One-shot-learning Rotation-Invariant Convolutional Neural Network for Digital Pathology Slide Segmentation

Lyndon Chan
University of Toronto
lyndon.chan@mail.utoronto.ca

***Abstract*—There is currently a need for efficient and robust computational tools to automatically annotate digital histopathological images for computer-aided diagnosis by human experts. Convolutional Neural Networks have lately become the state-of-the-art in image classification and have recently been applied to semantic segmentation. However, large, openly-available databases of digital pathology images with good quality ground-truth segmentation are scarce and tissues may be arbitrarily rotated. Convolutional neural networks usually require large training sets and have no in-built rotation invariance. Hence, I present PolarNet, a deep convolutional neural network architecture that incorporates spatial local polar histogramming in order to be rotation-invariant and require minimal training ("one-shot learning"). I implement the proposed neural network, train it on a minimal set of digital pathology images, and argue that, while it fails to out-perform an equivalent Fully-Convolutional Network (FCN) preliminary tests, it is a promising direction for designing shallow convolutional neural networks.**

I. Overview

Traditionally, the task of pathology diagnosis has been solely the domain of human experts, who would extract a tissue specimen from a patient, stain it, and examine it under a microscope. But this process was slow and error-prone, and once digital histopathology scanners were made widely available, pathologists could view high-resolution scans of entire physical tissue specimens (known as Whole Slide Images (WSI)) on a computer screen [8]. And with the proliferation of computing resources, automated tools for computer-aided diagnosis (CAD) were used by human experts in many tasks. In this paper, I will focus on one of these tasks - semantic tissue segmentation.

The task of semantic tissue segmentation is to assign a single semantic label to each pixel in a histopathological image. As such, the task is different from unsupervised segmentation, which assigns a non-semantic, usually appearance-based label to each pixel in an image; and the task is also different from image classification, which assigns a single semantic label to each image as a whole.

In general, image analysis techniques used in other problem domains can also be applied to digital pathological image analysis. However, histopathological images are usually very large in size, with a typical 40X scanned image requiring up to 14.5 GB [2], and there is a scarcity of digital pathology images openly available with good-quality ground-truth labels. As a result, if an effective semantic tissue segmentation algorithm is to be trained, it must learn with a very small set of good-quality training data and generalize robustly from this prior knowledge to segment images it has never seen before (known as "One-Shot Learning"). We can also note that, as digital pathology images form a specific subset of all possible images, certain assumptions can be made which make the problem simpler:

1) Tissues are generally formed of regularly-repeating patterns of elementary structural components (e.g. cells) - suggestive of texture modelling
2) Tissues can be subjected to various transformations and deformations (such as spatial translation, rotation, shear, shape deformation) without changing their true semantic label
3) Tissues are imaged under approximately uniform lighting conditions

In this report, I will focus on the problems of overall recognition accuracy and rotation invariance for semantic segmentation of digital pathology images. I discuss relevant research in Section II, my proposed method in Section III, experimental results in Section IV, and discussion and future directions in Section V.

II. Related Work

In this section, I will describe related work in rotation-invariant image representations and semantic segmentation methods, especially those applied to analysis of digital histopathological images.

A. *Hand-crafted Methods*

SIFT (Scale-Invariant Feature Transform) [5] is a feature detector and extractor algorithm which extracts a 128-feature vector composed of a magnitude and orientation histogram of the pixels in a local grid neighbourhood around localized keypoints.

GLOH (Gradient Location and Orientation Histogram) [7] is another feature detector and extractor algorithm similar to SIFT, but which utilizes a log-polar grid instead, which the authors argues offers greater translation and rotation invariance.

Generic descriptor [11] extends the idea of locally-pooled feature extractors by learning the optimal pooling arrangement, which the authors conclude closely resembles the GLOH local pooling arrangement.

FESI (Foreground Extraction from Structure Information) [1] is a hand-crafted feature extractor specially designed for digital pathology, and uses a Gaussian-blurred Laplacian filter to extract structural information from an image.

B. *Machine Learning Methods*

FCN (Fully Convolutional Network) [4] is an adaptation of other state-of-the-art convolutional neural networks for the semantic segmentation task. It applies successive convolutional layers followed by an up-sampling layer to reduce the input image dimensionality into a compressed feature representation, and then up-sample this into a pixel segmentation map.

U-Net [9] is a convolutional neural network specially trained for semantic segmentation of digital pathology images, which applies a succession of convolutional layers and transpose convolution layers for up-sampling, just like Long's FCN. However, U-Net includes skip connections to transfer information from earlier layers to later layers.

T-Net [6] is a convolutional neural network based on U-Net for segmenting digital pathology images, and has a similar architecture, except it adds additional convolutional layers in the skip connections between layers.

III. Method

In this section, I describe two proposed methods for semantic texture segmentation (depicted in 1). The first proposed method (called PolarNet-1) is a modification of a Fully Convolutional Network, with its first convolution layer replaced with: (1) a convolution layer with a fixed convolutional filter bank (using MaxPol filters [3]) for edge and bar extraction, (2) a convolution layer with a fixed convolution filter bank (using polar pooling masks), (3) an orientation, depth concatenation, and switching layer to re-arrange the previous features, and (3) a 1 × 1-kernel convolution layer to combine the previous features, followed by a non-linearity and pooling function. The second proposed method (called PolarNet-2) is similar, but omits the MaxPol filters and only uses the polar pooling convolutions without feature re-arrangement. Subsequent layers for both proposed networks is identical.

A. *Architecture*

In this part, I briefly describe the unique aspects of the architectures of PolarNet-1 and PolarNet-2.

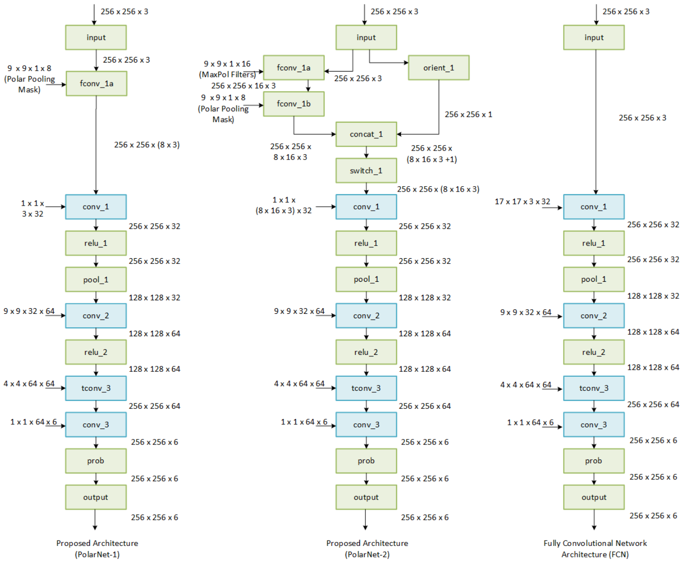

Fig. 1. The architectures of PolarNet-1 (left), PolarNet-2 (middle), and FCN (right). Note that the layers with learnable parameters are depicted in blue and fixed layers in green.

Polar Convolution Layers

Drawing on the local feature extractor work of [7], and [11], I propose replacing the rectangular-receptive-field $k_1 \times k_2 \times d_{in} \times d_{out}$ convolution layer normally used in convolutional neural networks with: (1) a $k_1 \times k_2 \times 1 \times n_{bins}$ convolution layer with polar pooling masks, then (2) a $1 \times 1 \times n_{bins} \times d_{in} \times d_{out}$ convolution layer. This is equivalent to performing a $k_1 \times k_2 \times n_{bins} \times d_{in} \times d_{out}$ convolution with a polar receptive field instead of a $k_1 \times k_2 \times d_{in} \times d_{out}$ convolution with a rectangular receptive field. Each of the n_{bins} masks consists of two unit-sum 2D Gaussian functions shifted to the dual centres of each numbered bin depicted in 3, and with a radius commensurate with the radial bin extent.

I hypothesize that this “polar convolution” arrangement is advantageous for semantic segmentation, because it:

- Enables more combinations of preceding information to be learned, i.e. $n_{bins} \times d_{in} \times d_{out}$ instead of $d_{in} \times d_{out}$
- Should be more invariant to translation and rotation, as each discrete value in the receptive field corresponds with an spatial bin instead of a single pixel
- Requires learning far fewer parameters; a rectangular convolution layer requires the learning of $k_1 \times k_2 \times d_{in} \times d_{out}$ parameters, but a polar convolution layer as proposed requires the learning of only $1 \times 1 \times n_{bins} \times d_{in} \times d_{out}$ parameters

Orientator Layer

In PolarNet-2, an orientator layer computes the gradient magnitude with a local sliding window $\mathcal{L}(x, y)$ at each pixel (x, y) of the incoming image (after first converting to grayscale) [10], extracts the gradient centroid (or dominant orientation), and then bins it in an angular histogram, as depicted in diagram 2:

$$\phi_d(x,y) = \tan^{-1} \frac{\sum_{(u,v)\in\mathcal{L}(x,y)} vG(u,v)}{\sum_{(u,v)\in\mathcal{L}(x,y)} uG(u,v)}$$

$$\phi_b(x,y) = \begin{cases} 1, & \text{if } -90^\circ \leq \phi_d(x,y) < -67.5^\circ, \\ & 67.5^\circ \leq \phi_d(x,y) < 90^\circ \\ 2, & \text{if } 22.5^\circ \leq \phi_d(x,y) < 67.5^\circ \\ 3, & \text{if } -22.5^\circ \leq \phi_d(x,y) < 22.5^\circ \\ 4, & \text{if } -67.5^\circ \leq \phi_d(x,y) < -22.5^\circ \end{cases} \tag{1}$$

Switching Layer

For PolarNet-2, after the dominant orientation is extracted at the pixel level, a switching level re-arranges the features outputted by the polar convolution layer `fconv_1`$binthediagram$1 to re-align the features with their dominant orientations. For example, in the diagram 3, a non-horizontally-aligned local patch is re-aligned to a canonically-arranged polar pooling mask by a circular shift counter-clockwise of 2 bins.

Convolution Layers

A $k_1 \times k_2 \times d_{in} \times d_{out}$ convolution layer applies d_{out} linear sums of each of the preceding d_{in} feature maps in a local neighbourhood of $k_1 \times k_2$. When a $1 \times 1 \times n_{bins} \times d_{in} \times d_{out}$ set of filters convolves an incoming feature map (as is done with the polar convolution discussed above), each incoming feature map's pixel is summed up independently of nearby pixels.

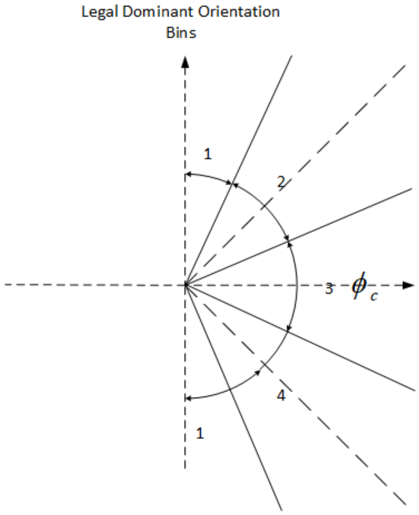

Fig. 2. The legal dominant orientation bins used in the Orientator Layer

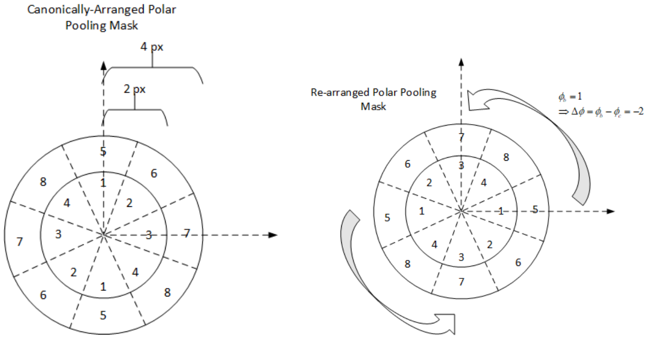
Canonically-Arranged Polar Pooling Mask
4 px
2 px
Re-arranged Polar Pooling Mask
$\phi_b = 1$
$\Rightarrow \Delta\phi = \phi_b - \phi_c = -2$

Fig. 3. The canonically-arranged polar pooling mask (left), and a re-arranged polar pooling mask aligned to the canonical arrangement (right)

Transposed Convolution Layers

Differently from convolutional neural networks used for image classification, semantic segmentation networks have to assign a label per pixel. Hence, it is necessary to up-sample the feature map back to the original image dimensions after down-sampling by the pooling layers. This is accomplished through the use of a transposed convolution layer which is identical to a convolution layer with fractional stride.In PolarNet-1 and PolarNet-2, only one such layer is used for 2X up-sampling.

B. Problem Formulation

The problem of learning the networks can be formulated as a non-linear and non-convex classic optimization problem of learning the optimal set of K network parameters $\Theta^* = \{\theta_1^*, \cdots, \theta_K^*\}$ which minimizes an objective function f_0 associated with predicting class labels $\hat{Y}$ for a given input image X and a set of network parameters $\Theta = \{\theta_1, \cdots, \theta_K\}$, and compared with ground-truth class labels Y:

$$\begin{aligned}
\Theta^* = \underset{\Theta}{\operatorname{argmin}} \quad & f_0(X;\Theta) = -\sum_{p=1}^{P} Y_p(X) \log \hat{Y}_p(X;\Theta), \\
\text{subject to} \quad & \sum_{p=1}^{P} \hat{Y}_p(X;\Theta) = 1, \\
\text{where} \quad & X \in \mathbb{R}^{h\times w\times d}, \\
& Y \in \mathbb{R}^{h\times w\times d\times P}, \\
& \Theta = \{\theta_1, \cdots, \theta_K\}
\end{aligned}$$

Due to the non-linear and non-convex nature of the problem, simplification schemes such as Lagrangian duality and KKT methods cannot be applied. Stochastic gradient descent and backpropagating the classification output to each layer's parameters in the network gives an effective learning heuristic which tends to converge to the global optimal network parameter configuration Θ^*:

$$\Theta^{(t)} := \Theta^{(t-1)} - \eta \frac{\partial f_0(X;\Theta^{(t-1)})}{\partial \Theta^{(t-1)}}$$

IV. Experiments

A. Setup

In this section, I will describe some preliminary results of my investigation into the suitability of polar convolution layers for semantic tissue segmentation. Both PolarNet-1 and PolarNet-2 were implemented on MATLAB R2017b, and built using the Neural Network library, then compared to an equivalent Fully Convolutional Network with an architecture identical to the sample semantic segmentation network provided by MathWorks, but specially trained on my custom data. The training and testing was conducted on a HP Z440 Workstation with 128 GB of RAM and an NVIDIA GeForce GTX 1080 Ti GPU.

B. Data

The data used in this paper consisted of a set of 18 histopathology images selected from a larger database of fluorescence microscopy scans specifically to maximize illumination uniformity and diversity of tissue structures, and were hand-labelled pixel-by-pixel with the 6 tissue classes, each exemplified by the single training images each depicted in Figure 4:

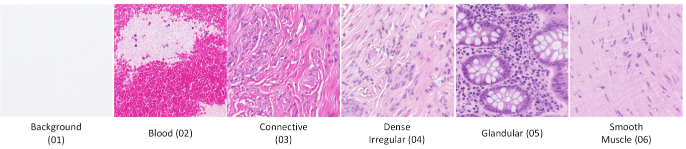

Fig. 4. The six training images used for the experiments

The training-test split was: 6 training images and 12 testing images, where each tissue class was represented by a training image predominantly consisting of that tissue class.

C. Experiment 1: Training Convergence

Each of the evaluated networks was trained using the above-mentioned training information on a mini-batch size of all 6 images and a maximum epoch count of 100. The plots of the training accuracy and loss are provided in 5. It is evident that FCN converges fastest and attains overall best training accuracy, while PolarNet-1 has not completed training by the maximum epoch point, and PolarNet-2 saturates in training after the first few epochs.

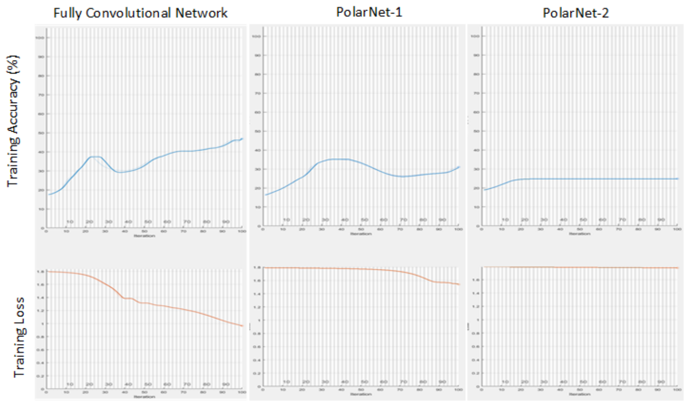

Fig. 5. The training convergence plots of all three networks: FCN (left), PolarNet-1 (middle), and PolarNet-2 (right)

D. Experiment 2: Pixel Classification Accuracy

Only preliminary experiments were completed at the time of writing, and they indicate that, although the polar convolution concept has a feasible theoretical basis for improving semantic segmentation performance, in fact, a simple, equivalent Fully Convolution Network trained on the same data is still more computationally-efficient and accurate than either PolarNet-1 and PolarNet-2. The qualitative and quantitative results of the semantic segmentation are presented respectively in 6 and I.

V. Discussion / Future Directions

It is evident from the qualitative and quantitative results presented above that, after training on the extremely

TABLE I

THE QUANTITATIVE SEMANTIC SEGMENTATION RESULTS, FROM ALL THREE EVALUATE NETWORKS, MEASURED WITH THE JACQUARD INDEX

	Test Image #												
Method	1	2	3	4	5	6	7	8	9	10	11	12	Avg
FCN	0.3908	0.7955	0.3471	0.2178	0.2102	0.2433	0.4758	0.5201	0.396	0.4829	0.001	0.0009	0.340117
PolarNet-1	0.4089	0.7706	0.1442	0.1711	0.0812	0.0529	0.0272	0.065	0.0056	0.1122	0	0	0.153242
PolarNet-2	0	0.3641	0.7712	0.7657	0	0	1	1	0.2304	0.3418	0	0	0.372767

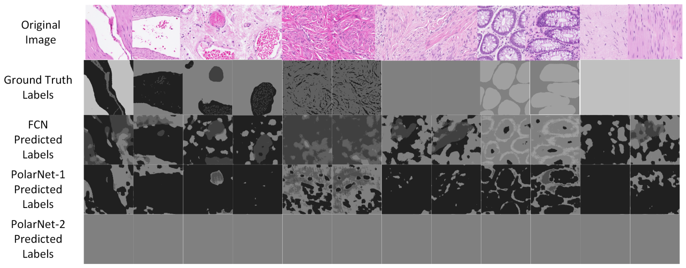

Fig. 6. The qualitative semantic segmentation results, from all three evaluated networks, along with the original images and ground-truth labels

small training set of digital pathology images, none of the three evaluated methods perform particularly well in semantic segmentation quality. The FCN does the best overall over all twelve test images, and PolarNet-1 still attains reasonable results, and in fact is both numerically and visually superior in isolating tissues from the background. On the other hand, there is clearly something failing in PolarNet-2, as it consistently assigns the label "Dense Irregular" to any pixel inputted to it.

The preliminary results presented in this paper do not support the hypothesis that the proposed polar convolution offers superior semantic segmentation performance over rectangular-receptive-field convolution. However, for many theoretical reasons, it offers an attractive alternative and further investigation is warranted. An obvious experiment to conduct would be to evaluate the rotation invariance of the semantic segmentation networks.

REFERENCES

[1] D. Bug, F. Feuerhake, and D. Merhof. Foreground extraction for histopathological whole slide imaging. In *Bildverarbeitung für die Medizin 2015*, pages 419–424. Springer, 2015. 2

[2] E. Chlipala, J. Elin, O. Eichhorn, M. Krishnamurti, R. E. Long, B. Sabata, and M. Smith. Archival and retrieval in digital pathology systems. Digital Pathology Association, Madison, WI, 2010. 1

[3] M. S. Hosseini and K. N. Plataniotis. Derivative kernels: Numerics and applications. *IEEE Transactions on Image Processing*, 26(10):4596–4611, 2017. 2

[4] 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. 2

[5] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110, 2004. 2

[6] G. Makarchuk, V. Kondratenko, M. Pisov, A. Pimkin, E. Krivov, and M. Belyaev. Ensembling neural networks for digital pathology images classification and segmentation. *arXiv preprint arXiv:1802.00947*, 2018. 2

[7] K. Mikolajczyk and C. Schmid. A performance evaluation of local descriptors. *IEEE transactions on pattern analysis and machine intelligence*, 27(10):1615–1630, 2005. 2

[8] L. Pantanowitz. Digital images and the future of digital pathology. *Journal of pathology informatics*, 1, 2010. 1

[9] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015. 2

[10] P. L. Rosin. Measuring corner properties. *Computer Vision and Image Understanding*, 73(2):291–307, 1999. 3

[11] S. A. Winder and M. Brown. Learning local image descriptors. In *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*, pages 1–8. IEEE, 2007. 2