DeepLab with adversarial loss for semi-supervised semantic segmentation of Medical Images.

Karan Lalwani
Computer Science Department
University at Buffalo
karanlal@buffalo.edu

***Abstract*—Convolution neural networks are very powerful at finding features in an image, thus solving the problem of semantic segmentation of an images at pixel level but requires large amounts of labeled data to train on. Among all the categories of images, Segmentation of Medical images(WSI) using supervised methods [1] [2] [3] [4] is difficult, due to the lack of labeled data. This paper proposes a semi-supervised approach to solve this problem by using an adversarial loss. Firstly the model trains on per-pixel loss along with adversarial loss for paired images. After that for unpaired images only adversarial loss is used, helping the model to train on data which otherwise couldn't be used by traditional supervised methods. The model was tested on Radiology dataset obtaining an F1-score of 0.953.**

***Index Terms*—Semantic Segmentation, Medical Images, DeepLab, Adversarial loss**

I. Introduction

Semantic segmentation [1] [2] [3] [4] is a task that labels each pixel in an image to the class of its enclosing object or region. It is a task that resolves the global information of what and the local information of where. Because of its dense predictions at the pixel level it can help with the problem of partitioning, thus having many applications in the area of medical imaging. But medical image segmentation using supervised deep learning models [1] is expensive. Because of its specialized nature, it requires experts to annotation medical images to train the model.

Semi-supervised learning can leverage this large amount of available unlabeled data. Using Semi-supervised learning, we can first identify some specific hidden structure so that given x, x', where x is original image and x' is segmented image, we can find the joint probability p(x, x'). The proposed architecture uses a Convolution Neural Network on top of segmentation model(DeepLab v3+) to construct p(x, x') and is trained just like a Generative Adversarial Network.

II. Related Works

A. An integrated iterative annotation technique for easing neural network training in medical image analysis

This model [1] deals with the lack of annotated data and quality required to train a segmentation network by employing Active Learning strategy. The model devise an interface for data annotation and the display of neural network predictions within a commonly used digital pathology whole-slide viewer. The segmentation is repeatedly improved with the humans interact throughout the training process.

B. Structured and Unstructured Loss

Usually image-to-image [5] translation problems use per-pixel loss which considers each output pixel as conditionally independent of each other, thus treating the image as a whole as unstructured. These type of losses - L1(Mean Absolute Error), L2(Mean Squared Error), Cross-entropy Loss - can lead to blurry image. Where as Structured Losses penalizes the joint configuration of output. Conditional Random Field, cGAN loss are examples of such losses.

C. Adversarial Learning for Semi-Supervised Semantic Segmentation

The discriminator in this model [6] is designed to be a fully convolutional network, which instead of classifying input images as real or fake on the image level, it differentiate the predicted probability maps from the ground truth segmentation distribution at spatial resolution. Coupling the Generator with adversarial loss the model enables semi-supervised learning through discovering the trustworthy regions in predicted results of unlabeled images, thereby providing additional supervisory signals.

D. PatchGAN

The patchGAN discriminator [5] takes two images, and outputs a probability patch of size nxm instead of a single in-class or out-class, probability output. Each cell or pixel in the output corresponds to some region in the input image.

$$receptivefield = (outputsize - 1)strides + kernelsize$$

III. Approach

The basic architecture of this semi-supervised segmentation model is same as conditional Generative Adversarial Network [7] fig. 1. It has two Convolutional Neural Networks, one for segmentation and another to provide adversarial loss. These networks are trained simultaneously for all the labeled images. For unlabeled images only generator is trained with the loss obtained from the discriminator, assuming the discriminator is trained enough to imitate p(x, x') and give decent loss.

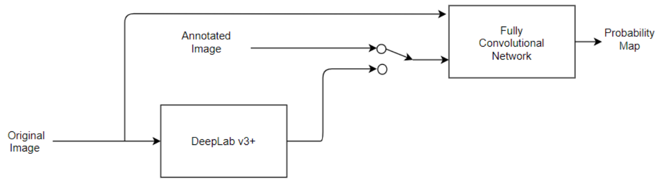


Fig. 1. Architecture

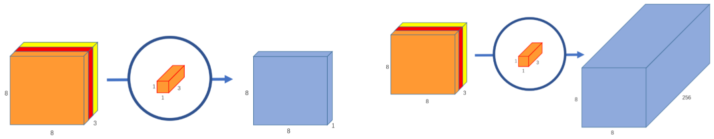
A. Generator

This model uses DeepLab v3+ [2] fig. 2 as a generator to segment the image which is an encoder-decoder architecture. DeepLab architecture has three main components : *Atrous Spatial Pyramid Pooling(ASPP)*, *Encoder-Decoder* and *Depth-wise Separable Convolution*.

Atrous Spatial Pyramid Pooling(ASPP) : It is group of Convolution applied at different dilation rate to capture the multi-scale information from the image, as the objects from the same class can have different scales in an image. In this architecture four convolution with dilation rate of 1, 6, 12, 18 are used in parallel and then concatenated along with global average pooling. The output of which is fed to the decoder.

Encoder-Decoder : Like any typical encoder-decoder architecture, DeepLab [2] also uses the same structure. The encoder which uses pre-trained Xception architecture as its backbone, gradually reduces the feature maps and captures higher semantic features from an image, and the decoder gradually up-sample the low resolution feature maps generated from the encoding layer above and eventually recover the spatial dimensions and object details.

Depth-wise Separable Convolution : It is a technique which is used in place of traditional convolution to reduce the number of parameters and the cost of computation by a factor of 20 [8] while maintaining similar performance accuracy, allowing us to create much deeper architectures. Depth-wise separable convolution is done by applying depth-wise convolution followed by a point wise convolution.



Depthwise convolution pointwise convolution

Fig. 2. Depth-wise Separable Convolution

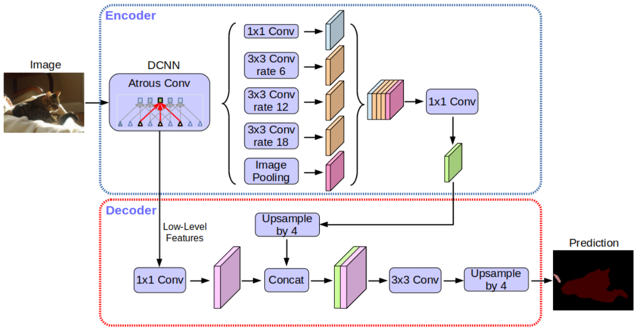


Fig. 3. Deeplab v3+

B. Discriminator

Convolution Neural Networks are good at approximating any function. Keeping that in mind the discriminator used in this model is a Fully convolution neural network which is an encoder-decoder type of architecture and gives the probability map as an output. The main purpose of this discriminator is to approximate the probability that given an original image and its corresponding segmented image, it has to determine, if the segmented pixel corresponding the original image are correct or not, p(x, x'). If the pixel is correctly classified then its probability is 1 and 0 otherwise.

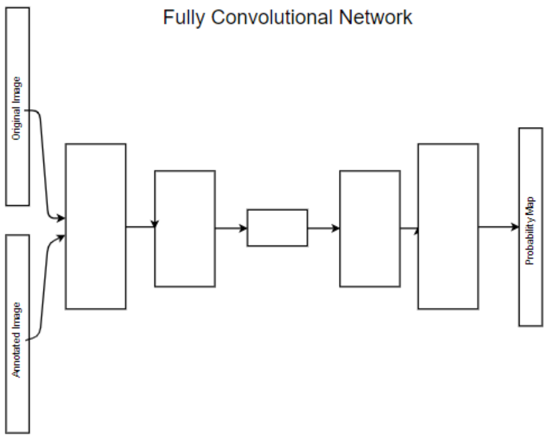


Fig. 4. Fully Convolutional Network

C. Training

1) Discriminator Training:

- Generator takes in a mini-batch of WSI images.
- It then segments the images into various classes and sends it to the discriminator.
- Discriminator alternately takes Annotated ground truth(in class) images and Annotated predicted(out class) images conditioned on the their respective input images and outputs a per pixel probability map.
- Binary cross-entropy loss obtained, is back-propagate to train the discriminator.

$$L_{BCE} = -y_i log(y_i') - (1 - y_i) log(1 - y_i')$$

2) *Generator Training:*

- Once the discriminator is trained on a mini-batch we again feed another mini-batch of images to the Generator.
- It then segments the images into various classes and sends it to the discriminator.
- Training of discriminator is set to False
- Discriminator then takes images Annotated by the generator conditioned on the their respective input images and outputs a per pixel probability map.
- Binary cross-entropy loss obtained from the discriminator along with the categorical cross-entropy loss obtained from the generator is used to train the generator.
- Training of discriminator is set to True

$$L_{CCE} = -\sum_{1}^{C} y_i log(y_i')$$

$$L_G = L_{CCE} + \lambda_1 L_{BCE}$$

3) *Supervised Training:*

- Train Discriminator on Binary cross-entropy loss
- Train Generator on weighted sum of Categorical cross-entropy loss and Binary cross-entropy loss.

$$L_D = L_{BCE}$$

$$L_G = L_{CCE} + \lambda_1 L_{BCE}$$

4) *Unsupervised Training:*

- Discriminator is not trained further
- Train Generator only on Binary cross-entropy loss obtained from the discriminator.

$$L_G = \lambda_2 L_{BCE}$$

IV. Experiments

The proposed model is evaluated on Radiology Data [9] which contains 3 foreground classes and 1 background class. The dataset contains 1000 images out of which 800 are used for training 100 validation and 100 for test. The performance is measured in terms of intersection-over-union averaged across all 4 classes and F1 score due to there being a class imbalance problem.

A. *Discriminator Choice*

Experiments with two different types of discriminators were done to generate adversarial loss, in which Fully Convolution Network worked better that PatchGAN.

1) PatchGAN-Discriminator: PatchGAN is very sensitive to the size of receptive field. Depending on the size it was giving tiling effect. The experiment was done using various sizes such as 16x16, 70x70, 256x256 out of which 70x70 gave the best results. Also the learning rate of PatchGAN discriminator was kept higher than the generator which used pre-trained Xception architecture as its backbone.

	IOU	**Precision**	**Recall**	**F1 score**
BackGround	0.964	0.967	0.971	0.969
Glomeruli	0.957	0.963	0.959	0.96
Podocyte	0.802	0.926	0.892	0.908
Non-Podocyte	0.837	0.903	0.927	0.914
mean	0.891	0.94	0.937	0.932

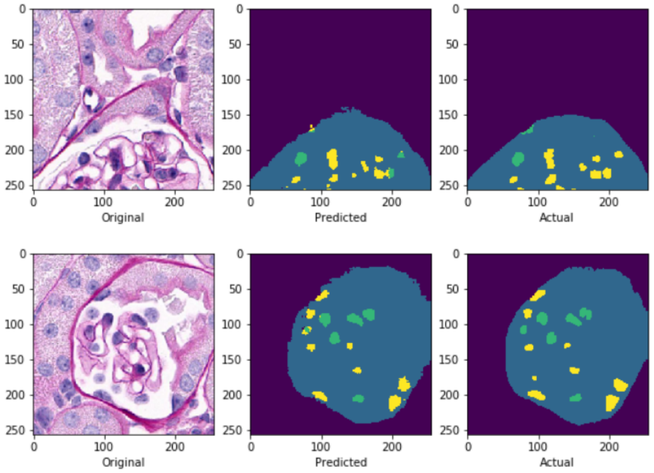


Fig. 5. PatchGAN Results

2) Fully-Convolutional-Network-Discriminator:

	IOU	**Precision**	**Recall**	**F1 score**
BackGround	0.989	0.992	0.996	0.994
Glomeruli	0.966	0.986	0.979	0.983
Podocyte	0.835	0.952	0.871	0.910
Non-Podocyte	0.856	0.912	0.932	0.922
mean	0.911	0.961	0.944	0.953

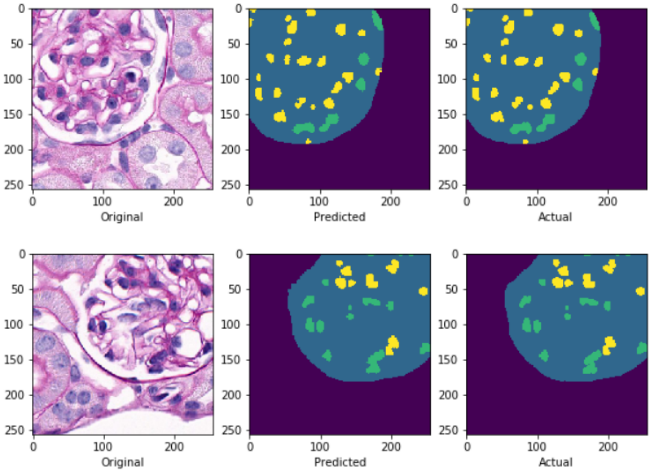


Fig. 6. Fully Convolutional Networ Results

V. Conclusion

This model propose an adversarial learning technique for semi-supervised semantic segmentation using the discriminator to train the Deeplab network on both labeled and unlabeled data. For supervised training, the discriminator is trained to learn joint probability of original image and its segmented image. For unsupervised training, the probability maps generated by the discriminator network is used for rening the segmentation network.

References

[1] Brendon Lutnick, Brandon Ginley, Darshana Govind, Sean McGarry, Peter Laviolette, Rabi Yacoub, Sanjay Jain, John Tomaszewski, Kuang-Yu Jen, and Pinaki Sarder. An integrated iterative annotation technique for easing neural network training in medical image analysis. *Nature Machine Intelligence*, 1:112–119, 02 2019.

[2] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation, 2018.

[3] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation, 2014.

[4] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation, 2015.

[5] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. Image-to-image translation with conditional adversarial networks, 2016.

[6] Wei-Chih Hung, Yi-Hsuan Tsai, Yan-Ting Liou, Yen-Yu Lin, and Ming-Hsuan Yang. Adversarial learning for semi-supervised semantic segmentation, 2018.

[7] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets, 2014.

[8] Chi-Feng Wang. A basic introduction to separable convolutions, 2018.

[9] Pinaki Sarder. Podocyte dataset.