Classification of Breast Cancer Histopathological Images using Convolutional Neural Networks with Hierarchical Loss and Global Pooling

Zeya Wang[1,2](✉), Nanqing Dong[1,3], Wei Dai[1], Sean D. Rosario[1], and Eric P. Xing[1]

[1] Petuum Inc., Pittsburgh, PA 15222, USA
{zeya.wang,nanqing.dong,wei.dai,sean.drosario,eric.xing}@petuum.com
[2] Rice University, Houston, TX 77005, USA
[3] Cornell University, Ithaca, NY 14850, USA

Abstract. Deep learning-based computer-aided diagnosis (CAD) has been gaining popularity for analyzing histopathological images. However, there has been limited work that addresses the problem of accurately classifying breast biopsy tissue with hematoxylin and eosin stained images into different histological grades. We propose a system which can automatically classify breast cancer histology images into four classes, namely normal tissues, benign lesion, *in situ* carcinoma and invasive carcinoma. Our framework uses a Convolutional Neural Network (CNN) with a hierarchical loss, where failing to distinguish between carcinoma and non-carcinoma is penalized more than failing to distinguish between normal and benign or between *in situ* and invasive carcinoma. The network also includes a patch-wise design with global pooling directly on input images. By incorporating the hierarchical and global information of the input images, our framework can outperform the previous system by a large margin.

Keywords: Convolutional Neural Networks · Image classification Histopathology · Breast cancer · Hierarchical loss

1 Introduction

Breast cancer is one of the leading causes of death by cancer in women, and early detection can give patients more treatment options. Breast cancer can be detected by microscopic analysis [1,2]. During a screening examination, breast tissue biopsies can be obtained from suspected patients, which pathologists analyze for tumor progression and type [2,3]. The tumor type is evaluated in terms of the extent of variation of structure from normal tissues, and how cancer spreads during detection. Benign lesion, *in situ* carcinoma, and invasive carcinoma are three types of tumors that can be determined from biopsy through histological analysis. Benign lesions lack the ability to invade neighbors, so they are

A. Campilho et al. (Eds.): ICIAR 2018, LNCS 10882, pp. 745–753, 2018.
https://doi.org/10.1007/978-3-319-93000-8_84

non-malignant. *In situ* and invasive carcinoma are malignant, hence spread to other areas. Invasive tissues, unlike *in situ*, invade the surrounding normal tissues beyond the mammary ductal-lobular system [2]. After the microscopic examination of biopsies at specific magnification levels, pathologists generate stained images by applying Hematoxylin-Eosin (H&E) staining to enhance the nuclei (purple) and cytoplasm (pinkish) for the purpose of diagnosis [3]. Stained images are labeled using manual methods based on the experience of pathologists, which is costly in terms of workload. Since the majority of biopsies are normal and benign, most of the work is redundant. CAD approaches for automatic diagnosis improve efficiency by allowing pathologists to focus on more difficult diagnosis cases [3,4]. CAD can reduce the workload of classifying histopathological images, using machine learning methods. Several existing machine learning approaches perform classification for two-class (malignant/benign) and three-class (normal, *in situ*, invasive) through extraction of nuclei-related information [3]. With the rise in computing power, deep learning algorithms are widely adopted for analysis of medical images [5]. In the Camelyon Grand Challenge 2016, several works demonstrated high accuracy for a similar four-class classification task on TNM breast cancer staging system [6]. These works follow a two-stage pipeline. In the first stage, patches that constitute the whole slide image are classified as tumor or normal. In the second stage, the tumor region features extracted from these classified patches are input into a random forest classifier in order to classify the cancer type [7]. That challenge provided pixel-wise annotation of tumors, which is expensive to collect, and not frequently available. In [3], the authors propose a CNN framework to solve the four-class classification problem (normal, benign, *in situ*, invasive) on H&E stained microscopic images by retrieving nuclei and tissue structure information. We believe there is scope for improving classification performance by using better network design. In this paper, we design a loss function that leverages hierarchical information of the histopathological classes. We also incorporate embedded feature maps with information from the input image to maximize grasp on the global context.

2 Data and Methods

2.1 Dataset

The dataset used in this paper is provided by Universidade do Porto, Instituto de Engenharia de Sistemas e Computadores, Tecnologia e Ciência (INESC TEC) and Instituto de Investigação and Inovação em Saúde (i3S), in TIF format. The dataset consists of 400 high resolution (2048 × 1536) H&E stained breast histology microscopic images with 200× magnification. Each pixel of an image corresponds to $0.42\,\mu\text{m} \times 0.42\,\mu\text{m}$ of the biopsy. These images are labeled with four classes: normal, benign, *in situ*, and *invasive*, and each class consists of 100 images.

Prior to the quantitative analysis, inconsistencies brought by the way of staining the histology slides should be minimized. We perform normalization on all images using the method proposed in [8]. This method first converts RGB values to their corresponding optical density (OD) values through a logarithmic transformation. Then singular value decomposition (SVD) is applied to find the two directions with higher variances of the OD tuples. All the OD transformed pixels are projected onto the plane created from the two SVD directions to find the robust extremes. These extremes are converted to the OD space and then used for deconvolving the original images to the H&E components. Concentrations for each stain are scaled to have the same pseudo-maximum. Finally, all images are recreated using normalized stain concentrations and the reference mixing matrix [9].

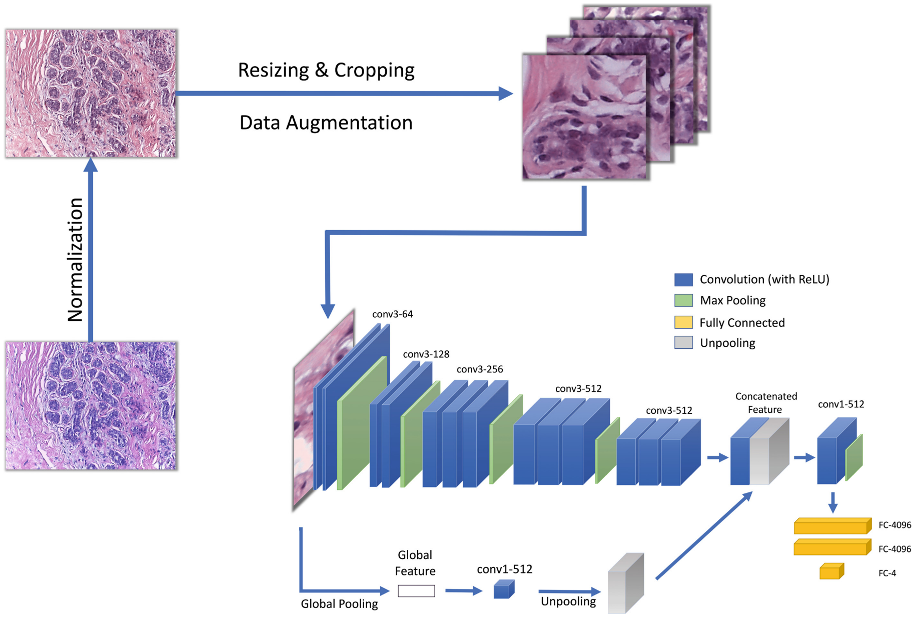

Fig. 1. Illustration of our classification pipeline.

2.2 Image-Wise Classification Framework

Our cancer-type classification framework consists of a data augmentation stage, a patch-wise classification stage, and an image-wise classification stage. After normalizing for staining inconsistencies, we rescale and crop each image to small patches with a size that can be fed as input to the CNN for patch-wise classification (See Fig. 1). The label of each patch is consistent with the label of the image which the patch is cropped from. During the training phase, the cropped patches are augmented to increase the robustness of the model as a method of regularization. A VGG-16 network with a hierarchical loss and global image pooling is trained for classifying these patches into four classes [10–12]. In the inference phase, we generate patches from each test image and combine patch classification results to classify the image [3]. We implement three patch probability score fusion methods for assigning the class label, namely, majority voting, maximum probability, and sum of probabilities [3]. In addition to these fusion techniques, we also adopt dense evaluation over the test image to get a class score map and select the class with the highest score [13]. This is detailed in Sect. 3.

2.3 Data Augmentation

We obtain 280 normalized images in the training set of size 2048 × 1536. The original images are too large to be fed into the network, so we crop images of size 224 × 224. Cropping small patches from a 2048 × 1536 image at the high magnification level of 200× can break the overall structural organization of the image, and therefore leave out important tissue architecture information. While training a CNN model, images are conventionally resized. However, for microscopic images, resizing could decrease magnification level. There is no consensus on the best magnification level, so we isotropically resize the whole image to a relatively small size, specifically, 1024 × 768 and 512 × 384 [14]. Each scaled image is then cropped to 224 × 224 patches with 50% overlap. In our experiments, the final choice for isotropic image resizing is 512 × 384, which generates a total of 3360 different patches from the original 280 training images. Contrast-limited adaptive histogram equalization (CLAHE) is then performed on the Lightness component after converting the RGB image to LAB format, and then the image is converted back to RGB, for enhancing the local contrast of cropped images [15]. Mean subtraction is performed by subtracting the average value from the R, G and B channels separately. The training set is augmented by image rotation with $\frac{k\pi}{2}$, where $k \in \{0, 1, 2, 3\}$, and vertical reflections. The patches after cropping and augmentation share the same label as the original stained image.

2.4 CNN Architecture for Patch-Wise Classification

The VGG-16 network is chosen to classify the 224 × 224 histology image patches, in order to explore the scale and organization features of nuclei and the scale features of the overall structure, which do not have complicated semantic information [10]. A 16-layer structure suffices for exploring these features. The VGG-19 network is also used for the sake of comparison in our experiments. To leverage the whole contextual information from the cropped images, we add global context to the last convolutional layer of the VGG network. Similar to ParseNet [12], the input images are passed to two independent branches, our VGG network and a global average pooling layer [12]. With a B × H × W × C input (B is batch size; W and H is the width and height; C is the number of channels), the output of the global pooling layer is B × 1 × 1 × C. One 1 × 1 convolutional layer will transform the last dimension of output to the desired number, which in our case is 512. The transformed output is unpooled to the same shape as that of the feature maps after the last convolutional layer of VGG network and is then concatenated with it. These two feature maps are fused by another 1 × 1 convolutional layer and then passed through three fully-connected (FC) layers for classification (See Fig. 1).

2.5 Hierarchical Loss

Hierarchical loss is a novel addition to this classification work. As mentioned before, we can further group normal/benign into non-carcinoma, and *in situ*/invasive into carcinoma. The classes have a tree organization, where normal/benign can be considered as leaves from the node non-carcinoma, and *in situ*/invasive as leaves from the node carcinoma. From the root, we have two nodes for carcinoma and non-carcinoma, respectively connected to two leaves normal/benign and *in situ*/invasive. This structure motivates us to apply a hierarchical loss for classification instead of the vanilla cross entropy loss. The hierarchical loss uses an ultrametric tree to calculate the amount of metric "winnings" [11]. Hence, failing to distinguish between carcinoma and non-carcinoma is penalized more than failing to distinguish between normal and benign or between *in situ* and invasive, which follows intuition. The amount of the "winnings" is calculated from the weighted sum of the estimated probability score of each node along the path from the first non-root node to the correct leaf. The probability score of each node is obtained by summing up the scores from its child nodes. The weights are given in Fig. 2. Finally, the loss (the negative of "winnings") uses the negative logarithm as in computing cross entropy loss.

$$L(\hat{\boldsymbol{y}}, \boldsymbol{y}) = -\sum_{i=1}^{N}\sum_{c=1}^{C} y_i^c \log\{\frac{1}{2}\hat{y}_i^c + \frac{1}{2}(\hat{y}_i^c + \sum_{c' \in siblings(c)} \hat{y}_i^{c'})\} \tag{1}$$

For class hierarchy with a height equal 2, the loss function is defined as shown in Eq. 1, where y_i^c is the binary label of sample i belonging to class c, $\hat{y}_i^c$ is softmax output for the labeled class channel c, C the number of channels, N the number of samples, and $siblings(\cdot)$ denotes the sibling set of classes for a specified class node.

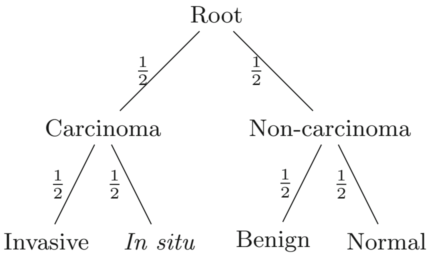

Fig. 2. Tree representation of "winnings" in the hierarchical loss.

3 Experiments

In this section, we present our evaluation results in terms of accuracy for patch-wise and image-wise classification as per [3]. We demonstrate the performance of our proposed framework through a series of ablation studies.

3.1 Experimental Setup

We first randomly split all images to 280 images for training, 60 images for validation and 60 images for testing. All four classes are distributed equally between splits. The objective of the model is to minimize the hierarchical loss using a momentum optimizer with momentum 0.9 and batch size 32 [16]. The weights of the network are initialized with the pre-trained weights of the VGG-16 model on ImageNet [17]. The learning rate is initialized to 0.0002 and decreases exponentially every 1000 mini-batch iterations with a decay factor of 0.9. The weights are regularized with weight decay with L2 penalty multiplier of 0.003. Dropout with ratio 0.5 is applied to the first two fully-connected layers. Training usually converges very quickly (after around 5 epochs). At test time, we first pre-process and resize test images, and then classify each test image using the aforementioned fusion methods. In the dense evaluation, we convert the last three FC layers in VGG networks to convolutional layers, and then densely apply the converted networks over the rescaled test image to get a class score map. The class with the highest score is selected [13]. Our final results are shown below in Table 1. For each performance reported, the patch-wise classification accuracy is calculated with the same model used in image-wise classification.

Table 1. Image-wise classification accuracy (%) of best model setting with different post-processing methods (patch-wise accuracy (%) given in the bracket).

	Majority vote	Sum probability	Maximum probability	Dense evaluation
Validation set	0.92 (0.87)	0.92 (0.87)	0.92 (0.87)	0.92 (0.87)
Test set	0.93 (0.85)	0.90 (0.85)	0.92 (0.85)	0.92 (0.85)
Average	0.93	0.91	0.92	0.92

3.2 VGG-19 vs. VGG-16

As described previously, a shallow CNN model is preferred by virtue of the content of semantic information in a histopathological image. To support this heuristic choice, we compare with the performance of the deeper VGG-19 (See Table 2 compared with Table 1) to demonstrate that a network with an appropriate depth is able to perform better.

Table 2. Image-wise classification accuracy (%) of using VGG-19 with different post-processing methods (patch-wise accuracy (%) given in the bracket).

	Majority vote	Sum probability	Maximum probability	Dense evaluation
Validation set	0.85 (0.76)	0.83 (0.76)	0.80 (0.74)	0.82 (0.76)
Test set	0.80 (0.72)	0.80 (0.73)	0.80 (0.71)	0.80 (0.74)
Average	0.83	0.82	0.80	0.81

Table 3. Image-wise classification accuracy (%) of removing hierarchical loss with different post-processing methods (patch-wise accuracy (%) given in the bracket).

	Majority vote	Sum probability	Maximum probability	Dense evaluation
Validation set	0.88 (0.82)	0.87 (0.82)	0.88 (0.82)	0.87 (0.81)
Test set	0.88 (0.83)	0.87 (0.81)	0.85 (0.81)	0.87 (0.83)
Average	0.88	0.87	0.87	0.87

3.3 Hierarchical Loss vs. No Hierarchical Loss

To emphasize the importance of using hierarchical loss, we include a comparison experiment that just uses the vanilla cross entropy loss. The results of the experiment are shown below in Table 3 (compared with Table 1).

3.4 Global Image Pooling vs. No Global Image Pooling

Global image pooling is another important feature we integrate into our network architecture. This structure, which is often used for global information extraction from high resolution feature maps, can pass global context from the input image to the last convolution layer in deep networks, thus improving the final performance. We also include an experiment without image pooling layers with its results (see Table 4 compared with Table 1), to demonstrate its improvement of model performance.

Table 4. Image-wise classification accuracy (%) of removing global average pooling with different post-processing methods (patch-wise accuracy (%) given in the bracket).

	Majority vote	Sum probability	Maximum probability	Dense evaluation
Validation set	0.93 (0.90)	0.93 (0.90)	0.93 (0.90)	0.93 (0.90)
Test set	0.87 (0.84)	0.87 (0.84)	0.90 (0.84)	0.87 (0.84)
Average	0.90	0.90	0.92	0.90

3.5 Different Scales

The microscopic images are obtained with a high magnification level of 200×, for capturing the nuclei-scale feature. We pre-process the large image into different

224 × 224 patches that can be fed into the VGG networks. However, this cropping could result in a loss of most of the structural information. Thus, we first isotropically resize the 2048 × 1536 image into a smaller scale before cropping. In our presented performance, the width and height are both down-scaled 4× to 512 × 384 respectively. This scale is used because it can maintain most of the nuclei structural information from the original whole image, while keeping most information of tissue structural organization for the cropped patches. We include a larger scale 1024 × 768 for comparison with results shown in Table 5 (compared with Table 1).

Table 5. Image-wise classification accuracy (%) of resizing images to 1024 × 768 with different post-processing methods (patch-wise accuracy (%) given in the bracket).

	Majority vote	Sum probability	Maximum probability	Dense evaluation
Validation set	0.82 (0.72)	0.82 (0.72)	0.83 (0.72)	0.83 (0.72)
Test set	0.75 (0.68)	0.75 (0.68)	0.78 (0.68)	0.78 (0.68)
Average	0.79	0.79	0.81	0.81

4 Discussion

In this work, we present a CNN-based approach with preprocessing and post-processing methods to classify of H&E stained histopathological images for breast cancer tissue classification. We propose to resize and crop images after considering the trade-off between capturing nuclei associated scale information and the overall structural organization. We utilize the VGG-16 network, which has been successful in general image recognition tasks [10,17]. Additionally, we apply a hierarchical loss based on the biological nature of the problem, and use global average pooling to incorporate the global information in an image. Our approach succeeds in classification of cancer types and shows competitive performance on the given dataset.

Magnification is an important factor for analyzing microscopic images for diagnosis. The most informative magnification is still debatable, so we compare two possible scales in our work. In future work, we will study the influence of other scales on the performance.

References

1. Siegel, R.L., Miller, K.D., Jemal, A.: Cancer statistics, 2016. CA Cancer J. Clin. **66**(1), 7–30 (2016)
2. American Cancer Society: Breast Cancer Facts & Figures 2017–2018. American Cancer Society, Inc., Atlanta (2017)
3. Araújo, T., Aresta, G., Castro, E., Rouco, J., Aguiar, P., Eloy, C., Polónia, A., Campilho, A.: Classification of breast cancer histology images using convolutional neural networks. PloS One **12**(6), e0177544 (2017)

4. Gurcan, M.N., Boucheron, L.E., Can, A., Madabhushi, A., Rajpoot, N.M., Yener, B.: Histopathological image analysis: a review. IEEE Rev. Biomed. Eng. **2**, 147–171 (2009)
5. Litjens, G., Kooi, T., Bejnordi, B.E., Setio, A.A.A., Ciompi, F., Ghafoorian, M., van der Laak, J.A.W.M., van Ginneken, B., Sánchez, C.I.: A survey on deep learning in medical image analysis. Med. Image Anal. **42**, 60–88 (2017)
6. Ehteshami Bejnordi, B., Veta, M., van Diest, P.J., et al.: Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. JAMA **318**(22), 2199–2210 (2017)
7. Wang, D., Khosla, A., Gargeya, R., Irshad, H., Beck, A.H.: Deep learning for identifying metastatic breast cancer. arXiv preprint arXiv:1606.05718 (2016)
8. Macenko, M., Niethammer, M., Marron, J.S., Borland, D., Woosley, J.T., Guan, X., Schmitt, C., Thomas, N.E.: A method for normalizing histology slides for quantitative analysis. In: IEEE International Symposium on Biomedical Imaging: From Nano to Macro, ISBI 2009, pp. 1107–1110. IEEE (2009)
9. Veta, M., van Diest, P.J., Willems, S.M., Wang, H., Madabhushi, A., Cruz-Roa, A., Gonzalez, F., Larsen, A.B., Vestergaard, J.S., Dahl, A.B., et al.: Assessment of algorithms for mitosis detection in breast cancer histopathology images. Med. Image Anal. **20**(1), 237 (2015)
10. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
11. Wu, C., Tygert, M., LeCun, Y.: Hierarchical loss for classification. arXiv preprint arXiv:1709.01062 (2017)
12. Liu, W., Rabinovich, A., Berg, A.C.: Parsenet: Looking wider to see better. In: CoRR. Citeseer (2015)
13. Sermanet, P., Eigen, D., Zhang, X., Mathieu, M., Fergus, R., Lecun, Y.: Overfeat: integrated recognition, localization and detection using convolutional networks. In: International Conference on Learning Representations (ICLR2014), CBLS, April 2014
14. Komura, D., Ishikawa, S.: Machine learning methods for histopathological image analysis. arXiv preprint arXiv:1709.00786 (2017)
15. Pizer, S.M., Amburn, E.P., Austin, J.D., Cromartie, R., Geselowitz, A., Greer, T., ter Haar Romeny, B., Zimmerman, J.B., Zuiderveld, K.: Adaptive histogram equalization and its variations. Comput. Vis. Graph. Image Process. **39**(3), 355–368 (1987)
16. Sutskever, I., Martens, J., Dahl, G., Hinton, G.: On the importance of initialization and momentum in deep learning. In: International Conference on Machine Learning, pp. 1139–1147 (2013)
17. Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., Fei-Fei, L.: Imagenet: a large-scale hierarchical image database. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2009, pp. 248–255. IEEE (2009)