Significance of Hyperparameter Optimization for Metastasis Detection in Breast Histology Images

Navid Alemi Koohbanani[1,2(✉)], Talha Qaisar[1], Muhammad Shaban[1], Jevgenij Gamper[1], and Nasir Rajpoot[1,2]

[1] Department of Computer Science, University of Warwick, Coventry, UK
n.alemi-koohbanani@warwick.ac.uk
[2] The Alan Turing Institute, London, UK

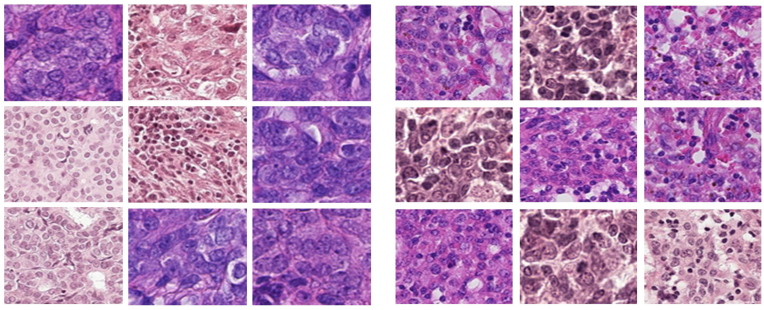
Abstract. Breast cancer (BC) is the second most leading cause of cancer deaths in women and BC metastasis accounts for the majority of deaths. Early detection of breast cancer metastasis in sentinel lymph nodes is of high importance for prediction and management of breast cancer progression. In this paper, we propose a novel deep learning framework for automatic detection of micro- and macro- metastasis in multi-gigapixel whole-slide images (WSIs) of sentinel lymph nodes. One of our main contributions is to incorporate a Bayesian solution for the optimization of network's hyperparameters on one of the largest histology dataset, which leads to 5% gain in overall patch-based accuracy. Furthermore, we present an ensemble of two multi-resolution deep learning networks, one captures the cell level information and the other incorporates the contextual information to make the final prediction. Finally, we propose a two-step thresholding method to post-process the output of ensemble network. We evaluate our proposed method on the CAMELYON16 dataset, where we outperformed "human experts" and achieved the second best performance compared to 32 other competing methods.

1 Introduction

Breast cancer (BC) is the second most common type of cancers and the primary cause of cancer mortality in women. Majority of deaths from BC are due to its metastasis to other organs in the body [1]. Therefore, early stage detection is important for the diagnosis and prognosis of BC. The sentinel lymph node (SLN) biopsy is the most pragmatic way of attaining BC metastasis. One of the challenging aspects of this problem is that a lymph node tissue contains some other cells (histiocytes) and regions (follicles and medullary sinus) having morpholgical resemblance to tumor cells and tumor regions as shown in Fig. 1.

Feature engineering to capture the discriminative attributes of each region is a non trivial task. Therefore, learning robust and discriminative features from the data in an automated manner using convolutional neural networks (CNNs) is a captivating choice for the problem at hand. Here, we proposed a framework

D. Stoyanov et al. (Eds.): COMPAY 2018/OMIA 2018, LNCS 11039, pp. 139–147, 2018.
https://doi.org/10.1007/978-3-030-00949-6_17



(a) (b)

Fig. 1. (a) Tumor regions, exhibiting non-uniformity in chromatin texture (b) histiocyte regions resembling tumor regions in terms of their morphological appearance.

for the detection of metastasis in whole-slide images (WSIs). Firstly, extracted patches from WSIs are categorized based on the presence of metastasis using CNNs. In this paper, modified CNNs are utilized and we show the high impact of these modification for boosting CNNs performance. To this end, we used Gaussian process for optimization of hyperparameters. Secondly, to enhance the final prediction on WSI level, cellular and contextual information are combined to get tumour likelihood map. This is achieved by fusing the outputs of two modified CNNs trained on patches of different resolutions and sizes. Finally, we propose a simple yet effective algorithm to convert the tumor likelihood map into a binary map. The proposed two-step thresholding method removes the highly uncertain regions from the likelihood map. Confidence score of each WSI for containing a metastases region is calculated using area of highly certain metastasis regions.

We have made three major contributions in this paper. First, we show that following a principled Bayesian approach to hyperparameter optimization can significantly improve performance for histology images with standard state-of-the-art CNNs instead of using new complex network architectures. Second, we propose an ensemble strategy to mimic the routine clinical practice where pathologist examines a WSI at different magnification levels (40×, 20×, etc.) under the microscope. Our proposed method combines the predictions of two different resolutions (40× and 20×) to make final prediction which integrates cellular and contextual information. Finally, with the above principled approach, we achieve competitive results; 2nd only to the current winner in the leaderboard and better than the pathologists' performance on one of the largest publicly available histology image datasets (CAMELYON16[1]), according to the criterion used by the challenge organizers.

[1] https://camelyon16.grand-challenge.org.

2 Related Work

Czerniecki *et al.* [2] proposed IHC biomarkers to assist the pathologist in breast metastasis detection but it requires more time, cost and increases the number of slides required for analysis. An automated computer-based system was also developed for detecting micro-metastasis in lymph node biopsies [3] based on cell detection. Instead of cell detection, our method detects metastasis using robust multi-resolution features automatically learned from the very large dataset.

Recently, deep learning techniques have been used for a variety of histology image analysis problems. Some early work using CNNs was done for mitosis counting for primary breast cancer [4]. Recently, Wang *et al.* [5] assign a prediction value to each patch using CNNs and then make decision based on probability map of WSIs. Overall, the CAMELYON16 challenge [6] has shown the utility of deep learning algorithms for automatic tissue analysis, outperforming the pathologists in terms of detecting tumors within the WSIs. Since feeding large high resolution patches into deep learning model is computationally infeasible; one should consider a trade off between patch size and the amount of information lying within that patch. Most existing methods, consider using patches at one resolution. Here we train two separate networks with high and low resolution patch size, then merge probability maps generated by these networks.

The rest of this paper is organized as follows. Section 3 introduces the methodology details, Sect. 4 demonstrate the experimental results and comparison. Finally, Sect. 5 draws conclusion about the paper.

3 Methodology

Our framework (Fig. 2) is based on ensemble of two networks where one network encodes on cellular structures and the other captures the context. We find optimal architecture of classification networks and training hyperparameters through Bayesian optimization method. WSI probability map is created by merging all probabilities of small patches.

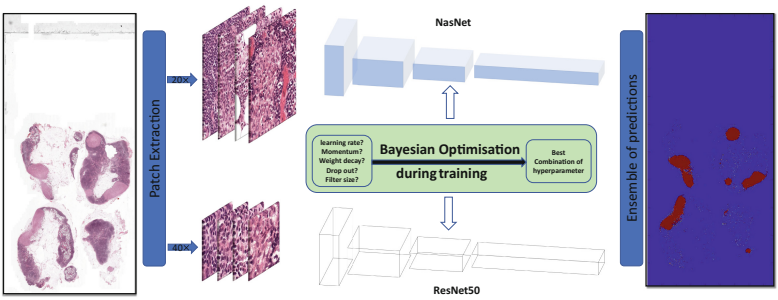


Fig. 2. An overview of the proposed approach.

3.1 Bayesian Optimization for Boosting Network Performance

Selection of the best network for a specific task is not trivial due to the stochastic nature of deep learning networks and their dependencies on various hyperparameters. These hyperparameters range from learning rate, momentum to more complicated variables like selection of number of layers, filter size, etc. Optimal selection of these hyperparameters is a known challenge in computer vision community. Non-optimal values of hyperparameters may lead to poor overall performance of a network. Search over all possible combination of hyperparameter values (grid search or random search) is computationally infeasible and extremely time consuming.

One way to overcome this problem is using Bayesian Optimization (BO) methods [7]. Gaussian process (GP) is one of BO approaches that can be used to predict optimal hyperparameter. A GP is a supervised learning method, which addresses the problem of leaning input-output mappings from training data. It utilizes kernel functions to learn these relations. In GP having observed N input vectors (hyperparameters) and their corresponding output variables (accuracies), we wish to make assumption about unobserved parameters. Acquisition function using this set of information suggest the next set of parameters. Here we used Expected Improvement [8] as an acquisition function to find the optimal settings for network and optimization level hyperparameters.

Convolutional filter size controls the receptive field for subsequent layers. Larger filter size for initial layers works well for natural images as classification of these images does not require high resolution information. However, tumor classification highly depends on high resolution images of tumor cells. Therefore, we consider filter size of first convolutional layer along with l2 weight regularization as hyperparameters to give flexibility to the network to choose the best receptive field. The standard architectures of ResNet and NASNet does not contains any dropout layer. However, dropout layer is useful to reduce over-fitting. Therefore, it is considered as another hyperparameters along with learning rate and momentum to be chosen by Bayesian optimization. The best hyperparameters are selected during training with early stopping. It means where the validation remains the same after two epochs, the new hyperparameters which are predicted by GP are replaced. This process continues until the best hyperparameters are found.

3.2 Pre-processing

Tissue Localization. WSIs contain large section of background (white) regions which should be ignored during processing. Therefore, tissue separation (ROI) should be performed beforehand to reduce computation time and efficiency of algorithm. Here, we follow the same approach as [5], firstly, we transfer image from RGB color space to HSV color space, then binary mask of tissue region is obtained by applying Otsu adaptive thresholding algorithm on S channel.

3.3 Network Selection and Ensembling

The proposed framework is an ensemble of two different CNNs. One network learns the cell level representation of tumour and normal patches at 40× resolution. However, since context information is also important along with high resolution cell appearance, cell level information is not enough to predict the label of a patch. We used second network to capture the context information from a larger patch at 20× resolution.

Different networks have been proposed in the literature for classification of natural images (ImageNet) including Inception-V3 [9], ResNet [10], DenseNet [11], and NASNet [12]. We choose ResNet for cellular feature based classification of patches due to its better performance. Moreover, use of smaller filter size in the first convolution layer and dropout before last fully connected layer results in improved performance. We explored these modification along with weight optimization level hyperparameters through BO.

We feed large patches(448×448) at lower resolution to the context based prediction models. Having the context information as input, NasNet has the best performance. It is based on seperable convolution with different filter sizes that helps to learn the representation at different resolution. Additionally, inclusion of dropout with hyperparameter tuning results in 5% increase in patch level performance as shown in Table 1.

Finally, we construct the probability map from each network (ResNet and NasNet) for each WSI. Then we fuse them together to build the final probability map that reflect both high and low resolution information. Afterwards, the probability maps are post-processed by morphological operations to achieve tumor localization and WSI classification. The post-processing procedure is described in Sect. 3.4.

Table 1. Validation accuracy of different networks trained on the CAMELYON16 dataset. Networks trained on both 448×448 at 20× resolution and 224×224 at 40× resolution with default hyperparameter settings and after applying Bayesian Optimization (BO)

Networks	40×	40× (BO)	20×	20× (BO)
ResNet [10] (%)	97.32	99.12	86.15	90.56
InceptionV3 [9] (%)	95.10	95.10	85.67	86.57
DenseNet [11] (%)	95.35	96.67	87.68	89.15
NasNet [12] (%)	96.28	97.01	86.75	91.91

3.4 Post-processing

Our careful inspection of the tumor probability shows that regions with high probability values but characterized by abrupt changes in the values (i.e. high uncertainty) generally correspond to false positive decisions. Therefore, to eliminate these high uncertainty regions, we use two different threshold values t_{low}

Table 2. Dropout, learning rate, momentum, Weight decay for 1st layer and filter size for 1st layer

Networks	Dropout	learning rate	momentum	weight decay	filter size
ResNet (40×)	✓	0.001	0.95	l2	3 × 3
ResNet (20×)	✓	0.001	0.97	l2	3 × 3
InceptionV3 (40×)	✓	0.010	0.80	-	3 × 3
InceptionV3 (20×)	✓	0.001	0.85	l2	3 × 3
DenseNet (40×)	-	0.001	0.90	l2	3 × 3
DenseNet (20×)	-	0.010	0.87	l2	3 × 3
NasNet (40×)	✓	0.001	0.95	l2	3 × 3
NasNet (20×)	-	0.010	0.99	l2	3 × 3

and t_{high}. First, we obtain sets of regions $\mathcal{B}(t_{\text{low}})$ and $\mathcal{B}(t_{\text{high}})$ by thresholding the tumor probability image at t_{low} and t_{high}, respectively. From the construction, each region in $\mathcal{B}(t_{\text{high}})$ is a shrink version of some region in $\mathcal{B}(t_{\text{low}})$, and there can be multiple regions in $\mathcal{B}(t_{\text{high}})$ that correspond to the same region in $\mathcal{B}(t_{\text{low}})$. For each region $C^{(i)} \in \mathcal{B}(t_{\text{low}})$, let $\left\{\tilde{C}_1^{(i)}, ..., \tilde{C}_{N_i}^{(i)}\right\} \in \mathcal{B}_{\text{high}}(t)$ be a set of N_i regions corresponding to $C^{(i)}$. We eliminate each region $C^{(i)} \in \mathcal{B}(t_{\text{low}})$ such that

$$\frac{\left|\cup_{j=1}^{N_i} \tilde{C}_{N_i}^{(i)}\right|}{|C^{(i)}|} < \alpha, \tag{1}$$

where α is the area threshold ratio, and $|\cdot|$ denotes the cardinality of a set. We set $t_{\text{low}} = 0.3$, $t_{\text{high}} = 0.9$, and $\alpha = 0.5$. For each remaining candidate region, we calculate the confidence of being tumor using the minimum probability found in that region weighted by its area. The candidate regions are further removed based on this weighted confidence.

After clearing probability map, we binarize all probability maps by applying a threshold (threshold = 0.6). We dilate the tumor regions as much as 75 um, since locations within the 75 um of tumor areas are also considered as true positive for FROC validation criteria that is used in CAMELYON 16 challenge. Finally, the coordinate of those regions along with the maximum value on probability map are recorded for FROC. For plotting ROC curve, the maximum probability of largest tumor area is reported as probability of WSI being tumor.

4 Experimental Results

Our experiments were conducted on the CAMELYON16 dataset to evaluate the proposed framework for cancer metastasis detection in WSIs. This dataset contains 400 WSIs: 270 WSIs for training and remaining 130 WSIs for evaluation purpose. The cancer metastasis regions were exhaustively annotated under the supervision of expert pathologists. The WSIs were stored at different magnification levels with the highest magnification level of 40× and 0.243 um/pixel

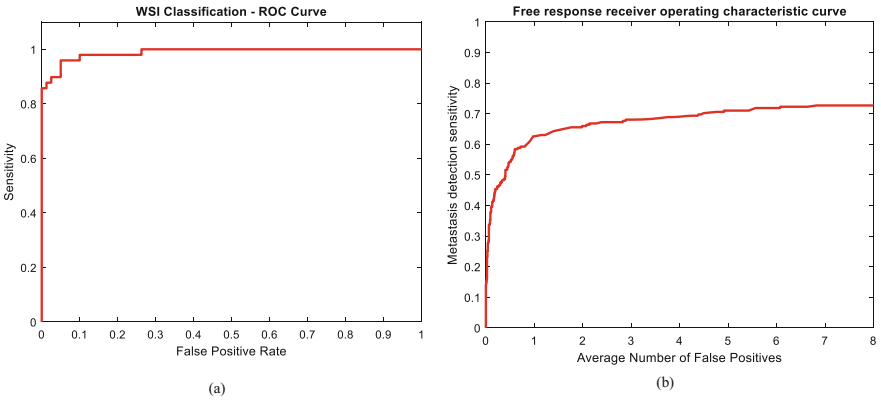


Fig. 3. (a) ROC curve and (b) FROC curve which shows sensitivity against average number of false positives

resolution. Our experiments were performed at both 40× and 20× magnifications to have both cellular and contextual information. Table 2 shows the values that have been chosen for hyper parameters after applying BO. Filter size should be 3 × 3 for most of networks as the objects in images are small and l2 regularization in first layer can lead to better performance. Moreover we use drop out in our settings according to BO estimation.

We experimented with different networks on our data in order to choose optimal achitecture. To improve the performance of the networks we used hyperparameter optimization method (Sect. 3.1) for increasing the accuracy. Table 1 shows the performance of different networks on our dataset. We achieve considerable increase in accuracy after hyperparameter optimization. To this end, we chose Resnet50 for patches of 40× and NasNet for 20× patches as they have higher accuracy on their corresponding dataset after BO optimization. The accuracy increased 2% and 5% for ResNet (with 40× patches) and NasNet (with 20× patches) respectively which is a high improvement with regarding high variability in dataset and low inter class variability.

After first iteration of training, we process the WSIs in patch based manner to generate the probability maps. We observed that the probability maps of WSIs have many false positive regions that affect the final result. To this end, by recognizing the false positive regions, corresponding patches from WSIs are extracted and fed back into the networks for fine-tuning. Therefore, we come up with cleaner probability map with very few false positive regions.

We followed the same evaluation criteria (ROC and FROC) as explained in [6]. As shown in Fig. 3, ROC curve clearly depicts that our method predicts large number of tumor slides with very few false positives. In fact, it classifies 83% of tumor slides without throwing any false positive. The FROC curve also shows that the algorithm is capable of localizing tumor regions with various mean number of false positives in per whole slide image. The values for area under

ROC curve (AUC) are shown in Table 3. Our method achieves a high AUC, which shows the privilege of ensemble of two networks at different resolutions and highlight the effect of parameter tuning in final prediction. The score obtained from AUC, put us on top of the Table 3 which means our algorithm is performing very well for classification of WSIs into two categories of tumor and normal.

Table 3. Comparison of AUC measure

Rank	Methods	AUC	Rank	Methods	AUC
1	HMS & MITII [5]	0.994	4	HMS & MGH III	0.976
2	Proposed Method	0.990	5	HMS & MGH I	0.965
3	Pathologist	0.966	6	CULab	0.940

5 Conclusions

In this paper, we investigated the impact of hyperparameter optimization on network performance. Tuning hyperparameters with the Gaussian process could increase the validation accuracy on average by 5%. Furthermore, a multi-resolution network for detecting breast cancer metastasis from sentinel lymph node WSIs was proposed. Therefore, with combined contextual and cell level information and also optimizing hyperparameter, we achieve AUC and average sensitivity of 0.990 and 0.6583 respectively. This results in competitive performance of our framework applied on the CAMELYON16 dataset.

References

1. American Cancer Society: Cancer Facts & Figures (2015)
2. Czerniecki, B.J., et al.: Immunohistochemistry with pancytokeratins improves the sensitivity of sentinel lymph node biopsy in patients with breast carcinoma. Cancer **85**(5), 1098–1103 (1999)
3. Weaver, D.L., et al.: Comparison of pathologist-detected and automated computer-assisted image analysis detected sentinel lymph node micrometastases in breast cancer. Mod. Pathol. **16**(11), 1159 (2003)
4. Cireşan, D.C., Giusti, A., Gambardella, L.M., Schmidhuber, J.: Mitosis detection in breast cancer histology images with deep neural networks. In: Mori, K., Sakuma, I., Sato, Y., Barillot, C., Navab, N. (eds.) MICCAI 2013 Part II. LNCS, vol. 8150, pp. 411–418. Springer, Heidelberg (2013). https://doi.org/10.1007/978-3-642-40763-5_51
5. Wang, D., et al.: Deep learning for identifying metastatic breast cancer. arXiv preprint arXiv:1606.05718 (2016)
6. Bejnordi, B.E., 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. Snoek, J., et al.: Practical Bayesian optimization of machine learning algorithms. In: Advances in Neural Information Processing Systems, pp. 2951–2959 (2012)

8. Bergstra, J.S., et al.: Algorithms for hyper-parameter optimization. In: Advances in Neural Information Processing Systems, pp. 2546–2554 (2011)
9. Szegedy, C., et al.: Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2818–2826 (2016)
10. He, K., et al.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
11. Huang, G., et al.: Densely connected convolutional networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, vol. 1 (2017)
12. Zoph, B., et al.: Learning transferable architectures for scalable image recognition. arXiv preprint arXiv:1707.07012 (2017)