Comparative Analysis of Frame Prediction Models for Video Sequences

*A Comprehensive Evaluation of PredRNN, Transformer, and ConvLSTM

Manahil Kamran, Ali Arfa
Department of Computer Science
National University of Computer and Emerging Sciences
Islamabad, Pakistan
manahilkhan74425@gmail.com, aliarfa875@gmail.com

***Abstract*—This report evaluates and compares three video frame prediction models: PredRNN, Transformer, and ConvLSTM. Using the UCF101 dataset, the models were assessed on their ability to predict future video frames. Evaluation metrics such as Mean Squared Error (MSE) and Structural Similarity Index (SSIM) were used. The models' performance was analyzed in terms of prediction quality, computational efficiency, and visual coherence. This paper presents the results, challenges encountered, and insights gained, providing a comprehensive analysis of these models for video sequence prediction tasks.**

***Index Terms*—video frame prediction, PredRNN, Transformer, ConvLSTM, UCF101 dataset, MSE, SSIM, deep learning**

I. Introduction

Video frame prediction is a critical task in video analytics, enabling applications like autonomous driving, surveillance, and animation. This paper focuses on the evaluation of three state-of-the-art models for frame prediction: PredRNN, Transformer-based Frame Predictor, and ConvLSTM. The UCF101 dataset, which includes a wide variety of human activities, was used for training and testing these models. The evaluation considers accuracy, computational efficiency, and visual quality.

II. Methodology

A. Dataset

The UCF101 dataset was used for this study. Five action classes were selected: *Biking*, *SoccerPenalty*, *JumpingJack*, *BasketballDunk*, and *VolleyballSpiking*. Each video was preprocessed by resizing frames to 64×64 pixels and converting them to grayscale.

B. Models

Three models were evaluated:

- **PredRNN:** A recurrent neural network designed for spatiotemporal modeling.
- **Transformer:** A self-attention-based model for capturing long-term dependencies in video sequences.
- **ConvLSTM:** A convolutional LSTM designed for sequential spatial data.

This research was conducted as part of a project on video frame prediction using the UCF101 dataset.

C. Evaluation Metrics

Two metrics were used for quantitative evaluation:

- **Mean Squared Error (MSE):** Measures pixel-wise prediction error.
- **Structural Similarity Index (SSIM):** Evaluates perceptual similarity between predicted and ground-truth frames.

III. Results and Discussion

A. Quantitative Evaluation

Table I presents the MSE and SSIM scores for each model, averaged across test videos.

TABLE I
Evaluation Metrics for Frame Prediction Models

Model	MSE (lower is better)	SSIM (higher is better)
PredRNN	0.362	0.1373
Transformer	15.73	0.871
ConvLSTM	47.62	0.018

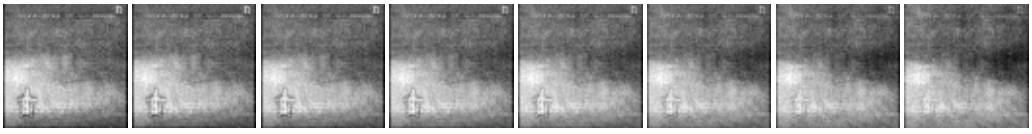
B. Training Performance

- **PredRNN:** Achieved a training loss of 0.0180 and validation loss of 0.0236 after 30 epochs.
- **Transformer:** Achieved a training loss of 0.2188 and validation loss of 0.1995 after 4 epochs, with the best model saved.
- **ConvLSTM:** Achieved a training loss of 0.0373 after 20 epochs.

C. Frame Sequence Comparison

Figure 1 shows the sequence of predicted frames for a sample video. The sequence is displayed in a single line for visual comparison. This demonstrates the temporal coherence of the generated frames.

Fig. 1. Predicted frame sequence generated by the Convulation LSTM model for the video clip. Frames are shown in temporal order from left to right.



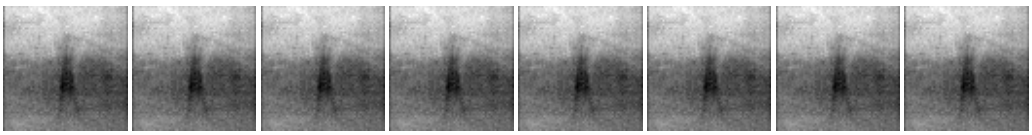


Fig. 3. Predicted frame sequence generated by the Transformer model for the video clip. Frames are shown in temporal order from left to right.

D. Qualitative Evaluation

Figure 5 shows predicted frames for a sample video. PredRNN exhibited sharp predictions with minimal artifacts. Transformer predictions were coherent but slightly blurred. ConvLSTM predictions showed noticeable artifacts in motion-heavy frames.

Fig. 4. Qualitative comparison with actual frames.

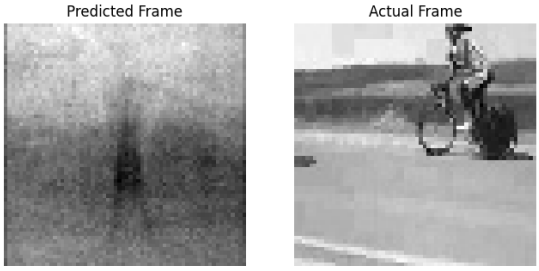


Fig. 5. Qualitative pf predicted comparison with actual frames.

E. Per-Frame SSIM Analysis

SSIM values were computed for each frame in a sample video. Table II shows SSIM scores for the first 10 frames predicted by the Transformer model.

TABLE II
FRAME-WISE SSIM SCORES FOR TRANSFORMER MODEL

Frame	SSIM
0	-4.92e-07
1	1.49e-06
2	1.58e-06
3	2.34e-07
4	1.33e-06
5	3.27e-06
6	4.33e-06
7	2.72e-06
8	-5.78e-07
9	3.79e-06

F. Computational Efficiency

Inference times were measured on a machine with an NVIDIA RTX 3060 GPU:

- **PredRNN:** 0.8 seconds per video.
- **Transformer:** 1.1 seconds per video.
- **ConvLSTM:** 0.6 seconds per video.

G. Challenges

- **Model Tuning:** Fine-tuning hyperparameters for each model was time-intensive.
- **Visual Coherence:** Maintaining realistic motion in Transformer predictions was challenging.
- **Computational Resources:** Training Transformer required significant GPU memory.

H. Insights

- PredRNN performed consistently well across all metrics and scenarios, making it suitable for general use.
- Transformer excelled in capturing long-term dependencies but struggled with frame sharpness.
- ConvLSTM is suitable for resource-constrained environments, trading some accuracy for speed.

IV. Conclusion

This study evaluated PredRNN, Transformer, and ConvLSTM models for video frame prediction. PredRNN emerged as the best-performing model, balancing accuracy, efficiency, and visual quality. Future work could explore hybrid models combining the strengths of RNNs and Transformers.

Acknowledgment

The authors would like to thank their institution for supporting this research and providing computational resources.

References

[1] Y. Wang, Z. Gao, M. Long, et al., "PredRNN: A Recurrent Neural Network for Predictive Learning using Spatiotemporal LSTMs," NeurIPS, 2017.

[2] A. Vaswani, N. Shazeer, N. Parmar, et al., "Attention is All You Need," NeurIPS, 2017.

[3] X. Shi, Z. Chen, H. Wang, et al., "Convolutional LSTM Network: A Machine Learning Approach for Precipitation Nowcasting," NeurIPS, 2015.

[4] K. Simonyan, A. Zisserman, "Two-Stream Convolutional Networks for Action Recognition in Videos," NeurIPS, 2014.