Innovative Approaches in Semantic Segmentation and Video Prediction

Ragnor Wu, Weikai Jin, Brandon Gao
Deep Learning SP24
ww2342@nyu.edu
wj2246@nyu.edu
bsg9679@nyu.edu

May 13, 2024

Abstract

This study introduces innovative models for semantic segmentation and video prediction, employing U-Net and Convolutional LSTM (ConvLSTM) to address the challenges of analyzing complex dynamic scenes. Our approach enhances semantic segmentation with U-Net, optimized for precise pixel classification, and leverages ConvLSTM for its superior capability in integrating spatial and temporal data for video prediction. Through extensive experimentation, our models have demonstrated high efficacy in predicting dynamic sequences and segmenting intricate scenes. We discuss the integration of attention mechanisms as a future enhancement to further improve the consistency and accuracy of predictions, highlighting the potential of our approach in advancing real-world applications in video analysis and autonomous systems..

1 Introduction

Semantic segmentation and video prediction are two complementary fields in computer vision that help systems analyze and anticipate visual information. Semantic segmentation focuses on labeling every pixel in an image according to its class, enabling detailed scene understanding vital for applications like autonomous driving and surveillance. It ensures autonomous vehicles can identify and differentiate between roads, pedestrians, and vehicles, improving navigation and safety. Meanwhile, video prediction forecasts future frames based on previous ones, enhancing motion tracking and behavior anticipation. This forecasting ability is essential for proactive responses in real-time surveillance, robotics, and traffic management systems. Together, these fields advance the broader goal of enabling intelligent systems to comprehend and anticipate visual environments.

2 Literature Review

Semantic segmentation and video prediction have rapidly advanced, propelled by increased computational capabilities and refined algorithms. These fields employ diverse methods to exploit temporal and spatial data features, enhancing scene understanding and future state predictions.

2.1 Past Semantic Segmentation Techniques

2.1.1 Keyframe-Based Approaches

Methods such as Temporal Feature Correlation (TFC) and Feature Alignment for Semantic Segmentation in Videos (FASSVid) utilize temporal data from preceding frames to stabilize and refine predictions, improving accuracy in dynamic environments.

2.1.2 CNN-Based Architectures

Fully Convolutional Networks (FCNs) replace fully connected layers with convolutional ones to enable end-to-end training and precise pixel-wise classification. Techniques like DeepLab utilize dilated convolutions to capture detailed features without losing perspective, while Mask R-CNN enhances object-level segmentation by integrating region proposals with segmentation maps.

2.1.3 Self-Supervised Learning

This approach reduces reliance on manual annotations by employing pseudo-labeling techniques, which generate labels from the inherent properties of the frames, facilitating training on unlabeled data.

2.2 Past Video Prediction Techniques

2.2.1 Optical Flow-Based Prediction

This method estimates pixel movement across frames to track motion trajectories, enhancing scene understanding.

2.2.2 RNN-Based Models

Recurrent Neural Networks, including Long Short-Term Memory (LSTM) networks, analyze temporal sequences to predict future frames accurately, learning from the dependencies present in video data.

2.2.3 GAN-Based Prediction

Generative Adversarial Networks (GANs) model video frame distributions to generate realistic future frames, significantly improving prediction realism through adversarial training.

3 Methodology

3.1 Overview

We present a dual-stage approach integrating semantic segmentation with video prediction, aiming to enhance frame sequence forecasting by leveraging the clarity of segmentation masks. This strategy benefits from simplified data complexity and improved efficiency in temporal processing.

3.2 Process Description

3.2.1 Semantic Segmentation

Figure 1: Prediction Result of U-Net

Using the U-Net architecture, we segment the initial 11 frames of each video. U-Net, recognized for its effective encoder-decoder structure, classifies each pixel in these frames into predetermined categories, preparing the data for subsequent predictive modeling.

3.2.2 Video Prediction

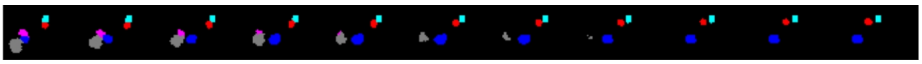

Figure 2: Predicted Result of ConvLSTM

Figure 3: Ground Truth

We employ a Convolutional LSTM (ConvLSTM) model with dense connections to predict future frames' segmentation masks up to the 22nd frame. This model choice is informed by its proficiency in capturing spatial and temporal dependencies within the data.

3.3 Data Preparation

Segmentation masks offer simplified visual data that facilitates more accurate and efficient predictions. We process these masks into one-hot encoded formats and resize them to reduce computational demands. The ConvLSTM model processes batches of these adjusted frames, predicting the evolution of segmented scenes.

3.4 Model Specifications

Our ConvLSTM framework features three encoder and three decoder layers, enhancing data flow through dense connections. This architecture supports comprehensive contextual understanding, essential for accurate future frame predictions.

3.5 Implementation Notes

Our approach is designed for efficiency, handling simplified masks without textual details to speed up the computation. It is adaptable for varying complexities and can be scaled for broader applications in real-time environments.

4 Conclusion

Our evaluation of the U-Net and ConvLSTM models on the task of predicting the 22nd frame from initial video segments resulted in an Intersection over Union (IoU) Loss Score of 0.34. This score indicates a reasonably good performance in predicting semantic segmentation in dynamic scenes.

4.1 Interpretation of Results

The achieved IoU score demonstrates that our models can effectively predict future frames with a significant level of accuracy. This suggests that the architectural choices, particularly the use of ConvLSTM for capturing temporal dynamics and U-Net for detailed segmentation, are well-suited for tasks involving complex scene dynamics.

4.2 Limitations and Future Work

One of the main constraints during this research was the unavailability of unlabeled data for training. Given the self-supervised nature of parts of our

methodology, incorporating unlabeled data could potentially enhance the model's learning capability and improve performance.

5 Visualization & Discussion

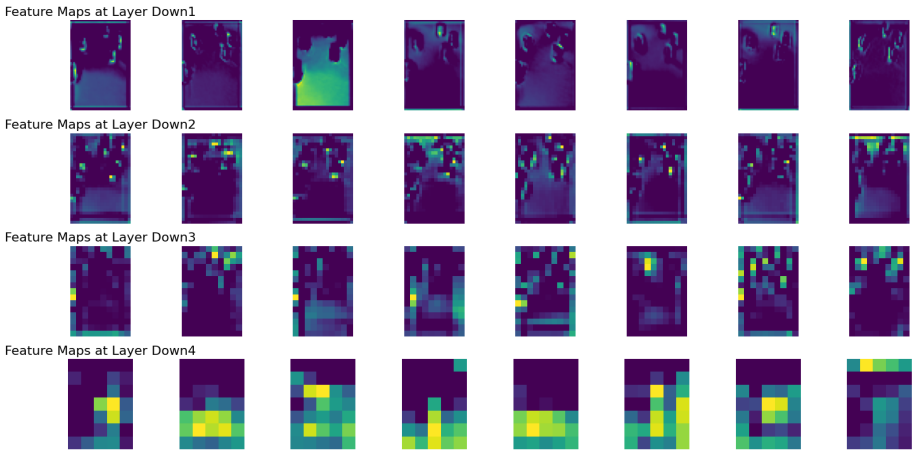

Figure 4: Feature Visualization

Based on the visualization of the feature maps, we make interpretations as below. For the Early Layers (Down1, Down2), they typically capture basic features such as edges, textures, and simple shapes. This is evident from the more detailed feature maps seen in the first two rows. For the Deeper Layers (Down3, Down4), they show more abstract and high-level representations. These layers aggregate the simple features into more complex patterns, which are less visually interpretable but are crucial for understanding complex scenarios in the image. This enables the model to identify objects and its shapes, thus identifying its class and segment the object from the background and other objects.

Our methodology achieved a decent accuracy since we adopted the way where we first do segmentation, and then do video prediction. Our pipeline is, first we do frame wise semantic segmentation, by a simple U-Net. Then we use a convolutional Long Short Term Memory model with dense connection to perform video prediction on segmentation masks. Which means that this model directly use segmentation masks as input, and directly generate a sequence of segmentation masks of future frames. We choose this method because we think that segmentation masks are clearer well-structured data, rather than the raw videos. It has no textures and useless visual details, and it could be easier for a model to learn. The result turns out to be good as expected.

References

[1] Gadde, R., Jampani, V., & Gehler, P. V. (2017). Semantic Video CNNs through Representation Warping. *Proceedings of the IEEE International Conference on Computer Vision.*

[2] Dosovitskiy, A., Fischer, P., Ilg, E., Hausser, P., Hazirbas, C., Golkov, V., van der Smagt, P., Cremers, D., & Brox, T. (2015). FlowNet: Learning Optical Flow with Convolutional Networks. *IEEE International Conference on Computer Vision (ICCV).*

[3] Lotter, W., Kreiman, G., & Cox, D. (2017). Deep Predictive Coding Networks for Video Prediction and Unsupervised Learning. *arXiv preprint arXiv:1605.08104.*

[4] Vondrick, C., Pirsiavash, H., & Torralba, A. (2016). Generating Videos with Scene Dynamics. *Advances in Neural Information Processing Systems (NeurIPS).*

[5] Ronneberger, O., Fischer, P., Brox, T. (2015). U-Net: Convolutional Networks for Biomedical Image Segmentation. *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI).*

[6] Shi, X., Chen, Z., Wang, H., Yeung, D.-Y., Wong, W.-K., Woo, W.-c. (2015). Convolutional LSTM Network: A Machine Learning Approach for Precipitation Nowcasting. *Advances in Neural Information Processing Systems (NeurIPS).*