Future Frame Mask Prediction using Diffusion & UNet

Amardeep Kumar
New York University
amardeep.kumar@nyu.edu

Ritika Saboo
New York University
rss9311@nyu.edu

Sujana Maithili
New York University
sc10648@nyu.edu

Abstract

This paper introduces a method for predicting future frame segmentation masks in video sequences. We utilized a dataset of 1,000 labeled training and validation videos and 13,000 unlabeled videos, each containing 22 frames. Our approach combines a conditional diffusion model, trained on unlabeled data, with a UNet model, trained on labeled data. The diffusion model sequentially predicts future frames up to the 22nd frame, using preceding frames as input (e.g., predicting frame 12 from frames 1-11). The UNet model then generates segmentation masks for these predicted frames. This methodology demonstrates an advancement in video processing by efficiently integrating frame prediction with accurate mask generation. [1]

1 Introduction

Diffusion models have gained prominence in image and video processing, especially for predicting future frames in video sequences. These models operate on a noising and denoising principle. Initially, they add noise to an image or video frame in a controlled manner, transforming it into a noisy state. This process, termed 'noising,' degrades the frame's content progressively. Remarkably, this transformation is reversible. The denoising process, which is the inverse of noising, involves the model learning to reconstruct the original frame from its noisy counterpart. In image and video prediction tasks, diffusion models excel by generating high-quality future frames. They achieve this by conditioning the denoising process on a sequence of preceding frames, allowing the model to predict the evolution of a scene.

For video prediction, diffusion models adapt to the temporal dynamics of video sequences, capturing and predicting changes across frames. This capability is crucial in applications like autonomous navigation and video editing, where understanding the temporal progression is essential.

2 Related Work

Future frame prediction involves forecasting subsequent frames in a video sequence, focusing on accurately capturing the evolution of visual content over time. In contrast, future mask prediction tasks extend this concept by not only predicting the next frames but also generating their corresponding segmentation masks, delineating specific objects or regions of interest within these future frames.

2.1 SimVP

SimVP represents a streamlined approach in spatiotemporal predictive learning, emphasizing simplicity and efficiency. Built entirely on convolutional networks without recurrent architectures, it is

[1] All authors contributed equally.

Preprint. Under review.

trained end-to-end using common mean squared error loss. This straightforward design allows SimVP to achieve superior performance on various benchmarks without complex strategies or auxiliary inputs. Its reduced training cost and strong generalization capabilities make it a solid baseline model, with extensions like gated spatiotemporal attention further enhancing its performance. SimVP's simplicity and effectiveness position it as an influential model in spatiotemporal predictive learning, particularly suitable for complex real-world scenarios.

2.2 MCVD

Masked Conditional Video Diffusion (MCVD) is a comprehensive framework designed to tackle various video synthesis tasks, including future frame prediction. It utilizes a probabilistic conditional score-based denoising diffusion model, conditioned on past and/or future frames. The unique aspect of MCVD is its training method, where past or future frames are independently masked, enabling a single model to perform diverse tasks like future/past prediction, unconditional generation, and interpolation. Built on non-recurrent 2D-convolutional architectures, MCVD excels in generating high-quality frames for diverse video types. Its block-wise, autoregressive approach allows for the creation of videos of arbitrary lengths. MCVD's efficiency in training and its state-of-the-art results across various benchmarks make it a significant advancement in video prediction and related tasks.

3 Method

Our methodology centers around the deployment of a conditional diffusion model, which is adept at predicting future frames in a sequential manner. Specifically, the model is designed to predict a given frame based on the preceding sequence of frames – for example, predicting the 12th frame using frames 1-11 as the conditional input, and so forth. This strategy is applied iteratively to predict up to the 22nd frame of each video. Notably, the diffusion model is trained exclusively on the unlabeled video data, leveraging its inherent structure for learning frame progression.

Subsequent to frame prediction, we introduce a UNet-based architecture, trained solely on the labeled training data, for the task of generating segmentation masks. This model is employed to produce the segmentation mask for the 22nd frame, which is initially forecasted by the diffusion model.

Our approach combines the strengths of conditional diffusion models for frame prediction with the robust segmentation capabilities of the UNet model.

3.1 Frame Segmentation

The UNet model is a convolutional neural network known for its U-shaped architecture, effectively combining downsampling and upsampling paths with skip connections for precise image segmentation. In our method, we train this model on training data with mask labels containing 49 classes, achieving a Jaccard Index of 0.966 after 15 training epochs.

3.2 Future Frame Prediction

In this methodology, the future frame is predicted conditioned on past frames using conditional diffusion. Once the 22^{nd} frame is predicted, a U-Net segmentation model is employed to generated a mask. For space constraints we would be omitting discussion on segmentation task using Unet.

3.2.1 Forward & Reverse Diffusion Process

Forward Diffusion Process:

- Starts with a data sample x_0 from a data distribution p_{data}.
- Corrupts x_0 over time from $t = 0$ to $t = T$ using the transition kernel:

$$q_t(x_t|x_{t-1}) = \mathcal{N}(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t I)$$

Reverse Diffusion Process:

- Begins from Gaussian noise x_T and reverses the FDP.

- Utilizes the transition kernel:

$$p_t(x_{t-1}|x_t, x_0) = \mathcal{N}(x_{t-1}; \tilde{\mu}_t(x_t, x_0), \tilde{\beta}_t I)$$

3.2.2 Video Prediction via Conditional Diffusion:

- The model directly predicts future video frames given past frames.
- It conditions the diffusion models on past frames to predict current frames using a specific loss function:

$$L_{vidpred}(\theta) = \mathbb{E}_{t,[p,x_0]\sim p_{data},\varepsilon\sim\mathcal{N}(0,I)}\left[\|\varepsilon - \varepsilon_\theta(\bar{\alpha}_t x_0 + (1-\bar{\alpha}_t)\varepsilon|p,t)\|_2^2\right]$$

3.2.3 Experiments:

In our method, we trained the diffusion model on 13,000 unlabeled videos. Each frame is resized from 160×240 into 128×128. Since the segmentation model is trained on the original size, the prediction is resize to the original size. We conducted multiple experiments with different training approaches:

- The first training experiment involved predicting the final 22^{nd} frame directly by conditioning on the first 11 frames. This resulted in a Jaccard Index of 0.198 on validation set. However, it was observed that the loss seemed to stagnate after 30 epochs. This experiment was treated as the baseline.
- The second training experiment involved predicting six future frames i.e., $12^{th}, 14^{th}, 16^{th}, 18^{th}, 20^{th}$ and 22^{nd} frame, conditioned on five past frames i.e., $2^{nd}, 4^{th}, 6^{th}, 8^{th}$ and 10^{th}. This resulted in a Jaccard Index of 0.136 on validation set. However, it was observed that loss seemed not to converge.
- The third training experiment involved autoregressively predicting all future frames conditioned on past 11 frames i.e., predicting the 12^{th} frame conditioned on frames 1-11, then predicting the 13^{th} frame conditoned on frames 2-12 utilzing the predicted 12th frame from the previous step. This resulted in a Jaccard Index of 0.308 on validation set, giving us the best result.

4 Results

In our study, we aimed to build a sequential pipeline of auto-regressively predict future frame conditioned on past frames by utilzing a sliding window over past frames and segmenting the final frame. This gave us the best Jaccard Index of 0.308 on validation set.

In this section, we will focus on the analysis of frame prediction results. Among the three experiments conducted to predict the 22nd frame, autoregressive prediction of the 22nd frame yielded the best result. The training and validation loss curve for this experiment is depicted in Figure 1. Our ambitious experiment aimed at predicting 6 alternate frames given 5 frames, but the loss did not converge. The training and validation curve for this experiment is shown in Figure 2.

Furthermore, We analyzed the frames predicted by the Diffusion models and present their strengths and weaknesses.

Strength: One of the notable strengths is that our model predicted the correct shape without any blur or other noise, a challenge observed in SimVP.

Weakness: First, our diffusion model predicted the exact same frame, but as it was generated in 120 x 120 pixels, resizing it to 160 x 240 pixels caused pixel loss and a lower Jaccard score. An example is shown in Figure 4, where this exactly predicted image only has a Jaccard score of 0.76. The second weakness we noticed is that during the autoregressive prediction of the 12th to 22nd frame, if any of the frames displays an incorrect shape, its effects are seen in the subsequent frames, causing the final frames to deviate significantly from the original frame.

5 Future Work

Some future ideas that we would like to suggest are:

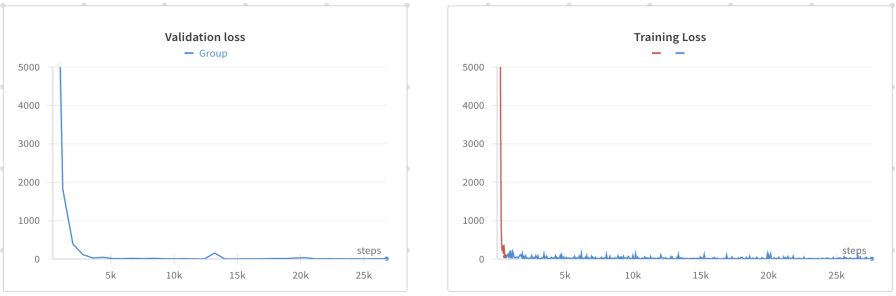

Figure 1: Training curves for auto-regressive prediction

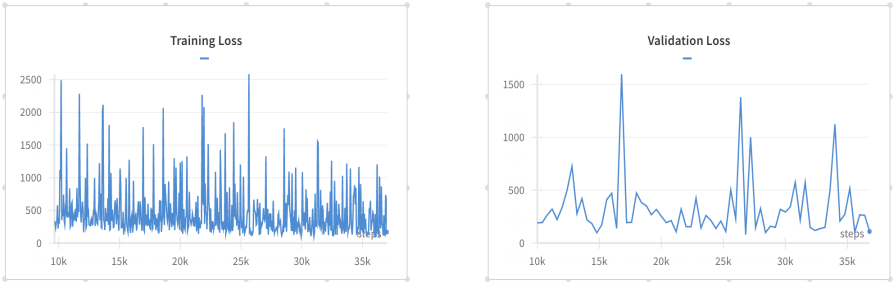

Figure 2: Training curves for predicting 6 alternate frames based on 5 frames

1. Perform these experiments in a different setting, where we will create segments of the frames first and then perform future frame generation on the segmented frames, instead of using raw video frames.
2. The diffusion model converged very well during training, but it produced the best results when the number of subsampling steps during diffusion was kept at 1000. This made the inference of the diffusion process very inefficient and time-consuming.
3. The appearance of new elements after the 11th frame was really confusing for the diffusion model. We would like to filter these datasets to improve the quality of frame prediction.

References

[1] Vikram Voleti, Alexia Jolicoeur-Martineau and Christopher Pal, "MCVD: Masked Conditional Video Diffusion for Prediction, Generation, and Interpolation", arXiv:2205.09853

[2] Zhangyang Gao, Cheng Tan, Lirong Wu and Stan Z. Li, "SimVP: Simpler yet Better Video Prediction", arXiv:2206.05099

[3] Olaf Ronneberger, Philipp Fischer and Thomas Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation", arXiv:1505.04597

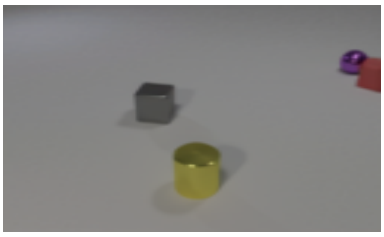

Figure 3: 22nd predicted frame

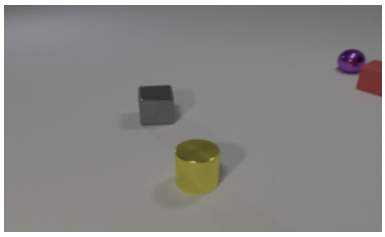

Figure 4: Ground truth