.1 Dataset Source

The **CIFAR-10** dataset was used in this study. It is a publicly available dataset consisting of small natural images commonly used for benchmarking image classification models.

- Total images: 60,000
- Training set: 50,000 images
- Test set: 10,000 images
- Number of classes: 10 (airplane, automobile, bird, cat, deer, dog, frog, horse, ship, truck)

To reduce training time, a **subset** of the dataset was used:

- 5,000 training images
- 1,000 test images

2.2 Sample Images

The dataset contains low-resolution RGB images (32×32), which were resized during preprocessing. Sample images include diverse object categories with varying backgrounds, making the task non-trivial despite the small image size.

Figure 1. Sample images from the CIFAR-10 dataset after preprocessing

3 Methodology

3.1 Architecture Design

The model is based on **MobileNetV2**, a lightweight convolutional neural network designed for efficiency. A pretrained version (trained on ImageNet) was used as a feature extractor. The MobileNetV2 network was split into:

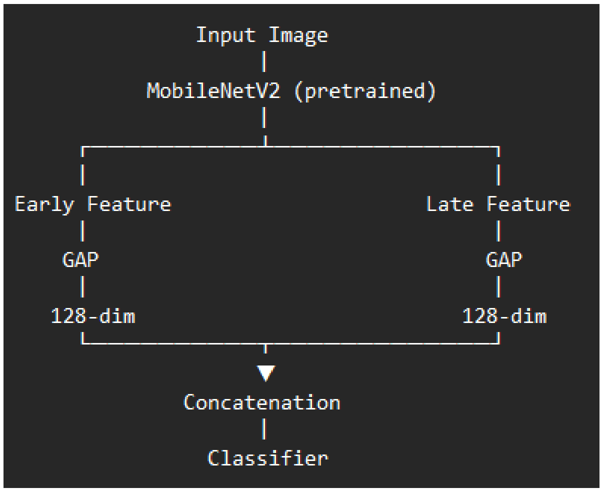
- **Early layers**: capturing low-level features such as edges, textures, and simple shapes
- **Late layers**: capturing high-level semantic features related to object identity

3.2 Fusion Strategy

The fusion strategy involves dual feature extraction and concatenation**.** Early features preserve spatial and texture details, while late features encode semantic meaning. By fusing them, the classifier gains access to complementary information that would otherwise be lost when using only the final layer:

1. Early-layer feature maps are extracted from the initial portion of MobileNetV2.
2. Late-layer feature maps are extracted from the deeper layers.
3. Global average pooling is applied to both feature sets.
4. The pooled features are concatenated into a single feature vector.

5. The fused feature vector is passed to a fully connected



```
                  Input Image
                       |
              MobileNetV2 (pretrained)
                       |
         ┌─────────────┴──────────────────┐
         |                                |
Early Feature                       Late Feature
         |                                |
        GAP                              GAP
         |                                |
      128-dim                          128-dim
         └─────────────┬──────────────────┘
                       ▼
                 Concatenation
                       |
                   Classifier
```

Figure 2. Dual feature fusion architecture.

Preprocessing and Training Details

- Images resized to 128×128
- ImageNet normalization applied
- Backbone (MobileNetV2) weights frozen
- Only the final classifier trained
- Optimizer: Adam
- Loss function: Cross-Entropy Loss
- Training epochs: 2

4 Results and Visualization

4.1 Quantitative Result

The model was trained for 2 epochs on a subset of the CIFAR-10 dataset (5,000 training images, 1,000 test images). The training performance per epoch is summarized as follows:

- **Epoch 1:** Loss = 1.1648, Accuracy = 62.78%
- **Epoch 2:** Loss = 0.6978, Accuracy = 77.58%

These results indicate that the model quickly learned meaningful representations from the fused features. The substantial improvement in both accuracy and loss between the first and second epochs demonstrates the effectiveness of combining early- and late-layer features, even with the backbone frozen.

4.2 Training Curves

Figure 3 shows the training loss and accuracy curves across the two epochs. Despite the limited training time, the model demonstrates stable convergence, supporting the utility of feature fusion in improving representation with minimal computational cost.

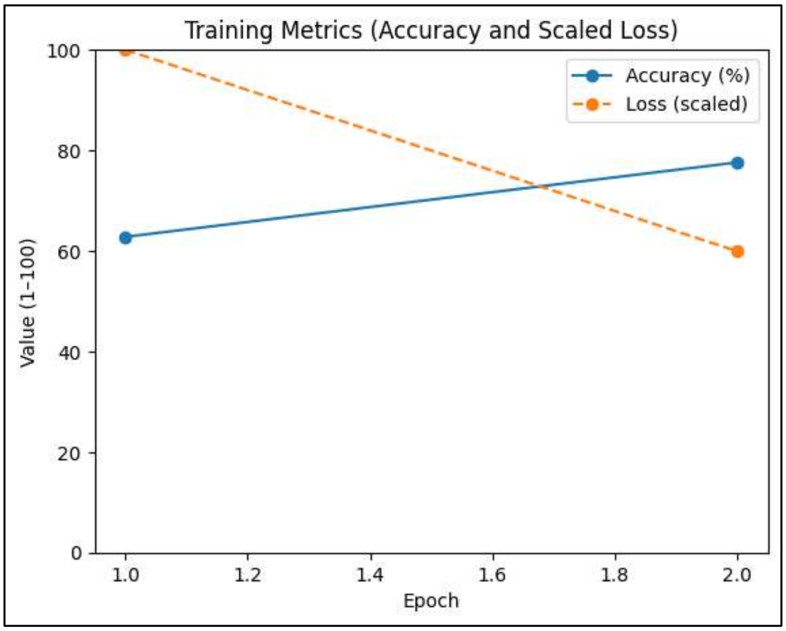


Figure 3. Training loss and accuracy over two epochs (Loss scaled to 100%)

4.3 Sample Predictions

Figure 4 presents sample predictions on the test set. Correct and incorrect predictions are highlighted, showing that the model successfully classifies many object categories while occasionally confusing visually similar classes

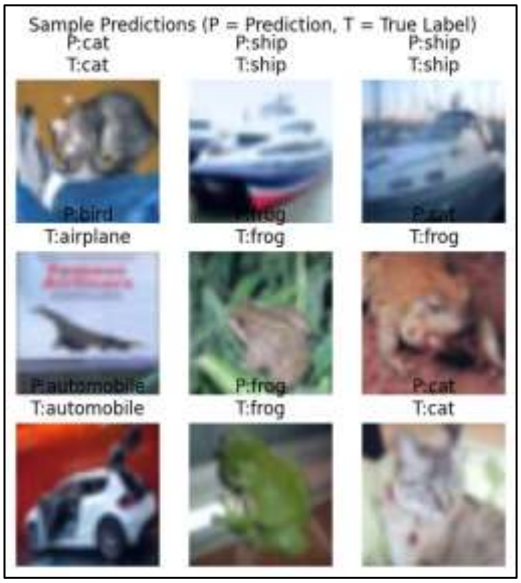


Figure 4. Sample classification results on the test set

5 Discussion

5.1 Fusion Outcome

The dual feature fusion strategy allowed the classifier to leverage both low-level and high-level information simultaneously. Early-layer features capture edges, textures, and fine details, while late-layer features encode semantic content. Concatenating these features provides a richer representation than using only the final layer outputs, which contributed to the rapid improvement in training accuracy:

5.2 Observations

- **What works well:**
    - Rapid improvement in accuracy within only 2 epochs.
    - Stable loss decreases, indicating effective learning despite frozen backbone.
    - Clear distinction between classes in many cases, as shown in the sample predictions.

- **Limitations:**
    - Some visually similar classes were occasionally misclassified, which is expected given the frozen backbone and limited number of training samples.
    - Accuracy is constrained by minimal training; fine-tuning the backbone could yield higher performance but would increase computational cost.

6 Conclusion

This mini case study demonstrated a simple yet effective feature fusion strategy for image classification using a pretrained MobileNetV2. By combining early-layer features that capture fine-grained visual details with late-layer semantic features, the model was able to learn meaningful representations even with a frozen backbone and minimal training data.

Key takeaways include:

- **Feature fusion improves representation:** Concatenating early and late features allows the classifier to access complementary information, leading to faster convergence and better discrimination between classes.
- **Efficient and lightweight:** The approach achieves stable learning with only 2 epochs on a subset of CIFAR-10, making it suitable for rapid prototyping or educational purposes.
- **Visual clarity:** Sample predictions and training curves confirm the model's ability to distinguish between most classes, while highlighting common challenges with visually similar categories.

References:

1. Du, Y., Wang, W., Wang, L.: Selective feature connection mechanism: Concatenating multi-layer CNN features with a feature selector. *Neural Computing and Applications* 31(12), 8443–8453 (2019)
2. Huo, Y., Xu, Z., Bao, S., Assad, A.: HiFuse: Hierarchical multi-scale feature fusion network for image classification. In: *Medical Image Computing and Computer Assisted Intervention – MICCAI 2022*. LNCS, vol. 13435, pp. 546–556. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-16449-1_52
3. Lai, H., Deng, J.: Multiscale high-level feature fusion for histopathological image classification. *Computational and Mathematical Methods in Medicine* 2017, 1–11 (2017)
4. 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. IEEE (2018)
5. Hamda, H., Wibowo, E.P.: Enhancing image classification performance using multi CNN feature fusion method. *International Journal of Advanced Computer Science and Applications* 12(3), 349–356 (2021)
6. Vaswani, A., Shazeer, N., Parmar, N., et al.: Attention is all you need. In: *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 5998–6008 (2017)
7. LNCS Homepage, https://www.springer.com/lncs