One Shot Learning for Video Object Segmentation using Fully Convolutional I3D Network

Purna Sowmya Munukutla
Robotics Institute
Carnegie Mellon University
spmunuku@andrew.cmu.edu

Siddhant Jain
Robotics Institute
Carnegie Mellon University
siddhanj@andrew.cmu.edu

Abstract

In this work, we aim to explore the transfer learning of spatio-temporal feature extractors learned by the I3D network [2] pre-trained on Kinetics [6] + ImageNet datasets to other tasks. Specifically, we are interested in designing a network for video object segmentation for the DAVIS challenge. We propose an end-to-end learning framework for segmenting video objects by converting I3D to a fully convolutional architecture that fuses features across layers to define a non linear local-to-global representation. We also show how to bootstrap weakly annotated videos with existing action classification datasets like Kinetics for pre-training. We show that our results on the dataset for the DAVIS 2016 challenge display promise for further investment in this system for other tasks.

1. Introduction

In the last few years, convolution networks are being effectively adopted for per-pixel tasks like segmentation [13, 1], tracking [4]. More recently, CNNs have also been adopted for video tasks like action classification [2]. One such network, I3D [2] introduces two-stream inflated 3D convolutional network that is based on inflating 2D convolutions. The filters and pooling layers of ConvNets are expanded into 3D, making it possible to learn spatio-temporal features from videos while leveraging successful ImageNet architecture designs. I3D is initialized with inflated ImageNet weights and further fine-tuned on Kinetics [6] to achieve state-of-the-art results for action classification. In this paper, we intend to leverage the spatio-temporal feature extractors from I3D to other related but dissimilar tasks like video object segmentation.

Conventional approaches for video object segmentation employ 2D segmentation networks like [13, 11, 9] for per frame segmentation mask predictions, that are subsequently fused by incorporating optical flow information. Traditional image classification networks are employed for segmentation by modeling it as per-pixel classification problem. In CNNs, the deeper stages of the network contain abstract and semantically meaningful information to make mask predictions. The initial stages of the network contain information on local features that are useful to extract accurate boundaries of segmentation mask. Segmentation networks [13, 11] are designed to enable reverse information flow from last layers to initial layers of the network to make predictions.

Video Object Segmentation for the 2016 DAVIS [10] challenge, YouTube Objects [5] is a non-trivial task. DAVIS dataset consists of 150 sequences with ground truth annotations for each frame for 90 videos. The dataset captures tougher examples of foreground segmentations with viewpoint, illumination variances, shape deformations, partially or fully occluded objects to be segmented. The challenge is set for semi-supervised video segmentation where only the first frame annotations are available for test sequence.

2. Related Work

One-Video Video Object Segmentation (OSVOS) [1] is the baseline model for DAVIS 2016 challenge. OSVOS is based on fully convolutional neural network architecture that is able to transfer semantic context from ImageNet pre-training. Subsequent works [8, 12] introduced additional networks like mask propagation network, mask re-identification network that are jointly trained alongside CNN based RGB, Flow networks for segmentation. To improve the performance on test sequences, [1, 14] use one-shot learning methods for fine-tuning the models. [7] proposes a novel framework to generate future sequences for fine-tuning the network to segment sparsely annotated objects in test data.

3. Method

The main motivation of our method has been to understand transfer learning across videos for different tasks. The

effectiveness of transfer learning methods has been seen to work across datasets and tasks for images. However, transfer learning for videos is yet to be proven effective, as is noted in [2]. To that motive, we intend to capitalize on the learning done via I3D network with Kinetics pre-training for action recognition and generalize the semantic information learned to other tasks like foreground video object segmentation and tracking. We started with visualizing the activations of the frames of a video in Figure 1. It is easy to observe that the features are activated at foreground objects in the video (for example moving humans). This suggests that the network extracts meaningful semantic information, and sets up the motive to convert I3D to a fully convolutional architecture that fuses features across layers to define a non linear local-to-global representation that can be tuned end-to-end.

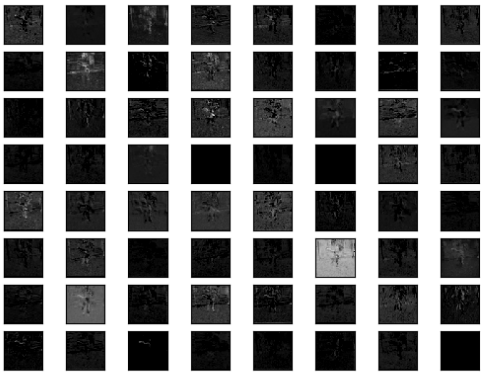

Figure 1. Visualization of activations of second inception layer of Kinetics pre-trained I3D on video (Parkour) from a new segmentation specific dataset (DAVIS)

In the first subsection, we discuss the architecture of fully convolutional I3D networks that has been shown effective for video object segmentation and describe the training process. In the later part, we focus on learning the appearance of single annotated frame from test sequence.

3.1. Fully Convolutional I3D Network

I3D or the Inflated 3D Convolutional Network has been introduced by *Carreira et al.*[2] for video classification. It has been trained inflating 2D convolutional layers on ImageNet data to bootstrap 3D convolutional layers. Each layer of data in the network has a size of c x f x w x h where c refers to the dimension of filters or channels, f refers to the number of frames and w x h refers to the image size. To ensure that the network can be trained for semantic segmentation, the network predictions are to be modified to output f x w x h segmentation mask. Drawing inspiration from [13] that modified VGG network to predict segmentation masks by adding upsampling layers, we apply a similar strategy for 3D convolutional networks like I3D. The I3D architecture consists of five pooling layers followed by fully connected layers and softmax to predict classification output. I3D network has been modified as shown in Figure 2 by introducing deconvolution layers for upsampling before each pooling step and discarding the last averaging pooling layer. The new transposed convolution layers facilitate skip connections from different resolutions of outputs. The skip connections capture local features obtained from lower layers with global semantic context derived from the last few layers. From each deconvolution layer, a segmentation mask is obtained after different stages of upsampling. These segmentation masks are of size f x w x h, which is the same size as the image, and are fused together to give final segmentation output. This end-to-end trainable network outputs five segmentation masks, with the first four masks taken directly from the upsampled layers and the last mask being derived by fusing the first four segmentation masks.

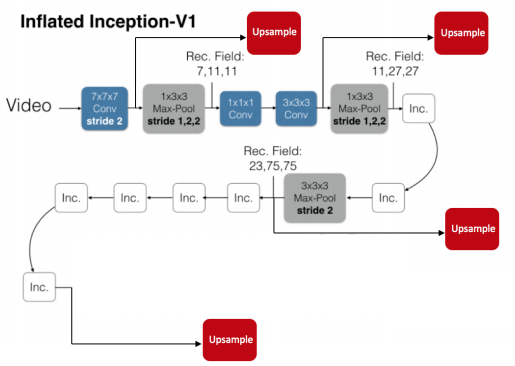

Figure 2. Network architecture of fully convolutional I3D network with last averaging pooling layer for classification outputs being removed and upsampling layers (red) added before every pooling stage. Figure adapted from [2].

The network in Figure 3 is initialized with pre-trained weights from Kinetics + Imagenet and it is fine tuned for DAVIS dataset. The deconvolution layers are initialized with 2D bilinear interpolation weights inflated along the third dimension that corresponds to the number of frames. The loss is computed to be per pixel-wise cross entropy for binary classification. In order to counter the imbalance between two classes, class balanced cross entropy loss as described in [1] is used. For each of the five segmentation masks outputted by the network, class balanced cross entropy loss is computed. We take a weighted average of the losses that we obtain from each of these masks. The weights are governed by the training epoch step. In the initial few epochs, the first four masks are given higher weights and this reduces linearly as the number of epochs progress.

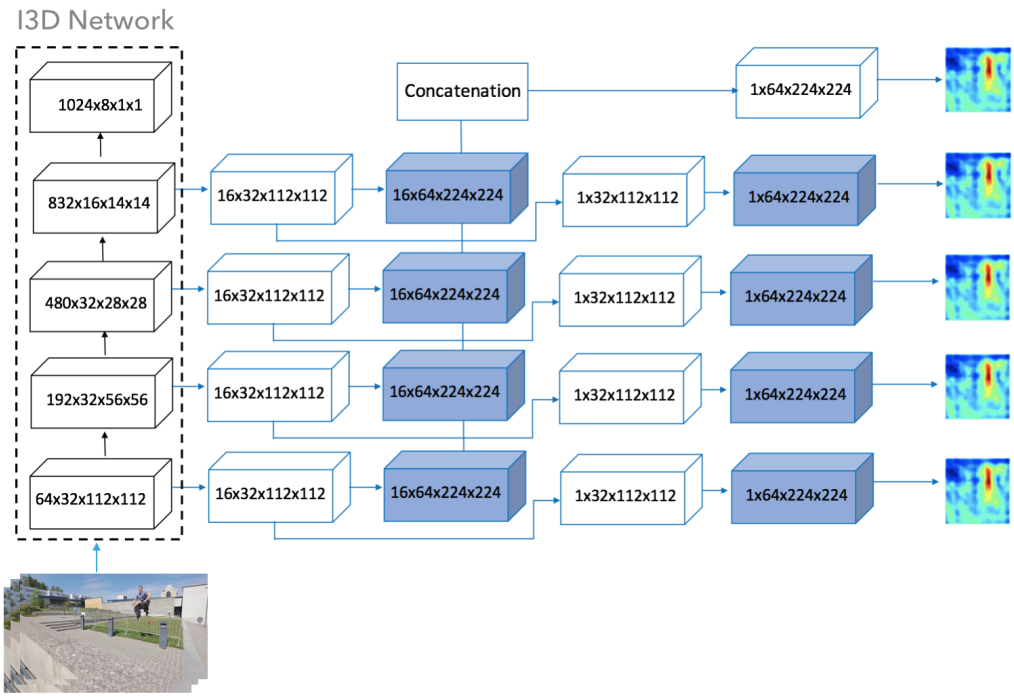

Figure 3. Visualization of fully convolutional I3D network architecture with upsmapling layers (blue) concatenated at four stages to produce output segmentation mask. Input to the network is number of channels x number of frames x image size and output to the network is 1 x number of frames x image size for segmentation masks

The deconvolution layers of the network are in general initialized with bilinear or nearest neighbor weights. The layers are expected to learn the upsampling weights over training iterations. In our initial experiments, we noticed that deconvolution layers in the initial epochs of training process produce checker-board like artifacts in the segmented output mask. These artifacts could be resolved to some extent by using overlapping deconvolutions. The deconvolution layers are overlapping if the kernel size is a divisible by stride. This is equivalent to sub-pixel convolution, a technique which has recently had success in image super-resolution [3].

3.2. Lucid Data Augmentation

During testing, only a single labeled training example (first frame) with object annotation is given and the trained model is expected to segment this object of interest over the test sequences. This can be achieved by fine-tuning the trained model separately for every video of test data. In this framework, one shot learning method allows us to adapt the network to a particular object instance given a single annotated object.

The first step of our one-shot learning framework is to create the 3D input data required to fine-tune the model for a particular test example. The data is hallucinated by generating future training sequences using lucid data augmentation technique proposed in [3]. We incorporate the first frame segmentation by back-propagating on a set of K frames, where the first frame is the given annotated frame of the video and the K-1 frames are generated by cutting-out the foreground object, in-painting the background, perturbing both foreground and background, and finally recomposing the scene.

3.2.1 Data Augmentation for training

Data augmentation techniques popular in the image domain, such as geometric transformations in the form of random crops, flips or rotations and photometric transformations fail to generalize the learning process when applied independently to each frame of 3D input. It is intuitive to reason that this could be because in the case of a video, context is important. Hence, for all the input frames that the network observes, we are expected to maintain strict temporal consistency. The geometric data augmentation technique is applied jointly to all the frames in a particular input batch. We also introduce a new style of applying random crops. For the first frame of the video, we find a random 224x224 crop and accept this crop if more than 10% of the pixels are foreground. This crop is applied on all subsequent frames

of the particular video.

3.2.2 Memory constraints

During training and fine-tuning, the inputs are randomly cropped to 224x224 to train fully convolutional I3D network described in Section 3.1. However, at test time, the segmentation outputs are to be determined for the full resolution of test sequence, which is FHD size for DAVIS dataset. Any particular test sequence of DAVIS dataset consists frames between 35 to 100. This ends up creating memory constraints in computing forward and backward weights on an input sequence on Nvidia GTX 1080 GPU given that I3D has a deep network architecture. To overcome this constraint, we divide the image into overlapping blocks of size 224x224 with a stride of 100 (configurable). The final segmentation masks are obtained by averaging the outputs for each of the overlapping blocks in every image.

4. Results

Dataset. We used DAVIS dataset to perform most of our experiments since it is one of the largest densely annotated datasets available with high quality, per pixel segmentation annotations for high resolution ground truth videos. The dataset has 150 sequences with ground truth annotations for each frame for 90 videos. The remaining 60 test videos have the segmentation for the object to be tracked in the first frame. DAVIS dataset has a good distribution of challenging videos with deformable objects, illumination and viewpoint changes etc. compared to YouTubeObjects [5] which makes it a benchmark for video object segmentation. Most videos contain one or two moving objects in the foreground, and in this particular paper, we are focused on foreground segmentation of moving objects. In the subsequent iterations of this paper, we plan on extending the method to instance segmentation of foreground objects.

Training. We initialize convolution layers of fully convolutional I3D network with pre-trained weights from Kinetics + ImageNet and deconvolution layers are inflated with bilinear interpolation weights. The parent model is then fine-tuned on DAVIS for 700 epochs. With the parent network available, we proceed to the given task of segmenting a particular entity in a video, given the image and the segmentation of the first frame. The parent model is further trained (fine-tuned) for the particular image/ground-truth pair for 200 epochs. The 3D data consisting of image/ground-truth pairs is generated by lucid data augmentation, and this model is then tested on the entire sequence, using the new weights. The model is optimized using stochastic gradient descent (SGD) with a learning rate of 0.001, momentum 0.9 and input batch to every layer in the network is normalized.

Evaluation. In the given task of video object segmentation, the sparsely annotated mask of the first frame is expected to be propagated to the rest of the video. As justified in [10], we use two evaluation metrics to compute the performance of the model. The first metric is based on region similarity which measures the number of misclassified pixels to evaluate the model. Jaccard Index J, computed as intersection-over-union of image versus ground truth annotation, is a commonly employed measure to evaluate performance of segmentation algorithms. The next metric to evaluate performance of segmentation algorithm is based on contour accuracy F. The precision and recall values for closed contours in the segmentation mask via bipartite graph matching is computed. Our model was able to obtain J-mean of 37.8 and F-mean of 40.7 on the validation set after training and fine-tuning for the particular image/ ground truth pair. These results are comparable to performance of baseline (OSVOS) and introduction of flow network would remove false positives for segmentation mask predictions.

The qualitative evaluation of segmentation outputs of DAVIS validation sequences is shown in Figure 4. It can be seen that the method is robust to viewpoint, illumination variations, non rigid body deformations.The false positives being detected as foreground can be eliminated by improving the fine-tuning pipeline which could mean improving the quality of frames generated by lucid data augmentation technique. Training a two stream network to separately learn RGB and flow features would certainly improve segmentation accuracy. Owing to resource constraints and GPU limitations, two stream architecture of fully convolutional I3D network with full resolution 3D inputs could not experimented with and we plan on using a shallower network for flow module in the subsequent iterations of this paper. It can also be seen in Figure 4 that the network is able to finely segment foreground objects even for highly deformable objects like in parkour video. The fourth row in Figure 4 shows the failure case where the network is not able to finely segment the foreground objects that are not captured in the first frame. Incorporating flow network would fuse additional context on applying similar segmentation masks to the entire object being tracked and help resolve under-segmentation. Fine tuning the network with better augmentation strategies to hallucinate subsequent frames would resolve over-segmentations seen in the output.

The code and trained models for fully convolutional I3D network can be accessed here.

References

[1] S. Caelles, K.-K. Maninis, J. Pont-Tuset, L. Leal-Taixé, D. Cremers, and L. Van Gool. One-shot video object segmentation. In *Computer Vision and Pattern Recognition (CVPR)*, 2017. 1, 2

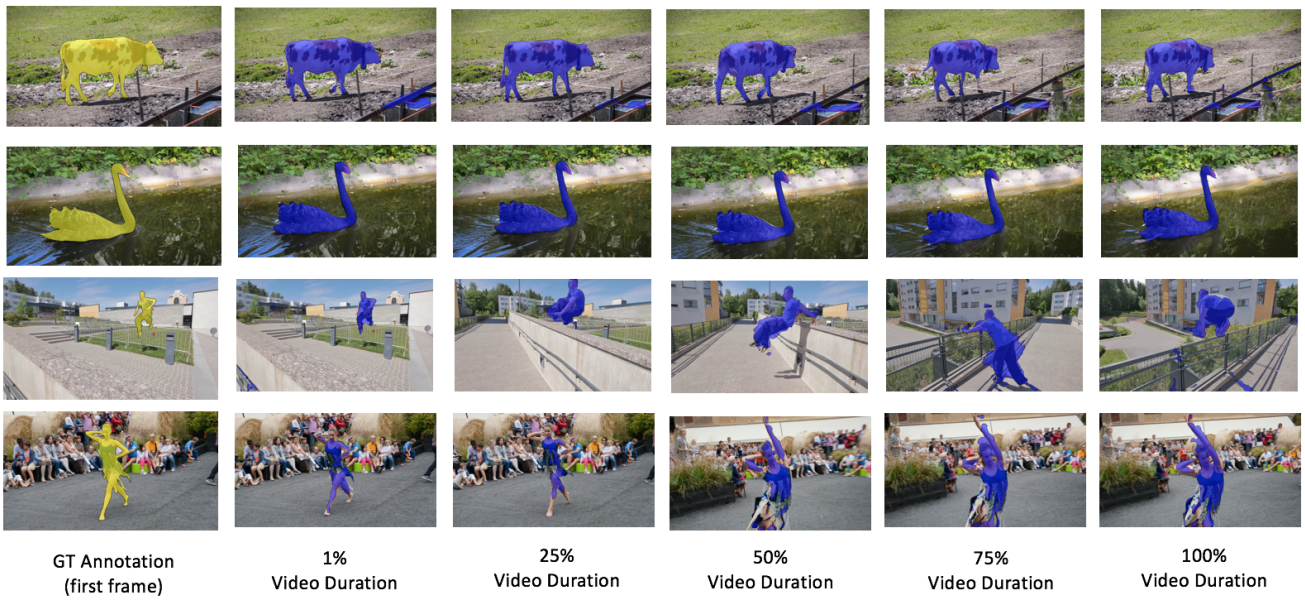

Figure 4. Qualitative evaluation of segmentation output for videos cow, blackswan, parkour, dance-twirl from the validation set. Over segmentation and under segmentation is observed for some frames in certain scenarios but the network learns to recover soon

[2] J. Carreira and A. Zisserman. Quo vadis, action recognition? A new model and the kinetics dataset. *CoRR*, abs/1705.07750, 2017. 1, 2

[3] C. Dong, C. C. Loy, K. He, and X. Tang. Image super-resolution using deep convolutional networks. *IEEE Trans. Pattern Anal. Mach. Intell.*, 38(2):295–307, Feb. 2016. 3

[4] A. Dosovitskiy, P. Fischer, E. Ilg, P. Hausser, C. Hazirbas, V. Golkov, P. v.d. Smagt, D. Cremers, and T. Brox. Flownet: Learning optical flow with convolutional networks. In *IEEE International Conference on Computer Vision (ICCV)*, Dec 2015. 1

[5] V. Kalogeiton, V. Ferrari, and C. Schmid. Analysing domain shift factors between videos and images for object detection. *IEEE transactions on pattern analysis and machine intelligence*, 38(11):2327–2334, 2016. 1, 4

[6] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev, M. Suleyman, and A. Zisserman. The kinetics human action video dataset. *CoRR*, abs/1705.06950, 2017. 1

[7] A. Khoreva, R. Benenson, E. Ilg, T. Brox, and B. Schiele. Lucid data dreaming for object tracking. In *The 2017 DAVIS Challenge on Video Object Segmentation - CVPR Workshops*, 2017. 1

[8] X. Li, Y. Qi, Z. Wang, K. Chen, Z. Liu, J. Shi, P. Luo, C. C. Loy, and X. Tang. Video object segmentation with re-identification. *The 2017 DAVIS Challenge on Video Object Segmentation*, 2017. 1

[9] P. H. O. Pinheiro, T. Lin, R. Collobert, and P. Dollár. Learning to refine object segments. *CoRR*, abs/1603.08695, 2016. 1

[10] 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. 1, 4

[11] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. *CoRR*, abs/1505.04597, 2015. 1

[12] A. Shaban, A. Firl, A. Humayun, J. Yuan, X. Wang, P. Lei, N. Dhanda, B. Boots, J. M. Rehg, and F. Li. Multiple-instance video segmentation with sequence-specific object proposals. *The 2017 DAVIS Challenge on Video Object Segmentation - CVPR Workshops*, 2017. 1

[13] E. Shelhamer, J. Long, and T. Darrell. Fully convolutional networks for semantic segmentation. *CoRR*, abs/1605.06211, 2016. 1, 2

[14] P. Voigtlaender and B. Leibe. Online adaptation of convolutional neural networks for the 2017 davis challenge on video object segmentation. *The 2017 DAVIS Challenge on Video Object Segmentation - CVPR Workshops*, 2017. 1