Final Project: Unguided Densely Annotated Video Segmentation

EECS 542 Advanced Topics in Computer Vision
Group members: Yuhao Tang, Haoyu Yu, Hong Moon, Cheng Ouyang

1 Problem Description

We chose the unguided DAVIS segmentation[1] as our problem. The objective of this problem is to give pixel-wise segmentation of one major moving object in a video without knowing the first frame of segmentation.

By observing the videos, we found that most of the videos have a single major object, which is believed to be easily segmented out by conventional single-image FCN[2]. However, in some videos, there are multiple distinguishable objects, with a moving object and several static objects (with respect to the background). In most cases, the segmentation object is spatially larger than other potential distinguishable objects.

Based on the observation, we believe the key to solving this problem lies in two directions 1. detecting a spatially major object in a video with pixel-wise segmentation against the background and other small objects. 2. Detecting the object which is moving, i.e. has different motion pattern against the background, and is always present throughout video.

2 Model Description and Training Strategy

To validate our analysis and solve this problem, we proposed two deep learning-based models, based on the foundation of fully convolutional network[2] for pixel-wise image segmentation. The first model, which is called Two-stream LSTM (TSL), tries to capture the "always present" nature of the target as well as the difference of motion pattern between the target and background (introduced by distance difference to the camera). The other model, Two-stream Masked CNN (TMC), tried to capture the spatially largest and most distinguishable nature of the target object. Both models were written using TensorFlow, and trained using the Nvidia K40m GPU on Flux and Nvidia K80 GPU on Google Cloud Computation Engine.

2.1 Two-stream LSTM

As discussed above, the core idea of TSL is to capture the always-present temporal nature and the different motion pattern of the target. A natural solution leveraging this temporal nature is using recurrent neural network (RNN)[3] or long-short term memory (LSTM) [4], which have the capability to "remember" the high-level representation of previous inputs. We hope the LSTM could "remember" and "focus" the object which is always present.

Meanwhile, since the moving object and the background has different motion pattern due to different distance with respect to the camera, we argue that optical flow, as a characterization of motions between two frame for each pixel, provides additional information for the network.

The structure of this model is shown as follows: During a forward pass, 5 consecutive images with size 480 × 854 are fed into the input network sequentially. An FCN-8 network resize the input and generate a primitive foreground-background segmentation of size 60 × 107. The primitive segmentation contains the segmented objects in the frame, regardless of their motion. The primitive segmented background is

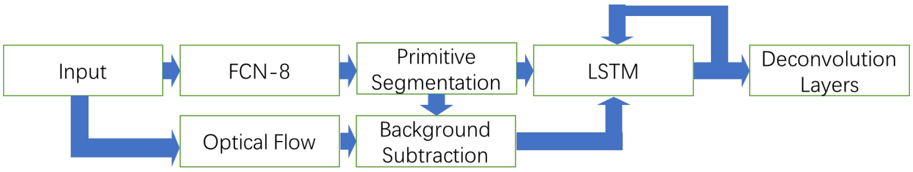

Figure 1: Structure of Two-stream LSTM

a subset of real background. Here we loosely assume the optical flow of background is generally different from that of foreground, and roughly homogeneous. Therefore, the primitive background is used as a mask for the optical flow to subtract its mean background value, and let the foreground stand out. Then, the optical flow is concatenated as additional feature layers to the primitive segmentation, and feed into the LSTM unit. LSTM ”remembers” what has been present in previous frames, and the output is deconvoluted in the same way as conventional FCN-8. Due to the limited computation resource and time, we didn’t finish the LSTM unit, and trained a ResNet[5] instead of FCN-8.

2.2 Two-stream Masked CNN

As discussed above, the core idea of TMC is to utilize the spatial size property of segmentation target. Therefore, if all distinguishable objects can be found in the image, this primitive segmentation could serve as additional information for segmentation and reduce the model search space. In the proposed model, this information is in the form a bounding box mask, generated by the *Single-shot MultiBox Detector* (SSD) proposed by *Liu et. al* [6]. The structure of the network is shown as follows:

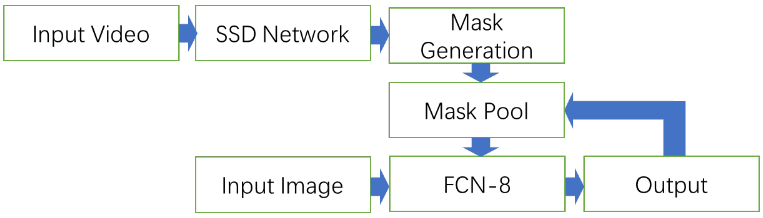

Figure 2: Structure of Two-stream Masked CNN

During a forward pass, every frame of input video is firstly fed into the SSD to generate initial multi-box annotation for each detected object. Then, the Mask Generation unit converts the largest bounding box into a binary mask to indicate the rough location of segment target. If multiple objects are detected, the mask generator will choose the spatially largest bounding box or overlapped bounding box union (e.g., a person riding a horse will generate overlapping bounding boxes of rider and the horse, these are regarded as one). Note that not all the frames have at least one bounding box. The temporally closest mask is concatenated with the input image, and then fed into the FCN-8 network. The segmentation result, in turn, is stored into the mask pool if there is no such a mask for this frame.

3 Result and Evaluation

3.1 Summery of Performance

We tested our training result on the evaluation set of the DAVIS 2016 dataset. We compared our result with the methods listed on the website. Our TMC model **beats the state-of-arts**.

	TSL	TMC	FST	KEY	MSG	CVOS	NLC
J Mean ↑	47.6	**69.2**	55.8	49.8	53.3	48.2	55.1
J Recall ↑	51.8	**82.4**	64.9	59.1	61.6	54.0	55.8
J Decay ↓	-0.0	**2.2**	0.0	14.1	2.4	10.5	12.6
F Mean ↑	40.2	**66.7**	51.1	42.7	50.8	44.7	52.3
F Recall ↑	35.5	**78.1**	51.6	37.5	60.0	52.6	51.9
F Decay ↓	-0.0	**2.0**	2.9	10.6	5.1	11.7	11.4

Table 1: Comparison of our results against other unsupervised methods

3.2 Sample Segmentations

3.2.1 Unfinished Two-stream LSTM

The results of unfinished TSL model are shown as in Figure 3 to Figure 4.

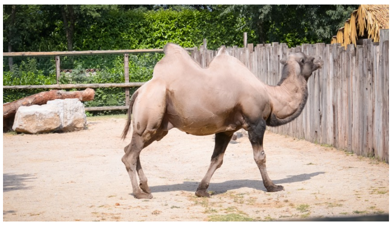

Figure 3: Original Image

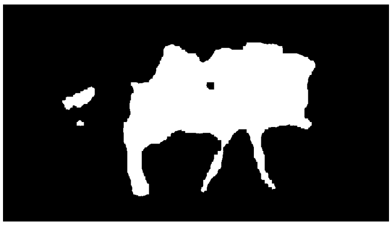

Figure 4: Primitive Result

3.2.2 Tow-stream Masked CNN

The results of TMC are shown in Figure 5 to Figure 10

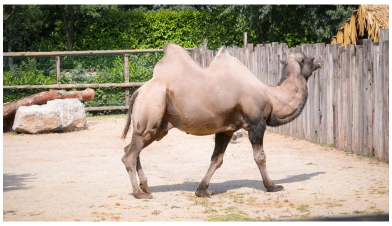

Figure 5: Original Image

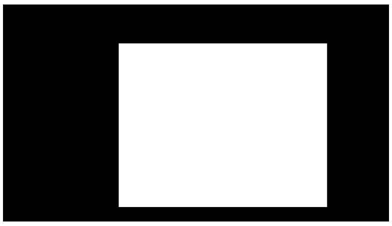

Figure 6: SSD Mask

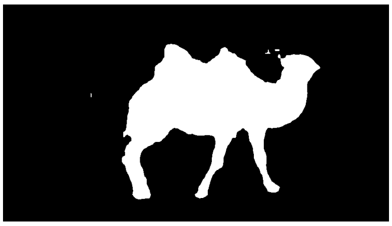

Figure 7: Good Result

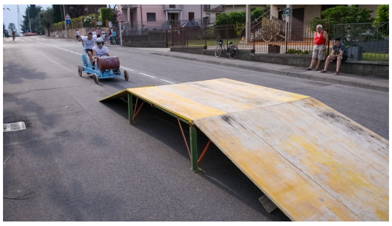

Figure 8: Original Image

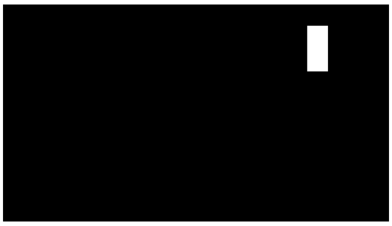

Figure 9: SSD Mask

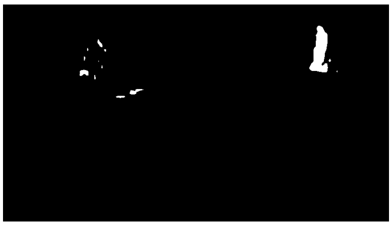

Figure 10: Imperfect Result

From the segmentation result we can observed that, due to the capability of FCN or ResNet on static semantic image segmentation, for videos with homogeneous background and single obvious foreground candidate, the model can give accurate segmentation. However, for cases with multiple candidates, or less distinction between foreground and background, the model didn't perfectly capture the always-present moving object.

4 Difficulties and Discussion

By observing the results, we concluded that there are 3 main difficulties. One is capturing the spatially always-present object. The optical flow in TSL didn't work as expected. This might be due to the roughness of optical flow itself and complexity of optical flow field given a complex background. Another possible reason is the lack of explicit mechanism to explicitly penalize the inconsistency of objects it tracks. Although the TMC used the spatial size assumption to by-pass this difficulty, a wrong mask assumption would mislead the model, as shown in 7. The second one is the difficult to differentiate objects from the adjacent background with similar texture or color. This might be due to the intrinsic structure of FCN architecture. Currently, contour-aware-networks[7] solved this problem.

4.1 Future Directions

Based on the analysis shown above, we can see the future improvements might lie in the following directions. 1. Finish the LSTM part for the TSL model. Since LSTM has its ability to ”memorize” the representations of previous frame, it might be promising to use LSTM to leverage the ”always present” nature. 2. Finetuning the part of SSD may be helpful to our results since in this case, any kind of objects may appear as the main moving part of the scene. In our experiments, we only have SSD with the ability to detect 20 classes(including useless classes such as chairs, tables and flowers). 3. More sophisticated mask generation strategy can be designed, for example, controlling the number of bounding boxes in the primitive mask generation, and taking the temporal consistence of masks into account.

References

[1] J. Pont-Tuset, F. Perazzi, S. Caelles, P. Arbeláez, A. Sorkine-Hornung, and L. Van Gool, “The 2017 davis challenge on video object segmentation,” *arXiv:1704.00675*, 2017.

[2] J. Long, E. Shelhamer, and T. Darrell, “Fully convolutional networks for semantic segmentation,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3431–3440, 2015.

[3] T. Mikolov, M. Karafiát, L. Burget, J. Cernockỳ, and S. Khudanpur, “Recurrent neural network based language model.,” in *Interspeech*, vol. 2, p. 3, 2010.

[4] S. Hochreiter and J. Schmidhuber, “Long short-term memory,” *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[5] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016.

[6] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, “Ssd: Single shot multibox detector,” in *European Conference on Computer Vision*, pp. 21–37, Springer, 2016.

[7] H. Chen, X. Qi, L. Yu, and P.-A. Heng, “Dcan: Deep contour-aware networks for accurate gland segmentation,” in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pp. 2487–2496, 2016.