Video Frame Prediction

Swapnil Sharma
swapnil.sh@nyu.edu

Nikita Anand
lna285@nyu.edu

Rishika Pabba
rp3608@nyu.edu

Abstract

This paper details an attempt to solve a video frame prediction task by utilizing various models and techniques, including ConvLSTM, Segformer, and AutoEncoders for semantic segmentation mask, ConvLSTM in an auto-regressive manner, ConvBiLSTM in a non-auto regressive manner by predicting all frames at once, and ultimately settling on VPTR Non Auto Regressive due to its convergence on a subset of data. The authors encountered significant challenges with training time and memory constraints, and attempted to address these issues by using mixed precision training and unsuccessfully attempting to train on multiple GPUs using PyTorch Distributed Data Parallel.

The ResNet AutoEncoders proved to be the most effective for the semantic segmentation mask with an impressive accuracy of 0.993 on the validation set. The VPTR model achieved a decent accuracy of 0.9 on a subset of the data for video frame prediction. Although the model was unable to converge on the whole dataset due to time and memory constraints, the authors believe that with more time and resources, they would have achieved better accuracy in predicting future frames.

1 Introduction

In recent years, predicting future frames in videos has become a challenging task that has attracted extensive research. This task has numerous applications, including video compression, editing, and surveillance. This report documents an attempt to solve a video frame prediction task presented by the NYU Deep Learning Course, which required predicting the semantic segmentation mask of the 22nd frame of a video based on the initial 11 frames.

The videos used in this project featured simple 3D shapes that interacted with each other based on basic physics principles. Each object in the videos was uniquely identified by a combination of three attributes: shape (cube, sphere, or cylinder), material (metal or rubber), and color (gray, red, blue, green, brown, cyan, purple, or yellow). No identical objects appeared in a single video.

To address the challenges of semantic segmentation mask translation and future frame prediction, we explored several approaches, including ConvLSTM, Segformer, ResNet AutoEncoders, convBiLSTM, and non-autoregressive VPTR. The ResNet AutoEncoders achieved the best results for the semantic segmentation mask translation task with an accuracy of 0.993 on the validation set. For the future frame prediction task, we achieved the best performance on a subset of the dataset using the VPTR model, although we were unable to converge the model on the entire dataset due to time and memory constraints.

2 Related Work

Most state-of-the-art (SOTA) models for video frame prediction use ConvLSTM-based AutoEncoders. These models were initially developed for predicting precipitation nowcasting, as introduced by Shi et al. [2015]. They have been later utilized for video prediction task, including Finn et al. [2016],

Preprint. Under review.

Lotter et al. [2016], Xu et al. [2016], Ballas et al. [2015]. According to Jing and Tian [2019], these models also work with self-supervised tasks.

Although the ConvLSTM-based models are flexible and efficient, they are generally slow due to recurrent prediction. To address this issue, standard CNNs or 3D CNNs and VAE-based methods have been proposed, such as those by Mathieu et al. [2015] and Babaeizadeh et al. [2017].

State-of-the-art models commonly rely on complex ConvLSTM models integrating attention mechanisms or memory-augmented modules. For instance, the Long-term Motion Context Memory model by Lee et al. [2021] stores long-term motion context through a novel memory alignment learning, and the motion information is recalled during the test to facilitate long-term prediction. Chang [2021] proposed an attention-based motion-aware unit to increase the temporal receptive field of RNNs.

Almost all the state-of-the-art (SOTA) VFFP models are based on ConvLSTMs, i.e. convolutional short-term memory networks,which are efficient and powerful. Nevertheless as per Ye and Bilodeau [2022], they suffer from some inherent problems of recurrent neural networks(RNNs), such as slow training and inference speed, error accumulation during inference, gradient vanishing, and predicted frames quality degradation. Researchers keep improving the performance by developing more and more sophisticated ConvLSTM-based models.

With the introduction of transformers, they have also been applied in the Vision domain, including video frame prediction. The ConvTransformer model by Liu et al. [2020] follows the architecture of DETR introduced in Meinhardt et al. [2021], a classical neural machine translation (NMT) Transformer architecture. DETR also inspired the development of the VPTR-NAR model by Ye and Bilodeau [2022], a non-autoregressive model for video frame prediction.

3 Methodology

3.1 Semantic Segmentation Mask

We settled on ResNet AutoEncoders to solve the semantic segmentation mask translation task after exploring multiple approaches. ResNet, a CNN architecture with skip connections, was used to avoid vanishing gradients. AutoEncoders, a type of neural network for dimensionality reduction, were employed to learn representations of the input using the encoder and to construct a semantic segmentation mask from the learned representation using the decoder. To simplify training, we removed any skip connections between the encoder and decoder layers, as suggested by Ye and Bilodeau [2022].

The encoder was composed of nine ResNet blocks, each with two convolutional layers. The decoder contained three deconvolutional layers with one filter and a sigmoid activation function at the end to produce a semantic segmentation mask with a channel dimension of 49. We used Mean Square Error Loss and the Adam optimizer with a learning rate of 0.0002 to train the model. Due to memory constraints, we trained the model with a batch size of 1 for 100 epochs. We trained the model on 1000 labeled videos and validated it on another 1000 labeled videos. The model took approximately 16 hours to train on a single GPU.

3.2 Future Frame Prediction

In the second task of predicting future frames, we experimented with various models, including ConvLSTM, convBiLSTM, and VPTR. Ultimately, we chose VPTR-NAR as it was the most recent and state-of-the-art model for video frame prediction. Ye and Bilodeau [2022] has two approaches one autoregressive and another non-autoregressive. As per their results they conculded that NAR works better and hence we decided to use NAR. NAR utilizes a transformer encoder and a transformer decoder in its architecture.

The architecture of VPTR is illustrated in Figure 1. The left part of the figure, denoted as $\mathcal{T}_E$, represents the encoder, which encodes all past frame features z_t ($t \in [1, L]$) into memory representations e_t ($t \in [1, L]$). The architecture of $\mathcal{T}_E$ comprises multiple VidHRFormer blocks, each of which contains a multi-head self-attention layer, a feed-forward layer, and a layer normalization layer.

On the right part of the figure, denoted as $\mathcal{T}_D$, the decoder is illustrated. It includes two additional layers compared to $\mathcal{T}_E$: a temporal multi-head attention (MHA) layer and an output convolutional

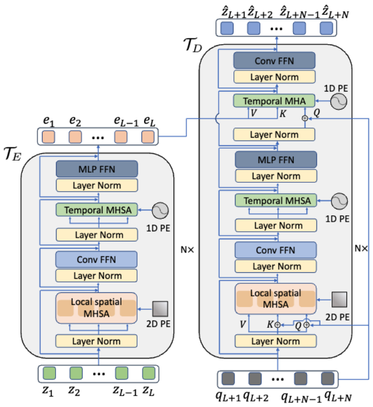

Figure 1: VPTR NAR model architecture

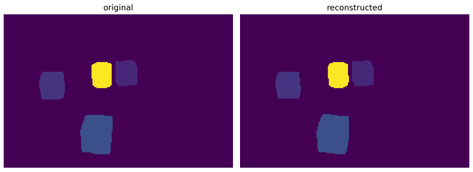

Figure 2: Reconstructed semantic segmentation mask

feed-forward network (Conv FFN) layer. The Temporal MHA layer is also known as the encoder-decoder attention layer, which takes the memories as value and key, and queries derived from the future frame query sequence $q_{L+1}, \cdots, q_{L+N}$.

4 Dataset

The dataset consists of 13,000 unlabeled videos, 2,000 labeled videos and 1000 test videos. The labeled videos are further divided into 1,000 training videos and 1,000 validation videos. Labeled and unlabeled videos have 22 frames each and the test videos have 11 frames each. Labeled videos also have the semantic segmentation mask for all the frames. All the frames and the semantic segmentation masks are of size 160×240.

We created two different datasets for the two tasks. For the first task of semantic segmentation mask translation, we used the labeled videos and constructed a dataset consisting of frames with their corresponding semantic segmentation masks. For the second task of future frame prediction, we utilized only the unlabeled videos containing 11 frames each along with the corresponding 11 future frames to construct the dataset.

5 Results

Our experiments started with predicting the semantic segmentation mask of the frames given normalized images of the frames. We tested ConvLSTM, Segformer, and ResNet AutoEncoders. Both

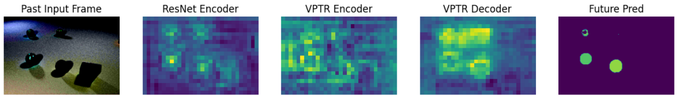

Figure 3: Frame prediction pipeline

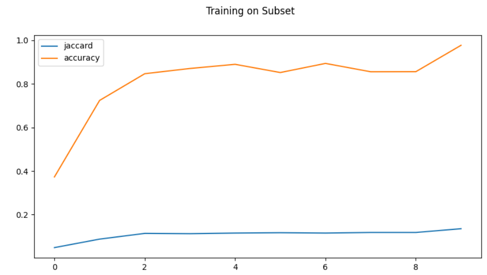

Figure 4: Training on subset of frames

ConvLSTM and Segformer showed good performance, achieving a decent accuracy of 0.93. However, ResNet AutoEncoders outperformed them, achieving a startling accuracy of 0.993 on the validation set. Figure 2 shows the reconstructed semantic segmentation mask for a frame predicted by the ResNet AutoEncoders.

For the second task of future frame prediction, we attempted to use ConvLSTM, convBiLSTM, and VPTR. However, we found that ConvLSTM and ConvBiLSTM were not effective, as they took a long time to converge and yielded poor-quality results. VPTR showed promising results on a subset of the data, as can be seen in Figure 4. However, we encountered memory constraints and were unable to train the model on the entire dataset, unable to exceed a batch size of 1. We explored using DataParallel and DataDistributedParallel to utilize multiple GPUs, but due to tight memory constraints, the overhead of using these libraries was too great. We were able to reduce per-epoch runtime from 6 hours to 3 hours by utilizing AMP (Automatic Mixed Precision) and Gradient Scaling.

Figure 3 shows the zeroth input, intermediate feature, and zeroth output of the VPTR-NAR model. From the ResNet Encoder feature, it can be observed that the model is able to see the objects in the frame. However, it is difficult to discern what the VPTR Encoder is learning, as it is supposed to be the attention layer output for each frame. The VPTR Decoder output shows that it vaguely understands where the objects should be, and then produces the future frame prediction.

6 Conclusion

In summary, we were able to achieve a very high accuracy of 0.993 on the semantic segmentation mask translation task using ResNet AutoEncoders. We achieved a decent accuracy of 0.9 on the future frame prediction task on a subset of the data, but the model did not perform as well on the test set, achieving only a 0.03 Jaccard score. Due to time and memory constraints, we were not able to converge the model on the entire dataset. With additional resources and time, we believe that we would have been able to achieve a better accuracy on the future frame prediction task.

References

Mohammad Babaeizadeh, Chelsea Finn, D. Erhan, Roy H. Campbell, and Sergey Levine. Stochastic variational video prediction. *ArXiv*, abs/1710.11252, 2017.

Nicolas Ballas, L. Yao, Christopher Joseph Pal, and Aaron C. Courville. Delving deeper into convolutional networks for learning video representations. *CoRR*, abs/1511.06432, 2015.

Zheng Chang. Mau: A motion-aware unit for video prediction and beyond. In *Neural Information Processing Systems*, 2021.

Chelsea Finn, Ian J. Goodfellow, and Sergey Levine. Unsupervised learning for physical interaction through video prediction. *ArXiv*, abs/1605.07157, 2016.

Longlong Jing and Yingli Tian. Self-supervised visual feature learning with deep neural networks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43:4037–4058, 2019.

Sangmin Lee, Hak Gu Kim, Dae Hwi Choi, Hyungil Kim, and Yong Man Ro. Video prediction recalling long-term motion context via memory alignment learning. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3053–3062, 2021.

Zhouyong Liu, Shun Nian Luo, Wubin Li, Jingben Lu, Yufan Wu, Chunguo Li, and Luxi Yang. Convtransformer: A convolutional transformer network for video frame synthesis. *ArXiv*, abs/2011.10185, 2020.

William Lotter, Gabriel Kreiman, and David D. Cox. Deep predictive coding networks for video prediction and unsupervised learning. *ArXiv*, abs/1605.08104, 2016.

Michaël Mathieu, Camille Couprie, and Yann LeCun. Deep multi-scale video prediction beyond mean square error. *CoRR*, abs/1511.05440, 2015.

Tim Meinhardt, Alexander Kirillov, Laura Leal-Taixé, and Christoph Feichtenhofer. Trackformer: Multi-object tracking with transformers. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8834–8844, 2021.

Xingjian Shi, Zhourong Chen, Hao Wang, D. Y. Yeung, Wai-Kin Wong, and Wang chun Woo. Convolutional lstm network: A machine learning approach for precipitation nowcasting. In *NIPS*, 2015.

Huazhe Xu, Yang Gao, Fisher Yu, and Trevor Darrell. End-to-end learning of driving models from large-scale video datasets. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3530–3538, 2016.

Xiutiao Ye and Guillaume-Alexandre Bilodeau. Vptr: Efficient transformers for video prediction. *2022 26th International Conference on Pattern Recognition (ICPR)*, pages 3492–3499, 2022.