A Joint Spatial and Magnification Based Attention Framework for Large Scale Histopathology Classification

Jingwei Zhang[1], Ke Ma[1], John Van Arnam[1], Rajarsi Gupta[1], Joel Saltz[1], Maria Vakalopoulou[2], Dimitris Samaras[1]
[1]Stony Brook University, USA [2]CentraleSupélec, University of Paris-Saclay, France
{jingwezhang, kemma, samaras}@cs.stonybrook.edu john.vanarnam@gmail.com
{Rajarsi.Gupta, Joel.Saltz}@stonybrookmedicine.edu maria.vakalopoulou@centralesupelec.fr

Abstract

Deep learning has achieved great success in processing large size medical images such as histopathology slides. However, conventional deep learning methods cannot handle the enormous image sizes; instead, they split the image into patches which are exhaustively processed, usually through multi-instance learning approaches. Moreover and especially in histopathology, determining the most appropriate magnification to generate these patches is also exhaustive: a model needs to traverse all the possible magnifications to select the optimal one. These limitations make the application of deep learning on large medical images and in particular histopathological images markedly inefficient. To tackle these problems, we propose a novel spatial and magnification based attention sampling strategy. First, we use a down-sampled large size image to estimate an attention map that represents a spatial probability distribution of informative patches at different magnifications. Then a small number of patches are cropped from the large size medical image at certain magnifications based on the obtained attention. The final label of the large size image is predicted solely by these patches using an end-to-end training strategy. Our experiments on two different histopathology datasets, the publicly available BACH and a subset of the TCGA-PRAD dataset, demonstrate that the proposed method runs 2.5 *times faster with automatic magnification selection in training and at least* 1.6 *times faster than using all patches in inference as the most of state-of-the-art methods do, without loosing in performance.*

1. Introduction

Deep neural networks have achieved significant success in a variety of medical tasks [18]. A particular domain of interest for deep learning methods has been histopathology, where deep learning approaches are already providing state of the art performance for classification tasks. Methods focusing on different cancer predictions [8], classification of tumor [17] or detection of cancer metastases in lymph node sections [2] have received a lot of attention recently. Moreover, methods tackling clinical endpoints such as survival analysis [31, 20] or prediction of prognostic mutated genes [3] focus on deep learning approaches indicating the importance of efficient, scalable and accurate automatic systems.

Even if current deep learning methods provide very promising directions in histopathology, we argue that the current way that they are used, exhaustively investigating all the different histology regions (patches) is costly and inefficient. The inefficiency is shown in two aspects: on one hand, not all the patches are informative for specific classification tasks. For example, a patch cropped at a fatty tissue region provides little information about grades of cancer. On the other hand, the informative patches may be a source of redundant information. For example, two patches that are spatially close to each other may have the same discriminative power. Based on this intuition, some methods consider patch selection strategies and attention mechanisms [32, 34, 11].

Among these patch-based methods, determining the optimal magnification to crop patches is also very important, but usually neglected. Given a large size image or a whole slide image (WSI) and a fixed patch size, cropping at too low magnification yields a small number of patches with adequate context. Additionally, low magnification can cause a loss of discriminative details that are useful for the specific classification task. However, cropping at too high magnification augments significantly the number of patches containing the finest scale of details at the loss of context, or at increased training cost. Currently, even if magnification seems to be very informative for a variety of tasks [23], few works investigated the effects of different patch sizes and magnifications [12]. Most of the time the selection of the

optimal magnification is based on medical practice without really ensuring the best performance of the deep learning models.

In this work, we propose an attention sampling framework that addresses these limitations introducing both spatial and magnification based attention for large histopathology images. The contributions of this work are: *(i)* we propose a novel spatial and magnification sampling mechanism for classification tasks adapting in an end-to-end fashion, *(ii)* we provide a fast (low training and inference time) and low-cost (based on small regions) framework that gives competitive results compared to other state of the art methods for two different classification tasks of histopathology. *To the best of our knowledge, this is the first attempt that proposes a fully automatic way for a joint spatial and magnification based attention mechanism. Our method performs comparably to the state of the art methods, using only a small subset of the initial input, reducing time and resources requirements.*

2. Previous Work

We roughly categorize the previous large size medical image classification methods presented in histopathology into patch exhaustive methods and patch selective methods. Exhausting all the patches in a large medical image is a straightforward way to classify the whole image. Korbar *et al*. [15] trained a convolutional neural network (CNN) to predict each patch label in a WSI in a supervised manner for colorectal polyp classification. The final whole image label is estimated by simple patch level majority voting. Similarly, Han *et al*. [16] proposed a deep learning based architecture to detect lymphocytes on large breast cancer images by exhaustively investigating all the available patches. Moreover, in [24] the authors present an analysis of different deep learning architectures in an extensive search of magnifications for the multi-classification of breast cancer histopathology images. Even though these methods achieve promising performance, most of the time, they require not only the image level labels, but also patch level annotations. Annotating each patch requires a significant amount of tedious work from pathologists since each large medical image contains hundreds and thousands of patches, while it is not clear on which magnification the patch level annotation should be provided.

To tackle this problem, Hou *et al*. [9] proposed to use image level annotation only. Thus, the image classification task becomes a Multiple Instance Learning (MIL) problem relying on different aggregation functions to combine information from large regions. In particular, Hou *et al*. proposed a patch-CNN with a two-level strategy to classify gigapixel glioma WSIs. The first level estimates the discriminative power of each patch by a CNN trained in an Expectation-Maximization manner. The second level aggregates the patch level features as a descriptor of the WSI and classifies the label of the WSI with conventional SVMs. Ilse *et al*. [11] extended [9] by proposing a deep attention based MIL framework. In this framework, patch level features extracted by a CNN are aggregated by an attention mechanism [5]. The whole framework is trained end-to-end with only image level labels. Takahama *et al*. [27] optimized the training procedures to process all patches in a WSI simultaneously with limited GPU memory capacity.

These patch exhaustive methods waste a lot of computational resources on not very informative and redundant patches, which is inefficient and possibly less accurate in both training and testing. Thus, patch selective methods have attempted to overcome this drawback. In [34, 32, 7] a small set of patches is randomly sampled, assuming a high probability that at least one patch in the sampled set is informative. Particularly, patches from multiple magnifications are included in the random sampled set in [7]. A heuristic strategy [6] samples patches based on nuclear density. Inspired by the attention based MIL [11], Katharopoulos *et al*. [13] proposed an attention sampling strategy: they computed an attention map on the low resolution image to localize the potential informative regions. Assuming the attention map represents a probability distribution, the authors sampled patches corresponding to high probability and cropped from the full resolution image. These patches are used to predict the image level labels. Compared to [11], this attention sampling method claimed a $25\times$ speed-up and consumed $30\times$ less memory, with comparable accuracy. Similarly [19] assumed some image labels could be determined by low resolution features, with no need to resort to high resolution. This assumption is partially supported by Jin *et al*. [12]. In this study, the authors demonstrated that the Gleason 3 regions in a prostate cancer WSI could be better determined by relatively low resolution. Thus a "switch" determined whether the network can assert the image label based on low resolution only, or the attention sampling pipeline needs to kick in to estimate the image label from high resolution patches. Overall, even if spatial attention mechanisms are quite popular, currently, magnification based attention is still under-explored [24].

Inspired by both attention sampling [13] and the insight that different cancer subtypes have different optimal recognition magnification [12, 7, 28], we propose a multi resolution attention sampling framework which learns to sample patches from not only spatial locations, but also different magnifications. Compared to the other attention based methods in the literature, our formulation offers an end to end formulation tested on two different and challenging classification tasks of histopathology, proving fast and comparable to state of the art results.

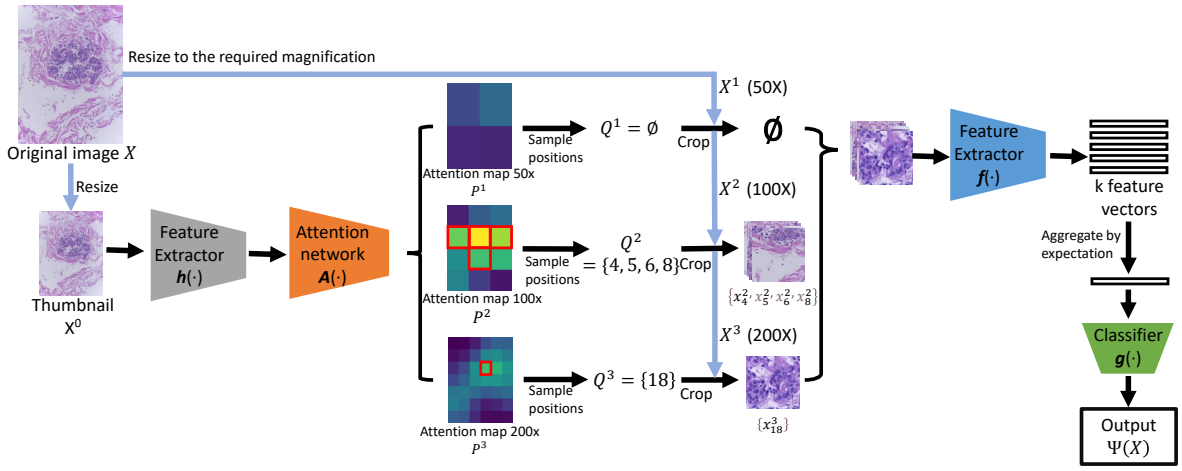

Figure 1. Overview of the proposed model. The first part of our architecture consists of the feature extractor $\boldsymbol{h}(\cdot)$ and the attention network $\boldsymbol{A}$ that highlights interesting spatial regions P_i^j on different magnification. For our setup, we used 3 different magnification levels (for visualization, we use $50X$, $100X$ and $200X$ in this figure). The second part of our architecture collects the selected patches and applies a second feature extractor $\boldsymbol{f}(\cdot)$ using an aggregation strategy for the final prediction $\Psi(X)$ of the classification task. The final prediction is performed by the classifier $\boldsymbol{g}(\cdot)$.

3. Method

3.1. Problem definition

Due to their large size, pathology images are usually divided into small patches to processes. Different from the models that need patch level annotations, which is rather expensive, we only use image level labels. Therefore, we formulate this classification problem as a MIL problem. We denote the original image as X. We aim to build a model $\Psi(\cdot)$ to predict the label of X: $Y = \Psi(X)$. $\Psi(\cdot)$ processes X in patches at different magnifications. Particularly, given m different magnifications, X is resized to X^j, $j = 1, 2, \ldots, m$. Each X^j contains a set of fixed size patches $\{x_i^j \mid i = 1, 2, \ldots, p_j, j \in [1, m]\}$, where p_j is the total number of patches in X^j at the magnification j, and i is the spatial location index.

During training, the image level label Y is annotated while the patch level label y_i^j is unknown. Thus our model $\Psi(\cdot)$ needs to aggregate the features from the patches to make a prediction $\Psi(X)$ on the image level. In our framework, instead of traversing all i, j, we propose to selectively process a small subset of i, j based on the attention mechanism of Sec. 3.2 to increase efficiency while maintaining a high classification accuracy.

3.2. Attention mechanism

In general, attention mechanisms aim to discover informative regions on the input image that could provide rich information for different tasks. Even if attention mechanisms could reveal informative regions, they still need to be trained using the entire input which could be memory and time demanding [11]. In our formulation, we address this issue by proposing a sampling strategy combining information from different magnifications.

Our overall framework is presented in Figure 1. Our framework takes as input a large size pathology image X. The original image X is then down-sampled to a thumbnail X^0 to reduce the memory and the computational complexity of our framework. Given a thumbnail X^0, we apply a feature extractor $\boldsymbol{h}(\cdot)$ and an attention network $\boldsymbol{A}$ to generate m attention maps $P^1, P^2, \ldots, P^m$ corresponding to the original image X at m different magnifications. These attention maps represent the probability distribution of informative patches at different spatial locations and magnifications. That is, P_i^j represents the probability that the patch at location i and magnification j is informative. Therefore, $\sum_{i,j} P_i^j = 1$.

Based on the probability distribution $P^j, j = 1, .., m$, we sample k patches at different i_s locations and j_s magnifications. We denote the set containing all (i_s, j_s), $s = 1, 2, \ldots, k$ as Q. The sampled patches are then inputted into a feature extractor $\boldsymbol{f}(\cdot)$ to get k feature vectors. These feature vectors are aggregated by expectation: $\sum_{i,j \in Q} P_i^j \boldsymbol{f}(x_i^j)$ based on the corresponding attention values $P_i^j, (i, j) \in Q$.

It has been proven that such sampling and aggregation by expectation is an approximation of using an attention mechanism on all patches [13]. The prediction on X using all

patches can be approximated by the following formulation,

$$\Psi(X) = g(\sum_{i,j} P_i^j \boldsymbol{f}(x_i^j)) \quad (1)$$

$$\approx g(\sum_{i,j \in Q} P_i^j \boldsymbol{f}(x_i^j)), \quad (2)$$

where $\boldsymbol{g}(\cdot)$ represents a neural network classifier and we use cross entropy as the loss function.

Eq. 2 is an unbiased approximation and the gradient regarding the network parameters can be approximated following the same strategy in [13]. By using such an attention expectation method, our whole framework is differentiable and can be trained end-to-end.

3.3. Attention Regularization

Our attention module estimates the most informative regions in the input image. However, such an attention estimation strategy encounters similar problems as other saliency detection models [33, 10, 25]. As pointed out in [11], the attention maps might be too sparse, and the learned attention distribution may focus on only one or two patches. This problem prevents the model from exploring patches at other positions and causes over-fitting, resulting in poor performance.

To tackle this problem, we applied two regularizations to the attention map. The first one is attention dropout, which is similar to [4]. It randomly resets a portion of the attention map to 0 before sampling in the training phase. Such a technique helps the network to explore more regions. Our second regularization is an entropy regularizer [13]:

$$\mathcal{R} = -\mathcal{H}(P_i^j) \quad (3)$$

$$= \sum_{i,j} P^j log(P_i^j) \quad (4)$$

where $\mathcal{H}$ is the entropy function on the attention distribution P^j. The entropy $\mathcal{R}$ encourages the attention maps to follow a more uniform distribution and penalizes the model if the attention maps focus on few patches.

3.4. Optimization

The final loss $\mathcal{L}$ is the sum of cross entropy and attention entropy regularizer $\mathcal{R}$:

$$\mathcal{L} = \mathcal{L}'(\Psi(X), Y) + \alpha\mathcal{R}, \quad (5)$$

where $\mathcal{L}'$ denotes the cross entropy loss between our prediction $\Psi(X)$ and the ground truth image label Y. α is a hyper parameter to adjust the regularization strength.

4. Experiments

In this section we evaluate our method on two different histopathology datasets.

4.1. Datasets

BACH Dataset. The ICIAR 2018 BreAst Cancer Histology (BACH) dataset [1] contains 400 images for training and 100 images for testing. The images were labeled as: normal, benign, in situ carcinoma, or invasive carcinoma based on the predominant cancer type in each image. All the classes have the same number of images. The images have the same size of 2048×1536 pixels. In our experiments, 320 out of 400 samples were used for training and the rest 80 for validation. We used the online submission system for testing.

TCGA-PRAD Large Patch Dataset. This dataset is based on the publicly available TCGA-PRAD dataset [35]. TCGA-PRAD contains 449 WSIs of prostate cancer. In our experiments, we extracted 285 2000×2000 pixel regions at $100X$ magnification, from 54 WSIs. We call this subset the TCGA-PRAD Large Patch Dataset. Our dataset contains 3 classes: benign, Gleason 3, and Gleason 4/5, annotated by a pathologist. Each class has 95 images. We performed a WSI-wise split and used 219 images from 44 WSIs for training/validation, and 66 images from 10 WSIs for testing. The training, validation, and test sets are also balanced.

4.2. Implementation Details

For all our experiments, we used two Inception V3 networks [26] pretrained on ImageNet as the feature extractors ($\boldsymbol{h}(\cdot)$ and $\boldsymbol{f}(\cdot)$) for thumbnails and sampled patches.

On the BACH dataset, we focused on three different magnifications: $50X$, $100X$ and $200X$. We set the input size of the thumbnail to 299×224. $\boldsymbol{h}(\cdot)$ takes a thumbnail as input and returns a feature vector of $2048 \times 8 \times 6$. The feature vector is the input to the attention module $\boldsymbol{A}$ which consists of 3 adaptive average pooling layers followed by 3 1×1 convolutional layers to output attention maps at different spatial resolutions. A sigmoid function is appended at the end to convert the logits in 3 attention maps into probability distributions. $\boldsymbol{A}$ outputs 3 attention maps of size: 2×2, 4×3, and 7×6, corresponding to the 3 input image magnifications: $50X$, $100X$ and $200X$

On the TCGA-PRAD large patch dataset, we focused on the magnifications of $25X$, $50X$ and $100X$ [1]. The experiment configuration is almost the same as that for the BACH dataset, except for the size of the thumbnails and attention maps. We set the thumbnail size to 299×299 since the input images are square. Thus $\boldsymbol{h}(\cdot)$ outputs a feature vector of $2048 \times 8 \times 8$. The spatial resolutions of the 3 attention maps are 2×2, 4×4, and 7×7, corresponding to the 3 input image magnifications: $25X$, $50X$ and $100X$.

We use a fully connected layer with a dropout layer as the final classifier. We set the regularization factor $\alpha = 0.01$

[1] We did not use $200X$ since usually it is not used by pathologists for the grading of prostate cancer.

Method	Accuracy
Roy *et al*. [22]	**87**%
Wang *et al*. [29]	83%
Wang *et al*. [30]	79%
Proposed single magnification ($50X$)	80%
Proposed single magnification ($100X$)	**85**%
Proposed single magnification ($200X$)	82%
Proposed multiple magnifications	**85**%

Table 1. Comparison of the overall accuracy on the BACH test dataset. Different state of the art methods are presented together with our proposed formulation using our magnification selection approach and three different single resolutions.

Method	Accuracy	AUC-ROC	AP
$25X$	89%	0.94	0.90
$50X$	**92**%	0.96	0.93
$100X$	89%	0.91	0.87
Multiple mag.	**91**%	**0.98**	**0.96**

Table 2. Overall accuracy on the TCGA-PRAD large patch dataset using our magnification selection method and using three single resolutions. $25X$ denotes Proposed single magnification of $25x$ method and so does for $50x$ and $100x$. *Multiple mag.* denotes Proposed multiple magnifications method. AUC-ROC denotes the area under the ROC curve. AP denotes the average precision

which was determined on the validation set.

Moreover, for all our experiments we used the PyTorch library [21] and an Nvidia Quadro RTX 8000 GPU. For optimization, we used the Adam [14] optimizer with an initial learning rate of $1e-4$ and decreased the learning rate by a factor of 0.1 when the validation accuracy did not improve for 25 epochs. The models on both datasets were trained for 100 epochs.

4.3. Results

In this section, we first present the performance of our method in comparison with other methods. We also demonstrate that our method identifies the optimal magnification by providing attention maps with higher attended distribution. We qualitatively show our model focuses more on the informative regions by comparing the attention map with the pathologist's interpretation. At last, we present comparisons on time efficiency.

4.3.1 Evaluation of overall performance

We choose overall accuracy as the main metric for the evaluation of our method and we perform an extensive study comparing our method with state-of-the-art methods. To demonstrate the efficiency of our method we also perform

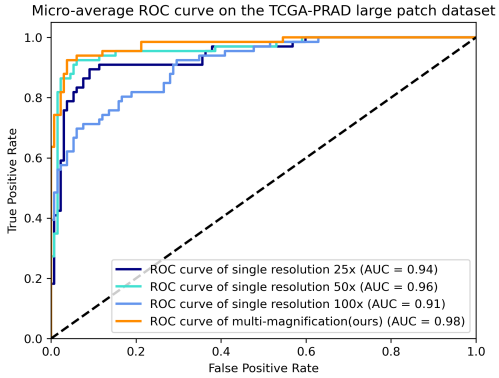

Figure 2. Receiver Operating Characteristic (ROC) curve on the test data of the TCGA-PRAD large patch dataset. Micro-average AUC is used for this multi-class classification problem. AUC stands for area under curve.

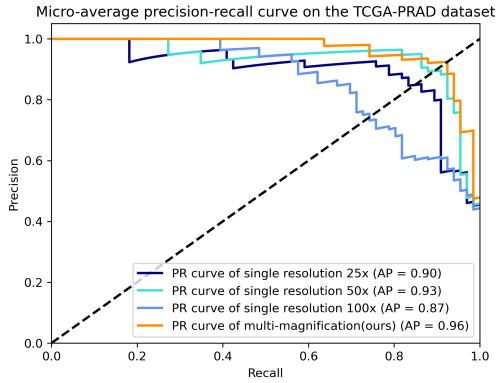

Figure 3. Precision-Recall (PR) curve on the test data of the TCGA-PRAD large patch dataset. Micro-average PR is used for this multi-class classification problem. AP stands for average precision.

experiments by training our formulation on a single magnification. On the BACH dataset, our method achieves an 85% accuracy on the test set (Table 1), which is comparable to other state of the art methods that did not use any external data nor ensemble models [22, 29, 30]. We need to highlight that the other methods extract information from the entire input image, while our method is able to use only a small subset (10%) of the initial input region. Moreover, our single magnification models reach lower or similar accuracy to our proposed method. That is, being completely agnostic on the appropriate magnification level, the conventional strategy needs to train 3 different models to achieve the same accuracy as our method. Such multiple model training could be very time and resource consuming.

The performance on the TCGA-PRAD large patch dataset is presented in Table 2. Our approach achieves 91% accuracy which is comparable to the performance of each

Magnification	Attention Percentage
$50X$	6%
$100X$	**78**%
$200X$	16%

Table 3. Attention percentage of the 3 selected magnifications on the BACH dataset.

Magnification	Attention Percentage
$25X$	1%
$50X$	**65**%
$100X$	33%

Table 4. Attention percentage of the 3 selected magnifications on the TCGA-PRAD large patch dataset.

independent magnification. Particularly, the model on the single $50X$ magnification yields a slightly better performance of 92% accuracy, which requires training of 3 different models as the optimal magnification is unknown before all 3 magnifications are evaluated. Also, Figure 2 and Figure 3 show the micro average Receiver Operating Characteristic (ROC) curve and Precision-Recall (PR) curve on this dataset. In particular, as shown in Table 2, our method achieves an area under curve of ROC (AUC-ROC) of 0.98 and an average precision of 0.96, which outperforms that of single magnification models. Such analysis on the BACH dataset is unavailable because the ground truth test labels are unknown.

4.3.2 Evaluation of the magnification based attention

We demonstrate our method is able to automatically focus more on the optimal magnification. On the BACH dataset, Table 1 shows that our model trained on a single magnification using $100X$ achieves a higher accuracy than $200X$ and $50X$. As summarised in Table 3, our magnification attention model also focuses mainly on the $100X$ magnification to make its final prediction. In particular, 78% of the overall attention distribution comes from the $100X$ magnification, while the other 2 magnifications contribute only 16% and 6% to the overall attention map. This means that our method selects $100X$ as the best magnification for this classification task, which is in accordance with our extensive single magnification experiments.

On the TCGA-PRAD large patch dataset, Table 4 highlights that our model is focusing on the $50X$ magnification. In particular, 65% of the overall attention distribution comes from the $50X$ magnification while 33% are from $100X$ and only 1% are from the $25X$. This is also comparable to our experiments summarised in Table 2 on which our single $50X$ magnification network reaches higher accuracy than that of $100X$ or $25X$.

4.3.3 Evaluation of time efficiency

One of the main advantages of our approach is the time efficiency compared to patch exhaustive methods. On the BACH dataset, Table 5 summarises the time efficiency comparing our proposed spatial and magnification attention method with the model trained on a single magnification. We show the time elapsed for each epoch in training as well as an average inference time for each image. For the training stage, our approach needs 25.2 seconds for one epoch, which is comparable to our formulation using a single magnification. Since the experiments on a single magnification need to process all 3 magnifications to find the best magnification, our approach is 2.5 times faster than running them one by one which requires a training time of 63.7 seconds per epoch. Our approach is 8.9 times faster in

Phase	Training		Inference	
#patches	5	all	5	all
$50X$	17.9s	17.9s	13.8ms	13.8ms
$100X$	22.9s	59.0s	18.4ms	50.3ms
$200X$	22.9s	148.1s	24.4ms	125.6ms
Combined	63.7s	225.0s	\	\
Ours	25.2s	\	30.6ms	316.1ms

Table 5. Average speed for training (per epoch) and average time (per sample) for inference on the BACH dataset. The reported times are in seconds for training and millisecond for evaluation. #patches correspond to two different schemes: using 5 sampled patches or the entire patches. Each cell denotes the running time of a single epoch under some conditions. Combined: running 50X, 100X and 200X one by one to determine the best resolution for this task.

Phase	Training		Inference	
#patches	5	all	5	all
$25X$	13.2s	13.2s	11.6ms	11.6ms
$50X$	15.1s	39.9s	14.4ms	39.6ms
$100X$	15.3s	115.0s	16.1ms	118.9ms
Combined	43.6s	168.1s	\	\
Ours	17.1s	\	23.2ms	288.0ms

Table 6. Average speed for training (per epoch) and inference phases on the TCGA-PRAD large patch dataset. The reported times are in seconds and correspond to two different schemes: using 5 sampled patches or the entire patches. Each cell denotes the running time of a single epoch under some conditions. #patches: the number of small patches we used to do prediction. Combined: running on 50X, 100X and 200X one by one to determine the best resolution for this task.

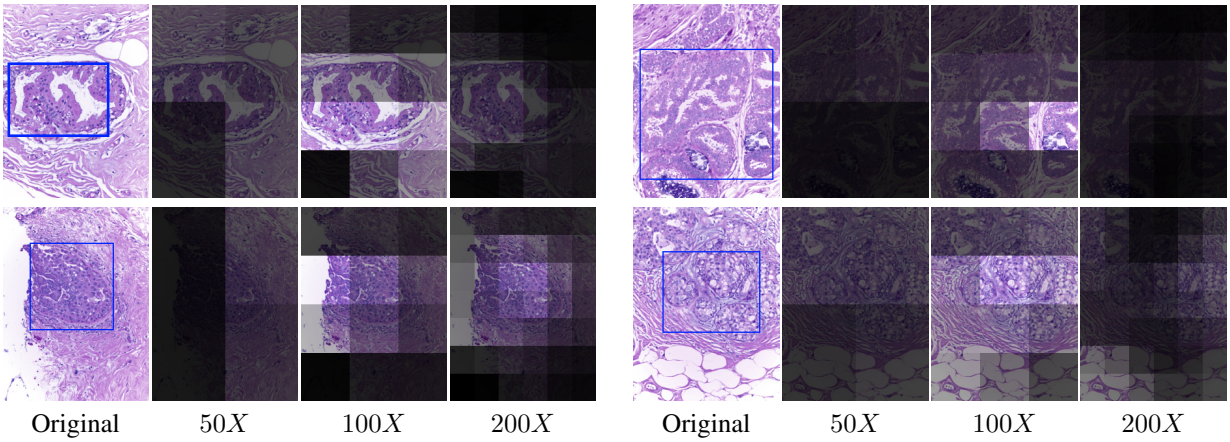

Figure 4. Visualizations of attention distribution over 3 magnifications on BACH dataset. Blue boxes in the original images denote the regions the annotating pathologist thinks are most informative

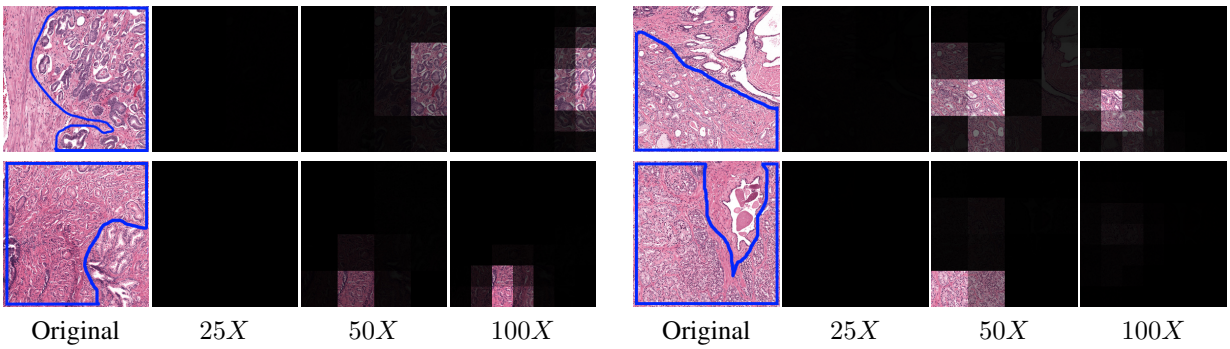

Figure 5. Visualizations of attention distribution over 3 magnifications on the TCGA-PRAD large patch dataset. Blue boxes in the original images denote the tumor regions annotated by an expert pathologist

training than traversing all magnifications and all patches to find the best one, which takes 225.0 seconds. We think that training time is especially important when only limited resources are available. In this scenario, reduced training time means that more tuning, data, and variations of networks can be tried, thus it leads to better performance of models. For inference, compared with the state of art method [22], which uses 12 patches of size 512×512 at $20X$ (similar to our 299×299 patch size at $10X$), our method runs at least 1.6 times faster. Compared with our spatial and magnification attention model using all patches, our approach is 10.3 times faster.

Table 6 shows the time efficiency on the TCGA-PRAD large patch dataset. Our approach runs at 17.1 seconds per epoch, which is 2.5 times faster than running our formulation on 3 resolutions one by one (43.6 seconds per epoch) in the training phase. Our approach is 10 times faster than the patch and magnification traversing method, which requires 172.7 seconds for each epoch. For inference, compared with our spatial and magnification attention model using all patches, our approach is 12.4 times faster than using all the patches.

4.3.4 Evaluation of attention maps

Another advantage of our approach is that the computed attention maps add interpretability to our model. Generally, our model focuses more on informative regions for most samples in the two datasets: Figure 4 shows 4 samples and the model's attention maps on magnification 50X, 100X, 200X on the BACH dataset. The regions that our model focuses on are consistent with the informative regions identified by the pathologist in the blue boxes.

Figure 5 shows the attention maps on the TCGA-PRAD large patch dataset. We show 4 samples and our model's attention maps on magnification 25X, 50X, 100X. Our model tends to focus more on the tumor regions that were identified by the pathologist within the blue lines.

5. Conclusion

In this paper, we presented a novel, end-to-end training strategy for joint spatial and magnification based attention.

Our method has been tested on two completely different classification tasks of histopathology problems proving its superiority and efficiency. In particular, using only 10% of the initial input our method is comparable or superior to other state of the art methods while it discovers completely automatically the optimal magnification and spatial locations for the specific task. Moreover, our method reports lower training and inference times, reducing the time complexity significantly, which is a very important consideration when it comes to medical applications and in particular application on pathology and large scale microscopy images. One limitation of our current formulation is the fixed number of patches and predefined magnifications that are required on the attention module. In the future, we aim to investigate ways that we can automatically select the magnification levels and the number of patches. Moreover, we aim to investigate additional attention formulations providing explainable tools on a variety of clinically relevant medical tasks.

Acknowledgements. This work was supported by NCI award UH3CA225021, ARC: Grant SIGNIT201801286, 4DVision project from the Partner University Fund and generous donor support from Bob Beals and Betsy Barton.

References

[1] Guilherme Aresta, Teresa Araújo, Scotty Kwok, Sai Saketh Chennamsetty, Mohammed Safwan, Varghese Alex, Bahram Marami, Marcel Prastawa, Monica Chan, Michael Donovan, et al. Bach: Grand challenge on breast cancer histology images. *Medical image analysis*, 56:122–139, 2019. 4

[2] Peter Bandi, Oscar Geessink, Quirine Manson, Marcory Van Dijk, Maschenka Balkenhol, Meyke Hermsen, Babak Ehteshami Bejnordi, Byungjae Lee, Kyunghyun Paeng, Aoxiao Zhong, et al. From detection of individual metastases to classification of lymph node status at the patient level: The camelyon17 challenge. *IEEE transactions on medical imaging*, 38(2):550–560, 2018. 1

[3] Mingyu Chen, Bin Zhang, Win Topatana, Jiasheng Cao, Hepan Zhu, Sarun Juengpanich, Qijiang Mao, Hong Yu, and Xiujun Cai. Classification and mutation prediction based on histopathology h&e images in liver cancer using deep learning. *NPJ precision oncology*, 4(1):1–7, 2020. 1

[4] Junsuk Choe and Hyunjung Shim. Attention-based dropout layer for weakly supervised object localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2219–2228, 2019. 4

[5] Yann N Dauphin, Angela Fan, Michael Auli, and David Grangier. Language modeling with gated convolutional networks. In *Proceedings of the International Conference on Machine Learning*. PMLR, 2017. 2

[6] Aditya Golatkar, Deepak Anand, and Amit Sethi. Classification of breast cancer histology using deep learning. In *International Conference Image Analysis and Recognition*. Springer, 2018. 2

[7] Noriaki Hashimoto, Daisuke Fukushima, Ryoichi Koga, Yusuke Takagi, Kaho Ko, Kei Kohno, Masato Nakaguro, Shigeo Nakamura, Hidekata Hontani, and Ichiro Takeuchi. Multi-scale domain-adversarial multiple-instance cnn for cancer subtype classification with unannotated histopathological images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3852–3861, 2020. 2

[8] Le Hou, Ayush Agarwal, Dimitris Samaras, Tahsin M Kurc, Rajarsi R Gupta, and Joel H Saltz. Robust histopathology image analysis: To label or to synthesize? In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019. 1

[9] Le Hou, Kunal Singh, Dimitris Samaras, Tahsin M Kurc, Yi Gao, Roberta J Seidman, and Joel H Saltz. Automatic histopathology image analysis with CNNs. In *2016 New York Scientific Data Summit (NYSDS)*, pages 1–6. IEEE, 2016. 2

[10] Qibin Hou, Peng-Tao Jiang, Yunchao Wei, and Ming-Ming Cheng. Self-erasing network for integral object attention. *arXiv preprint arXiv:1810.09821*, 2018. 4

[11] Maximilian Ilse, Jakub Tomczak, and Max Welling. Attention-based deep multiple instance learning. In *Proceedings of the International Conference on Machine Learning*. PMLR, 2018. 1, 2, 3, 4

[12] Chen Jin, Ryutaro Tanno, Moucheng Xu, Thomy Mertzanidou, and Daniel C Alexander. Foveation for segmentation of mega-pixel histology images. In *Proceedings of the International Conference on Medical Image Computing and Computer Assisted Intervention*. Springer, 2020. 1, 2

[13] Angelos Katharopoulos and François Fleuret. Processing megapixel images with deep attention-sampling models. In *International Conference on Machine Learning*, pages 3282–3291. PMLR, 2019. 2, 3, 4

[14] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 5

[15] Bruno Korbar, Andrea M Olofson, Allen P Miraflor, Catherine M Nicka, Matthew A Suriawinata, Lorenzo Torresani, Arief A Suriawinata, and Saeed Hassanpour. Deep learning for classification of colorectal polyps on whole-slide images. *Journal of pathology informatics*, 8, 2017. 2

[16] Han Le, Rajarsi Gupta, Le Hou, Shahira Abousamra, Danielle Fassler, Luke Torre-Healy, Richard A Moffitt, Tahsin Kurc, Dimitris Samaras, Rebecca Batiste, et al. Utilizing automated breast cancer detection to identify spatial distributions of tumor infiltrating lymphocytes in invasive breast cancer. *The American Journal of Pathology*, 2020. 2

[17] Marvin Lerousseau, Maria Vakalopoulou, Marion Classe, Julien Adam, Enzo Battistella, Alexandre Carré, Théo Estienne, Théophraste Henry, Eric Deutsch, and Nikos Paragios. Weakly supervised multiple instance learning histopathological tumor segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 470–479. Springer, 2020. 1

[18] Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen

Ghafoorian, Jeroen Awm Van Der Laak, Bram Van Ginneken, and Clara I Sánchez. A survey on deep learning in medical image analysis. *Medical image analysis*, 42:60–88, 2017. 1

[19] Sam Maksoud, Kun Zhao, Peter Hobson, Anthony Jennings, and Brian C Lovell. SOS: Selective objective switch for rapid immunofluorescence whole slide image classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2020. 2

[20] Pooya Mobadersany, Safoora Yousefi, Mohamed Amgad, David A Gutman, Jill S Barnholtz-Sloan, José E Velázquez Vega, Daniel J Brat, and Lee AD Cooper. Predicting cancer outcomes from histology and genomics using convolutional networks. *Proceedings of the National Academy of Sciences*, 115(13):E2970–E2979, 2018. 1

[21] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017. 5

[22] Kaushiki Roy, Debapriya Banik, Debotosh Bhattacharjee, and Mita Nasipuri. Patch-based system for classification of breast histology images using deep learning. *Computerized Medical Imaging and Graphics*, 71:90–103, 2019. 5, 7

[23] Mihir Sahasrabudhe, Stergios Christodoulidis, Roberto Salgado, Stefan Michiels, Sherene Loi, Fabrice André, Nikos Paragios, and Maria Vakalopoulou. Self-supervised nuclei segmentation in histopathological images using attention. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 393–402. Springer, 2020. 1

[24] Shallu Sharma and Rajesh Mehra. Conventional machine learning and deep learning approach for multi-classification of breast cancer histopathology images—a comparative insight. *Journal of digital imaging*, 33(3):632–654, 2020. 2

[25] Krishna Kumar Singh and Yong Jae Lee. Hide-and-seek: Forcing a network to be meticulous for weakly-supervised object and action localization. In *2017 IEEE international conference on computer vision (ICCV)*, pages 3544–3553. IEEE, 2017. 4

[26] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016. 4

[27] Shusuke Takahama, Yusuke Kurose, Yusuke Mukuta, Hiroyuki Abe, Masashi Fukayama, Akihiko Yoshizawa, Masanobu Kitagawa, and Tatsuya Harada. Multi-stage pathological image classification using semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10702–10711, 2019. 2

[28] Hiroki Tokunaga, Yuki Teramoto, Akihiko Yoshizawa, and Ryoma Bise. Adaptive weighting multi-field-of-view cnn for semantic segmentation in pathology. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12597–12606, 2019. 2

[29] Yaqi Wang, Lingling Sun, Kaiqiang Ma, and Jiannan Fang. Breast cancer microscope image classification based on cnn with image deformation. In *International Conference Image Analysis and Recognition*, pages 845–852. Springer, 2018. 5

[30] Zeya Wang, Nanqing Dong, Wei Dai, Sean D Rosario, and Eric P Xing. Classification of breast cancer histopathological images using convolutional neural networks with hierarchical loss and global pooling. In *International Conference Image Analysis and Recognition*, pages 745–753. Springer, 2018. 5

[31] Ellery Wulczyn, David F Steiner, Zhaoyang Xu, Apaar Sadhwani, Hongwu Wang, Isabelle Flament-Auvigne, Craig H Mermel, Po-Hsuan Cameron Chen, Yun Liu, and Martin C Stumpe. Deep learning-based survival prediction for multiple cancer types using histopathology images. *PLoS One*, 15(6):e0233678, 2020. 1

[32] Ellery Wulczyn, David F Steiner, Zhaoyang Xu, Apaar Sadhwani, Hongwu Wang, Isabelle Flament-Auvigne, Craig H Mermel, Po-Hsuan Cameron Chen, Yun Liu, and Martin C Stumpe. Deep learning-based survival prediction for multiple cancer types using histopathology images. *PLoS One*, 15(6):e0233678, 2020. 1, 2

[33] Xiaolin Zhang, Yunchao Wei, Jiashi Feng, Yi Yang, and Thomas S Huang. Adversarial complementary learning for weakly supervised object localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1325–1334, 2018. 4

[34] Xinliang Zhu, Jiawen Yao, Feiyun Zhu, and Junzhou Huang. Wsisa: Making survival prediction from whole slide histopathological images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 1, 2

[35] M Zuley, R Jarosz, B Drake, D Rancilio, A Klim, K Rieger-Christ, and J Lemmerman. Radiology data from the cancer genome atlas prostate adenocarcinoma [tcga-prad] collection. *The Cancer Imaging Archive. Available online: http://doi. org/10.7937 K*, 9, 2016. 4