VIDEO FUTURE FRAME PREDICTION

Vineeth S

M.Tech Artificial Intelligence, SR No. 16543

ABSTRACT

Video Frame Prediction is the task of predicting future frames given a set of past frames. This task is of high interest as it caters to many applications such as autonomous navigation and self-driving. We present a novel Adversarial Spatio-Temporal Convolutional LSTM architecture to predict the future frames of the Moving MNIST Dataset [1]. We evaluate the model on long-term future frame prediction and its performance of the model on out-of-domain inputs by providing sequences on which the model was not trained.

***Index Terms*—** video frame prediction, video prediction

1. INTRODUCTION

Despite the fact that humans can easily and effortlessly solve the future frame prediction problem, it is extremely challenging for a machine [2]. Complexities such as occlusions, camera movement, lighting conditions, or clutter make this task difficult for a machine [4]. Predicting the next frames requires an accurate learning of the representation of the input frame sequence or the video. We can even use such a model as a feature extractor for other performing tasks [3]. However, for the very same reason training such a model is difficult and requires significant amount of compute power.

2. TECHNICAL DETAILS

We can formally define the task of predicting future frames in videos as follows. Let $X_t \in \mathbb{R}^{w \times h \times c}$ be the t^{th} frame in the video sequence $\mathbf{X} = (X_{t-n}, ..., X_{t-1}, X_t)$ with n frames, where w, h, and c denote the width, height, and number of channels respectively. The goal is to predict the next frames $\mathbf{Y} = (\hat{Y}_{t+1}, \hat{Y}_{t+2}, ..., \hat{Y}_{t+m})$ from the input $\mathbf{X}$.

2.1. Architecture details

We use a Adversarial Spatio-Temporal Convolutional LSTM architecture. The frame predictor model takes in the first ten frames as input and predicts the future ten frames. The discriminator model tries to classify between the true future frames and predicted future frames, thereby increasing the quality of predicted frames. We train the model for 100 epochs with a training time of 100 GPU hours. For the first ten time instances, we use the ground truth past frames as input, where as for the future time instances, we use the past predicted frames as input.

3. RESULTS

We evaluate the performance of the model for long-term predictions to reveal its generalization capabilities. We provide the first 20 frames as input and let the model predict for the next 100 frames. The model is able to predict the frames until the numbers occlude in the video. From the time of occlusion, the model chooses one of the multiple future pathways, since occlusion of different two digits could look similar. Despite the prediction being wrong, we can still observe that the quality of the predicted frame is high with no significant deterioration. The results are presented in Figures 1 and 2

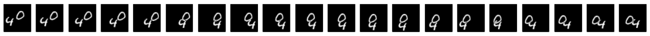

Fig. 1: Ground truth sequence for the first 20 frames

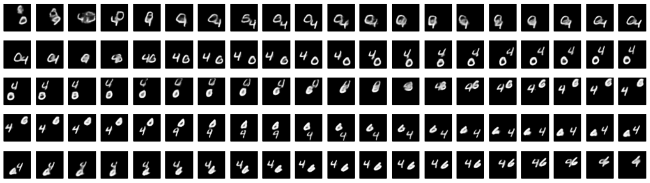

Fig. 2: Predicted 100 frames from time instance 2 onwards

We evaluate the performance of the model on out-of-domain inputs which the model has not seen during the training. We provide a frame sequence with one moving digit as input and observe the outputs from the model. The results are presented in Figures 3 and 4. We can observe that the model starts to hallucinate two numbers. The possible explanation for this might be that the model could be treating the single input as an overlapping or occluded image of two digits. From figure 4, we can see see this phenomenon that the model considers the 1 in the input frame to be an overlapped frame of two 1s and starts to generate frames with two 1s in the successive time instances.

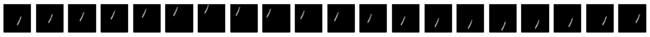

Fig. 3: Ground truth sequence for the first 20 frames

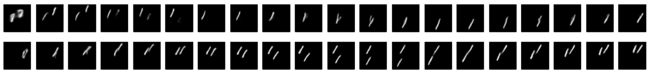

Fig. 4: Predicted 40 frames from time instance 2 onwards

4. CONTRIBUTIONS

We implement the whole architecture from scratch. The use of discriminator to improve frame quality in a Convolutional LSTM architecture has not been explored before.

5. RESOURCES

Dataset: Moving MNIST http://www.cs.toronto.edu/~nitish/unsupervised_video/

Toolkits: PyTorch, Torchvision, Scikit-image, PIL Image

References

[1] Nitish Srivastava, Elman Mansimov, and Ruslan Salakhutdinov. *Unsupervised Learning of Video Representations using LSTMs*. 2016. arXiv: 1502.04681 [cs.LG].

[2] Jun-Ting Hsieh et al. *Learning to Decompose and Disentangle Representations for Video Prediction*. 2018. arXiv: 1806.04166 [cs.LG].

[3] Wen Liu et al. "Future Frame Prediction for Anomaly Detection – A New Baseline". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2018.

[4] Sergiu Oprea et al. *A Review on Deep Learning Techniques for Video Prediction*. 2020. arXiv: 2004.05214 [cs.CV].