BRNO UNIVERSITY OF TECHNOLOGY FACULTY OF INFORMATION TECHNOLOGY

http://fit.vut.cz

POVa project 2023/2024

Semi-automatic image segmentation

Patrik Tiszai, Marek Bahník, Marek Hlavačka

Abstract

Object segmentation on a static image is a decades long problem that has already many functioning solutions and applications. Our goal is to apply the discovered knowledge and use it to create a CNN model, that can perform segmentation on frames based on previous frame segmentation and therefore segment the object in a video. We have applied the U-Net principles as well as pre-trained *Resnet* backbone to crate a functioning architecture. The skip connections have been upgraded from default with concatenation of skip connection data from two *Resnet* encoders - one encoding the current image and the other merge of previous frame segmentation and and image. Our model has reached an approximate F1 score of 0.25 which is not very good, but there are lot more places to improve on. The segmentation of the image is not precise and has a lot of fragments and noise. This may result from small and not very general data set or improper model training. This model can be used to track objects on a video, where precise segmentation is necessary. This is the first iteration of the problem solution and we can expect better results with more robust data sets and better training parameters.

Keywords: U-Net — Segmentation — Pytorch

Supplementary Material: Github link

Faculty of Information Technology, Brno University of Technology

1. Introduction

Our project focuses on building an application that can track moving objects, especially people, in videos. This app works by outlining the objects and following them as they move. We're testing it on various inputs to ensure it accurately keeps up with these objects as they change positions. This technology has potential applications in fields where precise monitoring in videos is essential.

Figure 1. Image merged with the mask

2. Datasets

For the purpose of the project we found the DAVIS (Densely Annotated Video Segmentation) data set the most suitable. It is commonly used in the sphere of video image segmentation. For the project the 480p version data set has been used with 60 training and 30 validation sequences. [2] Subsequently, we decided to use augmentation and thus expand our set to 120 training sequences. Next, we created a mask and frame join for each sequence to separate the segmented object from the background. The resulting image can be seen in figure 1. Thus, our data loader returns the past frame of the sequence modified according to image 1 together with the current frame and mask as ground Truth. This dataset however - as we will discuss further - is not big enough for our purposes. Therefore for further improvements of precision of this model we would need to highly increase the size of this data set as well as augment the existing data in proper way.

3. Model architecture

Our model has been created using Python mainly using **Pytorch** package. The source code can be downloaded from the link in the abstract. For model architecture we have chosen the principles of the U-Net architecture. [3] This is concept based on fully convolutional neural networks. It is mainly used for image segmentation, which is exactly our use case.

3.1 Encoder

As encoder, we have used the *Resnet34* encoder architecture as seen on figure 2. [1] Using pre-trained and already functioning backbone made the architecture a bit simpler. We have used the backbone two times. First backbone usage is for the merged image of the previous frame seen in 1. The second one is used for the current frame.

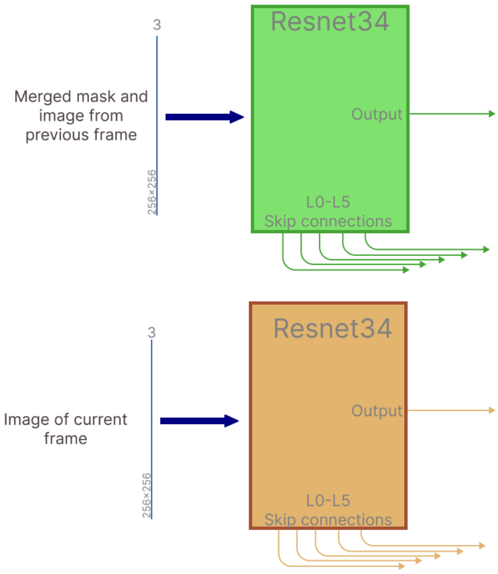

Figure 2. Resnet34 backbone inputs and outputs

The output of these two backbones is concatenated together as an input into a *bridge*. Our bridge consists of four 3x3 convolutions. The input and output channels of the bridge are the same.

3.2 Decoder

The decoder is manually created to fit the encoder skip connection dimensions. You can see the encoder architecture on the figure 3. The skip connections from both backbones are concatenated on every level and then again concatenated with appropriate connection from the previous convolution layer.

4. Conclusion

Our goal for this project was to create a CNN, that can predict object segmentation on a video frame given object segmentation from previous frame and current image. The results were underwhelming, which we suspect are because of poor data set as well as improper model training. But in general, this project shows, that this type of model and architecture can

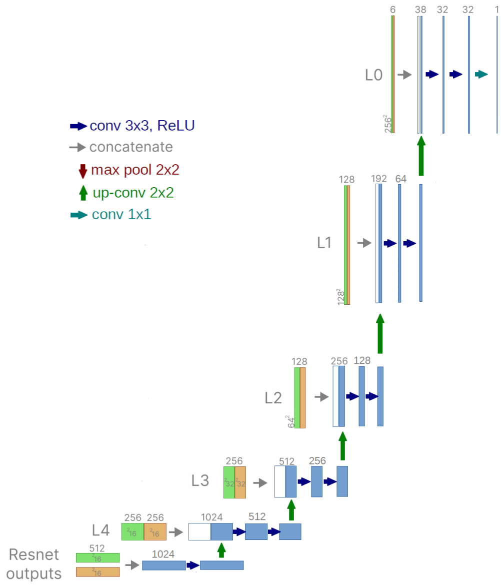

Figure 3. Decoder architecture

work with larger and more robust data sets as well as better training methods for the model. On figure 4 you can see the comparison of one of the better results of the model prediction to the ground truth. Given the fact that we are using the output of the first frame as the input to the next frame, the error accumulates and therefore in longer videos the mistake gets larger and larger.

Figure 4. On the left hand image you can see ground truth, right hand image is our model output.

References

[1] HE, K., ZHANG, X., REN, S. and SUN, J. Deep Residual Learning for Image Recognition. In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016, p. 770–778.

[2] PERAZZI, F., PONT TUSET, J., MCWILLIAMS, B., VAN GOOL, L., GROSS, M. et al. A Benchmark Dataset and Evaluation Methodology for Video Object Segmentation. In: *Computer Vision and Pattern Recognition*. 2016.

[3] RONNEBERGER, O., FISCHER, P. and BROX, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. In: NAVAB, N., HORNEGGER,

J., WELLS, W. M. and FRANGI, A. F., ed. *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*. Cham: Springer International Publishing, 2015, p. 234–241. ISBN 978-3-319-24574-4.