Project Proposal for EndoVisSub 2018 Challenge

Zhaoshuo Li, Hao Ding, Mingyi Zheng

I. Introduction

Robot-assisted surgeries are currently completely human teleoperated, given a prominent example of the da Vinci surgical robot developed by the Intuitive Surgical Inc. One future direction of surgical assisted surgeries is to increase the level of autonomy, which requires certain level of scene understanding. Therefore, this project will focus on semantic segmentation for various objects.

II. Methodology

A. Data

The training dataset is subsampled at 2Hz from a 60Hz video stream of da Vinci Xi robot performing partial nephrectomy in a pig. There are in total 2384 images available with 11 classes in total. The image is of size of 1920x1080 in RGB format. Ground truth is provided for the left frames of the scope. The dataset is publicly available known as EndoVisSub Challenge 2018.

B. Architectures

FCN [1] was the first successful network that combines classification and convolution for segmentation purpose. The idea is that each pixel will be assigned to a classification label. The network upsamples feature map at different resolution to the size of original input to produce from coarse to fine segmentation result. Later, U-Net [2] is developed using an encoder-decoder design, with the idea being low dimensionality latent representation can be "blown up" to high-level classification result. Currently, most of the architectures followed similar strategies and have produced great result in various challenges. In this proposal, three architectures with different designs are selected for this project.

TernausNet-16 [3]: This network follows closely to the design of U-Net, except that after down-sampling, the fully-connected layers in the bottleneck part of the network is replaced with a pre-trained VGG16 before the fully-connected layers. The VGG is used to enhance the capability of dimensionality compression.

PSPNet [4]: Pyramid Scene Parsing network (PSPNet) was motivated by the need of a strong global context to mitigate the problems of mismatched scene and object relationship, confusion between similar categories and inconspicuous objects in a large scene. To better solve the problems of detecting global context and local feature at the same time, a pyramid pooling module is proposed based on hierarchical prior. After pre-training the ResNet [5] with dilated convolution. The dilated convolution (or atrous convolution) is to increase the receptive field at an arbitrary given rate. The ResNet is also trained with auxiliary loss at the fourth stage of the residual block by adding an additional classification before the final output. After obtaining the feature map, average pooling at different kernel sizes are used to produce features at different level. Kernel of size 1x1 is followed to reduce the channel size. Up-samling with bilinear interpolation is then used to match the size of features of different scales as the size of the original input feature, and then all features are concatenated before a final global convolution before producing the result.

DeepLabV3+ [6]: While PSPNet has performed very well in the PASCAL VOC 2012 challenge, there is computation burden on the use of dilated convolution. On the contrast, U-Net has much faster computation without the dilated convolution. Therefore, the architecture used one tunable dilated convolution (named Atrous Spatial Pyramid Pooling, ASPP), which is not available in U-Net. On the other hand, it uses skip-connections to concatenate encoding and decoding features, which is not available in PSPNet. It is an improved version of DeepLabV3 [7]. Another key difference the network introduced is that the Xception modules in the ASPP and decoding path. Xception is an extreme case of Inception module, assuming spatial and channel correlation can be totally separated, and therefore applying the convolution kernel depth-wise, which can reduce the computation complexity. The network proposed of using DeepLabV3 as encoder, and applied ASPP on top of it. During the decoding process, two 4x up-sampling were performed instead of directly 8x up-sampling.

C. Training

We consider the following to facilitate training: 1. EndoVisSub 2017 dataset as pre-training. 2. Standard cropping and rotation can be used. 3. Self-supervised techniques such as coloring. 4. Auxiliary losses before the final layer by merging similar categories.

References

[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.

[2] Ronneberger, Olaf, Philipp Fischer, and Thomas Brox. "U-net: Convolutional networks for biomedical image segmentation." International Conference on Medical image computing and computer-assisted intervention. Springer, Cham, 2015.

[3] Iglovikov, Vladimir, and Alexey Shvets. "Ternausnet: U-net with vgg11 encoder pre-trained on imagenet for image segmentation." arXiv preprint arXiv:1801.05746 (2018).

[4] Zhao, Hengshuang, et al. "Pyramid scene parsing network." Proceedings of the IEEE conference on computer vision and pattern recognition. 2017.

[5] He, Kaiming, et al. "Deep residual learning for image recognition." Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.

[6] Chen, Liang-Chieh, et al. "Encoder-decoder with atrous separable convolution for semantic image segmentation." Proceedings of the European Conference on Computer Vision (ECCV). 2018.

[7] Chen, Liang-Chieh, et al. "Rethinking atrous convolution for semantic image segmentation." arXiv preprint arXiv:1706.05587 (2017)