A Novel Proof-of-concept Framework for the Exploitation of ConvNets on Whole Slide Images

A novel proof-of-concept framework for the exploitation of ConvNets on Whole Slide Images

A. Mascolini, S. Puzzo, G. Incatasciato, F. Ponzio*, E. Ficarra, S. Di Cataldo

Politecnico di Torino, Corso Duca degli Abruzzi 24, 10129 Torino, Italy
*Corresponding author francesco.ponzio@polito.it

Abstract. Traditionally, the analysis of histological samples is visually performed by a pathologist, who inspects under the microscope the tissue samples, looking for malignancies and anomalies. This visual assessment is both time consuming and highly unreliable due to the subjectivity of the evaluation. Hence, there are growing efforts towards the automatisation of such analysis, oriented to the development of computer-aided diagnostic tools, with a ever-growing role of techniques based on deep learning. In this work, we analyze some of the issues commonly associated with providing deep learning based techniques to medical professionals. We thus introduce a tool, aimed at both researchers and medical professionals, which simplifies and accelerates the training and exploitation of such models. The outcome of the tool is an attention map representing cancer probability distribution on top of the Whole Slide Image, driving the pathologist through a faster and more accurate diagnostic procedure.

Keywords: Digital Pathology, Whole Slide Imaging, Colorectal Cancer, Convolutional Neural Networks, Deep Learning, Attention Map

1 Introduction

The field of pathology often relies on the analysis of microscopic images to perform a diagnosis. Whole Slide Imaging represents a technology, through which glass slides are digitalized in the form of minimally compressed images, featuring a pyramid structure with various levels of magnification. This process enables microscopic images of tissues to be analyzed by advanced digital tools [1].

The Whole Slides Images (WSIs) have been used for a wide variety of both educational and clinical purposes, and several authors have reported good diagnostic concordance between the analysis of WSIs and glass slides [2]. It can be thus reasonably assumed that digital image classification, as well as deep learning techniques, can play a key role in the delicate and time consuming process of diagnosis. On one hand, they can serve as a way to double check and mitigate the extreme inter-operators variability. On the other hand, they can significantly reduce the evaluation time spent by the clinicians, by providing accurate and automatic information in a reasonably short time.

While literature features an extensive amount of papers in which machine and deep learning techniques are successfully applied to the field of Computer Aided Diagnosis (CAD) [3], [4], [5], there are very few instances of these techniques being used in a daily medical practice. This is mainly due to:

- the very long processing times and specialized hardware required to implement most of the newly developed architectures. For instance, ScanNet, a framework to analyze WSIs in a fully connected fashion, takes 15 minutes on a Titan X GPU to analyze a single Whole Slide Image [6], which is totally unfeasible in a everyday clinical scenario.
- The lack of easy-to-use interfaces for non-technical users, such as the medical staff.

With the aim of facing the above-mentioned challenges, we have thus developed a standalone, cross-platform, CPU based tool, featuring an easy-to-use Graphical User Interface to facilitate the user experience. Our tool features a staining normalization phase and an asynchronous sample pre-fetching to optimize computational time, and applies a dynamic resolution approach. Beside being a CAD oriented tool, our apparatus also allows researchers to train and prototype new WSIs segmentation models, without having to worry about tedious and extensive pre-processing, being the model easily embeddable in the framework.

The tool has been evaluated on a highly challenging dataset consisting of histological images of a specific type of tumor known as Colorectal Carcinoma (CRC). Nowadays, CRC is the third most frequent cancer that afflicts mankind with 1.8 million new cases in 2018 [7]. CRC is a type of epithelial cancer, coming from the colon or the rectum, which provokes the uncontrolled proliferation of mucosal cells covering the last part of the intestine. The initial diagnosis of CRC is performed by means of colonoscopy, i.e. the endoscopic inspection of the large and the distal part of the small intestine, during which the surgeon may perform a biopsy on the suspicious colorectal lesions. This surgical step is then usually followed by a diagnostic procedure carried out by the pathologist to determine the nature of lesions, studying the tissue sample under the microscope or through an analysis of the corresponding WSI. The importance of the early diagnosis of the tumour, crucial for the survival of a large number of patients, makes the CRC an interesting case study to test the feasibility of our method.

The rest of the paper is organized as follows. In section 2 we describe our dataset and then introduce the design characteristics of our proposed approach. In section 3, we report our experimental results and we discuss our findings. Finally, section 4 concludes the paper.

2 Materials and Methods

2.1 Dataset

Our case study dataset was extracted from a public repository of H&E stained whole-slide images (WSIs) of colorectal tissues, available on line at `http://www.virtualpathology.leeds.ac.uk/`. In order to obtain a

statistically significant dataset, in terms of inter-subjects and inter-class variability, 18 WSIs were selected among univocal subjects (i.e. one WSI per patient) and then split into regions containing either exclusively cancer or exclusively healthy tissue via a sliding windows approach, as shown in figure 1.

The cropped patches were separated into a training and a testing set with a 75%-25% split, ensuring that regions coming from a single patient always belong to the same set. These sets were fed to the network first to train it and then to evaluate the patch-wise predictions. A second independent cohort of 11 patients, never fed to the network model during training, were randomly selected to serve as the validation set for performance evaluation in terms of WSIs attention maps, which is ultimately a tissue segmentation task.

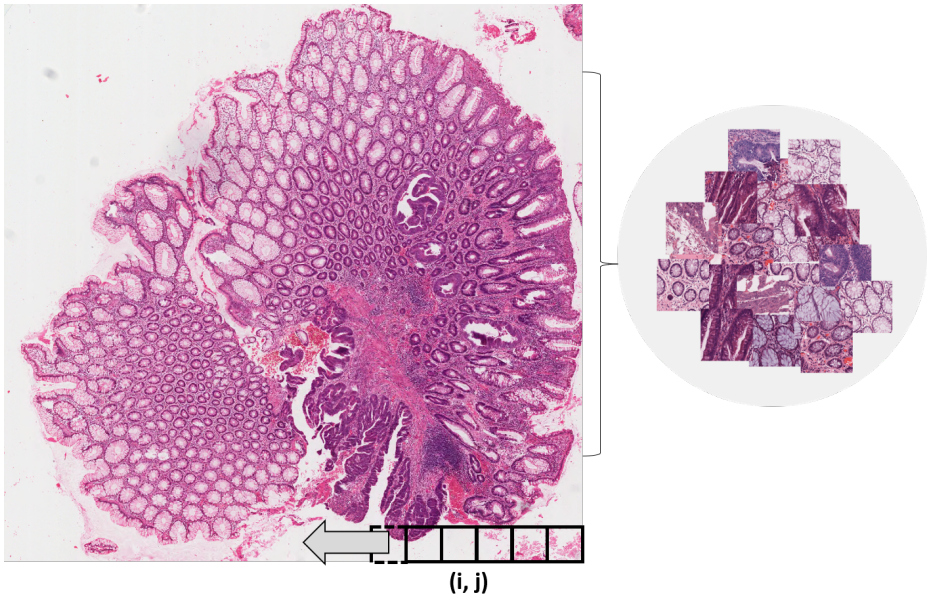

Fig. 1. A WSI example depicting the sliding window cropping technique.

2.2 Slide Analysis

As initial step, aimed at reducing the large inter-patients variability in terms of slide color (see figure 2), we perform staining normalization on the input H&E stained WSIs, using the well-known stain vector variation and correction method [8], providing the network with consistent colors. The slide is cropped using a sliding window, and references to every crop are stored in memory, while the crop itself is not loaded until necessary.

At runtime, two heuristics identify and remove crops which are either white (due to saturation of colors or absence of tissues) or contaminated by blood.

- To identify blood contaminated crops, the following value is calculated
$$\delta = 2 \cdot \frac{\overline{Red}}{\overline{Green} + \overline{Blue}}$$
where $\overline{Channel}$ represents the mean value of the channel in the image. If $\delta > 1.5$ the slide is discarded due to too much blood.

- To identify white crops, the mean luminance and standard deviation are checked to be respectively higher and lower than two fixed thresholds.

Every crop is taken from the slide at the magnification level which yields images closest in size to 200x200px and then resized to exactly 200x200px. This resolution is chosen to minimize both the amount of neurons and the amount of data to read from disk, significantly reducing the time it takes to train and generate predictions with the net, while providing a high enough resolution for the classifier to obtain good results.
When two neighbouring crops are assigned to different classes, they are recognized as a border. Additional windows are then added between border crops to increase the resolution of the classification where needed, in an iterative process which can be repeated as many times as the user requires. The crop loading process is asynchronous, allowing the CPU to process the previous sample while the next is fetched from the disk.

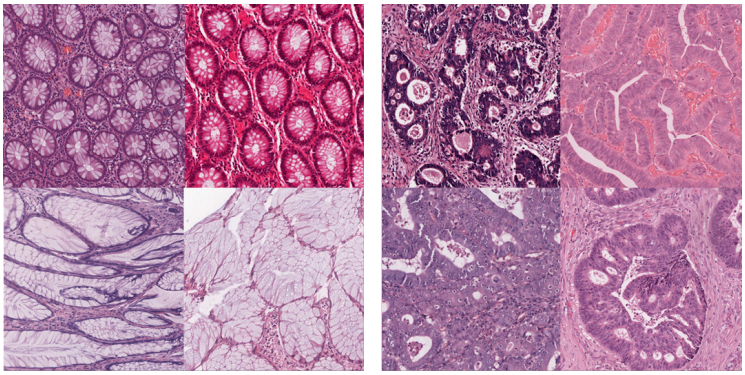

Fig. 2. Histological images of colorectal tissues (cropped patches) presenting a very different staining effect. (a) Healthy tissue; (b) Cancer.

The tool stores the location and classification of every crop. When an area of the slide needs to be shown with its corresponding attention map(s),

the tool averages the votes received by every pixel in the required area of the slide to create a series of attention maps, one for each class the network is capable of identifying. In other words, every crop containing that pixel votes for a specific class and the results are averaged, as show in figure 3.

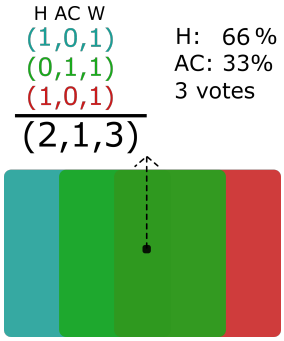

Fig. 3. Simplified example of probability image generation

Our tool offers a specific module to manage and crop WSIs, automatically generating patches, removing invalid areas, performing H&E stain normalization and producing data which is ready to be fed into a neural model. The implementation relies on CropList objects which can be joined and split at will to freely create datasets. The crops generated by our tool are asyncronously loaded in memory at runtime, allowing the employment of large datasets while making efficient use of the system resources.

2.3 Neural Network Architecture

Our tool provides an architecture-agnostic way to perform segmentation over a WSI; in order to test it, we created and trained a simple supervised neural network for patch classification of CRC, featuring AlexNet style CNN based features extraction (see figure 4). The architecture is made up of a base unit of 2 convolutional layers, with kernel size 3x3, stride 1 and no padding, which start from a 200x200 input size and progressively shrink, followed by batch normalization, ReLU activation and max pooling.
This unit is repeated 3 times, followed by a fully connected layer of 1000 neurons and an output layer of 2 neurons with softmax activation. Between the two fully connected layers, there is a dropout layer which randomly drops 40% of its inputs.
The network uses categorical Cross-Entropy as its loss function and is optimized using Adam optimizer with alpha=0.001, as suggested by the original paper[9]. Our accuracy showed little variability between epochs, we thus believe the learning rate does not need adjusting.

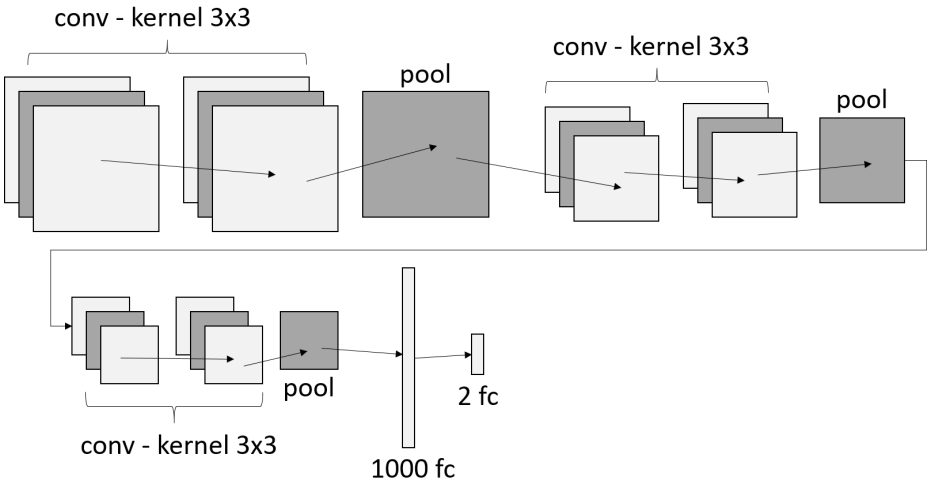

Fig. 4. Overview of the CNN architecture.

2.4 Graphical User Interface

The python back-end communicates using the Eel library with the javascript code in the front-end, providing a seamless user experience featuring (see figure 5):

- a file browser;
- the possibility of choosing the stride of the sliding window during the first step of the classification problem;
- a choice of which classes to show;
- an estimate of the time necessary to analyze the entire WSI;
- an area in which to annotate information about the slide.

3 Classification Accuracy

3.1 Performance metrics

To evaluate the quality of the predictions yielded by the architecture coupled with our tool, we had a group of professional pathologists annotating various WSIs containing cancer which the network had never seen before. These will be referred as the validation set.

3.2 Results and Discussion

The Dice coefficient is a metric used to evaluate the overlap between two discrete sets (X and Y), and is calculated as:

$$DSC = \frac{2|X \cap Y|}{|X| + |Y|}$$

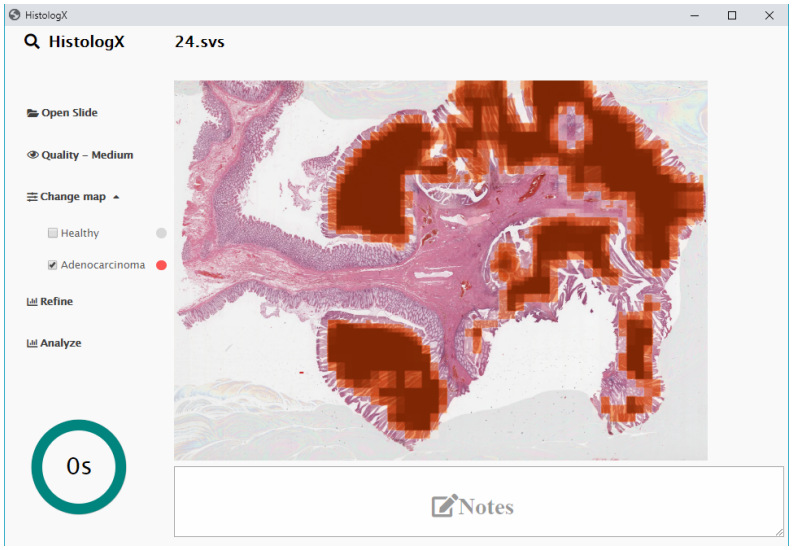

Fig. 5. Example of WSI analysis using the presented tool.

It is commonly employed to evaluate the similarity between segmentation masks [10]. The results obtained by our approach in terms of Dice coefficient are shown in table 1, alongside with the true positive (i.e. Sensitivity) and true negative rate (Specificity), as well as a the pixel-wise accuracy (Ac), defined as follows:

$$Ac = \frac{N_c}{N}$$

where N_c is the number of pixels which were correctly classified and N is the total number of pixels in an image. The two rows of the table respectively show mean and standard deviations of the figures of merit in the validation set.

Since our classifier is binary, we had to choose a threshold to indicate a pixel as belonging to the positive class. We did so by using the Receiver Operating Characteristics curve (see figure 6), which yield to a threshold of 27/255.

As reported in table 1, we obtained a mean Dice coefficient of 0.80 and an average per-pixel accuracy of 87%, with no bias towards any of the two classes. It must be noticed that a loss of performance in terms of Dice coefficient is possibly due to the tendency of the human pathologist to over-segment tumor-containing areas compared to our automatic algorithm, as it can be gathered from figure 7.

To fully assess our tool, we compared it to the most similar framework we found in literature to evaluate WSIs, which is described in [6]. To compare the segmentation performance, we used AUC (Area under ROC curve) on the entire validation set as the figure of merit, while the processing time was compared on a sample $200 \times 100 \times 10^3$ WSI, as reported

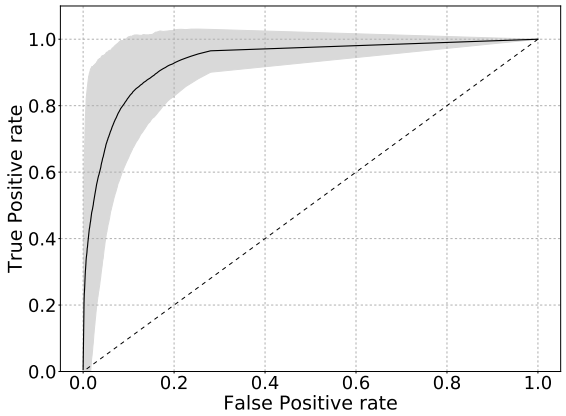

Fig. 6. Mean ROC curve over all validation subjects (n = 11). The grey band represents 95% confidence intervals.

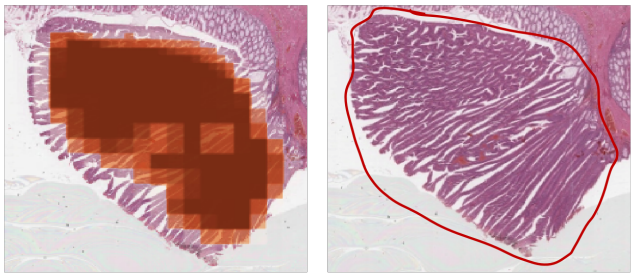

Fig. 7. Example showing the tendency of the pathologist to over-segment the lesion (right) with respect to out tool (left).

Table 1. Performance of our test architecture on the validation set.

	Dice	**Ac**	**Sensitivity**	**Specificity**
Mean	0.80	0.87	0.87	0.86
Std	0.07	0.06	0.10	0.08

in table 2. We tested our framework on the same hardware as the one used in [6], i.e. the Nvidia Titan X GPU, and on a widely available hardware consisting in a standard Intel i5 7200U CPU for personal computers. As it can be gathered from table 2, despite the slightly lower accuracy in terms of pixel-wise classification, our tool allows a quicker classification on cheaper, less powerful and widely available hardware.

Fig. 8. Patient 1 original WSI.

Fig. 9. Manual segmentation by pathologist.

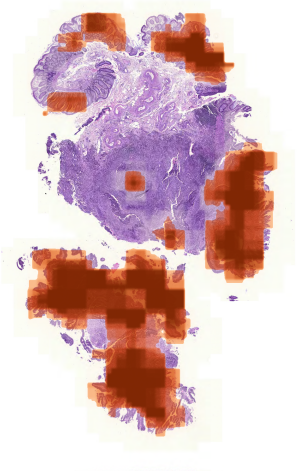

Fig. 10. Attention map generated by our tool.

The trade-off between accuracy and efficiency can be explained as follows. The proof-of-concept of our work was built on the idea of exploiting the classification potentials of some existing deep learning models, with the aim of achieving sufficiently accurate characterization of tissue regions with little or no extra effort in re-design of the model. The patch-based approach we used, which significantly differs from the dense per-pixel prediction implemented by the state-of-the-art tool, seems to be a promising choice, showing a good accuracy coupled with much lower processing time on a less powerful hardware.

Table 2. Comparison between our framework and the one described in [6]

	ScanNet	**Ours**	
Hardware	Nvidia Titan X	Nvidia Titan X	Intel i5 7200U
Type	Pixel-Based	Patch-Based	Patch-Based
AUC (mean ± std)	98.75 ± n.a.	93.44 ± 0.05	93.44 ± 0.05
Time (single WSI)	15 min	12 min	8 min
Asyncronous	Yes	Yes	Yes

While ScanNet features a highly parallelizable Fully Convolutional architecture which runs significantly faster on the GPU, our Patch Based approach proves to be I/O bound and as such more suitable to low hardware specifications, such as standard personal computers. The test slide is processed significantly faster on the CPU by saving on the overhead required to constantly move data between the system and the graphics

processor. Our architecture is thus best optimized for consumer devices such as laptops and tablets.

4 Conclusions and future work

In this work, we built and tested a novel framework which we believe could be useful in accelerating the development and adoption of deep learning techniques in the every-day digital pathology. We demonstrated our approach on WSI segmentation, showing that our easy-to-use framework can be run on cheap and widely available hardware with limited amount of processing time. Overall, the network we trained using our framework obtained results which agree with the manual segmentation performed by human pathologists.

References

[1] Mark Zarella et al. "A Practical Guide to Whole Slide Imaging: A White Paper From the Digital Pathology Association". In: *Archives of pathology & laboratory medicine* 143 (Oct. 2018). DOI: 10.5858/arpa.2018-0343-RA.

[2] Liron Pantanowitz, Navid Farahani, and Anil Parwani. "Whole side imaging in pathology: Advantage, limitations, and emerging perspectives". In: *Pathology and Laboratory Medicine International* 2015 (June 2015). DOI: 10.2147/PLMI.S59826.

[3] Yan Xu et al. "Large scale tissue histopathology image classification, segmentation, and visualization via deep convolutional activation features". In: *BMC Bioinformatics*. 2017.

[4] Francesco Ponzio et al. "Dealing with Lack of Training Data for Convolutional Neural Networks: The Case of Digital Pathology". In: *Electronics* 8 (Feb. 2019), p. 256. DOI: 10.3390/electronics8030256.

[5] Fuyong Xing et al. "Deep learning in microscopy image analysis: A survey". In: *IEEE transactions on neural networks and learning systems* 99 (2017), pp. 1–19.

[6] Huangjing Lin et al. *ScanNet: A Fast and Dense Scanning Framework for Metastatic Breast Cancer Detection from Whole-Slide Images*. 2017. eprint: arXiv:1707.09597.

[7] Freddie Bray et al. "Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries". In: *CA: A Cancer Journal for Clinicians* 68.6 (2018), pp. 394–424. DOI: 10.3322/caac.21492. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.3322/caac.21492. URL: https://onlinelibrary.wiley.com/doi/abs/10.3322/caac.21492.

[8] Marc Macenko et al. “A Method for Normalizing Histology Slides for Quantitative Analysis.” In: vol. 9. June 2009, pp. 1107–1110. DOI: 10.1109/ISBI.2009.5193250.

[9] Diederik P. Kingma and Jimmy Ba. “Adam: A Method for Stochastic Optimization”. In: *arXiv e-prints*, arXiv:1412.6980 (Dec. 2014), arXiv:1412.6980. arXiv: 1412.6980 [cs.LG].

[10] NJ Tustison and JC Gee. “Introducing Dice, Jaccard, and other label overlap measures to ITK”. In: *Insight J* 2 (2009).