Fall-2024
Course Project

Course Title: CS-4045 Deep Learning for Perception

Section: DS-D

Group Members:

1. 21I-2502 Mansoor Ali
2. 21I-0733 Irtiqa Haider
3. 21I-0411 Muhammad Sohaib

Department of Computer Science

National University of Computing & Emerging Sciences

Islamabad, Pakistan

Predicting Video Frames Using Deep Learning: A Comparative Analysis of ConvLSTM, PredRNN, and Transformer Models

Abstract

This paper presents our journey in developing and evaluating three deep learning models—ConvLSTM, PredRNN, and a Transformer-based model—designed to predict future video frames from short input sequences. Using a subset of the UCF101 dataset, we focused on predicting realistic and coherent video frames. Through meticulous preprocessing, experimentation, and evaluation using MSE and SSIM, we discovered that PredRNN significantly outperformed other models, demonstrating superior temporal coherence and visual fidelity. The findings, challenges, and insights from this work highlight the potential of advanced temporal modeling in video prediction tasks.

Index Terms

ConvLSTM, PredRNN, Transformer, Video Prediction, Deep Learning, UCF101 Dataset.

I. Introduction

The ability to predict future video frames from a given sequence has applications in diverse domains, from autonomous systems to video synthesis. As part of this project, we aimed to explore how well deep learning models can "imagine" and simulate future frames based on past ones.

Our approach was hands-on and iterative, where each stage brought new insights and learning. By working with the UCF101 dataset, which is rich in human activity sequences, we could test the capabilities of modern architectures in video prediction. Our goal was not just to build accurate models but also to understand the trade-offs between different approaches in terms of performance, computational efficiency, and visual coherence.

II. Dataset and Preprocessing

The UCF101 dataset, a comprehensive video collection of human activities, provided a perfect foundation for this task. From our experience, the dataset was both a challenge and an opportunity, given the variability in actions and motion patterns.

A. Selected Classes

To ensure consistency in motion patterns, we narrowed down our focus to the following classes:

- **PushUps**
- **PullUps**
- **BenchPress**
- **Lunges**
- **WallPushups**

These actions exhibited clear, predictable motion dynamics, which allowed us to assess the models effectively.

B. Preprocessing

We quickly realized that preprocessing would play a critical role in this project. Videos were resized to 64×6464 \times 6464×64 pixels to balance computational feasibility and sufficient detail. We converted frames to grayscale, simplifying the input while preserving essential motion information.

The data was split into input and target sequences:

- **Input:** 30 frames per sequence.
- **Target:** The subsequent 15 frames. Batch sizes of 16 were chosen after experimenting with memory constraints during training.

C. Data Statistics

We processed:

- **Training Batches:** 2697
- **Validation Batches:** 422

Each batch had a defined shape of (16,20,64,64,1) for input and (16,10,64,64,1) for target sequences.

III. Models

We implemented three architectures, each offering unique insights into video prediction.

A. ConvLSTM

ConvLSTM integrates convolutional operations within LSTM cells, allowing spatial and temporal dynamics to be modeled simultaneously. While training this model, we observed its ability to capture short-term dependencies effectively. It achieved an SSIM of 0.683 and a validation MSE of 0.011 after five epochs.

B. PredRNN

Our team found PredRNN exciting due to its recurrent memory flow mechanism, which improves temporal modeling. Training this model was particularly rewarding, as it quickly demonstrated its potential, achieving an SSIM of 0.802 and a validation MSE of 0.0078.

C. Transformer-based Model

The Transformer model posed unique challenges for video prediction due to its emphasis on attention mechanisms. Although its architectural sophistication was apparent, maintaining temporal coherence was challenging. This model achieved an SSIM of 0.463 and a validation MSE of 0.019 over ten epochs.

IV. Results

A. Training Performance

Each model displayed distinct characteristics during training:

- ConvLSTM: Showed consistent improvement, reducing loss from 0.012 to 0.009.
- PredRNN: Converged rapidly, with loss dropping to 0.0067 by the fifth epoch.
- Transformer: Improved steadily, with final training loss of 0.0099.

B. Quantitative Metrics

Model	MSE	SSIM
ConvLSTM	0.011	0.683
PredRNN	0.008	0.802
Transformer	0.020	0.463

C. Visualization

When we visualized the predicted frames, the differences were striking:

- **ConvLSTM:** Struggled with long-term dependencies, leading to motion artifacts in later frames.
- **PredRNN:** Produced the most realistic sequences, with smooth transitions and accurate motion representation.
- **Transformer:** While effective in certain scenarios, it failed to maintain temporal consistency.

V. Discussion

A. Comparative Analysis

Reflecting on the results, we found:

- **Performance:** PredRNN consistently outperformed ConvLSTM and Transformer models in SSIM and MSE metrics.
- **Efficiency:** ConvLSTM trained faster but lacked the prediction quality of PredRNN.
- **Complexity:** Transformers, though conceptually powerful, required significant tuning and computational resources.

B. Challenges

Throughout this project, we faced challenges such as:

- Handling computational constraints during training.
- Ensuring stable convergence for the Transformer model.
- Balancing input sequence length with model performance.

C. Insights

One key insight was the importance of memory mechanisms like those in PredRNN for capturing temporal dependencies. Additionally, we learned that preprocessing and data quality are critical for achieving high performance.

VI. Conclusion

Our exploration into video frame prediction underscored the strengths and limitations of each architecture. PredRNN emerged as the most effective model, achieving a balance between accuracy and computational efficiency. ConvLSTM, while simple and fast, was best suited for short-term dependencies. The Transformer model highlighted the potential of attention mechanisms but required further optimization for this task.

Future work will involve exploring hybrid architectures that combine the strengths of these models and scaling them to higher resolution datasets.

References

- [1] UCF101 Dataset: https://www.kaggle.com/datasets/matthewjansen/ucf101-action-recognition/data.