Video Frame Prediction and Semantic Segmentation

Sharad Dargan
Center for Data Science
New York University
sd5251@nyu.edu

Sargun Nagpal
Center for Data Science
New York University
sn3250@nyu.edu

Chaitali Joshi
Center for Data Science
New York University
cj2407@nyu.edu

Abstract

In this study, we used deep learning to model spatio-temporal interactions between diverse objects with varying shapes, materials, and colors based on basic physics principles, across an 11-frame video. The aim was to predict the 22^{nd} frame and generate a semantic segmentation mask for precise object classification. For frame prediction, we trained advanced deep learning models in a self-supervised manner, finding SimVP, a CNN-based architecture, particularly adept at capturing dynamic scene changes and excelling in predicting the 22^{nd} frame. In semantic segmentation, we implemented UNet, a widely-used CNN architecture, which proved effective in accurately classifying objects in the frame.

1 Introduction and Related Work

Video frame prediction revolves around predicting future frames through the comprehension of object interactions across prior frames, necessitating the grasp of temporal and spatial dependencies to extrapolate and forecast subsequent frames. Conversely, semantic segmentation involves the classification of individual pixels within an image or video frame into precise semantic categories, thereby offering a comprehensive breakdown of scene composition and detail.

The realm of video frame prediction has garnered significant attention within the deep learning community due to its diverse applications, such as autonomous driving, traffic flow estimation, human motion projection, and the facilitation of representation learning. For instance, a practical application involves the anticipation of a potential collision between a car and a pedestrian by analyzing their respective trajectories. Prior research endeavors in this domain have been categorized into four distinct groups based on the architectural composition involving encoders, translators, and decoders[1].

The first category, comprised of CNN-CNN-CNN architectures, relies solely on convolutional mechanisms. Notable models within this category, like SimVP[2] and DVF[3], demonstrate high efficiency in their predictive capabilities. The second category, CNN-ViT-CNN architectures, incorporates a Vision Transformer to capture spatiotemporal features. Examples of models falling under this category include Swin LSTM[4], Latent Video Transformer[5], and VideoGPT[6]. Recent efforts such as TAU[7] and VPTR[8] have been directed towards enhancing the computational efficiency of Transformer models for video prediction tasks. The third category, CNN-RNN-CNN models, leverages variations of the LSTM network to forecast future latent states via recurrent neural networks (RNNs). Distinguished models in this category, such as E3D-LSTM[9], and PhyDNet[10] exhibit predictive capabilities using RNN-based methodologies. Lastly, the fourth category, RNN-RNN-RNN architectures, relies entirely on recurrent networks for frame prediction. Models like ConvLSTM[11] and PredRNNv2[12] operate within this category, demonstrating the effectiveness of recurrent network structures in this domain.

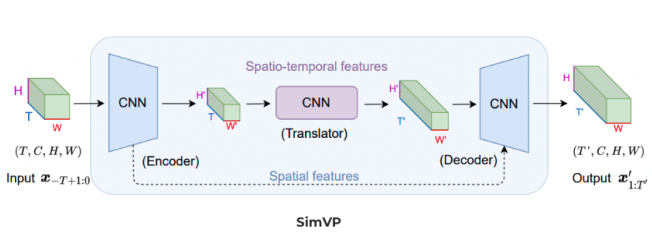
DS-GA 1008 Deep Learning Final Project Report.

2 Dataset

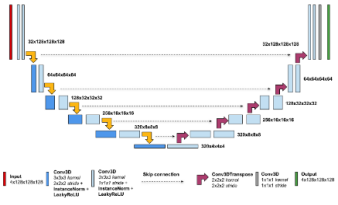
The dataset utilized in this project has three constituents: unlabeled videos, labeled training videos, and labeled validation videos. There are a total of 13,000 unlabeled videos that have 22 frames each and do not have a corresponding ground truth segmentation mask. The labeled training and validation videos both contain 1000 videos each, with each video having 22 frames. For the labeled videos, a full segmentation mask is provided for each frame. Each of the videos contains simple 3D shapes that interact with each other following the basic rules of physics. Objects have three shapes, namely cube, sphere, and cylinder, eight colors, namely gray, red, blue, green, brown, cyan, purple, and yellow and two materials, namely metal and rubber. No video contains duplicate objects. The task is to predict the semantic segmentation mask of 22^{nd} frame given the first 11 frames on a hidden set. Each of the frames has a dimension $(160 \times 240 \times 3)$.

3 Methods

We used a two-step approach for predicting the mask of the 22nd frame of the videos. First, we predicted the future 11 frames from the first 11 frames using a frame prediction model (SimVP) and then used these predictions as an input to the segmentation model (U-Net) to predict the final semantic segmentation masks. We trained each model independently.



SimVP



UNet

3.1 Video Frame Prediction

Architecture: We opted for the SimVP architecture to train our frame prediction model as it is the most commonly used architecture for this purpose. SimVP relies on a Convolutional Neural Network (CNN) design that consists of three components. The spatial extractor extracts spatial features from the input frames, the translator is responsible for extracting temporal features crucial for understanding the video's dynamic evolution and the decoder upsamples the frames generated by the translator, contributing to the final predicted frames, thus, combining the spatial and temporal patterns extracted by the previous modules. The encoder model has 4 blocks of Conv2D, GroupNorm, and LeakyReLU. The translator consists of 8 Inception modules with every module having a different kernel size from the set (3,5,7,11). Finally, the decoder model consists of 4 blocks of each made up of ConveTranspose2D, GroupNorm and LeakyReLU.

Training Details: We fed the first 11 frames of the video as input to the SimVP model and the model's task was to predict the next 11 frames as output, each having dimension $(160 \times 240 \times 3)$. The training process spanned 75 epochs. We utilized the Mean Squared Error (MSE) Loss as the standard loss function. To regulate learning during training, we employed an On-cycle Learning Rate (LR) scheduler.

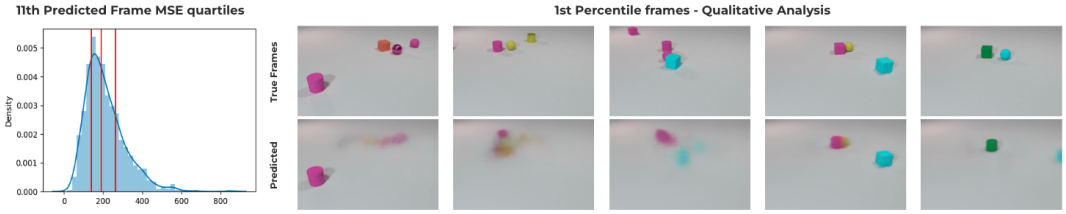
Enhancements: In addition to SimVP, we experimented with various Metaformer models as potential translators in SimVP, leveraging the OpenSTL framework. We also tried to replace the standard MSE Loss with a frame-weighted MSE. This involved assigning the highest weight to the 22nd frame in a harmonic progression. However, this modification did not yield practical improvements. Exploring alternative approaches, we experimented with predicting only the 22nd frame. However, both of these techniques proved impractical in practice.

3.2 Semantic Segmentation

The segmentation task was addressed using the UNet architecture, a proven and widely adopted convolutional neural network (CNN) design. UNet comprises three main components: an encoder, a bottleneck, and a decoder. The encoder captures spatial features from the input, the bottleneck consolidates these features, and the decoder reconstructs the spatial information. The encoder consists of 4 blocks each containing 2 stacks of Conv2D, LayerNorm and ReLU with MaxPooling applied after each block. The decoder consists of 4 blocks each containing a ConvTranspose2D, and 2 stacks of Conv2D, LayerNorm and ReLU. The output from decoder is passed through a Conv2D layer with kernel size$=1$. It was observed that the final predictions from SimVP were blurred. Hence, to train UNet on images that were similar to SimVP outputs, we adopted blur augmentation on images input to UNet. The model was trained for 20 epochs with cross-entropy loss as the classification involved classifying each pixel into one of the 49 classes possible.

3.3 Error Analysis

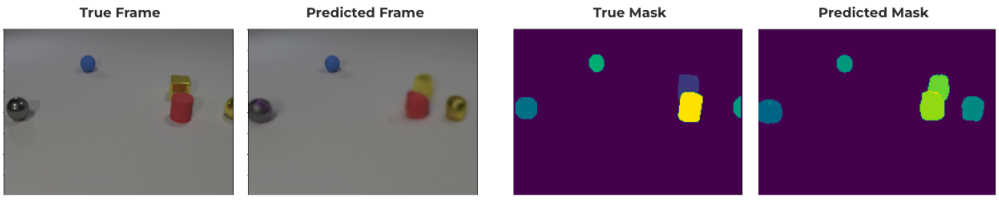
Since the segmentation model performed well with an Average IoU of 0.96 over the validation set, it was clear that the scope of improvement was higher in the Frame prediction models. We performed a qualitative error analysis to identify the scope of further improvement and found that the SimVP could only accurately predict static objects in the video frames. For moving objects, it predicted an average blurry prediction. Future work will address improving the frame prediction model.



4 Results

The segmentation model with UNet architecture performed well with an Average IoU of 0.96 over the entire validation set. For the Video Prediction task, Table 1 shows that sequential and transformer-based models exhibited approximately four times the training time per epoch compared to convolution-based models. SimVP2 with gated spatio-temporal attention, achieved the lowest MSE, but the original SimVP model outperformed in terms of the Jaccard index with the full pipeline. Note that, Table 1 metrics are reported after 20 epochs of each model.

The final pipeline resulted in a Jaccard Index of `0.2721` on the validation set and `0.2727` on the hidden set in the final leaderboard of the competition.



5 Conclusion & Future Work

This project successfully developed a deep learning pipeline capable of predicting segmentation masks with high accuracy, particularly for non-moving objects in video sequences. The core strengths of this system lie in its robust segmentation model and the individual efficiency of the Video Prediction and Segmentation models. However, challenges arise when these models are combined, leading to a noticeable distribution shift, which leads to a high error.

	Model	Train MSE	Val MSE	Wall Time (h) / Epoch	Jaccard Index
1	SimVP	51.29	67.92	0.6	**0.24**
2	SimVP + gSTA (SimVP2)	**48.89**	**47.55**	**0.5**	0.23
3	SimVP + Swin Transformer	127.25	126.76	2.2	0.21
4	SimVP + ViT	127.20	126.75	1.9	0.22
5	SimVP + PredRNNv2	121.61	180.31	2.3	-

Table 1: Metaformer models for SimVP

In future work, this pipeline could be significantly enhanced by incorporating labeled training data to improve the accuracy of the frame prediction model. Additionally, implementing a filtering mechanism to exclude videos where new objects appear after the initial 11 frames would allow the model to focus on more consistent scenarios and improve prediction accuracy. There is also scope for post-processing the model outputs that will improve the final Jaccard Index.

References

[1] Zhangyang Gao, Cheng Tan, Lirong Wu, and Stan Z. Li. Simvp: Simpler yet better video prediction, 2022.

[2] Cheng Tan, Zhangyang Gao, Siyuan Li, and Stan Z Li. Simvp: Towards simple yet powerful spatiotemporal predictive learning. *arXiv preprint arXiv:2211.12509*, 2022.

[3] Ziwei Liu, Raymond A Yeh, Xiaoou Tang, Yiming Liu, and Aseem Agarwala. Video frame synthesis using deep voxel flow. In *Proceedings of the IEEE international conference on computer vision*, pages 4463–4471, 2017.

[4] Song Tang, Chuang Li, Pu Zhang, and RongNian Tang. Swinlstm: Improving spatiotemporal prediction accuracy using swin transformer and lstm. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 13470–13479, October 2023.

[5] Ruslan Rakhimov, Denis Volkhonskiy, Alexey Artemov, Denis Zorin, and Evgeny Burnaev. Latent video transformer. *arXiv preprint arXiv:2006.10704*, 2020.

[6] Wilson Yan, Yunzhi Zhang, Pieter Abbeel, and Aravind Srinivas. Videogpt: Video generation using vq-vae and transformers. *arXiv preprint arXiv:2104.10157*, 2021.

[7] Cheng Tan, Zhangyang Gao, Lirong Wu, Yongjie Xu, Jun Xia, Siyuan Li, and Stan Z Li. Temporal attention unit: Towards efficient spatiotemporal predictive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18770–18782, 2023.

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

[9] Yunbo Wang, Lu Jiang, Ming-Hsuan Yang, Li-Jia Li, Mingsheng Long, and Li Fei-Fei. Eidetic 3d lstm: A model for video prediction and beyond. In *International conference on learning representations*, 2018.

[10] Vincent Le Guen and Nicolas Thome. Disentangling physical dynamics from unknown factors for unsupervised video prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11474–11484, 2020.

[11] Zhihui Lin, Maomao Li, Zhuobin Zheng, Yangyang Cheng, and Chun Yuan. Self-attention convlstm for spatiotemporal prediction. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 11531–11538, 2020.

[12] Yunbo Wang, Haixu Wu, Jianjin Zhang, Zhifeng Gao, Jianmin Wang, S Yu Philip, and Mingsheng Long. Predrnn: A recurrent neural network for spatiotemporal predictive learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(2):2208–2225, 2022.