Ensemble Learning for Gesture Recognition: A ResNet50 and EfficientNet-B0 Fusion

James Darwen B. Bañas, Carl A. Cañas, Ralph Rhey A. Lumigue, and Joesant Cordova

Department of Computer Science
james.banas@student.university.edu

Abstract. Hand gesture recognition is a fundamental component of human-computer interaction (HCI). This report presents a robust ensemble deep learning architecture for the classification of Rock-Paper-Scissors gestures. By fusing the feature extraction capabilities of ResNet50 and EfficientNet-B0, we aim to balance deep representational capacity with computational efficiency. The proposed model concatenates feature vectors from both backbones to form a dense representation before classification. Experimental results on a dataset of 325 unseen test images demonstrate an overall accuracy of 99%, with an F1-score of 0.99. The model exhibited perfect precision in identifying the 'Scissors' class and perfect recall for the 'Paper' class. Analysis of the confusion matrix reveals minor misclassification between geometrically similar gestures (Rock and Scissors), suggesting avenues for future data augmentation.

Keywords: Deep Learning · Ensemble Methods · ResNet50 · EfficientNet · Gesture Recognition.

1 Introduction

Hand gesture recognition serves as a critical bridge in non-verbal communication between humans and machines. While single-architecture Convolutional Neural Networks (CNNs) have achieved significant success, they often face tradeoffs between depth and efficiency. Furthermore, static gestures like "Rock" and "Scissors" often share similar geometric hulls, leading to classification ambiguity.

This study explores an ensemble approach, leveraging **ResNet50** for its ability to capture high-level semantic features through residual learning, and **EfficientNet-B0** for its optimized parameter efficiency. The objective is to maximize classification accuracy on the standard Rock-Paper-Scissors dataset while maintaining a streamlined inference pipeline.

2 Dataset Description

The model was trained and evaluated on the "Rock-Paper-Scissors" dataset sourced from Roboflow public repositories.

- **Source:** Roboflow Public Datasets.
- **Classes:** Three distinct classes: Rock, Paper, and Scissors.
- **Preprocessing:** All input images were resized to 224×224 pixels to match the input requirements of the pre-trained backbones. Pixel values were normalized using the standard ImageNet means ($[0.485, 0.456, 0.406]$) and standard deviations ($[0.229, 0.224, 0.225]$).
- **Splits:** The dataset was strictly divided into Training and Testing sets. The test set consisted of 325 images, ensuring evaluation was performed on completely unseen data.

3 Methodology

The core architecture utilizes a feature fusion strategy. The data is processed in parallel through two pre-trained backbones with their classification heads removed.

3.1 Architectures Used

1. **ResNet50:** A 50-layer Residual Network. It extracts a deep feature vector of size 2048.
2. **EfficientNet-B0:** A scalable CNN optimized for floating-point operations (FLOPs). It extracts a feature vector of size 1280.

3.2 Fusion Strategy

As illustrated in Fig. 1, the output feature vectors from both backbones are flattened and concatenated.

$$F_{total} = F_{ResNet} \oplus F_{EfficientNet} \tag{1}$$

The resulting vector has a dimension of $2048 + 1280 = 3328$. This fused representation is passed through a custom classifier consisting of a Linear layer ($3328 \rightarrow 512$), a ReLU activation function, a Dropout layer ($p = 0.3$) to prevent overfitting, and a final Linear output layer for the three target classes.

3.3 Training Details

The model was implemented in PyTorch and trained on a T4 GPU. The optimizer used was **Adam** with a learning rate of 0.001. The loss function was Cross Entropy Loss. Training was conducted for 5 epochs, which was sufficient for the pre-trained weights to adapt to the new task.

4 Results and Visualizations

The model was evaluated on the held-out test set of 325 images.

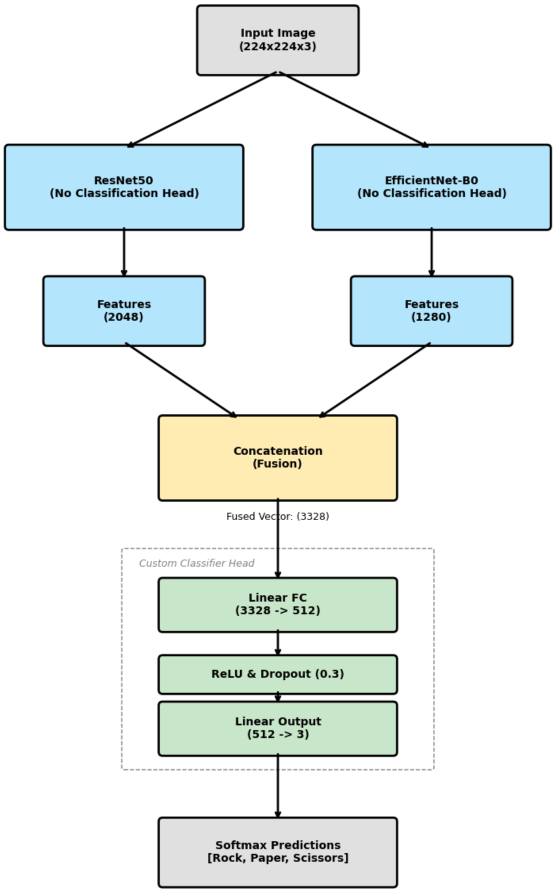
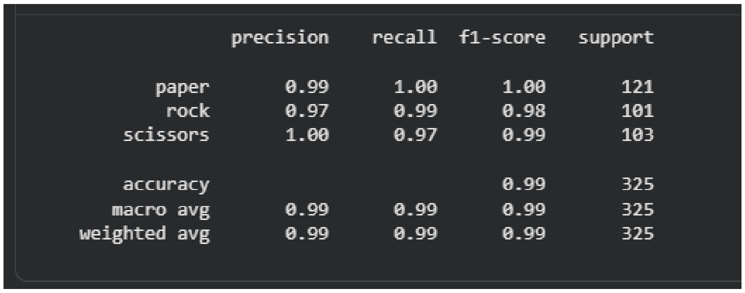


Fig. 1. Visual explanation of the Feature Fusion Architecture. Parallel backbones extract features which are concatenated before the final classification head.

4.1 Quantitative Analysis

The ensemble model achieved a final accuracy of **99%**. As detailed in Fig. 2, the weighted average F1-score was 0.99, indicating an exceptional balance between precision and recall.

- **Paper:** Achieved a Recall of 1.00, indicating the model correctly identified 100% of 'Paper' instances.
- **Scissors:** Achieved a Precision of 1.00, indicating zero false positives; every image predicted as 'Scissors' was correct.



```
              precision    recall  f1-score   support

       paper       0.99      1.00      1.00       121
        rock       0.97      0.99      0.98       101
    scissors       1.00      0.97      0.99       103

    accuracy                           0.99       325
   macro avg       0.99      0.99      0.99       325
weighted avg       0.99      0.99      0.99       325
```

Fig. 2. Classification Report showing Precision, Recall, and F1-Score.

4.2 Error Analysis

To investigate the remaining 1% error rate (4 misclassified instances), a Confusion Matrix was generated (see Fig. 3).

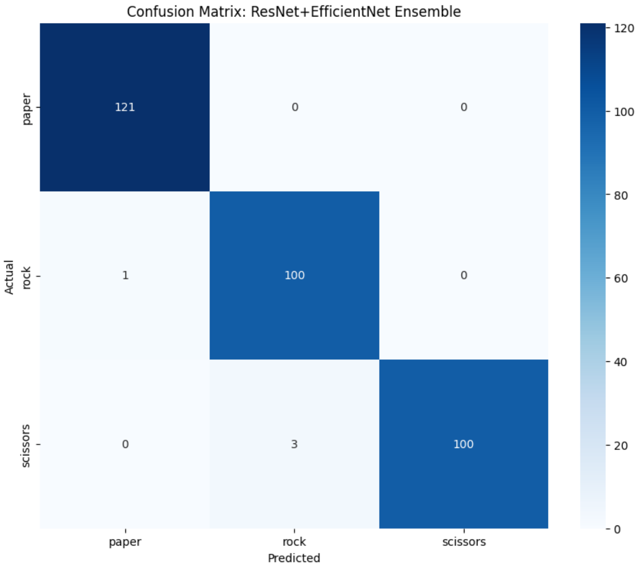


Fig. 3. Confusion Matrix revealing minor confusion between Rock and Scissors.

The matrix reveals that the primary confusion occurred between **Scissors** and **Rock** (3 instances where Scissors were predicted as Rock). Only one instance of Rock was misclassified as Paper.

4.3 Visual Verification

Qualitative assessment was performed by visualizing model predictions on random test batches. As shown in Fig. 4, the model demonstrated robustness against varying skin tones and hand orientations.

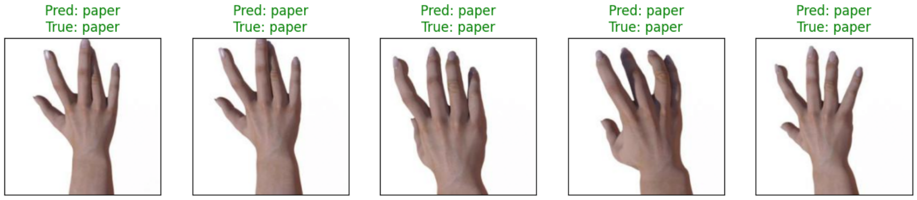


Fig. 4. Model predictions on test samples (Green indicates correct prediction).

5 Discussion

What the Fusion Contributed: The combination of ResNet and EfficientNet provided a "safety net" for feature extraction. While ResNet captured global shape information, EfficientNet focused on local textures. This allowed the model to maintain high confidence even when one network might have been uncertain.

What Worked vs. What Didn't: The fusion strategy worked exceptionally well for the 'Paper' class, which is visually distinct. However, the model struggled slightly with the 'Scissors' vs. 'Rock' distinction. This is attributed to geometric similarities: a closed fist (Rock) shares a similar convex hull to a fist with two fingers extended (Scissors), particularly under specific rotation angles or occlusion.

6 Conclusion

This study demonstrated that fusing ResNet50 and EfficientNet-B0 features yields a highly accurate classifier for gesture recognition. With a 99% accuracy rate and near-perfect precision on complex classes, the ensemble approach effectively mitigates the weaknesses of individual architectures. Future work will focus on targeted data augmentation for the 'Scissors' class to resolve geometric ambiguity with 'Rock'.

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. Tan, M., Le, Q.: Efficientnet: Rethinking model scaling for convolutional neural networks. In: International conference on machine learning. pp. 6105–6114. PMLR (2019)