Project Proposal for Semantic Segmentation

Yujie Liu Yumeng Liu Zetian Xiao
University of Rochester
{yliu134, yliu114, zxiao8}@u.rochester.edu

1. Team Information

1.1. Team name

29-Characters

1.2. Team members

- Yujie Liu
- Yumeng Liu
- Zetian Xiao

2. Related work

Very deep convolutional networks for large scale image recognition: In this work researchers investigate the effect of the convolutional network depth on its accuracy in the large-scale image recognition setting. Thorough evaluation of networks of increasing depth using an architecture with very small (33) convolution filters, they show that a significant improvement on the prior-art configurations can be achieved by pushing the depth to 1619 weight layers. These representations generalize well to other datasets, where they achieve state-of-the-art results. Their work (VGG net) are used by the 2015 Berkeley paper to build FCNs.[4]

Deep residual learning for image recognition: In this paper, researchers present a residual learning framework to ease the training of networks that are substantially deeper than those used previously. They explicitly reformulate the layers as learning residual functions with reference to the layer inputs, instead of learning unreferenced functions. Researchers provide comprehensive empirical evidence showing that these Resnets are easier to optimize, and can gain accuracy from considerably increased depth. On the ImageNet dataset they reached residual nets with a depth of up to 152 layers that is 8 times deeper than VGG nets but still having lower complexity. An ensemble of these residual nets achieves 3.57% error on the ImageNet test set. The ResNet can potentially be transformed to FCN for segmentation.[2]

Fully convolutional network for semantic segmentation: In their 2015 paper, researchers at UC Berkeley improved state-of-the-art performance of semantic segmentation by transforming pretrained neural networks used for classification into fully convolutional network. By replacing the fully-connected layers with convolutional ones, the network produces pixel-wise classification. They then define a skip architecture that combines semantic information from a deep, coarse layer with appearance information from a shallow, fine layer to produce accurate and detailed segmentations. Remarkably, their FCN gives a 20% relative improvement over the state-of-the-art on the PASCAL VOC 2011 and 2012 test sets and reduces inference time. [3]

3. The problem and our approach

In exploring the possibilities of object detection, advances are made through bounding-box, key-point detection and so on. To pursue finer result (pixel-level accuracy),scholars advance into the task of semantic segmentation, where each pixel is assigned an object class. Besides recognizing objects, semantic segmentation also delineates the boundary between objects, which requires dense pixel-wise predictions from models.

Popularized by the 2015 Berkeley paper, FCN builds fully convolutional networks that takes input of arbitrary size and produce correspondingly-sized output with efficient inference and learning. But usually, it takes great amount of time to train a model. For this Project, we want to implement FCN algorithm that is faster to train and with limited dataset, but also give reasonable results. Specifically, we want to use heuristic methods and choose subsets of the dataset to train our network. Also, we would try to improve state-of-art FCN approach by considering context information.

We want to adapt other new classification network(like ResNet) instead of VGG into fully convolutional networks and transfer its learned representations by fine-tuning to the segmentation task. Like the 2015 Berkeley implementation of FCN, we also plan to combine information from shallower layers with the coarse output from deeper layers to make use of both high-level features and low-level contex-

tual information. The authors of the Berkeley paper implemented FCN in Tensorflow, and we decide to adapt the FCN to Pytorch for the ease of implementation.

4. Potential dataset

PASCAL VOC (Visual Object Classes) 2012: a benchmark in visual object category recognition and detection, providing a standard dataset of images and annotation, and standard evaluation procedures. The training and validation contains 10,103 images while testing contains 9,637 images. A subset of images are annotated with pixel-wise segmentation of each object present, to support the segmentation competition. [1]

5. Timeline

- April 1st- April 14th: Study related works and implementing the original FCN with Pytorch.
- April 14th- April 21st: Brainstorm and tentatively add modifications to the algorithm based on contextual information and comparing the outcome. This process may consist in several stages of weighing different techniques and make modifications accordingly.
- April 21st- April 23rd: Choose one specific improvement with the best outcome and finalize all the details.
- April 23rd- April 25th: Prepare for the presentation, making slides and visualize the results.
- April 25th- May 15th: Finalize the paper.
- May 15th - TBD: possibly improve our implementation continually and adapting to new technologies.

References

[1] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2):303–338, 2010.

[2] 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*, pages 770–778, 2016.

[3] 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*, pages 3431–3440, 2015.

[4] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.