Fully Convolutional Networks Semantic segmentation

Group 19

Baikai Li, Tianhang Gao, Yu Liu, Zhenzhen Dai

1. Introduction

Several recent works have applied convolutional networks to dense prediction problems, including semantic segmentation which each pixel is labeled. A fully convolutional network (FCN) is a net with layers only computing a general nonlinear filter. In this project, we implemented a FCN based on a published paper which is the first work to train FCNs end-to-end for pixelwise prediction and from supervised learning and achieves state-of-art segmentation on PASCAL VOC, NYUDv2, and SIFT Flow image sets. Both training and testing are performed whole image at a time by dense forward and backpropagation computation. For our work, we trained and tested the FCN network on VOC 2011 and compared our results to the reference paper.

2. Network Components

The FCN network mainly consists of three types of layers:

a. Convolutional layer, ReLU layer and Pooling layer

One of the difference between CNN and FCN is that, instead of using fully connected network, utilizing a fully convolutional network can produce an efficient network for end-to-end dense learning as in case of image segmentation.

ReLU layer is a layer to produce nonlinear results. Additionally, ReLU costs less computation especially in back propagation compared with sigmoid.

Pooling layer can reduce the spatial size of the representation, which reduces computation in the network, meanwhile, pooling layer keeps the feature of the representation.

b. Deconvolution layer

Deconvolution layer performs upsampling operation, which enlarges the representation of high level features to the size in bottom layer.

c. loss layer

This is the layer produce the parameter for back propagation. It calculates the difference between the ground truth and the result from the network pixelwise using a certain function.

3.Methods

Caffe is a deep learning framework developed by BVLC which includes abundant model zoo and online sources for standard distribution format of Caffe models, and provides trained models.

a. Network architecture

5 * Conv--relu--conv--relu--pooling consist the base net, which is the structure in VGG 16 net.

2 * Conv--relu--drop consist two fully convolutional net (fcn), which is composed as FCN-32s net with a upsampling layer.

Upsampling the result of fcn and fusing it with pool4, then composing it as FCN-16s net with a upsampling layer. Detailed FCN-8s network structure is attached with our report.

See PDF for detailed network architecture!

b. Train

Transplanting the weights in vgg16.caffemodel into FCN-32s network and training it with dataset listed in train.txt and doing this on FCN-16s and FCN-8s network.

c. Test

Testing our FCNs with dataset listed in val.txt and we save the output images and scores.

4. Result and Evaluation

We used the same metrics as the paper: overall accuracy: $\sum_i n_{ii} / \sum_i t_i$, mean accuracy: $(1/n_{cl}) \sum_i n_{ii}/t_i$, mean IU: $(1/n_{cl}) \sum_i n_{ii}/(t_i + \sum_i n_{ji} - n_{ii})$ and frequency weighted IU: $\sum_k {t_k}^{-1} \sum_i t_i n_{ii}/(t_i + \sum_i n_{ji} - n_{ii})$, where n_{ij} is the number of pixels of class i predicted to class j and $t_i = \sum_j n_{ij}$ the total number of pixels of class i.

	Overall accuracy	Mean accuracy	Mean IU	fwavacc
FCN-32s	0.88121	0.68213	0.56273	0.80341
FCN-16s	0.89008	0.72467	0.61170	0.82198
FCN-8s	0.90236	0.74867	0.62069	0.83007

Table 1. results

There are 2 ways we can improve our results.

1) Iteration times. We believe the limited training time is the major reason. With limited time, we tried several times with different learning rates, which are fast at the beginning and slower afterwards. However, these learning rates are not small enough to be defined as fine-tuning. Thus, the results are not as good as them in the paper. Additionally, from one failed attempt, we can clearly see that, if the accuracy is not good enough in one level, the following level will be deeply affected which can't be recovered in following level training.

2) The training set. After going through our results, we find that the picture with human figure can achieve much better results. The reason is that, in the training sets, the number of images of human is much greater than other categories. If the dataset can provide more pictures of other categories, the results can be improved.

We also met several difficulties during the project:

1) Since all of our group members have never used Caffe before, it took us long time to learn how to use the platform. It is a painful process for all of us.

2) It is hard to get an both efficient and effective set of parameters. For example, only a small range of learning rates can work really well. We attempted some larger values, where the loss failed to converge, and also smaller values, where the loss function converges too slowly.

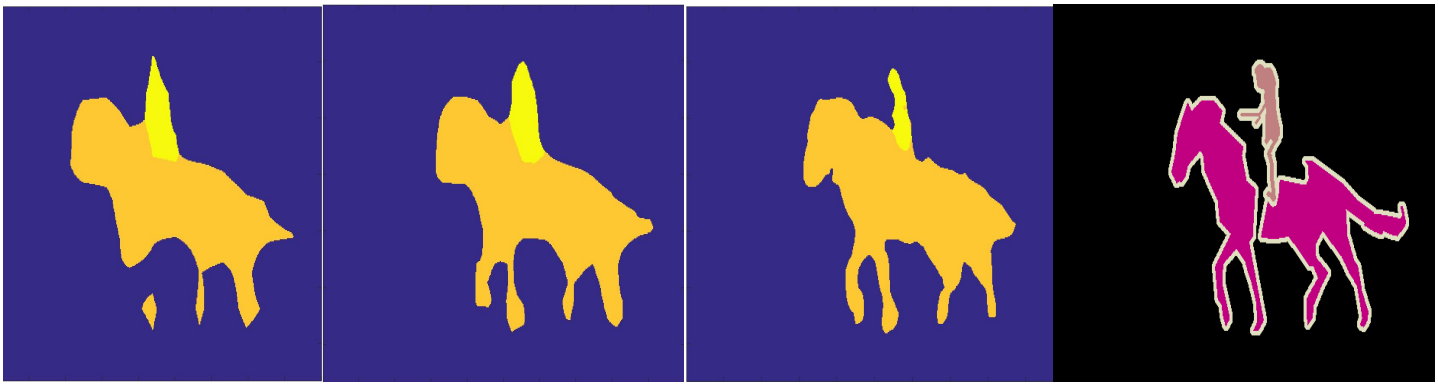

Figure 1. 32s, 16s, 8s, ground truth

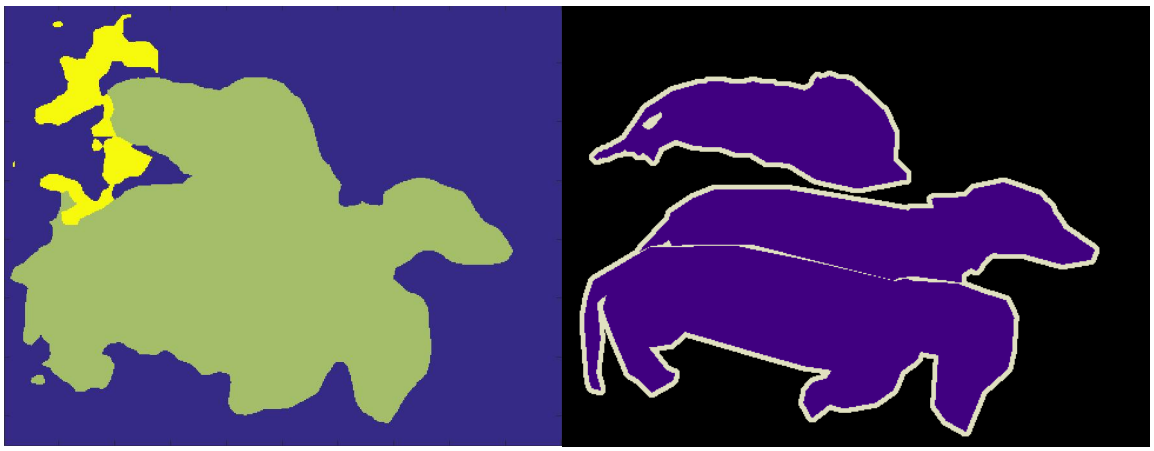

Figure 2. Failed sample(8s & ground truth)

Reference

[1] Long, Jonathan, Evan Shelhamer, and Trevor Darrell. "Fully convolutional networks for semantic segmentation." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2015.