Future Video Frame Prediction

Karan Kumar Gangadhar
New York University, Courant
kk5409@nyu.edu

Prithviraj Murthy
New York University, Courant
pkm5789@nyu.edu

Satyanarayana Chillale
New York University, Courant
sc9960@nyu.edu

Abstract

Video frame prediction is one of the challenging task in machine learning. Predicting high quality images is still an evolving area in the field of generative artificial intelligence. The task for this report was to predict the 22nd frame given only the first 11 frames of a video consisting on multiple small objects of different shapes. In this work, for video future frame prediction we used a framework called Masked Conditional Video Diffusion (MCVD)[10] which uses a probabilistic conditional score-based denoising diffusion model[4], conditioned on past 12 frames in a sliding window manner. For the image semantic segmentation, we used U-Net[8] architecture that consists of series of down and up convolutions.

1 Literature Review

In the realm of video frame prediction, recent advancements have spurred significant interest and innovation. Leveraging conditional score-based models, Grathwohl et al. (2020) [3] introduced a denoising diffusion probabilistic model, forming the basis for techniques like Masked Conditional Video Diffusion (MCVD). This approach, as employed in the present study, facilitates accurate future frame prediction by iteratively refining noisy inputs conditioned on past observations. Additionally, generative adversarial networks (GANs) have made notable strides in realistic video generation, with seminal works by Vondrick et al. (2016) [11] and Villegas et al. (2017) [9] showcasing their efficacy in synthesizing dynamic scenes. Moreover, probabilistic generative models, such as variational autoencoders (VAEs) and autoregressive models, have been explored extensively for video frame prediction tasks, as evidenced by Kalchbrenner et al. (2016) [5] and Babaeizadeh et al. (2017) [1], contributing to the diverse landscape of methodologies in this domain.

On the front of image segmentation, the U-Net architecture stands as a cornerstone in the field, initially introduced by Ronneberger et al. (2015) [8] for biomedical image segmentation. This architecture, characterized by its expansive contracting path followed by an equally expansive symmetric expanding path, has demonstrated remarkable efficacy across various segmentation tasks. Concurrently, advancements in semantic segmentation with deep learning have propelled the field forward, as showcased in the seminal work by Long et al. (2015) [6], which pioneered the application of fully convolutional networks (FCNs) for pixel-wise segmentation. Moreover, attention mechanisms and multi-scale feature fusion techniques have further refined segmentation accuracy. Oktay et al. (2018) [7] proposed an attention U-Net architecture tailored for medical imaging tasks, while Chen et al. (2017) [2] introduced DeepLab, incorporating atrous convolution and fully connected conditional random fields to effectively aggregate multi-scale features for precise segmentation. These advancements collectively highlight the diverse array of methodologies and approaches driving progress in image segmentation.

2 Methodology

In our project, we aimed to tackle the challenge of video frame prediction and subsequent segmentation. This involved two main components: generating future frames from existing video sequences, and performing image segmentation on those frames.

2.1 Future Frame Generation using MCVD

The first part of our methodology utilizes the Masked Conditional Video Diffusion (MCVD) model. MCVD is a powerful tool for video synthesis tasks, which includes the capability to predict future video frames. It works by learning a distribution of video data and then generating frames that are likely to succeed the given sequence. This is particularly useful in scenarios where understanding future states of a dynamic scene is crucial. The MCVD model uses a probabilistic approach, conditioning on past and/or future frames, to ensure that the generated frames are not only plausible but also coherent with the video context.

2.2 Segmentation using U-Net

After generating the future frames, the next step involves segmenting these frames to identify and classify different objects and regions within them. For this task, we employed the UNet model, a type of convolutional neural network that is highly effective for image segmentation tasks. UNet is designed to work well with fewer training images and to produce precise segmentations. It operates by using a contracting path to capture context and a symmetric expanding path that enables precise localization. This architecture is particularly adept at dealing with the nuances in the spatial hierarchy of images, which makes it ideal for segmenting the complex scenes depicted in the synthetically generated video frames.

The integration of these models showcases a significant stride in video processing technology, offering both enhanced predictive capabilities and detailed analytical insights into the generated video content.

2.3 Visualization

Below figures visualizes the feature map from the convolution network at different layers.

Figure 1: Second layer

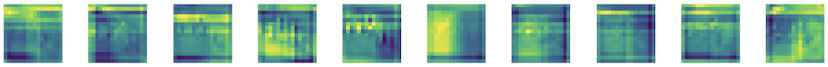

Figure 2: Twenty-third layer

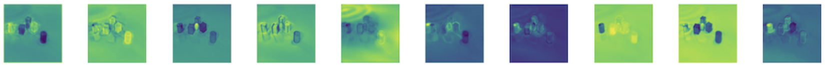

Figure 3: Thirty-fifth layer

3 Results

In this section, we present the evaluation results of our models applied in the tasks of future frame prediction and segmentation of those frames. Our results are indicative of the effectiveness of our approach in handling complex video data and achieving significant predictive and segmentation accuracy.

3.1 Segmentation Model Performance

Our UNet segmentation model demonstrated exemplary performance on the validation set, achieving an accuracy of over **96**%. This high level of accuracy underscores the model's ability to effectively delineate and classify various objects within the video frames, which is crucial for detailed scene understanding in numerous practical applications.

3.2 Future Frame Prediction

We evaluated the performance of our Masked Conditional Video Diffusion (MCVD) model in predicting future frames under different sampling scenarios:

- With **100 DDPM samplings**, the model achieved a Jaccard index score of approximately 0.12. This score reflects the initial capability of our model to approximate the future state of the video scenes, albeit with a broad margin for improvement.
- Increasing the sampling to **500** improved the Jaccard index score to around 0.20, indicating enhanced prediction accuracy with more extensive sampling. This result suggests that higher sampling rates may be beneficial for capturing more nuanced details in video frame prediction.
- A reduced model size, using approximately **two-thirds of the original model capacity**, resulted in a Jaccard score of about 0.06. This significant drop highlights the importance of model complexity in capturing the dynamics of video scenes effectively.
- Our final approach with **1000 samplings** achieved the best results with a Jaccard score of 0.36 on the validation dataset.

Figures 4 and 5 below displays some of the predictions. This shows some limitations of the models as not all objects are captured. The predicted image is of a difference scale compared to the actual image which is a limitation of our model.

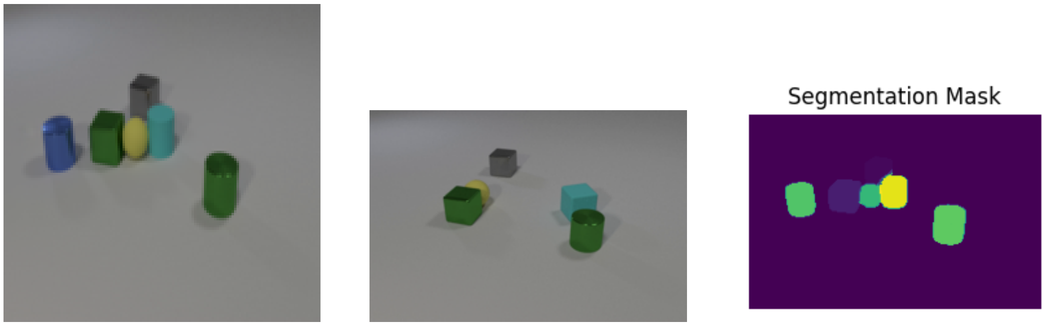

Figure 4: From left to right: Predicted frame, ground truth, Segmentation Mask of the 22nd Frame

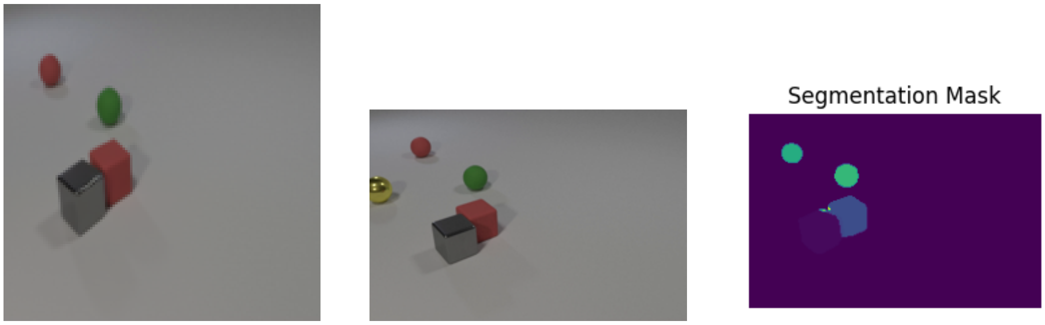

Figure 5: From left to right: Predicted Image, ground truth, Segmentation Mask of the 22nd Frame

4 Conclusion

We have shown that the diffusion model for video future frame prediction and the U-Net model for semantic segmentation generation works well in combination for the given dataset. We found that generating the masks for the frames and then training the diffusion model for frame prediction could give better results. The stength of our model is that it predicts the images without blur compared to models. A limitation of our model is that the images generated are small and the inference time is large to obtain clearer images. Another limitation is that the model misses some objects or adds objects that are not part of the initial video.

References

[1] Mohammad Babaeizadeh, Chelsea Finn, Dumitru Erhan, Roy H Campbell, and Sergey Levine. Stochastic video generation with a learned prior. In *International Conference on Learning Representations (ICLR)*, 2017.

[2] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(4):834–848, 2017.

[3] Will Grathwohl, Ricky Tang, and David Duvenaud. Improved techniques for training score-based generative models. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

[4] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *arXiv preprint arXiv:2006.11239*, 2020.

[5] Nal Kalchbrenner, Aaron van den Oord, and Karen Simonyan. Video pixel networks. In *Advances in Neural Information Processing Systems*, volume 29, 2016.

[6] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[7] Ozan Oktay, Jo Schlemper, Loïc Le Folgoc, Matthew Lee, Mattias Heinrich, Kazunari Misawa, et al. Attention u-net: Learning where to look for the pancreas. In *Medical Image Computing and Computer Assisted Intervention (MICCAI)*, 2018.

[8] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. *arXiv preprint arXiv:1505.04597*, 2015.

[9] Ruben Villegas, Jimei Yang, Yuliang Zou, Sungjin Sohn, Xunyu Lin, and Honglak Lee. Generating videos with scene dynamics. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

[10] Vikram Voleti, Alexia Jolicoeur-Martineau, and Christopher Pal. Mcvd: Masked conditional video diffusion for prediction, generation, and interpolation. *arXiv preprint arXiv:2205.09853*, 2022.

[11] Carl Vondrick, Jon Shlens, and Jian Wang. Video generation with deep generative models. In *Advances in Neural Information Processing Systems*, volume 29, 2016.