Project Report

Trang Le, Mohamed Reda Keddar, Senbai Kang

May 7, 2019

Contents

1 Literature review

1.1 Detecting Cancer Metastases on Gigapixel Pathology Images [1]

1.1.1 Introduction

In [1], the authors proposed a framework to automatically detect and localise local (lymph node) metastatic lesions in gigapixel pathology images, basing their models on the detection framework proposed in [4], and the convolutional neural network (CNN) architecture "Inception" (V3) presented in [5]. Moreover, a custom sampling strategy along with different data augmentation techniques were employed to address the imbalanced nature of the used dataset (Camelyon16).

1.1.2 Methods

Training set sampling The authors trained their models with 270 pathology slides (images) from the Camelyon16 dataset. Each slide represents H&E-stained (Hematoxylin and Eosin) human lymph node tissue scanned at a magnification of 40X. Given the size of the whole slide images (WSI), each slide was divided into smaller patches.

The number of patches per slide ranged from 10,000 to 400,000 (median 90,000), and tumour slides contained only 20 to 150,000 tumour patches (median 2,000). In order to avoid bias towards slides containing more patches (normal or tumour), the following sampling strategy was adopted: after selecting "normal" or "tumour" with equal probability, a slide containing that class of patches was selected uniformly at random, and then patches were sampled from that slide. Likewise, to mitigate the imbalance between tumour patches and normal ones, different data augmentations were applied on the tumour patches, such as rotations and image flips.

Training The training was done in two phases: the patch-based classification phase and the heatmap-based inference stage [4].

During the patch-based classification phase, the Inception V3 architecture is used, with input patches of size 299×299 pixels. For each input patch, the label of the centre 128×128 region is considered for prediction. A patch is labelled as tumour if at least one pixel in the centre region is labelled as such. Here, the influence of the number of Inception V3 parameters was investigated, and a multi-scale approach where the models were trained with down-sampled patches of 20X and 10X magnifications was also explored.

Next, inference was performed across the slide in a sliding window with a stride of 128 to match the centre region's size, thus generating a probability heatmap. The maximum value in the heatmap was reported as the slide-level tumour prediction.

During training, the following approaches were tested but did not yield improvements in the models' performance:

- Multi-scale detection approach (at 20X and 10X magnifications)
- Pre-training the model on ImageNet image recognition
- Colour normalisation

1.1.3 Results

The authors used the Camelyon16 dataset to train and test their model, with 270 pixel-level-annotated slides for training and validation (159 normal and 110 tumour) and 130 slides for testing. Background patches, representing biological tissues other than those of interest (such as adipocytes), were removed to reduce computation. Additionally, NHO-1, a 110-slide dataset, was digitised and labelled in order to be used as an independent evaluation set.

Two metrics were used to evaluate the models: area under the ROC curve (AUC), and FROC which is a metric used to evaluate tumour detection and localisation.

Regarding the slide-level classification, which is obtained by taking the maximum probability value of a slide's corresponding heatmap, the models achieved AUCs greater than 97%. For tumour-level classification, the models achieved FROCs ranging from 85.5% to 88.5%.

On the NHO-1 independent dataset, the models achieved an AUC of 97.6%.

1.2 U-net: Convolutional Networks for Biomedical Image Segmentation [2]

1.2.1 Introduction

U-net was a winning deep learning algorithm on the ISBI cell tracking challenge 2015 with overwhelming performance and efficiency on image segmentation. The authors mainly proposed solutions to three major problems in the field of biomedical image processing, and applied u-net to three segmentation tasks to test the performance.

1.2.2 Methods

It is agreeable that a large amount of images is essential to successful deep learning. In practice, however, only a limited, small number of images is usually available, which poses the first problem to the application of deep learning in biomedical image processing. The authors employed data augmentation by applying elastic deformations to the images, which not only enlarged the available data, but had practical meaning that deformation is the most common variation in tissue.

The second problem is that, although convolutional neural networks are commonly used in the classification of images, there is a need of segmenting different areas in an image into corresponding classes. In other words, instead of labelling the whole image of a specific class in classification, each pixel in an

image was labeled of its corresponding class in segmentation. In order to localize, the authors built a structure of two symmetrical processes, i.e., contraction and expansion. The former was a typical combination of convolutional layers and max pooling layers, which extracted features of the input image while ignoring the information of location. By applying unsampling operators, the latter recovered the information of location, and created high-resolution segmentation maps. Note that in the contraction process, all the convolutional layers only used the valid part of input data, which means that they were not padded and would shrink during convolution. Meanwhile, The authors employed a strategy of overlap-tile, where prediction of the segmentation in a specific area in the input image required a larger area with context information. If there is missing data in the larger area, it will be extrapolated by mirroring. Finally, dropout layers appeared at the end of the contraction process performing additional implicit data augmentation.

The last challenge is the separation of connecting objects from the same class, which is extensively required in biomedical segmentation problems. The authors suggested the application of a weighed loss $w(\boldsymbol{x})$ in 1, which gave more importance to pixels of separating background between connecting cells in the loss function.

1.2.3 Training setup

In the training process, the loss function of u-net was computed by a pixel-wise softmax over the final feature map:

$$E = \sum_{\boldsymbol{x} \in \Omega} w(\boldsymbol{x}) log(p_{l(\boldsymbol{x})}(\boldsymbol{x})) \tag{1}$$

where $l : \Omega \to \{1, \ldots, K\}$ is the true label of each pixel and:

$$w(\boldsymbol{x}) = w_c \boldsymbol{x} + w_0 \cdot \exp(-\frac{(d_1(\boldsymbol{x}) + d_2(\boldsymbol{x}))^2}{2\sigma^2}) \tag{2}$$

where $w_c : \Omega \to \mathbb{R}$ is the weight map to balance the class frequencies, $d_1 : \Omega \to \mathbb{R}$ is the distance to the border of the closest cell and $d_2 : \Omega \to \mathbb{R}$ is the distance to the border of the second closest cell. Setting $w_0 = 10$ and $\sigma = 5$ pixels.

1.2.4 Results

U-net was first applied to the task of the segmentation of neuronal structures in electron microscopic recordings. It achieved a warping error of 0.0003529, the new best score. In terms of rand error, u-net was higher than the model of IDSIA-SCI with 0.0382 to 0.0189, but the latter employed data set which was post-processed, while u-net used the original data set.

The second task u-net was applied to is a cell segmentation task in light microscopic images with two data sets. On the first data set "Phc-U373", u-net achieved an average intersection over union (IOU) of 92%, while the score of the second best model is 83%. On the second best data set "DIC-HeLa", u-net

acquired an average IOU of 77.5%, while that of the second best algorithm is 46%.

1.3 Weakly-Supervised Semantic Segmentation by Iteratively Mining Common Object Features [3]

1.3.1 Introduction

Bridging the gap between high-level semantic (segmentation label) to low-level appearance (image detail) is a challenging task, especially when training with weakly labeled images. Since no pixel-to-pixel mask is available for training, previous models have to rely only on image labels to localize objects. Inaccurate and coarse discriminative object region detection can harm the performance of these models. To solve this problem, the authors proposed an iterative bottom-up and top-down framework which alternatively expands object regions and optimizes segmentation network.

1.3.2 Methods

The authors proposed a Mining Common Object Features (MCOF) framework which contains RegionNet (top-down), PixelNet (bottom-up) and a saliency guide to iteratively produce more refined segmentation mask (details in Figure 1).

RegionNet Classification networks can produce initial localization of regions (hypothetically) containing significantly common features about objects. After each training round (epoch), the refined object regions (output of PixelNet) are used as training data to predict object masks (new initial localization, higher accuracy).

PixelNet From these initial localizations output from RegionNet, common object features were mined. The object regions were expanded with mined features.

Saliency-guide To segment non-discriminative regions, saliency maps are then considered under Bayesian framework to refine the object regions.

1.3.3 Experimental setup

Data Pascal VOC 2012 dataset: 20 class objects + 1 background class. For the segmentation task, it contains 1464 training, 1449 validation and 1456 test images. They used augmented data of 10,582 images as training data set.

Hardware Not mentioned!

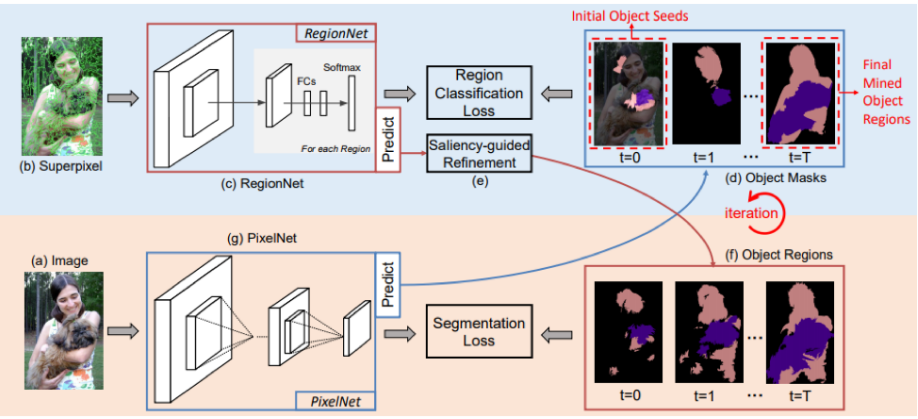

Figure 1: Pipeline of the proposed MCOF framework. At first (t=0), we mine common object features from initial object seeds. We segment (a) image into (b) superpixel regions and train the (c) region classification network RegionNet with the (d) initial object seeds. We then re-predict the training images regions with the trained RegionNet to get object regions. While the object regions may still only focus on discriminative regions of object, we address this by (e) saliency-guided refinement to get (f) refined object regions. The refined object regions are then used to train the (g) PixelNet. With the trained PixelNet, we re-predict the (d) segmentation masks of training images, are then used them as supervision to train the RegionNet, and the processes above are conducted iteratively. With the iterations, we can mine finer object regions and the PixelNet trained in the last iteration is used for inference. [3]

1.3.4 Results

The MCOF framework outperformed previous state-of-the-art weakly-supervised semantic segmentation models by a large margin. The performance metrics for comparison is mIOU (mean Intersection over Union of all 21 classes), and benchmark models including CCNN, EM-Adapt, MIL-sppxl, STC, DCSM, BFBP, AF-SS, SEC, CBTS and AE-PSL.

Conclusion The MCOF framework has two main advantages over previous approaches. First, the iterative bottom-up and top-down framework tolerates inaccurate initial object localization. By iteratively mining common object features, the model can progressively produce segmentation masks with improving accuracy. Second, saliency-guided refinement method can allow for non-discriminative regions in initial localization, hence increasing segmentation accuracy. This framework proved new state-of-the-art performance.

References

[1] Yun Liu, Krishna Gadepalli, Mohammad Norouzi, George E. Dahl, Timo Kohlberger, Aleksey Boyko, Subhashini Venugopalan, Aleksei Timofeev, Philip Q. Nelson, Gregory S. Corrado, Jason D. Hipp, Lily H. Peng, and Martin C. Stumpe. Detecting cancer metastases on gigapixel pathology images. *CoRR*, abs/1703.02442, 2017.

[2] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, 2015.

[3] Xiang Wang, Shaodi You, Xi Li, and Huimin Ma. Weakly-supervised semantic segmentation by iteratively mining common object features. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1354–1362, 2018.

[4] Dayong Wang, Aditya Khosla, Rishab Gargeya, Humayun Irshad, and Andrew H. Beck. Deep Learning for Identifying Metastatic Breast Cancer. *arXiv e-prints*, page arXiv:1606.05718, Jun 2016.

[5] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott E. Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–9, 2015.