Fully Convolutional Networks for Image Segmentation

Maria Camila Escobar
Universidad de los Andes
Bogotá D.C, Colombia
mc.escobar11@uniandes.edu.co

Laura Gongas
Universidad de los Andes
Bogotá D.C, Colombia
l.gongas10@uniandes.edu.co

1. Introduction

Long et. al developed a method to adapt classification networks such as AlexNet, VGG net and GoogLeNet into fully convolutional networks suitable for semantic segmentation. For this task, they propose an initial adaptation of classifiers for dense prediction. Classification nets produce non-spatial outputs because their fully connected layers do not take into account spatial coordinates. However, fully connected layers are also convolutions with kernels covering all their input regions. By viewing fully connected layers in this way, they get fully convolutional networks with input of any size and output classification maps. Nonetheless, the output of this networks is subsampled. Long et. al suggest a deconvolution layer to upsample the output [1]. It is important to note that for their network, they discarded the last classifier layer of the nets for classification. Also, they added a 1x1 convolution with channel dimension 21 for prediction of the PASCAL dataset.

Initially, they developed a network with 32 pixel stride in the final layer (FCN-32). Also, they developed FCN-16s which changes the stride of the predicting layer to 16 pixel and it is initialized with the parameters of FCN-32s. Additionally, the learning rate is decreased. A 1x1 convolution layer is added on top of pooling 4 which results in more class predictions. The output is combined with the output on top of convolutional layer 7 at stride 3 [1].

2. Results and discussion

Figures 1, 2 and 3 show results of semantic segmentation using FCN-16s from FCN-32s weights. The first two images have only one object in the image that occupies the majority of it. The jaccard index for the television and the cat segmentations were 0.8865 and 0.9348 respectively. On the other hand, the segmentation of the plant pots yielded a result of 0.6775. This image has more details which require a finner segmentation that could be obtained by using FCN-8s proposed by Long et. al. On the contrary, the television and the cat yield good results because a coarser segmentation is acceptable because they do not have as many details.

So, FCN-32s is better for images that lack fine details. On the other hand, FCN-16s does not limit as much the scale of detail in its output. Also, small scale objects are segmented better in FCN-16s while bigger scale objects yield acceptable results with FCN-32s.

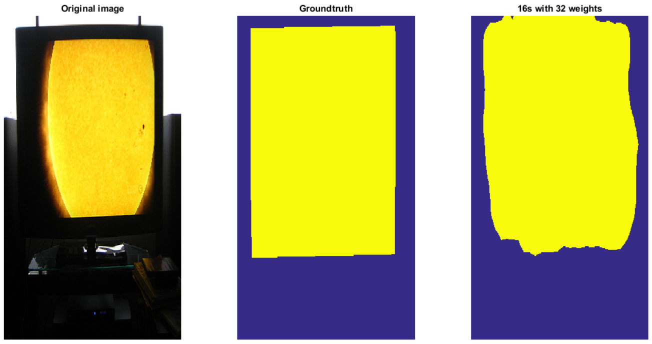

Figure 1. Results of television image segmentation training with FCN-16s from FCN-32s weights. Jaccard: 0.8865.

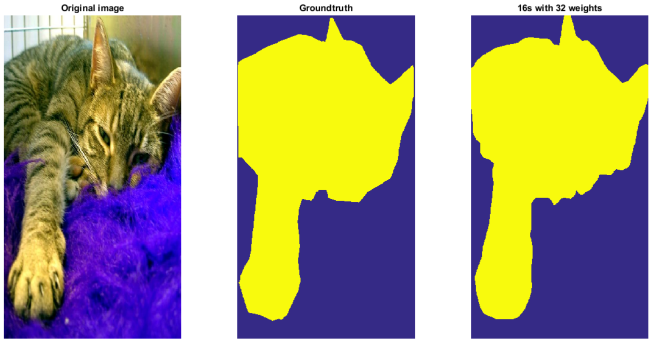

Figure 2. Results of cat image segmentation training with FCN-16s from FCN-32s weights. Jaccard: 0.9348.

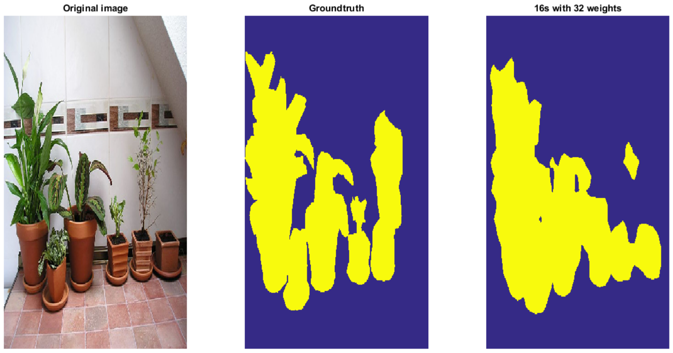

Figure 3. Results of plant pot image segmentation training with FCN-16s from FCN-32s weights. Jaccard: 0.6775.

References

[1] E. Shelhamer, J. Long, and T. Darrell. Fully convolutional networks for semantic segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(4):640–651, 2017.