Predictive Video Segmentation with PredNet and U-Net - Group 03

Kathleena Inchoco
Center For Data Science
New York University
ki2130@nyu.edu

Rod Aryan
Center For Data Science
New York University
ra2829@nyu.edu

Wei-Han Hsu
Center For Data Science
New York University
wh2405@nyu.edu

Abstract

This study explores the intersection of unsupervised learning and video frame prediction by integrating advanced deep-learning architectures. We present a novel method that combines the strengths of the PredNet architecture, rooted in predictive coding principles from neuroscience, with the capabilities of U-Net, a model highly effective in image segmentation tasks. Our approach harnesses PredNet's ability to predict future video frames and utilizes U-Net to create segmentation masks for these predictions. This synergistic application provides a detailed analysis of video content, capturing both spatial and temporal dimensions comprehensively. The paper outlines the implementation and assessment of this method, demonstrating its potential to enhance video frame prediction and offering a more nuanced understanding of video data.

1 Introduction

The advent of deep learning has revolutionized the field of video analysis, particularly in tasks related to frame prediction and scene understanding. Traditional methods have primarily focused on supervised learning paradigms, relying heavily on large labeled datasets. However, the reliance on extensive labeled data limits the applicability of these methods in scenarios where such resources are scarce. Unsupervised learning, particularly in the context of video frame prediction, offers a promising avenue for learning rich representations of visual data without the need for explicit labeling.

2 Literature Review

In the field of deep learning, unsupervised learning for video analysis has gained increasing attention, especially in scenarios with limited labeled data Lotter et al. [2016]. This shift is driven by the need for models that can learn from and interpret video data without extensive annotated datasets. PredNet, introduced by Lotter et al. [2016], stands out as a significant development in this domain. PredNet's architecture is inspired by predictive coding principles from neuroscience, focusing on the prediction of future video frames. It operates by continually updating its internal model based on the differences between predicted and actual video frames. This methodology has been effective in diverse scenarios, including the prediction of object movements in both synthetic environments and natural image streams, such as videos from car-mounted cameras. PredNet's internal representations have practical applications, like estimating vehicle steering angles, demonstrating its applicability in real-world situations.

Moreover, Lotter et al. [2016] stresses the importance of an architecture's capacity to implicitly understand object structures and their transformations. This perspective marks a departure from conventional static image-based training methods and underscores the significance of temporal dynamics in learning visual representations. Adding to these advancements, the work of Couprie et al.

Preprint. Under review.

[2018] on future instance segmentation in videos has furthered the capabilities of predictive video analysis. Using the Mask R-CNN framework, their model effectively predicts fixed-size, high-level features, enabling accurate instance segmentation for future video frames. This approach has shown to be superior to standard models in predicting variable numbers of instance segmentations.

3 Methodology

3.1 PredNet Implementation and Training

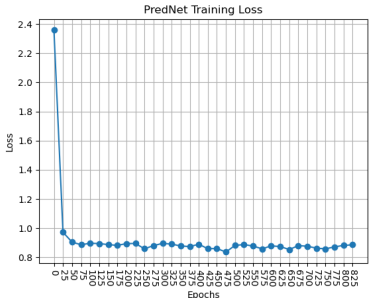

Figure 1: Loss over epochs during PredNet training.

At the heart of our approach lies the PredNet architecture, a model inspired by predictive coding principles in neuroscience (Watanabe [2021]) and tailored for video frame prediction. PredNet's design is particularly adept at capturing the temporal progression within video sequences, making it an ideal choice for predicting future frames. Central to our implementation of PredNet is the ConvLSTM structure, which forms the basis of our model's ability to recognize and predict spatial-temporal features. The training of PredNet, as depicted in Figure 1, demonstrated a notable trend in the loss values over epochs. It was observed that after approximately 100 epochs, the loss value plateaued, maintaining a relatively consistent value of around 0.9. This indicates a potential overfitting of the PredNet model, as no significant improvement in loss reduction was observed beyond this point.

3.2 U-Net Training Performance

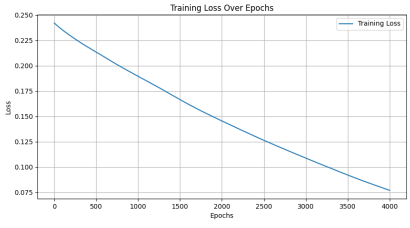

Figure 2: U-Net loss over epochs.

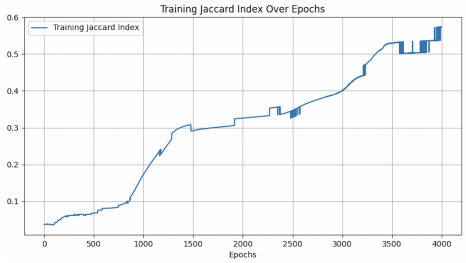

Figure 3: Jaccard Index over epochs.

With the temporal predictions of PredNet is the U-Net model, renowned for its effectiveness in image segmentation. In our study, U-Net's primary role was to generate segmentation masks for the 22nd frame of video sequences as predicted by PredNet. This integration of U-Net allowed us to add a layer of spatial detail to the frames predicted by PredNet, enriching our analysis with precise segmentation of objects within the frames.

To quantitatively assess the performance of the U-Net model, we employed the Jaccard Index, also known as Intersection over Union (IoU), as our primary evaluation metric. This metric provided a clear and quantifiable measure of the accuracy of the segmentation masks relative to the ground truth available in the validation set. The training process for U-Net, illustrated in Figures 2 and 3, showed a steady decline in training loss over time, suggesting a continual learning process. However, the loss never fully plateaued, implying that the model could have potentially benefited from additional training time or an enhanced training setup.

4 Evaluation

4.1 Validation Performance

To assess the performance of our integrated PredNet and U-Net models, we evaluated them on a validation set prior to their submission for the 2nd and final leaderboards. The evaluation metrics used were Validation Loss and Validation Jaccard Index. The results of these evaluations provided insights into the models' performance and their ability to generalize to unseen data.

Table 1: Validation Performance for 2nd and Final Leaderboards

Metric	2nd Leaderboard	Final Leaderboard
Validation Loss	3.9324	3.9811
Validation Jaccard Index	3.4733e-05	2.5247e-05

As shown in Table 1, there was a slight increase in validation loss from the 2nd to the final leaderboard, indicating a marginal reduction in model performance. Concurrently, the Validation Jaccard Index decreased, suggesting a decrease in the accuracy of segmentation predictions. These results highlight areas for potential improvement in model training and optimization for future iterations.

5 Results and Discussion

5.1 Training Outcomes and Model Behavior

The training outcomes for both PredNet and U-Net models reveal crucial insights into their learning capabilities and limitations. PredNet displayed a tendency to overfit after a certain number of epochs, suggesting the need for a more dynamic training approach. Techniques such as early stopping or regularization could be employed to prevent overfitting in future iterations. In contrast, the U-Net model showed continual improvement in both loss reduction and Jaccard Index, indicating ongoing learning. This suggests that extending the training period or adjusting the training parameters could further enhance its performance.

5.2 Leaderboard Review

On the 2nd leaderboard, our model achieved a performance score of 0.0192, which went down to 0.0099 in the final leaderboard. The change potentially reflects some difficulty in the model understanding the correct color in the predicted images. Specifically, in the images predicted by the model submitted to the final leaderboard, the frames look more brown. This could be due to it "averaging" all colors rather than distinguishing between them. This alteration in the model output reflects the changes in the training set-up made to the model between these two stages (severe over-training). A comparative analysis of the prediction images from the models responsible for the 2nd and final leaderboard submissions provides a visual representation of the model's evolution. The 2nd leaderboard model prediction, shown in Figure 4, exhibits a light gray color grading, while the final leaderboard model prediction, as seen in Figure 5, displays a dark shade of brown.

Figure 4: Prediction from the 2nd leaderboard model.

Figure 5: Prediction from the final model.

Although the objects within the dark brown image of the final leaderboard model appear more defined, the predominance of brown coloration likely contributed to the lower score. This outcome suggests a need for further refinement in the color grading and object distinction capabilities of the model.

5.3 U-Net Segmentation Analysis

Figure 6 showcases an instance of U-Net's segmentation capabilities. This example highlights U-Net's proficiency in accurately segmenting objects, especially those with distinct color profiles and non-chrome colors in the foreground. Such effectiveness indicates that U-Net, with its robust segmentation abilities, can perform exceptionally well when provided with an optimal training cycle.

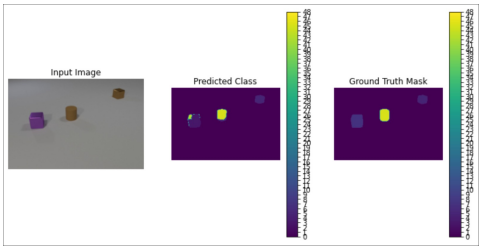

Figure 6: Example segmentation image from U-Net.

The strong performance of U-Net in this context underscores the significance of proper training setup and parameter tuning. Our study suggests that the limitations observed in the overall project performance are likely attributable to factors related to the training regimen of PredNet or the integration of the two models, rather than the capabilities of U-Net itself. Future work could focus on refining the training process and exploring more effective ways to harness U-Net's segmentation strengths, potentially leading to further improvements in model accuracy and generalization.

6 Conclusion and Future Work

Our method leverages the predictive coding capabilities of PredNet for future video frame prediction and the segmentation efficiency of U-Net. The results demonstrated that this combination provides a comprehensive analysis of video content, capturing both spatial and temporal dimensions. However, challenges such as the overfitting tendency in PredNet and the need for optimal training cycles in U-Net were identified. Looking forward, we aim to refine our approach by incorporating strategies discussed in the literature review, particularly the recommendation of predicting on segmentation masks instead of raw frames. This strategy could address the issues observed in our model, such as the 'brown filter' effect on our predicted images, by providing a more focused and relevant training target for the network. By optimizing the training cycle, we anticipate improvements in model performance, particularly in terms of segmentation accuracy and the ability to handle a wider range of visual scenarios.

Further research will also delve into enhancing the integration of PredNet and U-Net, ensuring that the combined strengths of both models are utilized more effectively. This may involve fine-tuning the interplay between the predictive and segmentation components of the system to better handle complex video sequences. Ultimately, our goal is to develop a more robust and versatile system for video frame prediction and segmentation, one that can adapt to diverse visual environments and offer more accurate and detailed insights into video content.

References

Luc Couprie, Yann LeCun, and Jakob Verbeek. Predicting future instance segmentation by forecasting convolutional features. *arXiv preprint arXiv:1803.11496*, 2018.

William Lotter, Gabriel Kreiman, and David Cox. Deep predictive coding networks for video prediction and unsupervised learning. *arXiv preprint arXiv:1605.08104*, 2016.

Eiji Watanabe. Predictive coding deep neural network in pytorch. figshare. Software, 2021. URL `https://doi.org/10.6084/m9.figshare.16910767.v1`.