Deep Learning Using ConvLSTM, PredRNN, Transformer based Model

Deep Learning Project

Muti Ur Rehman
Bachelor's (Computer Science)
FAST NUCES
Islamabad, Pakistan
i210872@nu.edu.pk

AbuBakar
Bachelor's (Computer Science)
FAST NUCES
Islamabad, Pakistan
i211379@nu.edu.pk

Osama Ali
Bachelor's (Computer Science)
FAST NUCES
Islamabad, Pakistan
i210587@nu.edu.pk

***Abstract*—This documentation outlines the implementation and evaluation of three video prediction models: ConvLSTM, PredRNN, and a Transformer-based model. These models are designed to predict future video frames based on input sequences, enabling applications in areas such as surveillance, robotics, and video compression. The document provides a detailed description of the data preprocessing pipeline, training methodologies, and evaluation metrics used to compare the models. Key insights into the strengths and limitations of each model are discussed, alongside implementation details and practical guidelines for replication and further development.**

I. Introduction

The advancement of deep learning has enabled significant progress in video prediction, where models can predict future frames based on past observations. This capability has applications in various fields, including video synthesis, anomaly detection, robotics, and autonomous systems. In this project, we explore three state-of-the-art approaches to video prediction: ConvLSTM, PredRNN, and a Transformer-based model.

The primary objective is to develop and compare these models to predict multiple consecutive video frames, creating a smooth and coherent sequence of actions. Using the UCF101 dataset, which comprises videos of diverse human activities, we preprocess the data and train these models to generate future frames. The document also delves into the challenges faced during implementation, the metrics used to evaluate performance, and the insights gained from comparative analysis of the models. This work aims to contribute to the field by providing a comprehensive understanding of these approaches and their practical applications.

II. Conducting a Comparative Analysis of Video Prediction Models

In this section, we present a detailed comparative analysis of the three video prediction models: ConvLSTM, PredRNN, and the Transformer-based model. This analysis evaluates the models based on prediction quality, computational efficiency, and visual coherence of the generated frames. Additionally, we document the results, challenges encountered, and key insights for each model to provide a comprehensive understanding of their performance.

A. Prediction Quality

Prediction quality is a crucial factor in video prediction models, determining the accuracy of the generated future frames. The evaluation includes metrics such as Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index (SSIM), and Mean Squared Error (MSE).

ConvLSTM: ConvLSTM performs well on short-term frame prediction due to its ability to capture spatiotemporal dependencies. However, its performance tends to degrade when tasked with predicting longer sequences. PredRNN: PredRNN improves upon ConvLSTM by introducing a memory flow mechanism that enhances its ability to model long-term dependencies, resulting in higher prediction quality for extended sequences. Transformer-based Model: The Transformer-based model exhibits exceptional performance for complex scenes due to its ability to capture global dependencies. However, its performance can vary for highly dynamic scenes with fine-grained motion details.

B. Computational Efficiency

Computational efficiency directly impacts the practical applicability of the models, particularly for real-time scenarios.

ConvLSTM: ConvLSTM is computationally efficient for small to medium-sized datasets, making it a suitable choice for resource-constrained environments. PredRNN: PredRNN, while more computationally intensive than ConvLSTM, offers a favorable trade-off between efficiency and improved prediction quality. Transformer-based Model: The Transformer-based model demands significant computational resources due to its attention mechanism, making it less efficient for large datasets or resource-constrained systems.

C. Visual Coherence of Generated Frames

Visual coherence assesses the ability of the models to generate smooth and realistic frame transitions.

ConvLSTM: ConvLSTM produces coherent frames for simple motion patterns but struggles with complex, non-linear

movements. PredRNN: PredRNN excels in maintaining visual coherence, even in scenarios involving intricate motion dynamics. Transformer-based Model: The Transformer-based model generates visually appealing frames, capturing intricate spatial details. However, its performance might be inconsistent when processing high-frequency motion.

D. Challenges and Insights

Challenges:

ConvLSTM: Limited scalability for complex video sequences. PredRNN: Increased computational overhead due to memory mechanisms. Transformer-based Model: High resource consumption and sensitivity to hyperparameter tuning. Insights:

ConvLSTM serves as an efficient baseline for straightforward tasks. PredRNN provides a balanced approach, excelling in long-term prediction tasks. Transformer-based models are promising for applications requiring high precision and complex motion analysis.

E. Documented Results

The analysis concludes with a comprehensive report summarizing the evaluation metrics, performance graphs, and qualitative comparisons through visualization of predicted frames. The documented results highlight the strengths and limitations of each model, offering insights for future research and practical deployment in video prediction tasks.

F. Evaluation Framework

To ensure a fair and consistent comparison, the evaluation framework includes the following:

Datasets:

UCF101 dataset is used for testing the models due to its diverse range of actions and dynamic scenes. Data is split into training, validation, and test sets to avoid overfitting. Metrics:

Quantitative Metrics: Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index (SSIM), and Mean Squared Error (MSE). Qualitative Metrics: Visual inspection for smoothness, clarity, and realism of the predicted frames. Environment:

Experiments are conducted on hardware with consistent computational resources (e.g., GPU and memory specifications). All models are trained and tested with the same hyperparameters wherever applicable.

G. Transformer-based Model (Continued):

The Transformer-based model demonstrates exceptional visual coherence for scenes with high spatial complexity, as it effectively captures long-range dependencies. However, for scenarios involving rapid or erratic motion, its predictions may exhibit temporal discontinuities, potentially requiring additional fine-tuning.

III. Experimental Setup

To ensure a fair and systematic evaluation, the experiments were conducted under the following setup:

A. Dataset

We utilized the UCF101 dataset, which contains videos of diverse human activities, as the primary dataset for training and evaluation. The dataset was split into 80% training, 10% validation, and 10% testing sets. Each video was resized to a resolution of 128×128 pixels, and frames were normalized to a range of [0, 1].

B. Training Details

All models were implemented using PyTorch and trained on an NVIDIA Tesla V100 GPU. The hyperparameters were tuned using grid search, with the following configurations: - **ConvLSTM**: Learning rate = 0.001, Batch size = 16, Optimizer = Adam - **PredRNN**: Learning rate = 0.0005, Batch size = 16, Optimizer = Adam - **Transformer-based Model**: Learning rate = 0.0001, Batch size = 8, Optimizer = AdamW

Each model was trained for 50 epochs, with early stopping based on validation loss to prevent overfitting.

C. Evaluation Metrics

The following metrics were used to evaluate model performance: - **Peak Signal-to-Noise Ratio (PSNR):** Measures the quality of the reconstructed frames compared to the ground truth. Higher values indicate better quality. - **Structural Similarity Index (SSIM):** Assesses the structural similarity between predicted and ground truth frames. - **Mean Squared Error (MSE):** Quantifies the pixel-wise difference between predicted and actual frames. - **Inference Time:** Measures the average time taken to predict a single frame.

IV. Results and Discussion

This section summarizes the results obtained from the experiments and provides a detailed discussion of the findings.

A. Quantitative Results

The table below presents the performance of each model on the test set:

TABLE I
Performance Comparison of Models

Model	PSNR (dB)	SSIM	MSE	Inference Time (ms)
ConvLSTM	27.5	0.78	0.012	32
PredRNN	29.2	0.84	0.009	45
Transformer-based Model	**30.8**	**0.89**	**0.007**	72

B. Qualitative Results

Qualitative analysis revealed that ConvLSTM generated accurate predictions for videos with simple motion patterns but introduced blurriness for complex sequences. PredRNN offered significant improvements in clarity and motion continuity, particularly for videos with long-term dependencies. The Transformer-based model provided the most realistic and visually coherent frames, albeit at the cost of higher computational demand.

C. Challenges

During the implementation, several challenges were encountered: - **Overfitting:** Especially in ConvLSTM, which required extensive regularization techniques. - **Computational Bottlenecks:** Training the Transformer-based model demanded substantial memory and computation time, necessitating optimization strategies such as mixed precision training. - **Motion Complexity:** Highly dynamic scenes posed challenges for all models, indicating room for improvement in temporal modeling.

D. Key Insights

- PredRNN strikes a balance between quality and efficiency, making it a practical choice for many applications. - The Transformer-based model is ideal for scenarios requiring high-quality predictions, provided computational resources are sufficient. - Future research could focus on hybrid models that combine the strengths of these approaches to address their respective limitations.

V. Conclusion and Future Work

This work conducted a comprehensive evaluation of ConvLSTM, PredRNN, and Transformer-based models for video prediction tasks. The Transformer-based model achieved the best performance in terms of prediction quality and visual coherence, while PredRNN offered a good trade-off between accuracy and computational efficiency. ConvLSTM, though computationally efficient, struggled with complex scenarios.

Future work will explore hybrid architectures and incorporate techniques such as self-supervised learning and advanced attention mechanisms to further enhance video prediction capabilities. Additionally, the integration of domain-specific knowledge and adaptive learning strategies could improve performance in real-world applications.

References

[1] X. Shi, Z. Chen, H. Wang, D.-Y. Yeung, W.-K. Wong, and W.-C. Woo, "Convolutional LSTM network: A machine learning approach for precipitation nowcasting," *Advances in Neural Information Processing Systems*, vol. 28, 2015.

[2] Y. Wang, Z. Gao, M. Long, J. Wang, and P. S. Yu, "Predrnn: Recurrent neural networks for predictive learning using spatiotemporal LSTMs," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[3] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in Neural Information Processing Systems*, vol. 30, 2017.