2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)
Venice, Italy, April 8-11, 2019

LOOK, INVESTIGATE, AND CLASSIFY: A DEEP HYBRID ATTENTION METHOD FOR BREAST CANCER CLASSIFICATION

Bolei Xu[⋆]*, Jingxin Liu*[⋆]*, Xianxu Hou*[⋆]*, Bozhi Liu*[⋆]*, Jon Garibaldi*[†]*, Ian O. Ellis* [†]*,*
Andy Green [†]*, Linlin Shen*[⋆]*, Guoping Qiu*[⋆†]

[⋆] Shenzhen University, Shenzhen, China
[†] University of Nottingham, Nottingham, United Kingdom

ABSTRACT

One issue with computer based histopathology image analysis is that the size of the raw image is usually very large. Taking the raw image as input to the deep learning model would be computationally expensive while resizing the raw image to low resolution would incur information loss. In this paper, we present a novel deep hybrid attention approach to breast cancer classification. It first adaptively selects a sequence of coarse regions from the raw image by a hard visual attention algorithm, and then for each such region it is able to investigate the abnormal parts based on a soft-attention mechanism. A recurrent network is then built to make decisions to classify the image region and also to predict the location of the image region to be investigated at the next time step. As the region selection process is non-differentiable, we optimize the whole network through a reinforcement approach to learn an optimal policy to classify the regions. Based on this novel Look, Investigate and Classify approach, we only need to process a fraction of the pixels in the raw image resulting in significant saving in computational resources without sacrificing performances. Our approach is evaluated on a public breast cancer histopathology database, where it demonstrates superior performance to the state-of-the-art deep learning approaches, achieving around 96% classification accuracy while only 15% of original image pixels are required for computation.

Index Terms— Deep Learning, Reinforcement Learning, Breast Cancer Classification, Visual Attention

1. INTRODUCTION

Breast Cancer is a major concern among women for its higher mortality when comparing with other cancer death [1]. Thus, early detection and accurate assessment are necessary to increase survival rates. In the process of clinical breast examination, it is usually fatigue and time-consuming to obtain diagnostic report by pathologist. Thus, there is large demand to develop computer-aided diagnosis (CADx) to relieve workload from pathologists.

In recent years, deep learning approaches are widely applied to the histopathology image analysis for its significant performance on various medical imaging tasks. However, one issue with deep learning approaches is that the size of raw image is large. By directly inputting raw images to the deep neural network, it would be computational expensive and requires days to train on GPUs. Some previous approaches address this problem by either resizing raw images to low resolution [2, 3, 4] or randomly cropping patches [5] from raw images. However, both approaches would lead to information loss and the detailed features of abnormality part could be missing, which might cause the misdiagnosed result. Another approach is to use sliding-window to crop image patches. However, there would be a large number of patches that are not related to the lesion part, since in some cases the abnormality part is usually in small portion.

One property of human visual system is that it does not have to process the whole image at once. In clinical diagnose, pathologist would first selectively pay attention to the abnormality region, and then investigate the region for details. In this paper, we formulate the problem as a Partially Observed Markov Decision Process [6], and we propose a novel deep hybrid attention model to mimic human perception system. We build a recurrent model that is able to select image patches that are highly related to abnormality part from raw image at each time step, which so-called the "hard-attention". Instead of directly working on the raw image, we could thus learn image features from the cropped patch. We further investigate the cropped patch through a "soft-attention" mechanism that is to highlight pixels most related to the lesion part for classification. It should be noticed that our approach does not directly access to the raw image, and thus the computation amount of our approach is *independent of the raw image size*. The patch selection process is non-differentiable, we regard the problem as a control problem, and thus could optimize the network through a reinforcement learning approach.

The contribution of this paper could be summarized in three-fold: (1) A novel framework is introduced to the classification of breast cancer histopathology image based on the hybrid attention mechanism. (2) The proposed approach can automatically select useful region from raw image, which is able to prevent information loss and also to save com-

978-1-5386-3640-4/19/$31.00 ©2019 IEEE

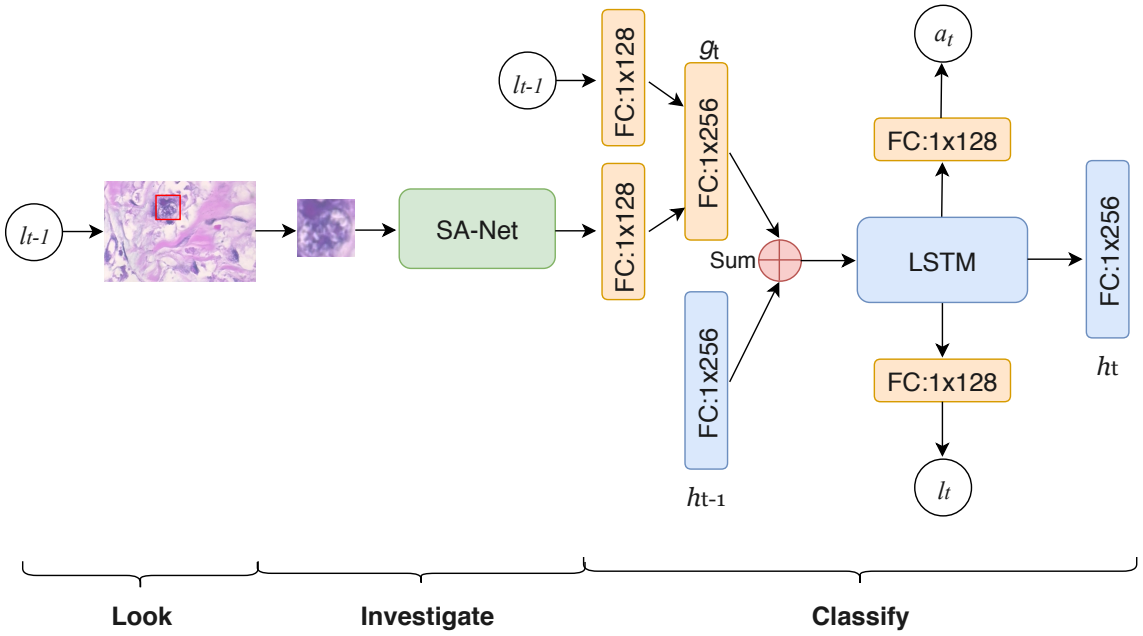

Fig. 1. The overall framework of our deep hybrid attention network. "FC" denotes fully-connected layer with ReLu activation. In each time step, the network has three stage to classify image. In the "Look" stage, a patch is cropped by hard-attention. Then in the "Investigate" stage, the abnormal features of image patch are extracted by the SA-Net as shown in Figure. 2. Finally, in the "Classify" stage, a LSTM is employed to process the image features and also to classify image and to predict region for the next time step. For each raw image. the network crops five patches for classification.

putational cost. (3) Our approach demonstrates superior performance to previous state-of-the-art methods on a public dataset.

2. METHODOLOGY

2.1. Network Architecture

We formulate the histopathology image classification problem as a Partially Observable Markov Decision Process (POMDP), which means at each time step, the network does not have full access to the image and it has to make decisions based on the current observed region. It takes three stages including "Look", "Investigate" and "Classify" stages as shown in Figure. 1.

Look Stage: At each time step t, a *hard-attention sensor* receives a partial image patch x_t based on the location information l_{t-1} (l_{t-1} is the center of patch), which has smaller image size than the raw image x. It is a coarse region that might be related to abnormality part.

Investigate Stage: The *soft-attention* mechanism $f_s(x_t;\theta_f)$ that is parameterized by θ_f encodes the observed image region x_t to a soft-attention map where the valuable information is highlighted. It is achieved by a soft-attention network (SA-Net) as shown in Figure.2. In the SA-Net, it contains a mask branch and a trunk branch. The soft mask branch aims to learn a mask $M(x_t)$ in range of $[0,1]$ by a symmetrical top-down architecture and a sigmoid layer to normalize the output. The trunk branch outputs the feature map $T(x_t)$ and the final attention map is computed by:

$$\mathcal{A}(x_t) = (1 + M(x_t)) * T(x_t), \tag{1}$$

and the soft-attention features $f_s(x_t;\theta_f)$ are then learned by a global average pooling over the attention map $\mathcal{A}(x_t)$. In order to fuse both learned attention features and location information, we build a fusion network $\mathcal{H}$ to finally produce fused feature vector $g_t = \mathcal{H}(f_s(x_t;\theta_f), l_{t-1};\theta_g)$ based on a fully-connected layer with ReLu activation.

Classify Stage: We further use a LSTM to process the learned fused feature g_t. The advantage of LSTM is that it is able to summarize the past information, and to learn an optimal classification policy $\pi((l_t, a_t)|s_{1:t};\theta)$, where a_t is decision to classify image at time step t and $s_{1:t}$ represents the past history $s_{1:t} = x_1, l_1, a_1, ..., x_{t-1}, l_{t-1}, a_{t-1}, x_t$. The internal state is formed and updated by the hidden unit h_t in LSTM [7]: $h_t = f_h(h_{t-1}, g_t;\theta_h)$. The recurrent LSTM network then has to choose actions including how to classify image and where to look at in the next time step based on the internal state. In this work, both actions are drawn stochastically from two distributions. The classification action a_t is drawn from classification network by softmax output at step t: $a_t \sim (\cdot|f_a(h_t;\theta_a))$. Similarly, the location l_t is also drawn from a location network by $l_t \sim (\cdot|f_l(h_t;\theta_l))$.

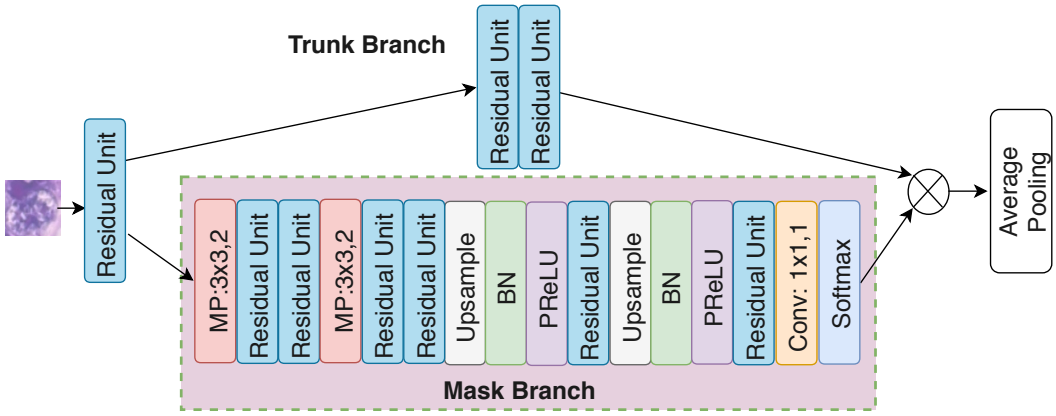

Fig. 2. The structure of SA-Net. Here Conv($1 \times 1, 1$) denotes a convolutional layer with kernel size of 1 and stride of 1. We use 64 convolutional filters for the last Conv layers. 'BN' denotes batch normalization. MP($3 \times 3, 2$) means max-pooling size is set to 3 and stride is 2. 'PReLU' refers to the activation function PReLU is applied. 'Upsample' denotes upsampling by bilinear interpolation. The sturcture of residual unit is shown in Figure.3

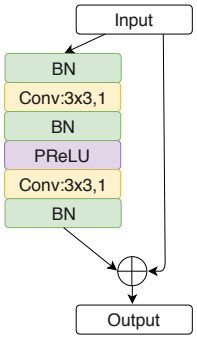

Fig. 3. The structure of residual unit in SA-Net. We use 64 convolutional filters in each Conv layer.

When executing the chosen actions, we could receive a image patch x_{t+1} and also a *reward* r_{t+1} referring to whether we have correctly classified image. The total reward could be written as: $R = \sum_{t=1}^{T} r_t$. In this paper, we set reward to 0 for all other time steps except the last time step. In the last time step, the reward is set to 1 if the image is classified correctly and 0 if not.

2.2. Network Optimization

As the hard-attention mechanism is non-differentiable, we optimize the whole network through policy gradient approach. In this paper, we aim to maximize the reward as:

$$J(\theta) = \mathbb{E}_{p(s_{1:T};\theta)}[\sum_{t=1}^{T} r_t] = \mathbb{E}_{p(s_{1:T};\theta)}[R]. \tag{2}$$

In order to maximize J, the gradient of J could be approximate by:

$$\begin{aligned} \nabla_\theta J &= \sum_{t=1}^{T} [\nabla_\theta \log \pi_\theta(a_t, l_t | s_{1:t}) R] \\ &\approx \frac{1}{M} \sum_{i=1}^{M} \sum_{t=1}^{T} \nabla_\theta \log \pi_\theta(a_t^i, l_t^i | s_{1:t}^i) R^i \end{aligned} \tag{3}$$

where $i = 1...M$ is the running epochs [8]. Equation. 3 encourages network to adjust parameters for the chosen probability of actions that would lead to high cumulative reward and to decrease probability of actions that would decrease reward. To achieve this, we could update the network by:

$$\theta \leftarrow \theta + \alpha \nabla_\theta J(\theta). \tag{4}$$

At the meanwhile, we could also combine Equation. 4 with the supervised classification training approach, i.e. to also train the network by the cross-entropy loss with ground-truth label. Thus, the network could be learned by minimizing the total loss:

$$\mathcal{L}_{total} = -J(\theta) + \mathcal{L}_c(y, \hat{y}), \tag{5}$$

where y is the ground-truth classification label, $\hat{y}$ is predicted label from network, and $\mathcal{L}_c$ is the cross-entropy classification loss.

3. EXPERIMENT

3.1. Datasets and Parameters Setting

We evaluated our approach on a public dataset BreakHis [9]. The dataset contains 7,909 images collected from 82 patients including 58 for malignant and 24 for benign. These tumor

Table 1. Performance comparison of magnification specific system (in %)."Ours w/o SA" denotes the SA-Net is removed. n/a denotes the authors did not report the corresponding data.

Methods	Magnification			
	40×	100×	200×	400×
Spanhol [9]	83.8 ± 4.1	82.1 ± 4.9	85.1 ± 3.1	82.3 ± 3.8
Spanhol [10]	90.0 ± 6.7	88.4 ± 4.8	84.6 ± 4.2	86.1 ± 6.2
Gupta [11]	86.7 ± 2.3	88.6 ± 2.7	90.3 ± 3.7	88.3 ± 3.0
Sequential [12]	94.7 ± 0.8	95.9 ± 4.2	96.7 ± 1.1	89.1 ± 0.1
FV+CNN [13]	90.0 ± 3.2	88.9 ± 5.0	86.9 ± 5.2	86.3 ± 7.0
MIL+CNN [14]	81.3± n/a	80.4± n/a	77.6± n/a	79.1± n/a
MIL [15]	89.5± n/a	89.0± n/a	88.8± n/a	87.7± n/a
S-CNN [3]	94.1 ± 2.1	93.2 ± 1.4	94.7 ± 3.6	93.5 ± 2.7
Ours w/o SA	88.6 ± 1.9	87.0 ± 1.8	86.6 ± 2.8	85.2 ± 1.9
Ours	97.5 ± 1.6	96.2 ± 1.3	97.4 ± 2.5	95.4 ± 1.5

tissue images are captured at four kinds of optical magnifications of $40\times$, $100\times$, $200\times$, and $400\times$.

In the experiment, we randomly select 58 patients (70%) for training and 24 patients (30%) for testing. Before training, we augmented raw image by applying rotation, horizontal and vertical flips, which results in 3 times the original training data. The raw image size in the dataset is 740×460. The size of five cropped images in our network is set to 112×112, which means we only have to process around 15% pixels of raw image. We choose Adam optimizer with a learning rate of 0.01 that exponentially decay over epochs. In the training stage, it usually takes around 200 epochs to convergence. The experiment is conducted on a workstation with four Nvidia 1080 Ti GPUs.

The performance of our approach is evaluated by the Patient recognition rate (PRR), in order to be comparable with previous work. PRR aims to calculate a ratio of correctly classified tissues to all the number of tissues. It could be formulated as:

$$PRR = \frac{\sum_{i=1}^{N} ACC_i}{N}, ACC = \frac{N_{rec}}{N_p} \quad (6)$$

where N is the total number of patients in the testing data. N_{rec} is the correctly classified tissues of patient p and N_p is total tissue number from patient p.

3.2. Comparison with other approaches

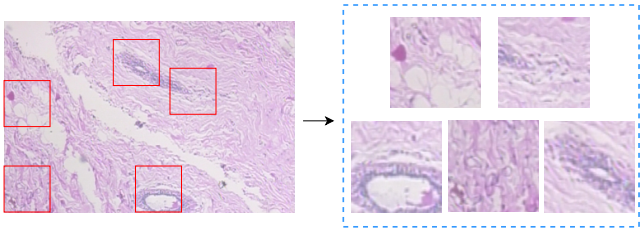

Fig. 4. An example of how hard-attention mechanism selects image patches.

To evaluate the performance of our approach to histopathology image classification, we compare our proposed deep learning framework with the state-of-the-art approaches. The results is shown in Table.1 which demonstrates our approach outperforms all previous approaches. It should be noticed that our approach achieves much higher accuracy rate than most CNN approaches [13, 14, 15]. It is achieved by the well-designed attention mechanisms to select useful regions for the decision network (Figure.4). The hard-attention mechanism finds out the regions most related to abnormality part and the soft-attention mechanism highlight those abnormal features. Apart from the superior performance to the previous approaches, our approaches prevents to resize raw image which might leads to information loss, and also enables network to process image in the small size image patch in order to save computational cost.

We also conducted an ablation study to evaluate the effectiveness of the soft-attention. We remove SA-Net to test the performance of rest network. It could be seen that classification accuracy dropped down by around 10%. The decreasing of performance is due to some redundant features are also processed by the network, which might contains noise features that leading to misclassification. Thus, it is essential to apply soft-attention mechanism to highlight useful features and also encourage network to neglect those unnecessary image features.

4. CONCLUSION

In this paper, we introduce a novel deep hybrid attention network to the breast cancer histopathology image classification. The hard-attention mechanism in the network could automatically find the useful region from raw image, and thus does not have to resize raw image for the network to prevent information loss. The built-in recurrent network can make decisions to classify image and also to predict region for next time step. We evaluate our approach on a public dataset, and it achieves around 96% accuracy on four different magnifications while only 15% of raw image pixels are used to make decisions to classify input image.

5. REFERENCES

[1] American Cancer Society, *Cancer facts & figures*, The Society, 2008.

[2] Fabio A Spanhol, Luiz S Oliveira, Paulo R Cavalin, Caroline Petitjean, and Laurent Heutte, "Deep features for breast cancer histopathological image classification," in *Systems, Man, and Cybernetics (SMC), 2017 IEEE International Conference on*. IEEE, 2017, pp. 1868–1873.

[3] Zhongyi Han, Benzheng Wei, Yuanjie Zheng, Yilong Yin, Kejian Li, and Shuo Li, "Breast cancer multi-

classification from histopathological images with structured deep learning model," *Scientific reports*, vol. 7, no. 1, pp. 4172, 2017.

[4] Nima Habibzadeh Motlagh, Mahboobeh Jannesary, HamidReza Aboulkheyr, Pegah Khosravi, Olivier Elemento, Mehdi Totonchi, and Iman Hajirasouliha, "Breast cancer histopathological image classification: A deep learning approach," *bioRxiv*, p. 242818, 2018.

[5] Alexander Rakhlin, Alexey Shvets, Vladimir Iglovikov, and Alexandr A Kalinin, "Deep convolutional neural networks for breast cancer histology image analysis," in *International Conference Image Analysis and Recognition*. Springer, 2018, pp. 737–744.

[6] Volodymyr Mnih, Nicolas Heess, Alex Graves, et al., "Recurrent models of visual attention," in *Advances in neural information processing systems*, 2014, pp. 2204–2212.

[7] Sepp Hochreiter and Jürgen Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[8] Ronald J Williams, "Simple statistical gradient-following algorithms for connectionist reinforcement learning," *Machine learning*, vol. 8, no. 3-4, pp. 229–256, 1992.

[9] Fabio A Spanhol, Luiz S Oliveira, Caroline Petitjean, and Laurent Heutte, "A dataset for breast cancer histopathological image classification," *IEEE Transactions on Biomedical Engineering*, vol. 63, no. 7, pp. 1455–1462, 2016.

[10] Fabio Alexandre Spanhol, Luiz S Oliveira, Caroline Petitjean, and Laurent Heutte, "Breast cancer histopathological image classification using convolutional neural networks," in *Neural Networks (IJCNN), 2016 International Joint Conference on*. IEEE, 2016, pp. 2560–2567.

[11] Vibha Gupta and Arnav Bhavsar, "Breast cancer histopathological image classification: is magnification important?," in *IEEE Conf. on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2017.

[12] Vibha Gupta and Arnav Bhavsar, "Sequential modeling of deep features for breast cancer histopathological image classification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2018, pp. 2254–2261.

[13] Yang Song, Ju Jia Zou, Hang Chang, and Weidong Cai, "Adapting fisher vectors for histopathology image classification," in *Biomedical Imaging (ISBI 2017), 2017 IEEE 14th International Symposium on*. IEEE, 2017, pp. 600–603.

[14] Jiajun Wu, Yinan Yu, Chang Huang, and Kai Yu, "Deep multiple instance learning for image classification and auto-annotation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3460–3469.

[15] Kausik Das, Sailesh Conjeti, Abhijit Guha Roy, Jyotirmoy Chatterjee, and Debdoot Sheet, "Multiple instance learning of deep convolutional neural networks for breast histopathology whole slide classification," in *Biomedical Imaging (ISBI 2018), 2018 IEEE 15th International Symposium on*. IEEE, 2018, pp. 578–581.