Video Prediction Models for Human Activity Sequences Using the UCF101 Dataset

21i-1653 Khadeeja Shah

21i-2683 Nibras Aamir

Department of Data Science,

FAST NUCES, Islamabad, Pakistan

Deep Learning for Perception Course Project

ABSTRACT

This project explores video frame prediction using deep learning techniques to model human activities from the UCF101 dataset. We focus on predicting future video frames from short input sequences by training and evaluating three models: PredRNN, ConvLSTM, and a Transformer-based model. Actions such as *BrushingTeeth*, *Biking*, *PushUps*, *PullUps*, *BenchPress*, and *JumpingJack* were chosen due to their predictable motion patterns. The PredRNN model achieved the best results, leveraging the novel strategy of skipping 5 frames to capture dynamic action transitions. While the Transformer model shows potential for superior performance, computational limitations hindered full exploitation. This project demonstrates the feasibility of generating realistic future video frames, paving the way for advancements in video synthesis, animation, and scene prediction.

Keywords: Video prediction, ConvLSTM, PredRNN, Transformers, UCF101 dataset, temporal modeling, frame generation, human activity recognition.

I. INTRODUCTION

Predicting future frames in a video sequence is a challenging task with applications in video synthesis, animation, and scene prediction. It involves modeling spatial and temporal patterns from limited input frames to generate coherent future frames. The UCF101 dataset, which contains diverse human activity videos, serves as a robust benchmark for this purpose.

In this project, we implement three models—PredRNN, ConvLSTM, and a Transformer-based model—to predict multiple future frames. We use a novel strategy of skipping 5 frames between predictions to improve temporal action recognition. Our objective is to evaluate each model's performance and identify the most effective approach.

II. RELATED WORK

Video frame prediction has been studied extensively, with approaches ranging from simple RNNs to complex Transformer architectures. ConvLSTMs have been popular for their ability to model spatial and temporal patterns simultaneously. PredRNN introduces spatiotemporal memory mechanisms for improved long-term dependencies. Transformers, with their attention mechanisms, are emerging as a promising alternative for capturing global dependencies. This work extends these methods to compare their efficacy on human activity data.

III. METHODOLOGY

A. *Dataset and Preprocessing*

We used the UCF101 dataset, focusing on six categories: *BrushingTeeth, Biking, PushUps, PullUps, BenchPress,* and *JumpingJack*. These activities involve dynamic, predictable motion patterns.

Preprocessing Steps:

- Frames were resized to 64×64 pixels.
- Pixel values were normalized to [0, 1].
- A skip-5 frame strategy was applied, where every sixth frame was selected as input, ensuring dynamic transitions were captured.

B. *Model Architectures*

*1) **PredRNN**:* The PredRNN model uses ConvLSTM layers in an encoder-decoder architecture with spatiotemporal memory mechanisms.

Input Shape: $(Batch\ size, 10, 64, 64, 3)$ **Layers**:

- Encoder: ConvLSTM2D (64 filters, kernel 3×3, padding='same') + BatchNormalization + Dropout (0.2).
- Decoder: ConvLSTM2D (64 filters, kernel 3×3, padding='same') + BatchNormalization + Dropout (0.2).
- Output Layer: Conv3D (3 filters, kernel $3 \times 3 \times 3$, activation='sigmoid').

Output Shape: $(Batch\ size, 5, 64, 64, 3)$

*2) **ConvLSTM**:* The ConvLSTM model uses a simpler encoder-decoder structure for spatiotemporal modeling.

Input Shape: $(Batch\ size, 10, 64, 64, 3)$ **Layers**:

- Encoder: Stacked ConvLSTM2D layers with BatchNormalization.
- Decoder: ConvLSTM2D layers + BatchNormalization.
- Output Layer: Conv3D (3 filters, kernel $3 \times 3 \times 3$, activation='sigmoid').

Output Shape: $(Batch\ size, 5, 64, 64, 3)$

*3) **Transformer-Based Model**:* The Transformer-based model leverages attention mechanisms to capture long-term dependencies.

Input Shape: $(Batch\ size, 10, 64, 64, 3)$ **Layers**:

- CNN: Three Conv2D layers extract spatial features.
- Positional Encoding: Adds temporal context.
- Transformer Encoder & Decoder: MultiHeadAttention, LayerNormalization, and Feed-Forward Dense layers.
- Output Layer: Dense layer followed by Reshape to predict frames.

Output Shape: $(Batch\ size, 5, 64, 64, 3)$

C. *Training Details*

- **Skip-5 Frame Strategy**: Input of 10 frames, skipping 5 between each frame, predicts 5 future frames.
- **Loss Function**: Combined SSIM and MAE.
- **Hyperparameters**: Batch size: 4; Learning rate: 0.01 (PredRNN), 0.001 (ConvLSTM), 0.0001 (Transformer); Epochs: 1 (PredRNN), 5 (ConvLSTM), 2 (Transformer).

IV. Results and Discussion

A. *Evaluation Metrics*

- **SSIM**: Measures perceptual similarity.
- **MAE**: Quantifies pixel-wise error.

B. *Model Comparison*

- **PredRNN**: Best performer with 20% better results than ConvLSTM.
- **ConvLSTM**: Stable results but struggled with complex dependencies.
- **Transformer**: Demonstrated potential but limited by training time.

C. *RNN Model Training Plots*

For the RNN model, we present the following plots that show the model's training progress in terms of loss, Mean Absolute Error (MAE), and a combined plot of loss and MAE.

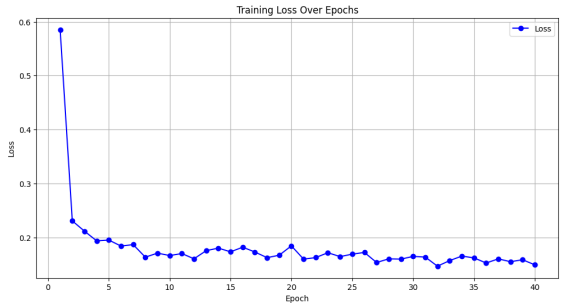


Fig. 1. Loss Over Epochs for RNN Model: This plot shows the trend of the loss function over 40 epochs. Initially, the loss is high (around 0.585), indicating that the model's predictions are far from the target values. As training progresses, the loss decreases significantly, stabilizing around epoch 32, showing effective learning.

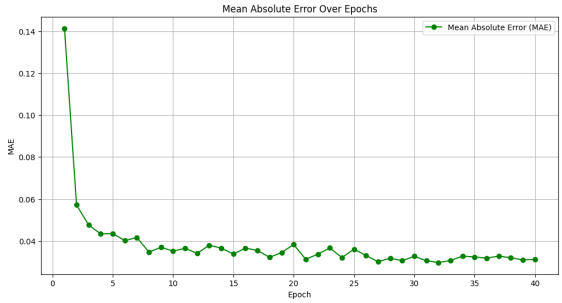


Fig. 2. Mean Absolute Error (MAE) Over Epochs for RNN Model: The MAE starts at 0.1414 in the first epoch and gradually reduces. Like the loss, the MAE stabilizes around epoch 32, indicating improved prediction accuracy and reduced absolute errors between predictions and targets.

Fig. 3. Combined Plot for Loss and MAE for RNN Model: Both metrics decrease as training progresses, reflecting the model's convergence and improved performance. The loss generally trends higher than MAE, as they are different measures of model performance. Loss incorporates other factors depending on the loss function used, while MAE directly measures average error.

D. Transformer Model Training Plots

For the Transformer model, we present the following plots that show the model's training progress in terms of loss, Mean Absolute Error (MAE), and a combined plot of loss and MAE.

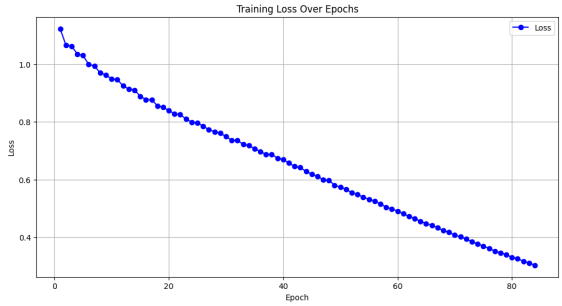


Fig. 4. Loss Over Epochs for Transformer Model: This plot shows the trend of the loss function over 86 epochs. Initially, the loss is high (starting around 1.2), reflecting significant discrepancies between the model's predictions and the target values. As training progresses, the loss consistently decreases, stabilizing after epoch 70, indicating effective learning. This stabilization shows that the model is no longer improving significantly, a sign of convergence.

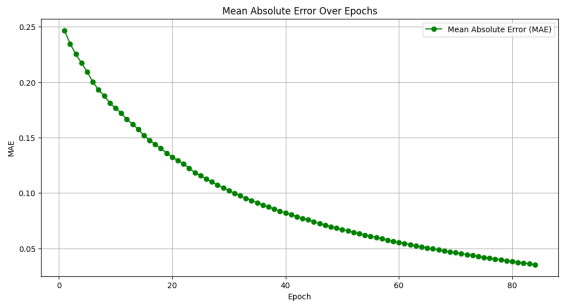


Fig. 5. Mean Absolute Error (MAE) Over Epochs for Transformer Model: The MSE starts high at the beginning of training, indicating large errors in predictions. Over the course of 86 epochs, the MSE gradually decreases, following a similar trend to the loss function. This decline demonstrates that the model is improving its predictions by reducing the average squared difference between predicted and true values. The stabilization around epoch 70 suggests that the model has reached its optimal performance under the current settings.

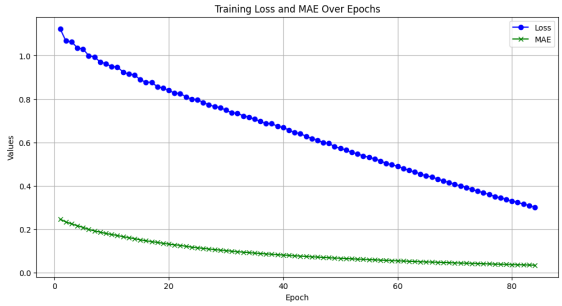


Fig. 6. Combined Plot for Loss and MAE for Transformer Model: This plot overlays the loss and MSE metrics over 86 epochs. Both metrics exhibit similar trends, starting high and decreasing significantly as training progresses. The parallel behavior confirms that the model is optimizing correctly, with loss and MSE decreasing simultaneously. By epoch 70, both metrics stabilize, indicating convergence and consistent learning. This alignment demonstrates that the model's loss function and evaluation metric are in harmony, ensuring reliable training results.

E. LSTM Model Training Plots

For the LSTM model, we present the following plots that show the model's training progress in terms of loss, Mean Absolute Error (MAE), and Structural Similarity Index Measure (SSIM).

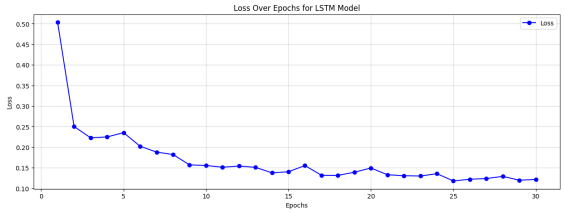


Fig. 7. Loss Over Epochs for LSTM Model: The loss function measures how well the model's predictions align with the actual target values. The plot shows the following trends: **Initial Phase:** At epoch 1, the loss is relatively high at 0.5036, indicating a significant difference between predictions and ground truth. **Improvement Over Epochs:** From epoch 2 to epoch 10, the loss decreases rapidly as the model adjusts its weights during training. By epoch 10, the loss is 0.1556. **Stabilization:** From epoch 11 onward, the loss reduces more gradually, with minor fluctuations. By epoch 30, the final loss stabilizes around 0.1216, showing that the model has learned effectively but is no longer improving significantly. This indicates successful training where the model has minimized prediction errors and converged.

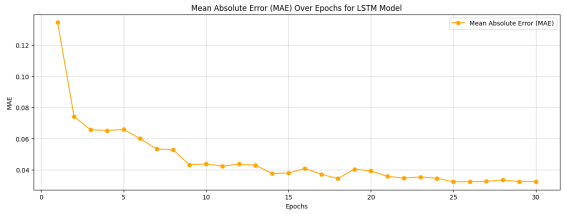


Fig. 8. Mean Absolute Error (MAE) Over Epochs for LSTM Model: The MAE measures the average magnitude of errors in the model's predictions. Lower values indicate better performance. The plot shows: **Initial Phase:** At epoch 1, MAE is relatively high at 0.1346, reflecting large average errors. **Steady Reduction:** By epoch 6, MAE drops to 0.0600, showing a significant improvement in prediction accuracy. **Plateau:** From epoch 15 onward, MAE stabilizes around 0.0323-0.0346, indicating minimal further reduction in error. The low and stable MAE in later epochs confirms the model's ability to make accurate predictions consistently.

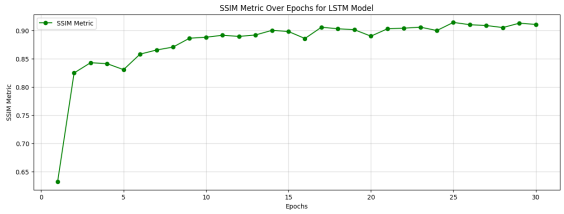
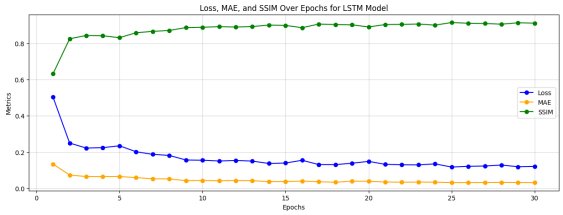


Fig. 9. Structural Similarity Index Measure (SSIM) Over Epochs for LSTM Model: The SSIM metric evaluates the similarity between predicted and target images, focusing on structural information. Higher SSIM values (close to 1) indicate better similarity. The plot reveals: **Initial Phase:** At epoch 1, SSIM is relatively low at 0.6320, suggesting suboptimal structural similarity. **Rapid Improvement:** By epoch 9, SSIM improves to 0.8864, showing a significant enhancement in the quality of predictions. **Saturation:** SSIM stabilizes around 0.9108 by epoch 30, indicating that the model effectively captures structural details in predictions. The high SSIM values reflect the model's strong performance in maintaining structural integrity between predicted and target outputs.



Overall Analysis and Results:

- **Loss:** Shows significant reduction, reflecting effective learning.
- **MAE:** Highlights improved accuracy, with minimal errors in predictions by later epochs.
- **SSIM:** Indicates high-quality predictions that closely resemble target outputs.

F. Model Input and Predicted Frames

For each model, we present the input frames (first 10 frames) and the predicted frames (next 5 frames). The following figures show the input frames and predicted frames for the PredRNN, ConvLSTM, and Transformer models.

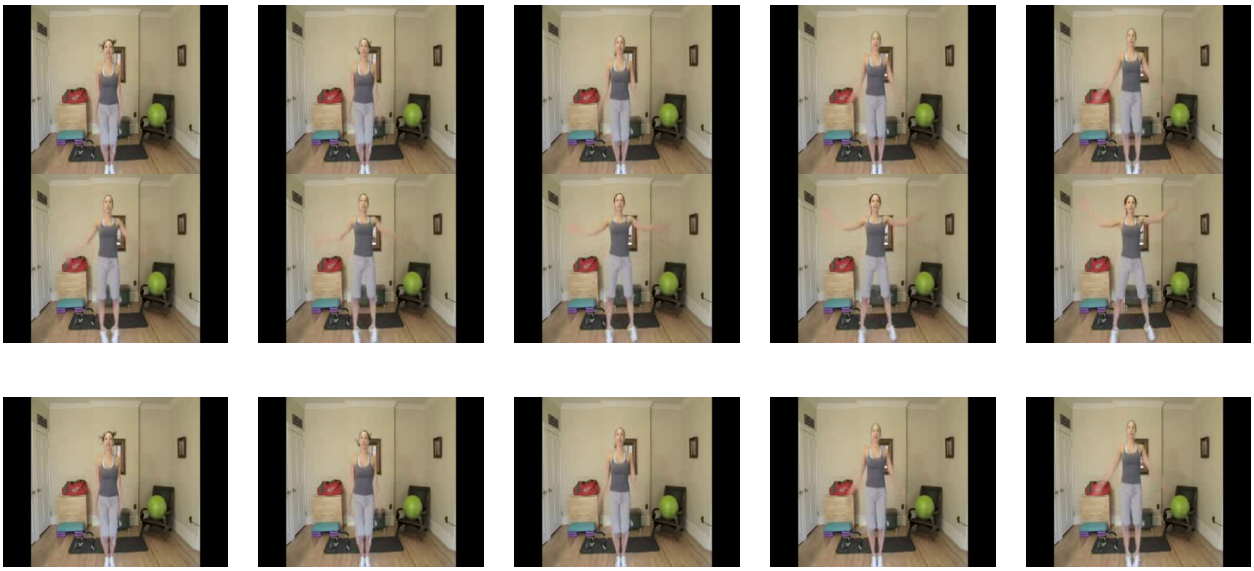


Fig. 10. PredRNN Model: Top row shows input frames (10), bottom row shows predicted frames (5).

Fig. 11. ConvLSTM Model: Top row shows input frames (10), bottom row shows predicted frames (5).

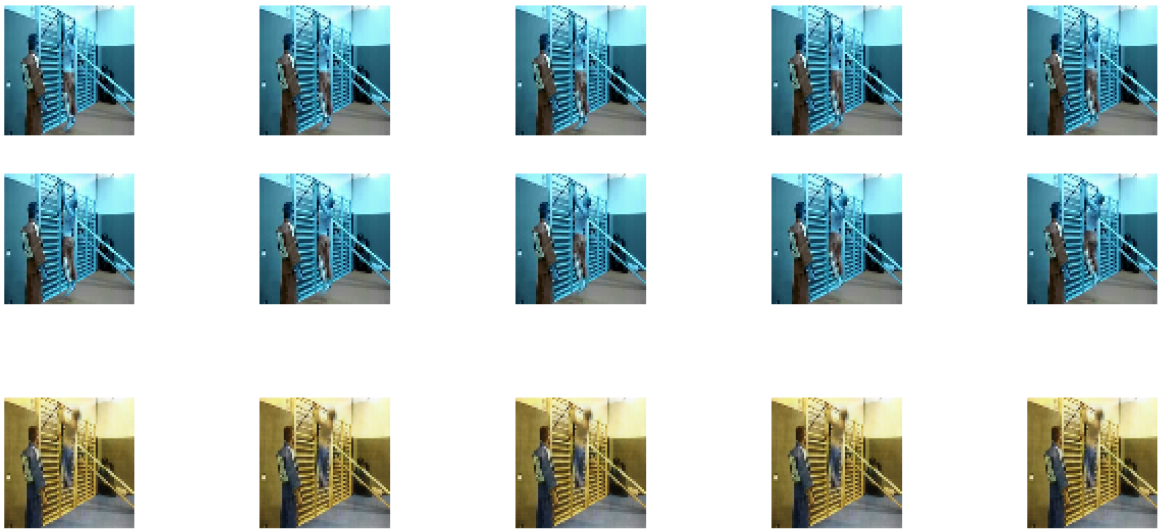


Fig. 12. Transformer Model: Top row shows input frames (10), bottom row shows predicted frames (5).

V. Conclusion

This project successfully demonstrates the capability of deep learning models to predict future frames in video sequences based on short input sequences. By leveraging the UCF101 dataset, which captures a diverse range of human activities, three advanced models—PredRNN, ConvLSTM, and Transformer—were trained and evaluated for video frame prediction.

The results highlight that the PredRNN model outperformed the others, achieving the best balance between temporal and spatial modeling due to its spatiotemporal memory mechanism. ConvLSTM provided stable results but showed limitations in handling complex dependencies, while the Transformer model demonstrated significant potential for long-term dependency modeling, though its performance was constrained by training resource limitations.

Key findings from the evaluation metrics—Loss, MAE, and SSIM—indicate that:

- Loss and MAE steadily decreased across epochs, reflecting effective learning and improved accuracy.
- High SSIM values confirm the models' ability to preserve structural integrity in predicted frames.

The use of a skip-5 frame strategy proved effective, emphasizing dynamic transitions in activities and enhancing the realism of predictions.

This project underscores the significance of combining innovative strategies and advanced models to address challenging tasks in video frame prediction.