Team 2 - Preliminary Project Report

Pushkal Mishra
ee20btech11042@iith.ac.in

Ankur Kumar
ee20btech11057@iith.ac.in

Sourabh Somnath Gholap
ee20btech11047@iith.ac.in

Perambuduri Srikaran
ai20btech11018@iith.ac.in

Abstract

With the rising trend of autonomous vehicles and intelligent agents, the crucial part in their functioning is their ability to make precise decisions. For this, sometimes it is helpful to predict anomalies in the near future before hand for error-free operations. One of the ways we can do it is by predicting the future frames from the past video frames. Many models exist which perform the above task but blurry predicted frames is a major issue. To solve this problem, we aim to explore image deblurring techniques using deep learning models in addition to the current state of the art prediction models to increase its performance. Furthermore, the problem of relevance arises when predicting more number of frames; so we will also explore methods of using input frames together with our predicted frames.

1. Introduction

In a broad sense, the prediction and anticipation of future events is a key component of intelligent decision making systems which we humans can perform seamlessly. But from the machine's point of view, the task of predicting future frames is extremely challenging due to factors such as occlusions, camera movement, lighting conditions, cluttering, or object deformations.

The problem of video frame prediction has become highly intriguing in recent times due to its importance in a wide variety of computer vision applications such as autonomous vehicles, intelligent agents [3], precipitation nowcasting [4], and many more. Also, recent advances in deep learning has improved the performance of video prediction as it is naturally a good self-supervised learning task.

Various deep learning techniques exist which extract meaningful spatio-temporal correlations from video data in a self-supervised learning fashion, such as in CNN-RNN-CNN based architectures the models generate predictions frame by frame with the previous output for capturing temporal evolution of frames. This paper [2] implements a completely CNN based architecture that generates predictions in a one-shot manner and potentially employ UNET connections between the convolutional layers.

The major problem we observed in the current models is that the predicted frames are slightly blurry for dynamic objects in the video which causes inaccuracies in reconstruction. In this project, we explore image deblurring techniques for the predicted frames to improve the performance of existing models.

2. Problem Statement

Current video frame prediction models use the predicted images to generate new predictions. It was observed that the model outputs slightly blurry images. This causes a cascading effect which makes the future frames to be blurrier.

To solve this problem, we look to deblur the predicted images before sending it to the model for predicting the next frames. We deblur the images using deep learning techniques.

But, predicting too many frames in continuum will cause loss of relevance in the original video. We look to explore the approach of using the predicted images and the input images together to predict future frames accurately.

This would be especially useful in autonomous driving. We would be predicting the future positions of obstacles accurately and also maintain the relevance by taking inputs from the environment at appropriate times. This also calls for having models with lesser inference time for this approach to work.

3. Literature Review

3.1. Video Frame Prediction

3.1.1 Next frame prediction using ConvLSTM

This paper [1] introduced a method for predicting future frames based on a series of prior input frames using ConvLSTM. The input video was first segmented into frames which are stored in a 2D array. Then each frame is sent

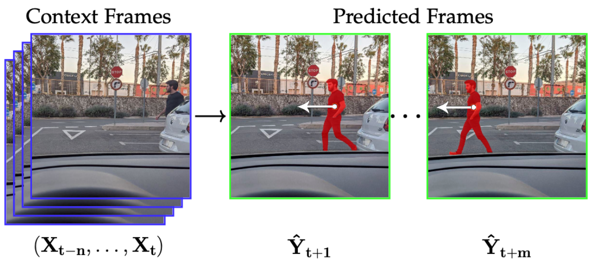


Figure 1. Video Frame Prediction for Autonomous driving

one at a time to the ConvLSTM module which extracts the featues, stores it and then takes in the next input frame. Using these learned features in the LSTM memory cells they make a prediction for the next frame. This predicted output is then compared with ground truth data.

The ConvLSTM model is divided into two sections: LSTM Encoder and Decoder. The first section reads the input sequence, extracts a **fixed-length** motion vector and a predicted frame with the help of previous reconstructed frame. Then the residual frame and motion vectors are passed to the LSTM decoder wherein a frame is predicted using these vectors and previous reconstructed frame. The final system's output is the sum of residual frame and predicted frame. To assess the prediction accuracy, this output frame is compared to the ground truth data.

This paper uses SSIM index and Perceptual Similarity as parameters to evaluate the model's performance.

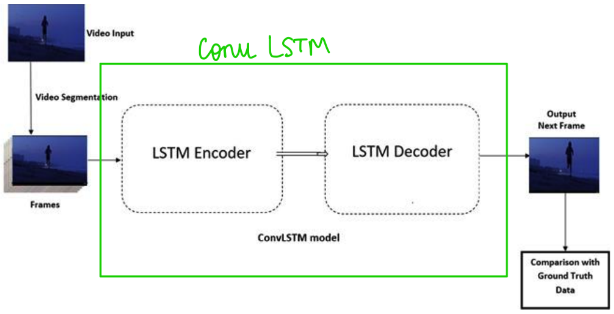


Figure 2. Architecture of the ConvLSTM model

3.1.2 VPTR

This [6] is a transformer block for video future frames prediction based on an efficient local spatial-temporal separation attention mechanism. To avoid prediction of similar frames, a contrastive feature loss is applied to maximize the mutual information between predicted and ground-truth future frame features.

This proposed transformer block is efficient for spatio-temporal feature learning by combining spatial local attention and temporal attention in two steps. The new Transformer block successfully reduces the complexity of a standard Transformer block with respect to same input spatio-temporal feature size, specifically, from $\mathcal{O}(T^2H^2W^2)$ to $\mathcal{O}(H^2W^2P^2 + T^2)$.

It has been demonstrated that the VPTR models, VPTR-NAR and VPTR-FAR, which utilize a straightforward attention mechanism, have the capability to achieve better results than the state-of-the-art VFFP models based on ConvLSTM that are more complex.

A comparison was formally carried out between two variants of VPTR, and the findings indicate that VPTR-NAR has a quicker inference rate and a lower error accumulation during inference compared to its counterpart. However, the process of training VPTR-NAR is more challenging. To address this issue, they utilized a contrastive feature loss that aims to increase the mutual information of the predicted and ground-truth future frame features.

VPTR-NAR is different from ConvTransformer with respect to the fundamental attention mechanism. ConvTransformer proposed a custom hybrid multi-head attention module based on convolution, but VPTR-NAR uses the standard multi-head dot product attention.

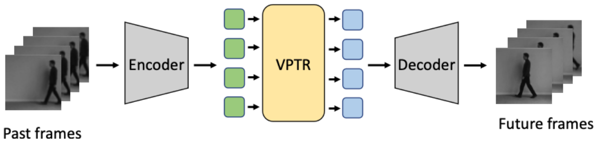
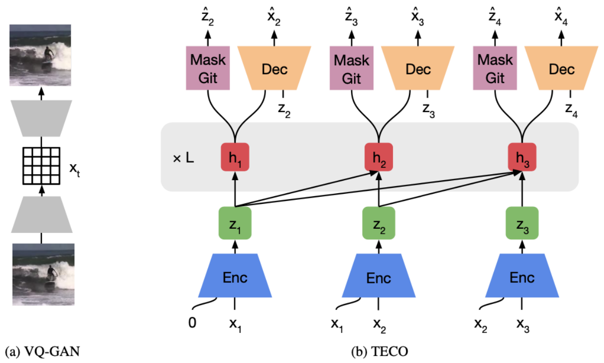


Figure 3. Overall framework of VPTR

3.1.3 Temporally Consistent Video Transformer

Temporally Consistent Video Transformer (TECO) [5] is a vector-quantized latent dynamics video prediction model that learns compressed representations to efficiently condition on long videos of hundreds of frames during both training and generation. TECO works in two broad steps: first the model learns temporal representations using encoder and temporal transformers and then it predicts future frames using some dynamics prior and a decoder. This model was applied to 3D environments such as DMLab, Minecraft and Habitat.

To spatially compress the video data, the encoder of the VQ-GAN encodes the current frame x_t conditioned on the previous frame by channel-wise concatenating x_{t-1} and then quantizes the output using codebook C to produce $z_t = E(x_t, x_{t-1})$. Then a single strided convolution was applied to downsample each discrete latent z_t to reduce the losses and lower the spatial resolutions. Afterwards, a large transformer model was learned to model temporal dependencies, and then apply a transposed convolution to upsample the representation back to the original resolution z_t. Then a decoder was used which is an upsampling CNN that reconstructs $\hat{x}_t = D(z_t, h_t)$. Lastly to model the dynamics prior, they used MaskGit which allows for faster and higher quality sampling compared to an autoregressive prior.



(a) VQ-GAN

(b) TECO

Figure 4. Architecture design of TECO

3.1.4 SimVP

SimVP [2] is a simple yet effective spatio-temporal predictive learning model using CNNs for frame prediction. It aims to design an autoencoder-like architecture that inputs the past frames and outputs the future frames while preserving the temporal dependencies. The CNN autoencoder is used to reconstruct a frame by borrowing voxels from nearby frames.

SimVP learns a mapping $\mathcal{F}_\theta : \mathcal{X}^{t,T} \rightarrow \mathcal{Y}^{t+1,T'}$ to encode the past frames $\mathcal{X}^{t,T}$ and decode the future frames $\mathcal{Y}^{t+1,T'}$.

The **spatial encoder** is employed to encode the high dimensional past frames into the low-dimension latent space. The **translator** learns both spatial dependencies and temporal variations from the latent space. The **spatial decoder** ultimately decodes the latent space into the predicted future frames.

A mapping is introduced between the spatial encoder and the spatial decoder to preserve the spatial features.

The training time and efficieny is better than ConvLSTM, PredRNN, PredRNN++, MIM, E3D-LSTM and PhyDNet. The predicted images are much clearer than other models.

The error is relatively large when the scene changes dramatically but remains low in most cases. SimVP accurately predicts the future trend to a large extent and unexpectedly finds the sudden traffic jam from the observations of placid transportation

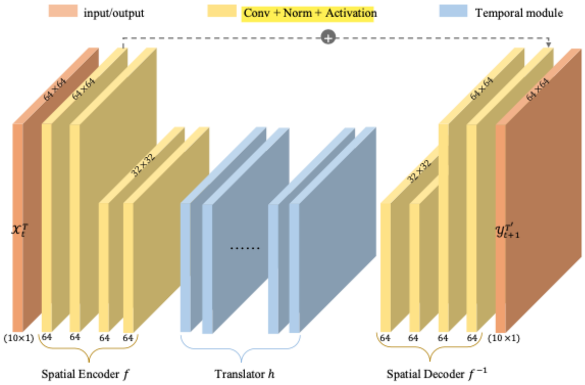


Figure 5. Architecture of SimVP

3.2. Image Deblurring

3.2.1 Efficient Transformer for Image Restoration

This paper [7] developes an efficient Transformer model that can handle high-resolution images for restoration task by introducing key design changes to the multi-head SA layer and a multi-scale hierarchical module that has lesser computing requirements than a single-scale network.

Say $H \times W$ is the spatial dimension and C is the number of channels, so firstly the model applies a convolution to obtain low-level feature embeddings $F_0 \in \mathrm{R}^{H \times W \times C}$. Then these shallow features F_0 pass through a 4-level symmetric encoder-decoder and transformed into deep features. From the input, the encoder hierarchically reduces spatial size, while expanding channel capacity whereas the decoder takes low-resolution latent features as input and progressively recovers the high-resolution representations. To assist the recovery process, the encoder features are concatenated with the decoder features via skip connections which is followed by a 1×1 convolution to reduce the channels at all levels.

Further, the deep features were enriched in the refinement stage operating at high spatial resolution which yields quality improvements. Finally a convolution layer is applied to the refined features to generate a residual image which will be added to the original degraded image to obtain the refined image. To validate their model, they performed various experiments for image processing tasks such as image deraining, image deblurring, defocus deblurring and image denoising which showed some promising results.

References

[1] Padmashree Desai, C Sujatha, Saumyajit Chakraborty, Saurav Ansuman, Sanika Bhandari, and Sharan

Kardiguddi. Next frame prediction using ConvLSTM. *Journal of Physics: Conference Series*, 2022. 1

[2] Zhangyang Gao, Cheng Tan, Lirong Wu, and Stan Z. Li. SimVP: Simpler yet better video prediction, 2022. 1, 3

[3] Wen Liu, Weixin Luo, Dongze Lian, and Shenghua Gao. Future frame prediction for anomaly detection – a new baseline, 2017. 1

[4] Xingjian Shi, Zhourong Chen, Hao Wang, Dit-Yan Yeung, Wai-kin Wong, and Wang-chun Woo. Convolutional lstm network: A machine learning approach for precipitation nowcasting, 2015. 1

[5] Wilson Yan, Danijar Hafner, Stephen James, and Pieter Abbeel. Temporally consistent video transformer for long-term video prediction, 2022. 2

[6] Xi Ye and Guillaume-Alexandre Bilodeau. VPTR: Efficient transformers for video prediction, 2022. 2

[7] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Restormer: Efficient transformer for high-resolution image restoration, 2021. 3