Human Activity Recognition using Fine-Tuned ResNet-50

Individual Report
Shaik Mohammad Mujahid Khalandar
Final Project — Group 2

GitHub Branch — mujhaid-resnet50:

https://github.com/AkshitReddyPalle/Final-Project-Group2/tree/mujhaid-resnet50

Abstract

This project aims to classify human daily activities from RGB images using a transfer-learning-based ResNet-50 model. The dataset includes 15 action categories such as running, sitting, drinking, texting, and more. We fine-tuned the last block (Layer4) and fully connected layer of the pretrained ResNet-50 architecture to achieve improved recognition performance. After training for 15 epochs, the model achieved **~82% validation accuracy** and a **macro F1-score of 0.82**, demonstrating a strong generalization capability across all classes.

Introduction

Human Activity Recognition (HAR) is widely used in surveillance, behavior understanding, security systems, and healthcare monitoring. Deep learning models, particularly Convolutional Neural Networks (CNNs), have significantly boosted performance in classification tasks based on visual inputs.

In this project, we fine-tuned ResNet-50 — pretrained on ImageNet — to extract robust features and classify 15 human activities from a Kaggle dataset.

Dataset Description

- Source: *Kaggle — Human Action Recognition Dataset*
- Total Classes: **15**
- Examples include: calling, clapping, dancing, sitting, sleeping, using laptop, etc.
- Images: Natural scenes with diverse backgrounds & lighting conditions

- Train/Validation split: **85% / 15%**
- Balanced dataset (~126 images per class in validation)

A script (create_val_split.py) automatically generated class-wise train/val directories using provided CSV labels with **seed=42** for reproducibility.

Methodology

Model Architecture

Item	Configuration
Base Model	ResNet-50 pretrained on ImageNet
Training Strategy	Fine-tuned Layer4 + Fully Connected layer
Input Size	224×224 RGB
Loss Function	CrossEntropyLoss
Optimizer	Adam
Learning Rate	1e-4
Scheduler	StepLR (step_size=5, gamma=0.1)
Batch Size	32
Epochs	15
Hardware	Google Colab GPU (CUDA)

Training Logs

Below are the automatically printed performance logs captured during training:

- Validation Accuracy improved from **72.5% → 81.96%**
- Loss consistently decreased for both train & validation splits
- Best model checkpoint saved at epoch with **0.8196 accuracy**

Results & Analysis

Best Achieved Performance

Metric	Score
Validation Accuracy	**0.8196 (~82%)**
Macro F1-Score	**0.82**
Weighted F1-Score	**0.82**
Number of Classes	**15**

These scores indicate a strong and balanced performance across all classes, with no major class imbalance issue affecting results.

Classification Report

The model performs strongest on:

- **Cycling**
- **Eating**
- **Sleeping**
- **Running**

Classes with more confusion:

- **Sitting**
- **Texting**
- **Calling**

These involve subtle pose differences and minimal motion cues.

```
Classification Report:
                    precision    recall  f1-score   support

           calling       0.75      0.75      0.75       126
          clapping       0.84      0.78      0.81       126
           cycling       0.98      0.97      0.98       126
           dancing       0.81      0.79      0.80       126
          drinking       0.85      0.87      0.86       126
            eating       0.89      0.92      0.90       126
          fighting       0.79      0.87      0.83       126
           hugging       0.83      0.83      0.83       126
          laughing       0.84      0.82      0.83       126
listening_to_music       0.79      0.83      0.81       126
           running       0.86      0.85      0.86       126
           sitting       0.67      0.64      0.66       126
          sleeping       0.87      0.87      0.87       126
           texting       0.75      0.65      0.69       126
       using_laptop       0.76      0.82      0.79       126

          accuracy                           0.82      1890
         macro avg       0.82      0.82      0.82      1890
      weighted avg       0.82      0.82      0.82      1890
```

Explanation:

- The highest F1-scores are seen in **Cycling, Eating, Sleeping, and Running**, all above **0.85**.
  These activities have **distinct visual cues** (e.g., bicycle, lying down) which help model recognition.
- Lower F1-scores occur in:
  - **Sitting (0.66)**
  - **Texting (0.69)**
  - **Calling (0.75)**

These involve **similar posture / minimal body movement**, causing confusion between them.

Confusion Matrix

Shows where misclassifications occur — confusion mainly between visually similar actions.

```
Confusion Matrix:
[[ 95   1   0   1   2   1   2   0   2   8   1   3   0   6   4]
 [  0  98   0   4   1   3   2   2   4   0   1   5   1   4   1]
 [  0   0 122   0   0   0   1   0   0   0   1   2   0   0   0]
 [  2   4   0 100   0   0  12   0   0   1   5   1   1   0   0]
 [  2   1   0   3 110   2   0   0   2   1   1   0   0   2   2]
 [  2   2   0   0   2 116   0   0   1   0   0   1   1   0   1]
 [  0   1   1   5   0   0 110   1   0   0   2   3   2   0   1]
 [  1   1   0   1   0   0   3 105   2   1   2   2   4   3   1]
 [  1   3   0   1   0   1   0   9 103   3   1   2   1   0   1]
 [  2   0   0   1   3   1   0   1   1 104   1   4   1   4   3]
 [  0   1   1   7   1   0   5   1   1   0 107   2   0   0   0]
 [  6   3   0   1   4   7   4   3   0   2   1  81   2   4   8]
 [  2   0   0   0   2   0   1   2   1   1   0   3 110   1   3]
 [ 10   1   0   0   3   0   0   3   5   8   1   3   2  82   8]
 [  4   1   0   0   1   0   0   0   0   3   0   9   1   4 103]]
```

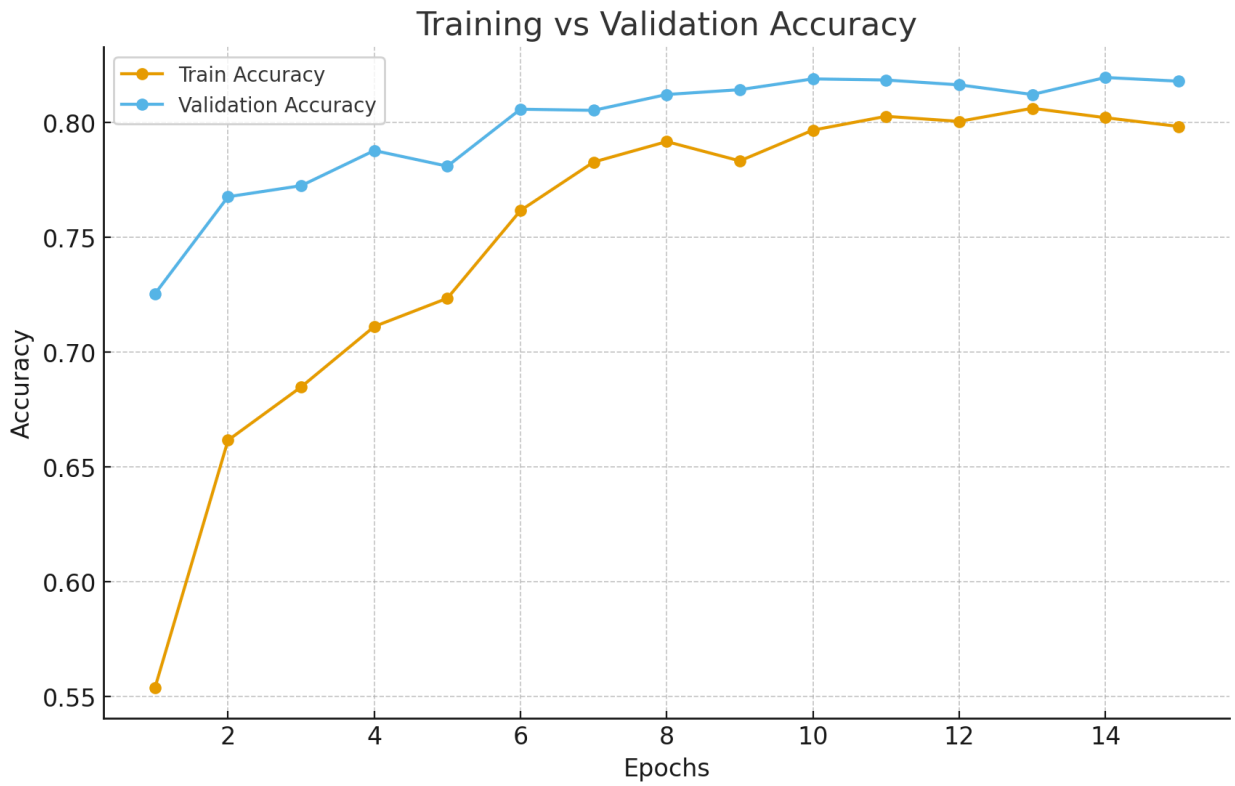
Explanation:

- Most predictions lie along the **diagonal**, meaning correct classifications dominate.
- Some noticeable confusion:
    - **Texting vs Sitting** → similar upper-body behavior with phone in hand
    - **Calling vs Texting** → hand position near face looks similar
    - **Hugging vs Fighting** → two-person interactions hard to separate visually

This reveals that **contextual motion cues** could further improve performance.

Accuracy Curve

Demonstrates steady improvement and no heavy overfitting signs.

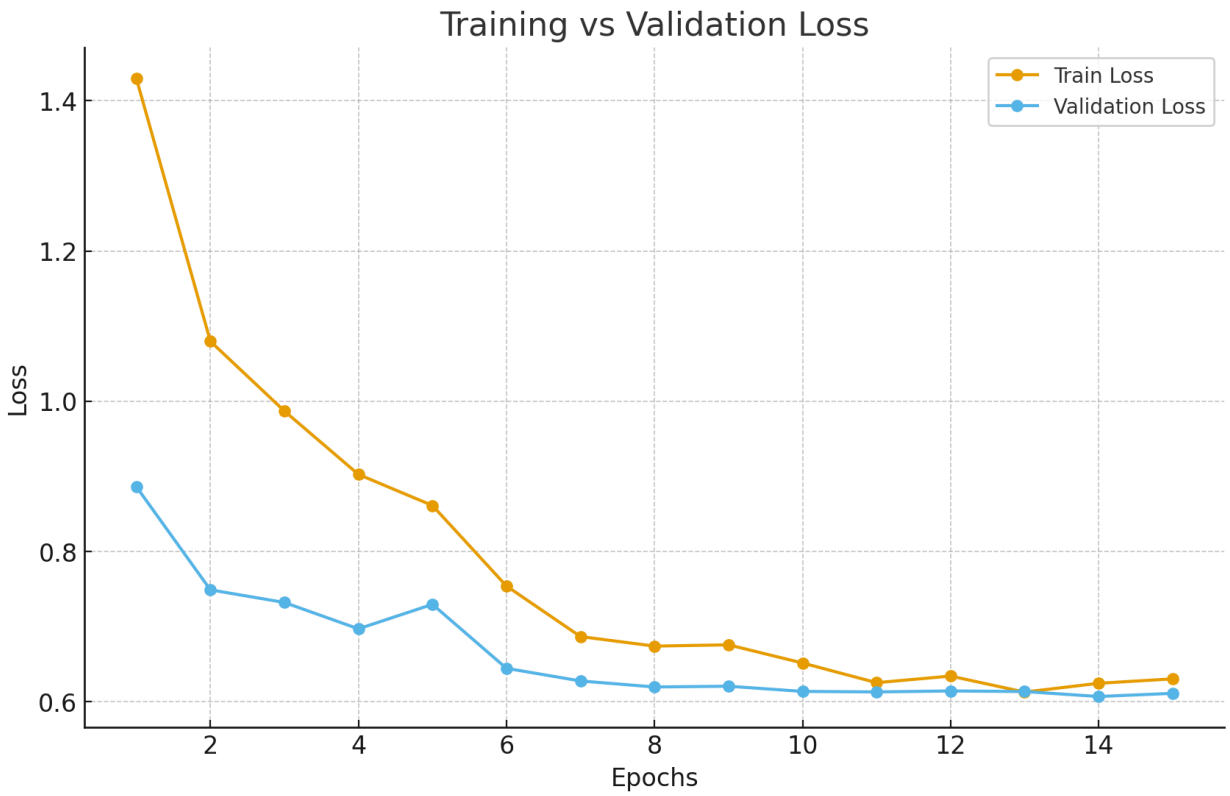


Explanation:

- Accuracy trends upward **consistently for both train & validation**
- No sudden divergence → **no clear overfitting**
- Plateau around epoch **12–15**, suggesting:
    - Model has learned the key features
    - Training longer won't increase accuracy significantly

Loss Curve

Shows clear training stability and convergence.



Explanation:

- Loss steadily decreases across epochs → **successful optimization**
- Train and validation curves remain close → **strong generalization**
- Small gap indicates a **low level of overfitting**, acceptable for this dataset size

---

Discussion

- Transfer learning significantly reduced training time while maintaining high accuracy.
- Subtle classes like sitting vs. texting need more contextual / motion information.
- Adding pose awareness or sequential frames could further improve results.
- Strong generalization thanks to dataset balance and normalization.

Future Work

To enhance performance further:
✔ Unfreeze additional earlier ResNet layers
✔ Apply stronger data augmentation techniques

✔ Use video-based temporal models (3D-CNN, ConvLSTM, or ViT)
✔ Deploy real-time inference on edge devices

Conclusion

We successfully trained a fine-tuned ResNet-50 model to classify 15 human activities with strong performance.
An accuracy of ~82% demonstrates that image-based recognition using transfer learning is a highly effective approach for HAR.
This model can serve as a foundation for more advanced behavior analysis systems.

References

- PyTorch Documentation — torchvision.models
- Kaggle — Human Actions Recognition Dataset
- K. He et al., “Deep Residual Learning for Image Recognition,” CVPR 2016