Supplemental Material A: Details about the feature extraction

Yushan Zheng

January 17, 2018

1 Motivation

In the body of this paper, the feature extraction framework is an extension of work Zheng et al. [1]. In Zheng et al. [1], histopathological features are extracted from nucleus locations using a designed neural network. The key steps of encoding for a region of interest (ROI) are illustrated in Figure 1. The nuclei in a ROI are first detected. Then, the patches centered on the nuclei are extracted and flattened into column vectors. These patches are encoded by a neural network to obtain patch features. Finally, these features are quantified by max-pooling to generate the feature of the ROI.

To apply the algorithm to WSI analysis proposed in this paper, there are two issues that need to be solved. (1) the algorithm is designed for square ROIs and the unit to analysis in this paper is superpixel, which is non-square. 2) the patches in the algorithm are defined by a nuclei detection method, which is only effective for images in high resolutions (e.g. under a 20× lens). However, in this paper, multiple resolutions (under 2×, 5×, 10×, 20× lens) are required to analysis. To utilize the feature extraction algorithm to low resolutions, an effective key points detection method is required.

According to research [2], scale-invariant feature transform (SIFT) [3] is effective in locating key regions in histopathological images. Therefore, we proposed applying SIFT points to define patches and using these patches to construct the feature extraction framework proposed in [1].

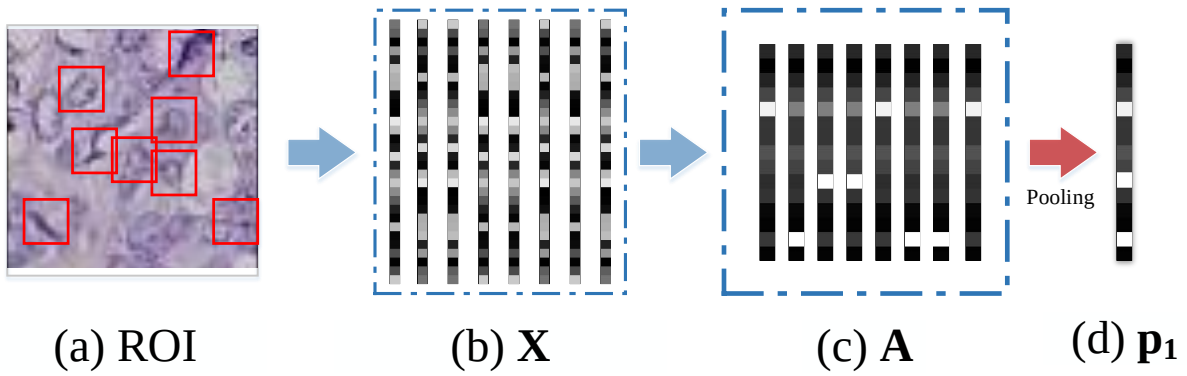
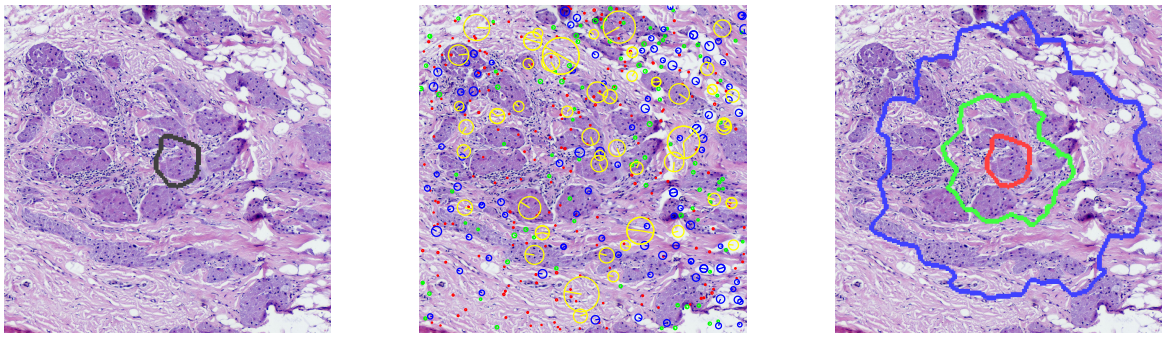


Figure 1: Feature extraction of the nucleus-guided neural network, where (a) shows a region of interest (ROI) in a histopathological image and nuclei detected in the ROI, (b) is the flattened column vectors of nucleus patches, (c) denotes nucleus-level features extracted by neural network, and (d) represents the ROI-level feature obtained by max-pooling operation.

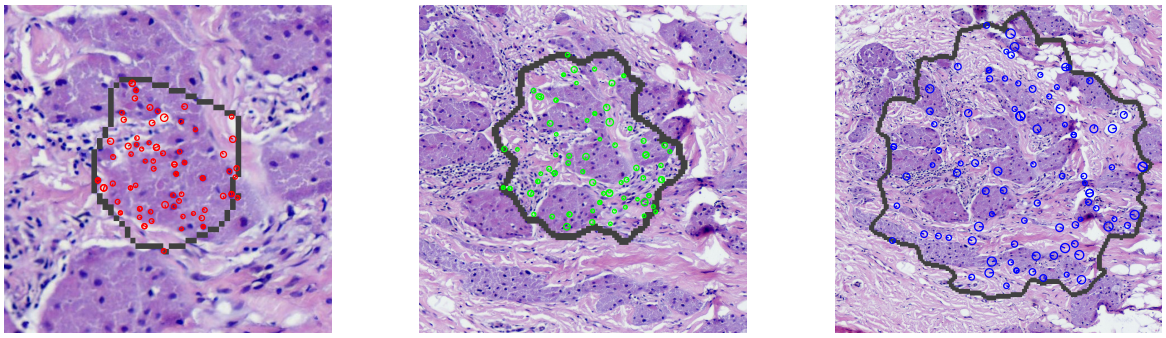
In this paper, the feature extraction is in connection with context definition and SIFT points detection. The details and relationships about these algorithms are presented as follows.

2 Context and SIFT points

SIFT points are detected in different octaves (i.e., different resolutions). In general, the first octave represents the original resolution of the image and the next octave concerns the resolution that is half of the previous



(a) Superpixel (b) SIFT points (c) Context regions



(d) $C_k^{(0)}$ & $Octave = 0$ (e) $C_k^{(1)}$ & $Octave = 1$ (f) $C_k^{(3)}$ & $Octave = 2$

Figure 2: Allocation of SIFT points for contextual regions ($L = 2$) for a superpixel, where (a) shows the superpixel, (b) displays the SIFT points detected in the images and points detected within $Octave = 0, 1, 2$, and ≥ 3 are respectively drawn in red, green, blue, and yellow (For clearance, only a part of SIFT points are displayed.), (c) displays the three context regions for the superpixel, and (d),(e), and (f) separately present the three context regions and the SIFT points these regions concerned.

octave [4]. The relationship between octaves is essentially consistent with the relationship between resolutions (under 20×, 10×, 5×, 2× lens) considered in our framework. Therefore, the SIFT points are detected from images under a 20× lens, and then assigned into four groups according to octave. Corresponding to the four selected resolutions, four context regions are defined (section II.B.3 in the paper). The feature of each context region is extracted from the SIFT points involved in the context region. Letting $octave = 0$ denote the first octave, the relationships between context regions, resolutions and scale of SIFT points are presented in Table 1. And Figure 2 gives an instance with a superpixel, where the first three context regions ($l = 1, 2, 3$) and the SIFT points in different octaves are displayed.

Table 1: Relationships between context regions, resolutions and scale (Octave) of SIFT points.

Context index	Context region	Magnification of lens	Resolution	Octave
$l = 0$	$C_k^{(0)}$	20×	1.2μm/pixel	0
$l = 1$	$C_k^{(1)}$	10×	2.4μm/pixel	1
$l = 2$	$C_k^{(3)}$	5×	4.8μm/pixel	2
$l = 3$	$C_k^{(7)}$	2×	12μm/pixel	$\geq$3

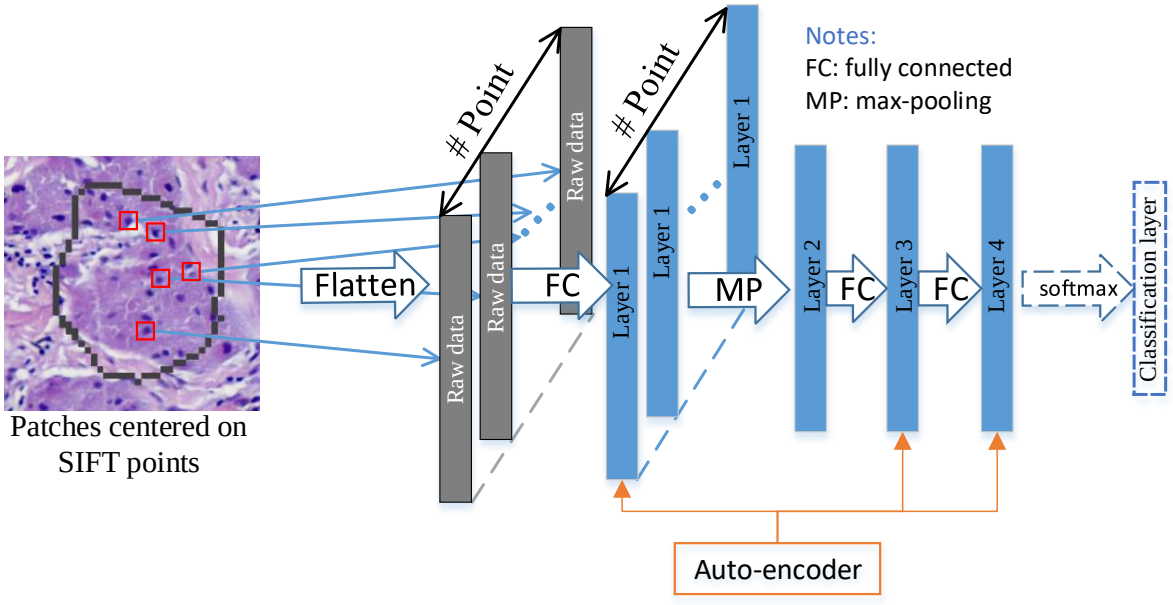


Figure 3: Structure of the neural network, where the patches data are flattened into column vectors, and then fed to a network consisting of three fully connected layers and a max-pooling layer.

3 Feature Extraction Neural Network

3.1 Structure

In this paper, four neural networks (NNs) corresponding to the four context regions are constructed. The layer structures of the four neural networks are the same, which is shown in Figure 3. For a certain context region, the patches centered on the corresponding SIFT points are extracted, and the pixel data in the patches are flattened into a set of column vectors. Letting $\mathbf{x}_i$ denote the column vector of the i-*th* patch, the encoding of layer 1 can be represented as

$$\mathbf{a}_i^{(1)} = \sigma(\mathbf{W}^{(1)\mathrm{T}}\mathbf{x}_i + \mathbf{b}^{(1)}), \tag{1}$$

where $\mathbf{W}^{(1)} = [\mathbf{w}_1^{(1)}, \mathbf{w}_2^{(1)}, \ldots, \mathbf{w}_K^{(1)}]$ and $\mathbf{b}^{(1)} = [b_1^{(1)}, b_2^{(1)}, \ldots, b_K^{(1)}]^{\mathrm{T}}$ are the weights and bias, K is the neuron number of layer 1, and σ is the activation function. In this paper, σ denotes the sigmoid function $\sigma(t) = 1/(1+e^{-t})$. To fit the input of sigmoid function, the pixel data of the patches is normalized. In this paper, the raw pixel data is truncated ± 3 times standard deviations and then squashed to $[0, 1]$.

Then, the activations of patches in a superpixel are quantified into one representation (layer 2 in Figure 3). Letting matrix $\mathbf{A}^{(1)} = (\mathbf{a}_1^{(1)}, \mathbf{a}_2^{(1)}, ..., \mathbf{a}_N^{(1)})$ denote the activations of the N patches in the superpixel. Then, the quantified representation is defined as the max-pooling result of $\mathbf{A}^{(1)}$:

$$\mathbf{a}^{(2)} = (\|\mathbf{A}_1^{(1)}\|_\infty, \|\mathbf{A}_2^{(1)}\|_\infty, ..., \|\mathbf{A}_K^{(1)}\|_\infty)^{\mathrm{T}}, \tag{2}$$

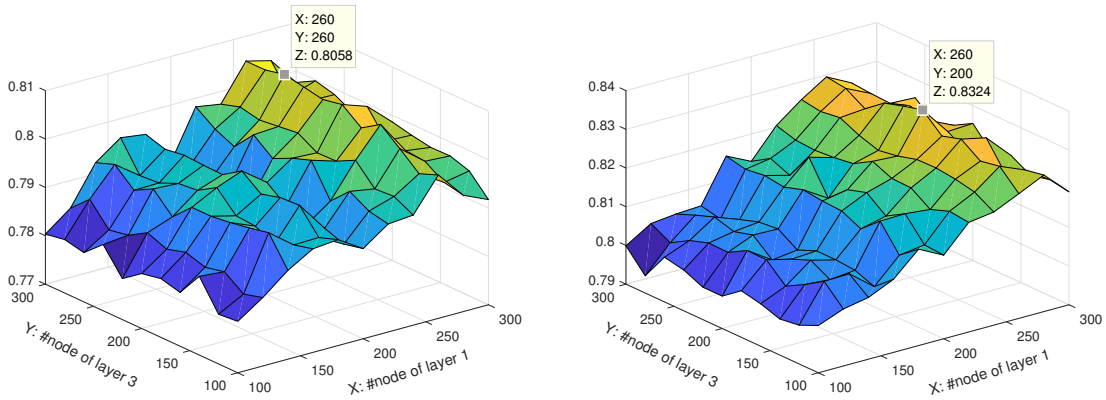
where $\mathbf{A}_k^{(1)}$ denotes the k-*th* row of $\mathbf{A}^{(1)}$, and $\|\cdot\|_\infty$ is the infinite norm.

Afterward, two other layers (layers 3 and 4 in Figure 3) are used to encode the representation of the patch $\mathbf{a}^{(2)}$, which can be represented as

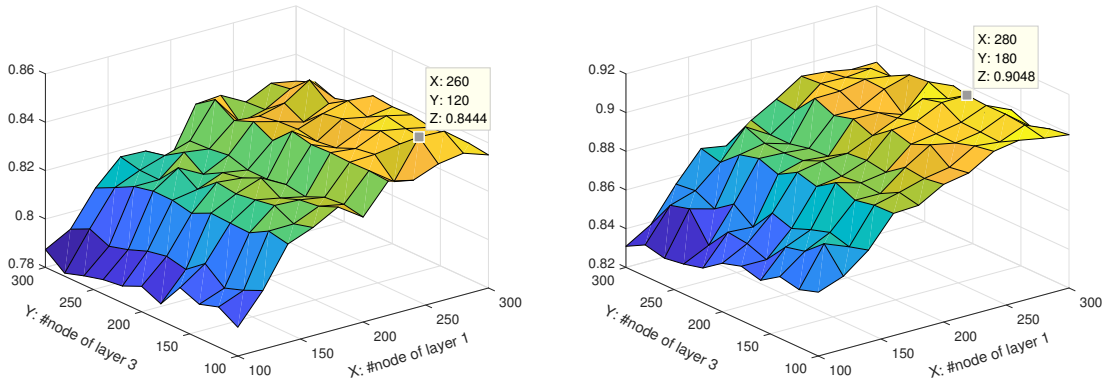
$$\mathbf{a}_i^{(3)} = \sigma(\mathbf{W}^{(3)\mathrm{T}}\mathbf{a}_i^{(2)} + \mathbf{b}^{(3)})$$
$$\mathbf{a}_i^{(4)} = \sigma(\mathbf{W}^{(4)\mathrm{T}}\mathbf{a}_i^{(3)} + \mathbf{b}^{(4)}),$$

where $\mathbf{W}^{(3)}, \mathbf{b}^{(3)}$ are weights and bias of the third layer, and $\mathbf{W}^{(4)}, \mathbf{b}^{(4)}$ are for the fourth layer.

In this paper, all the fully connected layers (Layers 1,3,4) are pre-trained by sparse auto-encoder (SAE) networks [5]. Then, these layers are stacked, and a softmax layer is connected to layer 4, using the supervised information to fine-tune the network. Finally, the output of layer 4 is considered as the feature of the superpixel.



(a) $l = 0$ (b) $l = 1$



(c) $l = 2$ (d) $l = 3$

Figure 4: Classification accuracies with different nodes for the four context regions, where the best results are located in the grids.

3.2 Choice of the network structure

The input of the network is flattened from pixel data in HE-space and the size of patch is 13×13. Then, the dimension of the input is $13 \times 13 \times 2 = 338$. To limit the computation of the networks and the following binarization stage, the dimension of the output layer (layer 4) was set 150. Then, the node number in layers 1 and 3 were searched through cross-validation in the training set according to the mean classification accuracy. Because layer 2 is a max-pooling layer, the node number of layer 2 is the same with layer 1. The result of the validation is shown Figure 4. In general, the classification performance is sensitive to the node number of layer 1, which should be finely selected. In contrast, the performance is relatively robust to the nodes of layer 3. The optimized number of nodes are presented in grids in Figure 4, and are summarized in Table 2.

Table 2: Classification performance for different number of nodes in the feature extraction networks.

Context index	Layer 1	Layer 2	Layer 3	Layer 4
$l = 0$	260	260	260	150
$l = 1$	260	260	200	150
$l = 2$	260	260	120	150
$l = 3$	260	280	180	150

The deep of the neural network is also validated. The first two layers are the bases of the network. Then, the layer number between the max-pooling layer (layer 2) and the output layer were tested from 1 to 4. The

performance for different number of layers between the two layers are given in Table 3, where the number of nodes for each layer is set 300. In general, the performance becomes better when the layer number increases from 1 to 2. While, the accuracies can be hardly improved when the layer number is adjusted from 2 to 4. Therefore, two layers between the max-pooling layer and the output layer are sufficient for the feature extraction networks used in our framework.

Table 3: Classification Accuracies for different number of layers between the max-pooling layer and the output layer.

Context index	**Number of layer**			
	1	2	3	4
$l = 0$	0.798	0.805	0.795	0.794
$l = 1$	0.825	0.831	0.827	0.826
$l = 2$	0.829	0.839	0.840	0.839
$l = 3$	0.890	0.899	0.900	0.899

References

[1] Y. Zheng, Z. Jiang, Y. Ma, H. Zhang, F. Xie, H. Shi, and Y. Zhao, "Feature extraction from histopathological images based on nucleus-guided convolutional neural network for breast lesion classification," *Pattern Recognition*, vol. 71, pp. 14–25, 2017.

[2] X. Zhang, W. Liu, M. Dundar, S. Badve, and S. Zhang, "Towards large-scale histopathological image analysis: Hashing-based image retrieval," *Medical Imaging, IEEE Transactions on*, vol. 34, no. 2, pp. 496–506, 2015.

[3] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 60, pp. 91–110, 2004.

[4] ——, "Object recognition from local scale-invariant features," in *International Conference of computer vision*, vol. 2, 1999, pp. 1150–1157.

[5] A. Coates, A. Y. Ng, and H. Lee, "An analysis of single-layer networks in unsupervised feature learning," in *International conference on artificial intelligence and statistics*, 2011, pp. 215–223.