Segmentation Mask Prediction of Future Video Frame

Charvi Gupta[1*] cg4177@nyu.edu
Anoushka Gupta[1*] ag8733@nyu.edu
Anisha Bhatnagar[1*] ab10945@nyu.edu

Courant Institute of Mathematical Sciences
New York University

Abstract

In this work, we evaluate the performance of Conv-LSTM-based model and U-Net Segmentation model for predicting the segmentation mask of the 22^{nd} frame of a video. The future frame prediction model receives the first 11 frames of a video and outputs the next 11 frames. We pass the predicted 22^{nd} frame as the input to the segmentation model, which generates its segmentation mask. We conducted experiments on Adversarial Conv-LSTM model as frame prediction model and U-Net as the segmentation model. The performance of the pipeline is evaluated using Jaccard index. A score of 0.2437 was obtained for the segmentation masks on the validation set.

1 Introduction

Predicting future frames of a video is an important task for an intelligent vision model. It has a wide variety of applications in abnormal video detection and autonomous driving. The main components of video frame prediction include capturing details of the interaction of different objects present in the video, along with their future dynamics. There is a lot of ambiguity in predicting the future frame of a video, due to complex factors at play, like filming factors such as foreground and background lighting and uncertainty of the interaction of objects due to entry and exit of different objects. This makes the future frame prediction task challenging and computationally demanding.

Despite these challenges, a variety of works have shown promising results in the past years. A video is essentially an ordered collection of spatio-temporal sequences. Hence many variations of Recurrent Neural Networks (RNNs) and Long Short Term Memory (LSTM) networks have been applied successfully in variations of video prediction problems. But their predictions are blurry. ConvLSTMs (introduced by SHI et al. [1]) have been found to be better at these kinds of tasks. Furthermore, Generative Adversarial Networks (GANs) (introduced by Goodfellow et al. [2]) have also been shown to improve performance. We aim to leverage these models and test their performance on the task at hand: Given a sequence of the first 11 frames of a video, predict the segmentation mask of the 22^{nd} frame.

2 Related Work

Future Frame Prediction Many approaches have been proposed for future frame prediction. Initial works include using GRU's [3] that produce competitive results with less computation. SHI et al. [1] found Convolutional LSTM to outperform LSTMs for precipitation nowcasting. Some other works like [4] use adversarial training, along with other measures to produce high SSIM on the UCF101 dataset. We found these works to be particularly interesting and applied them to our moving objects dataset.

Image Segmentation Works like MaskRCNN (introduced by He et al. [5]) have been successfully shown to identify objects in images and generate high quality segmentation masks. Additionally models like U-Net (introduced by Ronneberger et al. [6]) are able to achieve similar performance levels with much less data. Since we have few labeled data points with ground truth masks, we levarage this model for our use case.

3 Methodology

For Future Frame Prediction, a Generative Adversarial Network with a ConvLSTM generator was used. A ConvLSTM model consists of a series of convolutional layers. The output of these layers is used to compute the intermediate cell states and the activation vectors for input, output and forget gates that are typically present in an LSTM model. Due to the combination of convolutional the recurrent operations in a ConvLSTM model, they are excellent at capturing the spatio-temporal features and correlations in sequential image data such as videos.
A Unet model consists of contracting and expanding paths. It first downsamples an image using a series of convolution operations and then upsamples an image using matching deconvolution operations. Skip connections are used to pass the outputs of convolution layers to deconvolution layers. The downsampling allows the model to learn feature maps. Skip connection provide object localization information. The upsampling path uses the learned features and localizations to recreate the image and generate the mask. A solution combining the ConvLSTM-GAN architecture and the UNet segmentation model is proposed. Using a ConvLSTM as a base for the generator module for a GAN, the 22nd frame of the input sequence was generated by using the first 11 as an input. The predicted frame was then passed to a Unet model for generation of segmentation masks. The pipeline is depicted in Figure 1.

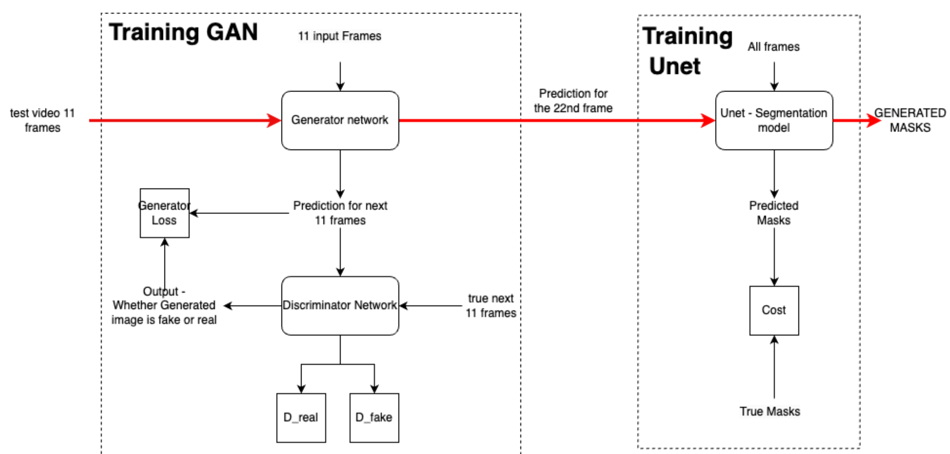

Figure 1: **Model Pipeline**. The Generator network was trained to predict the future frames of the sequence, of which the 22nd frame was extracted during inference(depicted in red) and passed as input to the segmentation model to generate the segmentation mask

3.1 Dataset

The dataset consists of moving objects of varying shapes colours and materials. Objects in the video could be of 3 shapes - cubes, spheres, or cylinders. They could be made of two materials - metal or rubber. They could be of 8 possible colours - gray, red, blue, green, brown, cyan, purple, or yellow. Each object can be uniquely identified using a combination of the three attributes. Therefore, 48 possible objects exist. (Additionally there is one extra class for the background). The dataset consisted of videos with 22 frames. The unlabeled set had 13000 videos, the labeled train and test sets had 1000 videos each. The labels were available in the form of segmentation masks for each frame.

3.2 Frame prediction using ConvLSTM

The Generator was made of 4 ConvLSTM cells. The model expected the first 11 frames as input and returned the prediction for the next 11 frames along with a learned representation of the 11 input frames. The generator module was trained using a combination of Binary Cross Entropy (BCE) loss and L1-L2 loss and Adam optimizer. The L1-L2 loss regularized the model to enhance the quality of generated images. The discriminator module was a linear model with four linear layers. 3 of them were followed by LeakyRelu activations. Sigmoid activation was used in conjunction with the final linear layer due to the binary nature of classification into fake or real samples. The discriminator module was trained using BCE loss and Adam optimizer. Training was performed for 50 epochs using two NVIDIA Tesla V100 GPUS.

3.3 Segmentation using UNet

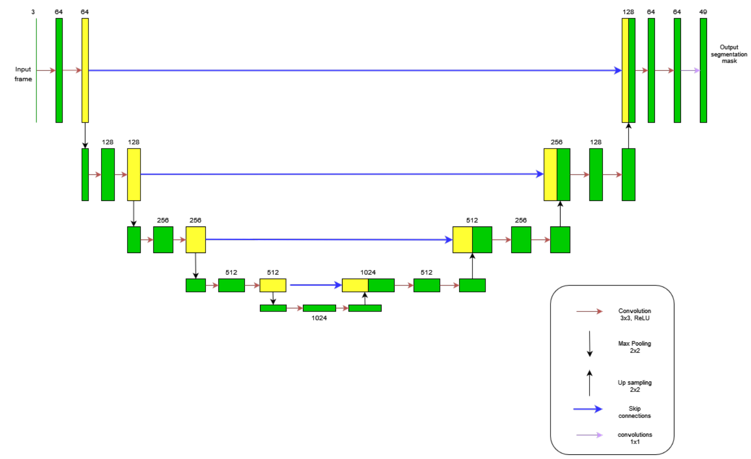

Figure 2: UNet model architecture

For generating the segmentation masks, the UNet encoder-decoder type architecture, depicted in Figure 2 was used. The model consists of 4 down-sampling layers and 4 corresponding up-sampling layers. The down-sample layers are connected to the up-sample layers using 4 skip connections. The input was an image of size 160×240 and output was a tensor of size $49 \times 160 \times 240$, which was the one hot encoding of every class in the segmentation mask. This segmentation model was trained on each frame of the 1000 labelled training videos for 10 epochs using Cross Entropy Loss and Adam Optimizer. The learning rate was set to $1e-4$.

4 Results

The segmentation model was able to generate masks with a $\mathbf{91.83}\%$ accuracy on the 22$^{\text{nd}}$ frames extracted from the validation set. Table 1 shows the performance of the inference pipeline on the validation and hidden sets. It achieved a Jaccard score of $\mathbf{0.2437}$ on validation set. The ConvLSTM GAN was able to efficiently extract the relations between distinctly visible objects and materials, as depicted in Figure 3. However, the model fails on examples where objects are overlapping or on examples where new objects appear post the 11$^{\text{th}}$ frame.

Table 1: Results on validation and hidden dataset

Dataset	Jaccard Score
Validation	0.2437
Hidden	0.1455

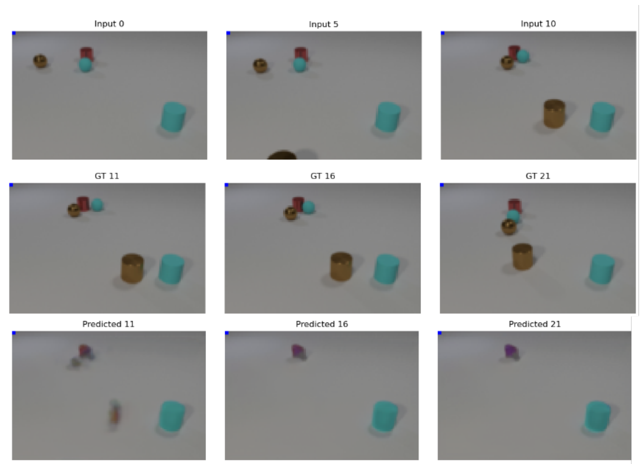

Figure 3: Input Frames $0, 5, 10$, Ground truth(GT) frames $11, 16, 21$ and Predicted frames $11, 16, 21$

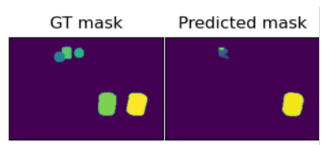

Figure 4: Ground truth(GT) mask and Predicted mask of final frame

5 Conclusions and Future Work

The ConvLSTM GAN + UNet model can infer the positions of some of the objects already present in the input frames in the predicted 22$^{\text{nd}}$ frame accurately. However, it is unable to predict new objects that enter the environment, resulting in lower performance on the hidden test. This is because the architecture fails to recognize the changes in the environment due to physical interactions between the objects. For future improvements on this dataset, the model can be trained on the full unlabelled data set so that it can learn to infer the presence of new objects. Various pre-processing and image normalization techniques can also be tried. Using a one-shot training technique where the frame prediction and segmentation mask prediction happen simultaneously can also be explored.

References

[1] 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. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015.

[2] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014.

[3] Marc Oliu, Javier Selva, and Sergio Escalera. Folded recurrent neural networks for future video prediction. In *European Conference on Computer Vision*, 2017.

[4] Michael Mathieu, Camille Couprie, and Yann LeCun. Deep multi-scale video prediction beyond mean square error. January 2016. 4th International Conference on Learning Representations, ICLR 2016 ; Conference date: 02-05-2016 Through 04-05-2016.

[5] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn, 2018.

[6] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In Nassir Navab, Joachim Hornegger, William M. Wells, and Alejandro F. Frangi, editors, *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, pages 234–241, Cham, 2015. Springer International Publishing. ISBN 978-3-319-24574-4.