Video Action Recognition with UCF50 using LRCN

Stephen Kakuda

March 2025

1 Introduction

In this assignment, we train and evaluate the performance of a Long-term Recurrent Convolutional Network (LRCN) for video action recognition. LRCN is a deep learning architecture that combines a Convolutional Neural Network (CNN) with a Recurrent Neural Networks (RNN). The initial input to the LRCN is passed through a CNN to extract spatial features, the output of which is then passed through several recurrent layers that capture temporal relationships. This unique architecture allows the model to learn from longer sequences and remember important information over time.

The UCF50 dataset was used to train and test the model, which contains over 6,000 realistic videos sourced from YouTube across 50 different action categories. Within each category, videos are organized based on common features into 25 different groups, each consisting of at least 4 action clips. The dataset was divided using a 75-15-10 split for train, test, and validation sets. This split ensured that the model had enough data to learn all of the different action categories while maintaining a large enough test set. Each video file was converted into a sequence of separate image files by extracting 16 frames using uniform random sampling. A smaller batch size of 4 was used for both training and testing due to the significant size of the model.

The goal of this assignment was to demonstrate the LRCN's ability to accurately classify action within videos, and the majority of this report will focus on the performance metrics.

2 Training

For this particular assignment, the LRCN model used a ResNet model pretrained on ImageNet as the backbone for 2D feature extraction. Due to the large size of the problem set, the model was trained with 25 epochs to allow the model to sufficiently learn each video action. Adam was used for the optimizer, with an initial learning rate of 0.00003. Torch's ReduceLROnPlateau was used as the learning rate scheduler to further optimize model training by lowering the learning rate when model improvement stagnates during training. Figure 1 shows how the model's accuracy changed over each epoch.

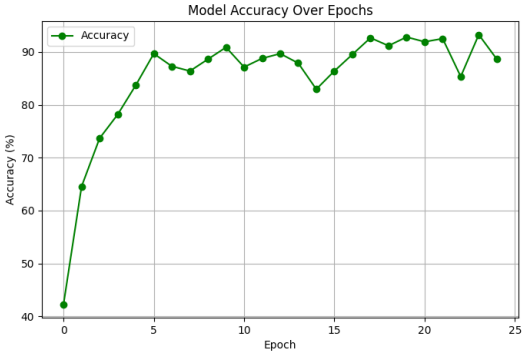


Figure 1: Accuracy vs Epochs

Looking at the validation accuracy over epochs, we can see that the model experienced significant improvement over the first 5 epochs before beginning to plateau. The model actually lowered in accuracy on the 23rd epoch and ended with a final accuracy of 88.64%. We can see a corresponding increase in the

loss, suggesting that the model updated the weights too aggressively at that particular step. The model would likely continue to improve if given more time to train but was limited due to time constraints of the assignment.

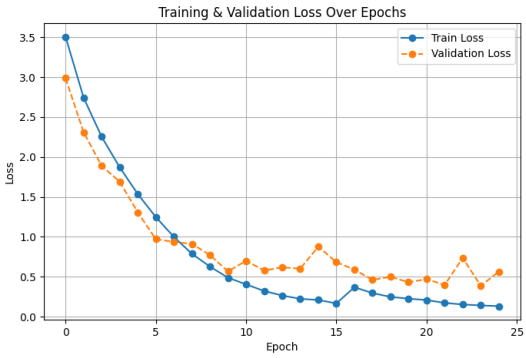


Figure 2: Loss vs Epochs

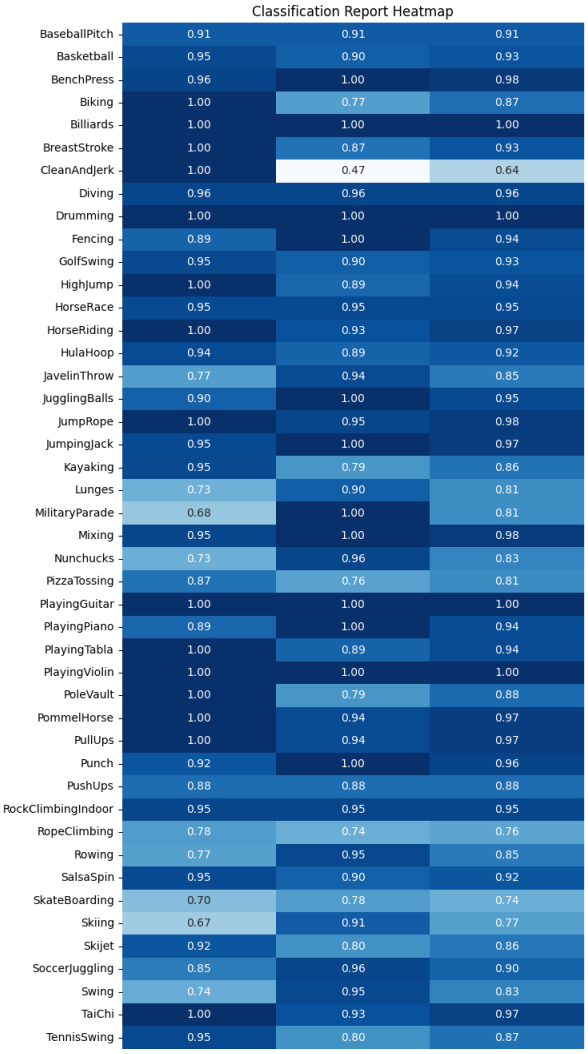
3 Results

We evaluated LRCN's performance on the video action recognition task across several standardized metrics. Looking at the percent of correctly predicted classifications using the test set, LRCN achieved an overall accuracy of 92.6221%. To gain a deeper understanding of the LRCN's performance, we also considered the precision, recall, F1-Score of the model. Precision measures how many predicted positives were actually correct, where a higher score indicates fewer false positives. Recall measures how many actual positive detections were correct. A higher recall score means that the model had fewer missed detections. F1-score is a combination of precision and recall, and indicates the balance between the two. The overall values were calculated as the weighted average across all 50 different classes.

Precision	**Recall**	**F1-Score**
50.5%	68.2%	58.5%

The model's ROC AUC score represents the area under the curve of the Receiver Operating Characteristic curve, which plots the true positive rate against the false positive rate at different classification thresholds. Our LRCN had a test ROC AUC score of 0.9987, indicating extremely high classification performance. Overall, the model achieved a high degree of accuracy across the 50 different classes.

To evaluate the per-class performance of the model, the heatmap below visualizes the precision, recall, and F1-Score for each class using a heatmap. The darker areas of the heatmap indicate a better score and the lighter areas indicate a weaker score.



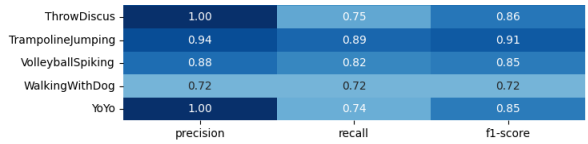


Figure 3: Confusion Matrix

Noticeably, the model struggled the most in correctly classifying the "Clean and Jerk" action, with a high precision of 1 but a significantly low recall of 0.47. This indicates that the model had few false positives when it did classify a "Clean and Jerk" but had many false negatives. There are several factors that could have led the model to struggle with this particular action. The "Clean and Jerk" involves several specific motions combined in a precise sequence that may have been difficult for the model to learn over the 16 frames. It is also possible that the dataset was split in a way that the training set did not contain a sufficient proportion of the "Clean and Jerk" action.

4 Conclusion

Overall, LRCN demonstrated remarkable success in classifying video actions with the UCF50 dataset. With further fine-tuning of the hyperparameters and more training time, the model has the potential to achieve even greater performance. Future work would also include comparing performance across other video datasets, including the larger UCF101 dataset.