Predicting Cancer with a Recurrent Visual Attention Model for Histopathology Images

Aïcha BenTaieb(✉) and Ghassan Hamarneh

School of Computing Science, Simon Fraser University, Burnaby, Canada
{abentaie,hamarneh}@sfu.ca

Abstract. Automatically recognizing cancers from multi-gigapixel whole slide histopathology images is one of the challenges facing machine and deep learning based solutions for digital pathology. Currently, most automatic systems for histopathology are not scalable to large images and hence require a patch-based representation; a sub-optimal solution as it results in important additional computational costs but more importantly in the loss of contextual information. We present a novel attention-based model for predicting cancer from histopathology whole slide images. The proposed model is capable of attending to the most discriminative regions of an image by adaptively selecting a limited sequence of locations and only processing the selected areas of tissues. We demonstrate the utility of the proposed model on the slide-based prediction of macro and micro metastases in sentinel lymph nodes of breast cancer patients. We achieve competitive results with state-of-the-art convolutional networks while automatically identifying discriminative areas of tissues.

1 Introduction

Cancers are primarily diagnosed from the visual analysis of digitized or physical histology slides of tumor biopsies [4]. Growing access to large datasets of digitized histopathology images has led to the emergence of computational models where the aim is to reduce experts workload and improve cancer treatment procedures [6]. Recently, convolutional neural networks (CNN) have become the state-of-the-art for many histopathology image classification tasks. However, CNNs are not the best suited for large scale (i.e. millions of pixels) multi-resolution histopathology whole slide images (WSI). Finding adequate and computationally efficient solutions to automatically analyze WSI remains an open challenge.

A standard approach for analyzing WSI consists of sampling patches from areas of interest and training a supervised model to predict a desired output (e.g., a class label) for each patch independently [6]. The trained model can then be applied to patches densely extracted from an unseen WSI where the final slide prediction is the result of an aggregation of all patch predictions. Such patch based representation comes with different shortcomings: (i) processing

A. F. Frangi et al. (Eds.): MICCAI 2018, LNCS 11071, pp. 129–137, 2018.
https://doi.org/10.1007/978-3-030-00934-2_15

all patches of a WSI is computationally inefficient (as most tissue areas are diagnostically irrelevant) and almost always unfeasible; (ii) randomly sampled patches can result in the loss of relevant information and often involve using finer-level annotations (i.e. segmentation masks) to guide the patch extraction; and (iii) using independently analyzed patches implies a loss of context.

Different works were proposed to improve patch-based representations. Mainly, these works present different aggregation strategies and encode global context. For instance, weakly-supervised models based on multiple instance learning [7] or structured latent representations [3] have been proposed to show the importance of identifying discriminative regions when training a prediction model. To capture context (without increasing patch size), pyramid representations where patches are extracted at different magnifications can be leveraged. For instance, Bejnardi et al. [2] proposed a patch-based model consisting of a cascaded CNN architecture where features from patches extracted at increasing scales are aggregated to classify breast cancer tissue slides. Another strategy for capturing spatial context from patch-based representations is to use recurrent networks. Agarwalla et al. [1] used 2D LSTMs to aggregate features from neighbouring patches in a WSI. While these works indirectly impose more context in the training of a patch-based prediction model, they rely on an initial random selection of patches that does not prevent from an eventual loss of information and most importantly requires processing all patches independently. In this work, we attempt to leverage spatial context while selecting discriminative areas. Studies on experts visual diagnostic procedure [4] showed that over time, experts make fewer fixations and perform less examinations of non-diagnostic areas. We hypothesize that patch-based analysis of tissue slides should be a sequential process in which a prediction model identifies where to focus given the context of the entire tissue and the history of previously seen regions without other forms of annotation than the slide level class.

To design such system, we take inspiration from visual attention models [8]. A number of recent studies have demonstrated that visual content can be captured through a sequence of spatial ‘glimpses’ [9] describing parts of an image. Focusing computational resources on parts of a scene has the interesting property of substantially reducing the task complexity as objects of interest can be placed in the center of the glimpse. Existing visual attention systems were introduced for analyzing natural scene images [9] but their utility for large scale images has not been demonstrated yet.

We propose a system to analyze whole slide histopathology images and predict the presence of cancer while automatically learning to focus on discriminative areas (Fig. 1). We assume the system should be able to predict normal vs abnormal slides from a limited set of observations or glimpses. Locations and scales at which glimpses are extracted should be automatically inferred. Decisions about the central locations of glimpses should be based on the global context of a given tissue slide as well as the memory of all previously observed glimpses. The slide level class prediction should be based on information integrated from all observed glimpses as well as the global context. Finally, through time, the system should learn to make decisions about the class of a tissue slide using a limited set of glimpses.

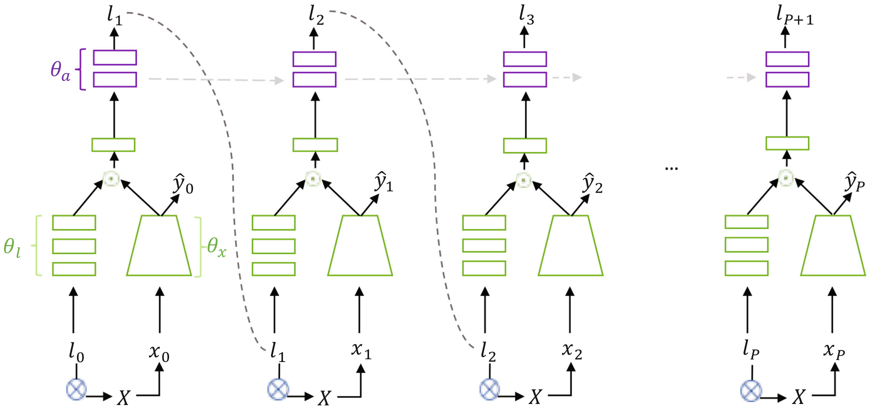

Fig. 1. Proposed recurrent visual attention model for classifying histopathology images. Grey dashed lines represent temporal connections while solid black lines describe the information flow between components within one time-step. The model includes three primary components composed of dense (rectangular boxes) or convolutional (trapezoid) layers. X is an input whole slide image, $\{x_0, \ldots, x_P\}$ is the sequence of glimpses with their corresponding location parameters $\{l_0, \ldots, l_p\}$. The system contains three main components parameterized by θ_x, θ_l and θ_a. $\odot$ represents the Hadamard product and $\bigotimes$ is a matrix multiplication. The model sequentially predicts a class label $\hat{y}$ for the tissue slide given the sequence of glimpses.

2 Method

Given a whole slide histopathology image X, our goal is to identify a set of locations $\{l_0, l_1, \ldots, l_P\}$ from which to extract glimpses $\{x_0, x_1, \ldots, x_P\}$ that are discriminative of a given class Y (e.g. presence or absence of metastatic cancer). To this end, we propose a sequential system structured around a recurrent neural network equipped with an attention memory and an appearance description of the tissue at different locations.

At each time step, the system receives a location l_p that defines the extraction of a corresponding glimpse x_p. A location network θ_l forms a feature representation of a given location and an appearance network θ_x generates a feature representation for a given glimpse. These feature representations are aggregated to form part of the input to the attention network θ_a. Given a sequence $\{x_0, x_1, \ldots, x_P\}$ of P extracted glimpses, the system parameterized by $\theta = \{\theta_l, \theta_x, \theta_a\}$ predicts a probability score $Q(Y|\{x_0, x_1, \ldots, x_P\}; \theta)$ for the slide-level label Y. The attention network is the recurrent component of the model and uses information from the glimpses and their corresponding location parameters to update its internal representation of the input and outputs the next location parameters. Figure 1 is a graphical representation of this sequential procedure.

Spatial Attention: The spatial attention mechanism consists of extracting a glimpse x_p from a tissue slide and is a modification of the read mechanism introduced in [8]. Given an input tissue slide $X \in \mathcal{R}^{H\times W\times 3}$ of size $H \times W$, we apply two grids (one for each axis of the image) of two-dimensional Gaussian filters, where each filter response corresponds to a pixel in the resulting glimpse $x_p \in \mathcal{R}^{h\times w\times 3}$ of size $h \times w$. The attention mechanism is represented by parameters $l = \{\mu_w, \mu_h, \sigma_w^2, \sigma_h^2, \delta_w, \delta_h\}$ that describe the centers of the Gaussians (i.e. the grid center coordinates), their variances (i.e. amount of blurring to apply), and strides between the Gaussian centers (i.e. the scale of the glimpse). Parameters l are dynamically computed as an affine transformation of the output of the recurrent network θ_a. Formally, the glimpse is defined by $x_p = A_p^x X A_p^{y^T}$, where A_p^x and A_p^y are the Gaussian grid matrices applied on each axis of the original image X. To integrate the entire context of a given tissue slide, we initialize the first location parameters l_0 such that the resulting glimpse x_0 corresponds to a coarse representation of the tissue slide (i.e. lowest magnification) re-sized to the desired glimpse size $h \times w$.

Combining Appearance and Spatial Information: Given a glimpse x_p and its corresponding location parameters l_p, we construct a fixed-dimensional feature vector comprising appearance and spatial information about the current glimpse. We denote the appearance-based features obtained for a given glimpse by $f_x(x_p; \theta_x)$ and the features computed for the corresponding location parameters by $f_l(l_p; \theta_l)$. We used a CNN to represent f_x and a fully connected layer for f_l. The outputs of both networks are fused to obtain a joint representation that captures spatial and appearance features using $g_p = \sigma(f_l(l_p; \theta_l) \odot f_x(x_p; \theta_x))$, where g_p is the output joint feature vector, σ corresponds to the logistic sigmoid function, and $\odot$ is the Hadamard product. By combining appearance and spatial features, the system integrates features related to "where" and"what" patterns to seek for when predicting the next glimpse location parameters.

Recurrent Attention: The recurrent component of the system aggregates information extracted from all individual glimpses and their corresponding locations. It receives as input the joint spatial and appearance representation (i.e. g_p) and maintains an internal state summarizing information extracted from the sequence of past glimpses. At each step p, the recurrent attention network updates its internal state (formed by the hidden units of the network) based on the incoming feature representation g_p and outputs a prediction for the next location l_{p+1} to focus on at time step $p + 1$. The spatial attention parameters l_p are formed as a linear function of the internal state of the network.

Objective Function: The system is trained by minimizing a loss function comprised of a classification loss term and auxiliary regularization terms that guide the attention mechanism. The total loss $\mathcal{L}(.)$ is given by:

$$\mathcal{L}(\mathcal{D}; \theta) = \mathcal{L}_c(\mathcal{D}; \theta) + \mathcal{L}_p(\mathcal{D}; \theta) + \mathcal{L}_a(\mathcal{D}; \theta) + \mathcal{L}_l(\mathcal{D}; \theta) \qquad (1)$$

where $\mathcal{D} = \{(X^{(i)}, Y^{(i)})\}_{i=1}^{N}$ is a training set of N tissue slides $X^{(i)}$ and their corresponding labels $Y^{(i)}$ and $\theta = \{\theta_a, \theta_x, \theta_l\}$ represent the system's parameters.

Tissue Slide Classification: The slide-level classification loss $\mathcal{L}_c$ is defined as the cross entropy between the final slide-level predicted label $\hat{Y}$ and the true label $Y^{(i)}$. To obtain a slide-level prediction, we combine the feature representations of all glimpses $f_x(x_{[1:P]}; \theta_x)$ using a non-linear function represented by a fully connected layer. This layer is then fed to another linear layer that generates final predictions $Q(Y^{(i)}|x_{[1:P]}^{(i)}; \theta)$. The slide-level loss is computed using $\mathcal{L}_c(\mathcal{D}; \theta) = \sum_{i=1}^{N} \log Q(\hat{Y} = Y^{(i)}|x_{[1:P]}^{(i)}; \theta)$.

Discriminative Attention and Selective Exploration: We observed that adding a patch-level classification loss facilitates training by enforcing the model to attend to discriminative tissue areas. $\mathcal{L}_p$ corresponds to a classification cross entropy loss between each predicted patch-level label $\hat{y}_p$ and the ground truth slide label $Y^{(i)}$. The goal here is not to leverage other forms of annotations but to encourage finding discriminative regions in a weakly supervised setting. Feature representations of each attended patch $f_x(x_p; \theta_x)$ are used to compute the patch-level loss by $\mathcal{L}_p(\mathcal{D}; \theta) = \sum_{i=1}^{N} \sum_{p=1}^{P} \log Q(\hat{y}_p = Y^{(i)}|x_p^{(i)}; \theta)$, where $Q(\hat{y}_p^{(i)}|x_p^{(i)}; \theta)$ represents the probabilities obtained from a fully-connected layer applied to the patch-level features $f_x(x_p; \theta_x)$ with the sigmoid activation.

We also observed that after seeing the coarse image representation x_0, it becomes harder to attend to other areas as the rich contextual representation is often enough to discriminate between simple cases (e.g. benign vs macro-metastases). To encourage the system to explore different locations and scales, we introduce a regularization term that serves two ends. First, we encourage the system to gradually approach the most discriminative regions and scales by favouring glimpses with high prediction probabilities for the ground truth class using $\mathcal{L}_a$. Second, we encourage exploration by enforcing large differences between successive predicted centers μ_w and μ_h using $\mathcal{L}_l$. Formally, we define:

$$\mathcal{L}_a(\mathcal{D}; \theta) = -\sum_{i=1}^{N} \sum_{p=2}^{P} Q(y_p^{(i)}|x_p^{(i)}; \theta) - \left(\frac{1}{p-1} \sum_{k=1}^{p-1} Q(y_k^{(i)}|x_k^{(i)}; \theta) \right) \tag{2}$$

$$\mathcal{L}_l(\mathcal{D}; \theta) = \gamma \sum_{i=1}^{N} \sum_{p=1}^{P} \exp(-|l_p - l_{p+1}|), \tag{3}$$

where the hyper-parameter γ enables us to control how much exploration the system performs by being linearly annealed from one to zero during training. At inference, given an unseen tissue slide, the model extracts a sequence of glimpses to attend to the most discriminative regions. The final prediction score for the slide is computed using the aggregated features $f_x(x_{[1:P]}^{(i)}; \theta_x)$.

3 Experiments

We tested the system on the publicly available Camelyon16 dataset [5] where the task is to predict benign from metastatic cases of lymph nodes tissue slides.

Table 1. Evaluation of different patch-based models for WSI classification. Columns represent: patch level (ACC-P) and WSI level (ACC-WSI) accuracy, area under ROC curve (AUC), precision (PREC) and recall (REC).

Method	ACC-P	ACC-WSI	AUC	PREC	REC
Wang et al. [10]	0.98	–	**0.96**	–	–
LSVM [3]	0.89 ± 0.2	0.84 ± 0.2	0.75	0.87	0.69
Dense Patches	0.84 ± 0.2	0.76 ± 0.3	0.72	0.94	0.66
Proposed 1 Glimpse - $\mathcal{L}_c$	0.81 ± 0.2	0.79 ± 0.2	0.68	0.60	0.48
Proposed 3 Glimpses - $\mathcal{L}_c$	0.84 ± 0.1	0.80 ± 0.2	0.85	0.69	0.64
Proposed 5 Glimpses - $\mathcal{L}_c$	0.81 ± 0.1	0.78 ± 0.2	0.83	0.65	0.62
Proposed 3 Glimpses - $\mathcal{L}_c + \mathcal{L}_p$	0.87 ± 0.2	0.86 ± 0.2	0.84	0.81	0.78
Proposed 3 Glimpses - $\mathcal{L}_c + \mathcal{L}_p + \mathcal{L}_a$	**0.97 ± 0.1**	**0.95 ± 0.2**	0.95	**0.98**	**0.82**

The dataset contains a total of 400 WSI and we used the same dataset splits as the ones released by the challenge organizers for training (270) and test (130).

Typically histopathology images contain billions of pixels but only a few portion of the slide contains biological tissues. To reduce the amount of computation, we remove all unnecessary background pixels using a simple threshold on the pixel intensity values and crop all slides around the tissue. Although the total size is reduced, in practice, performing the matrix multiplication for the spatial attention at the highest magnification level of a slide, is computationally unfeasible with standard resources. Instead, we opt for processing images at the intermediate 20x magnification using tiles covering as much context as possible. A tile size of 5000 × 5000 pixels (Fig. 2) was the largest we could process. To predict a class label for a slide, we apply the system on all 20x tiles and let it decide at which scale and location to attend. We use the average of the probabilities obtained after attention to get a final slide prediction. The total run-time was on average less than 4s per slide.

Table 1 reports the performance of the model against different baselines. Wang et al. [10], the winners of the challenge, used the Inception CNN architecture to train a patch-based classifier on randomly sampled patches at 40x magnification. To obtain slide level predictions, the output probabilities of the patch-based CNN are used to predict a heatmap. Statistical features are extracted from the resulting heatmap (e.g. morphology and geometry features) and used to train a random forest classifier that outputs the final predicted slide label. We also compared against the latent structured SVM model presented in [3]. To train this model, we extracted patches at two magnification levels (20x and 40x) and used a pre-trained Inception CNN model to extract features for each patch. The latent structured model uses a hierarchical representation of patches at both magnifications to identify the most discriminative patches while training the classifier. We also trained the Inception CNN model using densely sampled patches from each whole slide image at magnification 20x. Given the high ratio of positives to negatives, we leveraged the segmentation masks of tumors to train this baseline and dynamically sampled tumor patches. Finally, we tested the following configurations of our system: (i) different number of glimpses and (ii) different combinations of the proposed loss terms in Eq. (1).

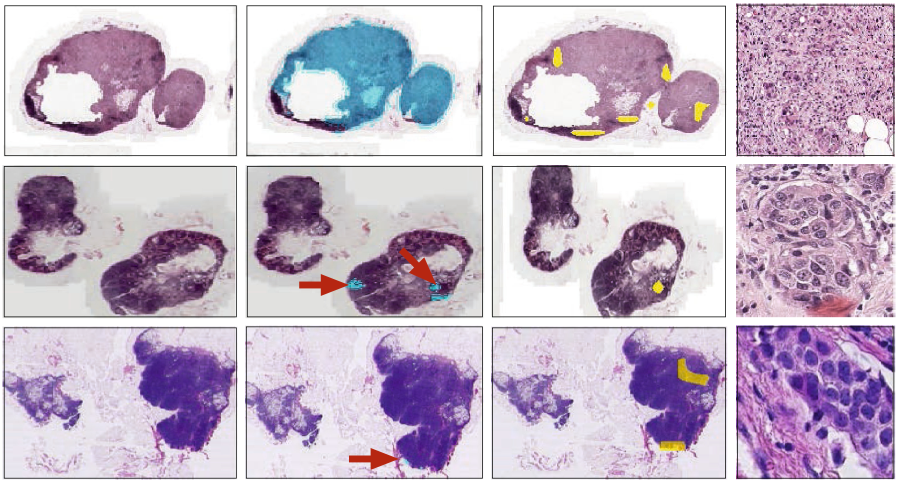

Fig. 2. Qualitative evaluation of the attention model. Rows represent different cases of macro to micro metastases. Columns from left to right are the downsampled WSI, the cyan overlay of the ground truth tumor mask with red arrows pointing at micro-metastasis, the yellow overlay of the attended glimpses and the glimpse with highest prediction score showing how glimpses are automatically extracted at different scales.

We tested the performance of the system using different numbers of glimpses (i.e., 1, 3 or 5 glimpses per tile). On average, after background removal, we obtain ∼14 tiles per tissue slide. Thus, the final performance results reported in Table 1 correspond to an aggregation of 14 (case of 1 glimpse per tile) to 70 glimpses. In contrast, all other automatic systems were trained with thousands of patches. We obtained best results using 3 glimpses (i.e., 85% AUC vs 68% and 83% for 1 and 5 glimpses when training with $\mathcal{L}_c$ only). We also observed that using 1 glimpse (i.e., 14 attention patches per slide) resulted in a 4% drop in AUC only. Note that this is most likely specific to this particular dataset in which macro-metastatic tissues contain large amounts of abnormality and are thus easily discriminated from benign tissues. However, this also shows the utility of identifying discriminative locations when training prediction systems.

We also tested the impact of the different loss terms in Eq. (1). In general, the patch-level loss $\mathcal{L}_p$ resulted in improving the attention on positive cases which is reflected by the improved recall scores (i.e., from 64% to 78% with 3 glimpses). Finally, adding the attention regularization terms $\mathcal{L}_a$ and $\mathcal{L}_l$ primarily helped facilitate convergence (i.e. reduced the convergence time by ∼15%) and improved the final AUC, precision and recall. Note that our final AUC is 1%

lower than [10], however, our aim in this work is to demonstrate how attention can be leveraged in histopathology by selectively choosing where to focus.

In Fig. 2 we show examples of glimpses. Comparing the attended areas to the ground truth masks of metastatic tissues (columns 3 and 2 respectively) shows that the attention mechanism is able to identify discriminative patterns and solely focus on those regions. The last column in Fig. 2 shows glimpses with the highest prediction score for each WSI class and demonstrates that the system learns patterns from different scales. The last row in Fig. 2 shows a failure example on a challenging case of micro-metastases. In this case, the model was correctly able to identify discriminative patterns (the yellow overlay on images of column 3 shows the attention areas used to predict the slide label) but unable to predict the correct slide level class. Given the high ratio of negative to positive tissue in micro-metastatic patches, this may indicate that a more complex aggregation strategy (instead of the simple linear aggregation) for the different attended glimpses may be necessary.

4 Conclusion

We hypothesized that enforcing a selective attention mechanism when predicting the presence of cancer within a tissue slide would enable the prediction system to identify discriminative patterns and integrate context. To test our hypothesis, we proposed a prediction model that integrates a recurrent attention mechanism. Experiments on a dataset of breast tissue images showed that the proposed model is capable of selectively attending to discriminative regions of tissues and accurately identifying abnormal areas with a limited sequence of visual glimpses.

Acknowledgements. We thank NVIDIA Corporation for GPU donation and The Natural Sciences and Engineering Research Council of Canada (NSERC) for funding.

References

1. Agarwalla, A., Shaban, M., Rajpoot, N.M.: Representation-aggregation networks for segmentation of multi-gigapixel histology images. arXiv preprint arXiv:1707.08814 (2017)
2. Bejnordi, B.E., et al.: Context-aware stacked convolutional neural networks for classification of breast carcinomas in whole-slide histopathology images. MedIA **4**(4), 044504 (2017)
3. BenTaieb, A., et al.: A structured latent model for ovarian carcinoma subtyping from histopathology slides. MedIA **39**, 194–205 (2017)
4. Brunye, T.T., et al.: Eye movements as an index of pathologist visual expertise: a pilot study. PloS one **9**(8), e103447 (2014)
5. Golden, J.A.: Deep learning algorithms for detection of lymph node metastases from breast cancer: helping artificial intelligence be seen. JAMA **318**(22), 2184–2186 (2017)
6. Komura, D., Ishikawa, S.: Machine learning methods for histopathological image analysis. arXiv preprint arXiv:1709.00786 (2017)

7. Mercan, C., et al.: Multi-instance multi-label learning for multi-class classification of whole slide breast histopathology images. IEEE TMI **37**(1), 316–325 (2018)
8. Mnih, V., et al.: Recurrent models of visual attention. In: NIPS, pp. 2204–2212 (2014)
9. Sermanet, P., Frome, A., Real, E.: Attention for fine-grained categorization. arXiv preprint arXiv:1412.7054 (2014)
10. Wang, D., et al.: Deep learning for identifying metastatic breast cancer. arXiv preprint arXiv:1606.05718 (2016)