DETERMINING THE SCALE OF IMAGE PATCHES USING A DEEP LEARNING APPROACH

Sebastian Otálora[1], *Oscar Perdomo*[2], *Manfredo Atzori*[1]
Mats Andersson[3], *Ludwig Jacobsson*[3], *Martin Hedlund*[3], *Henning Müller*[1]

[1] University of Applied Sciences Western Switzerland (HES-SO) TechnoPôle 3,
[2] MindLab Research Group, Universidad Nacional de Colombia,
[3] ContextVision AB, Linköping, Sweden.

ABSTRACT

Detecting the scale of histopathology images is important because it allows to exploit various sources of information to train deep learning (DL) models to recognise biological structures of interest. Large open access databases with images exist, such as The Cancer Genome Atlas (TCGA) and PubMed Central but very few models can use such datasets because of the variability of the data in color and scale and a lack of metadata. In this article, we present and compare two deep learning architectures, to detect the scale of histopathology image patches. The approach is evaluated on a patch dataset from whole slide images of the prostate, obtaining a Cohen's kappa coefficient of 0.9897 in the classification of patches with a scale of 5×, 10× and 20×. The good results represent a first step towards magnification detection in histopathology images that can help to solve the problem on more heterogeneous data sources.

Index Terms— Scale, Magnification, Deep Learning, DenseNet, Digital Pathology, Prostate.

1. INTRODUCTION

In recent years digital histopathology has become a major area of research and fully digital clinical workflows are now starting to become a reality. Larger storage capacities make image archives of whole slide images (WSI) feasible in fully digital format. This fact will likely create large archives that can be used for decision support in the future.

When analyzing images, pathologists usually look for sub-cell, cell and gland–scale tissue patterns (such as nuclei and gland deformations, for instance) to diagnose types of cancer and describe the structures included in the images in the pathology reports. These structural patterns are traditionally inspected through a light microscope and increasingly through digital biopsy slides (whole slide images, WSI) that are of a size of up to 100,000 × 100,000 pixels. One of the main patterns examined in breast cancer, for instance, are nuclei mitoses that are usually inspected at a 40× magnification. In prostate cancer, the ensemble of gland patterns is better observed at a 10× to 5× magnification.

Detecting the scale of histopathology images allows to better exploit large unannotated databases by adding scale as training parameter and by allowing to filter the relevant training images. For instance, given a particular biological structure (e.g. a gland ensemble), magnification detection allows to select the images of a scale that can potentially include it (e.g. 10×) and to have a scale-homogeneous training set. Leaving out the images where there is no relevant visual content available, can also lead to a better performance (due to less noisy data) in classification models and in similarity search for retrieval systems [1, 2].

In the last decade, deep learning (DL)-based techniques were successfully developed to classify, segment and localize biological structures for several types of cancer in digital biopsies, providing powerful techniques for computer aided decision support for digital pathology [1]. Convolutional neural networks (CNNs) are supervised DL models inspired by the human visual system. Janowczyk and Madabhushi [1] evaluate a single CNN architecture for classification, segmentation, and detection in a variety of scenarios and use cases, showing that a single architecture generalizes well across several histopathology tasks. The authors also discuss the importance of selecting the right scale (or magnification) for training CNN models. Gupta et al. [2] also evaluate the performance of hand–crafted features at different training-testing scales for breast cancer image classification. They observe that extreme magnifications usually harm the performance. In Bayramoglu et al. [3], the authors present a magnification-independent model for classification of breast cancer images, unifying in a unique loss function the magnification and the class learning parameters. The authors show that training with several scales in their model increases the performance of individual scale models. Given the availability of open access data repositories like the cancer genome atlas (TCGA[1]), the cancer imaging archive (TCIA[2]), digital teaching files and PubMed Central (PMC[3]), an open question is how to effectively use these datasets for leveraging knowledge from them

[1] https://cancergenome.nih.gov/
[2] http://www.cancerimagingarchive.net/
[3] https://www.ncbi.nlm.nih.gov/pmc/

and solving concrete medical inquires. In PMC, the number of papers and the amount of the indexed content including images that are made available has grown exponentially since the early 90's[4]. The scientific research community has been working on several tasks towards exploiting the content of the images included in these resources. For instance, the ImageCLEF benchmark made available manually annotated images in order to identify the image types [4] and a hierarchy of images types that is used for the classification and that includes histopathology images [5]. Identifying image types has been done many times [6, 7] but so far it has not allowed to leverage machine learning on a very large scale to our knowledge. Such a fully automatic workflow is not trivial due to the large variety of the image data in the biomedical literature but can potentially serve for medical imaging in the way that ImageNet served for general object recognition [8]. The main challenges for using histopathology images from open access data repositories to train deep learning models are the large variety of species (humans but also macaques, mice and rats), the variety of staining procedures and slide preparation methods and the unknown scale levels of the images. All these factors can vary strongly among digital pathology images [9, 10] and even more after figure editing, for example when writing a scientific publication or after the editing of an article by the publisher. Raw data of the WSIs are basically never available.

The objective of this paper is to tackle the variability in scale in order to allow using the images for training at the correct scale for a given biological structure and also to perform similarity search at the right scales. The paper compares two neural network architectures for detecting the scale of histopathology patches. The first architecture is a shallow 4-layer CNN and the second architecture is the state–of–the–art DenseNet [11]. DenseNet has a stronger connectivity pattern among all the layers without drastically increasing the number of parameters used. We test our approach on a dataset of patches with a scale of 5×, 10× and 20×, extracted from prostate WSI biopsies. The accuracy of the models measured with Cohen's kappa coefficient reaches 0.9897 using the fine–tuned DenseNet architecture with ImageNet pretrained weights and 0.9617 using the 4-layer CNN, showing that a large pre–trained network helps to recognize the real scale of the images better than an architecture trained from scratch. The main contribution of this paper is to show that current deep learning architectures can be optimized for classifying the scale of an input patch of a histopathology image, opening the possibilities for using these models to automatically classify images of the right scale levels from WSIs of the biomedical literature or from teaching files. This opens a large body of image data for training deep learning models and also for similarity search.

Table 1. Number of patches at each scale level.

Scale/Partition	Train	Test
5×	1652	670
10×	2000	1000
20×	2000	1000

2. METHODS

The approach is shown in Figure 1. First, we extract a set of patches for the three different scales from manually annotated ROIs of prostate biopsy WSIs. Second, we standardize the staining of all patches using a histogram normalization; then the dataset is partitioned and a stratified percentage of the training set is used for validation of the DL models. Finally, we test the two DL architectures using the optimized hyperparameters over the unseen test set.

2.1. Prostate WSI patch dataset

The dataset consists of patches extracted from 50 WSIs obtained at a resolution of 0.25μm per pixel of prostate biopsies. 16 of the 50 WSIs are classified as benign biopsies and the remaining 34 are from confirmed prostate cancer cases with local ROI (region of interest) annotations and Gleason grades ranging from 3 to 5. From the manually annotated ROIs, we extracted in total 8304 non-overlapping patches for all the scale levels and the annotations. Due to the relatively small number of patches that could be extracted at lower magnification (5×) we set an upper bound on the number of patches extracted at 10× and 20×, as seen in table 1. The size of the patches is 224×224 pixels with RGB channels. The patches were separated into a train and test set, separating all patches of a single WSI and patient into either test or training. A stratified sample of 30% of the training patches was used as validation set in the training of the DL models. The performance measures are computed over the unseen 2652 testing patches. For all the patches we normalize the color histogram using contrast limited adaptive histogram equalization to limit the influence of staining differences.

2.2. Deep Learning Architectures

We compared two DL architectures: the first is a very deep state–of–the–art CNN network with 121 layers; the second is a custom designed 4-layer CNN. Because a desirable feature of a scale detector is its capacity for predicting the scale of arbitrary large scale histopathology content in real time, a smaller network with few parameters is a natural choice for this task. We wanted to measure the tradeoff between a larger and potentially more discriminative but slow-to-infer network versus a shallower but faster architecture. DenseNet [11] is a DL architecture designed for having a dense connectivity pattern among the layers, introducing direct connections be-

[4] Computed from the downloaded PMC dataset

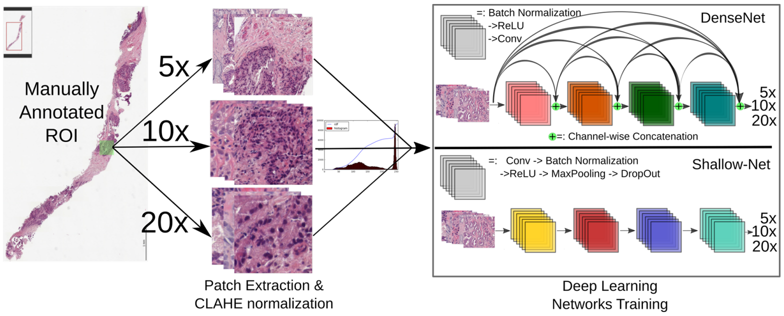

Fig. 1. Flowchart for scale classification for the two DL networks. Both networks receive as input 224 × 224 × 3 patches from annotated ROIs. For illustrative purposes, the schema for DenseNet only shows 4 blocks of the 121 that it is composed of.

tween any two layers with the same feature map size. It reuses information at multiple levels without drastically increasing the number of parameters in the model thanks to bottleneck and compression layers. The bottleneck layer applies a 1×1 convolution before each 3×3 convolution to reduce the number of input feature maps. The compression layer uses part of the feature maps in each transition layer [11]. The performance of DenseNet on image classification was compared with other very deep architectures of 100+ layers, such as ResNet. DenseNet obtains comparable performance with only one third of the parameters. For our experiments we use the variation DenseNet-BC 121. The details of the architectures are the following:

- DenseNet-BC 121: The 121-layer variation of DenseNet with 8 million parameters is used. We used the Adam optimizer with a learning rate of 0.000001. We performed experiments using both, fine–tuning all layers from pretrained ImageNet weights and training from scratch.

- ShallowNet: We designed a 4 layer CNN consisting of 32 3×3 convolutional kernels, followed by batch normalization, ReLU activation, dropout of 0.25 and max–pooling of a 2×2 neighborhood, ending in a dense layer with a Softmax activation connecting to the three classes. Dropout aims to reduce overfitting as the dependency of weights gets regulated by the units that are randomly turned off. In total, the number of parameters is 3.7 million. The best learning rate found was 0.00001 after logarithmic exploration between 0.01 and 10^{-9}

Both networks were trained using the Adam optimizer with a categorical crossentropy loss, monitoring accuracy during 10 epochs. All our experiments were developed using the Keras DL framework with TensorFlow backend. We ran our experiments using a Titan Xp GPU.

Table 2. Cohen's Kappa coefficient over the test set.

Architecture	**Kappa**
ShallowNet	0.9617
DenseNet	0.9477
fine–tuned DenseNet	**0.9897**

Table 3. Confusion matrices for the three architectures: Fine–tuned DenseNet/DenseNet/ShallowNet. Best diagonal result in bold.

	5X	10X	20X
5X	663/579/**665**	7/91/5	0/0/0
10X	11/0/8	989/**1000**/992	0/0/0
20X	0/0/0	0/0/54	**1000**/**1000**/946

3. RESULTS

Cohen's kappa coefficients are reported for test patches in Table 2 for the 3 different setups: ShallowNet, DenseNet and fine–tuned DenseNet. Interestingly, the shallow architecture achieves a similar performance to the DenseNet architecture trained from scratch, suggesting that for this problem it is not neccessary to have very deep architectures. Nevertheless, the performance increases when the weights of ImageNet are used for DenseNet. This is consistent with what other authors have experimented in other histopathology tasks [1, 4]. The confusion matrices for the three architectures are shown in Table 3. Only few patches at 5X and 10X scales get confused. This result can be due to the visual similarity between the two classes. While 20X patches mostly contain few large nuclei, in 5X and 10X patches the nuclei arrangements can look similar, as represented in Figure 2.

4. DISCUSSION AND CONCLUSIONS

The results described in this paper suggest that the exact scale of patches of histopathology WSIs can be determined with a very high confidence using DL. The simple network architecture in ShallowNet seems sufficient for the task of scale

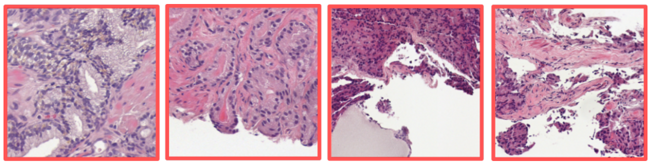

Fig. 2. Sample misclassifications for fine–tuned DenseNet Network: The first two patches on the left are at 10X but predicted as 5X, the other two were classified as 10X but are 5X.

detection, even though with ImageNet pre-training DenseNet was able to reach a better and almost perfect result. The good performance of ShallowNet can be attributable to the network size because it is likely to fit better due to the relatively small size of the dataset. Based on the results of the paper we suggest that scale can be detected also from varied histopathology images, as they appear in the biomedical literature and in medical teaching files. These images could then be harvested for visual similarity retrieval or for creating a focused training collection for deep CNN models. The data set used here is relatively homogeneous and contains images from one laboratory and one anatomic region, so complexity could be higher in real-world teaching files or when using images from the medical literature, when multiple organs and different stainings can be used. As future work, we are currently training our model to recognize a spectrum of scales as a regression problem. Any intermediate scale can indeed be determined given a set of labeled physical areas (e.g. nuclei masks), thus allowing to assess the performance of the scale detector on unlabeled data from open access databases like PMC. Despite the good results obtained, the classification of patch scales could be influenced by several factors, including the influence of disease. This should be studied in further studies.

Acknowledgements

This work was partially supported by the Eurostars project E! 9653 SLDESUTO-BOX and by Nvidia.

5. REFERENCES

[1] Andrew Janowczyk and Anant Madabhushi, "Deep learning for digital pathology image analysis: A comprehensive tutorial with selected use cases," *Journal of Pathology Informatics*, vol. 7, 2016.

[2] Vibha Gupta and Arnav Bhavsar, "Breast cancer histopathological image classification: Is magnification important?," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2017, pp. 17–24.

[3] Neslihan Bayramoglu, Juho Kannala, and Janne Heikkilä, "Deep learning for magnification independent breast cancer histopathology image classification," in *Pattern Recognition (ICPR), 2016 23rd International Conference on*. IEEE, 2016, pp. 2440–2445.

[4] Jayashree Kalpathy-Cramer, Alba García Seco de Herrera, Dina Demner-Fushman, Sameer Antani, Steven Bedrick, and Henning Müller, "Evaluating performance of biomedical image retrieval systems: Overview of the medical image retrieval task at ImageCLEF 2004–2014," *Computerized Medical Imaging and Graphics*, vol. 39, no. 0, pp. 55 – 61, 2015.

[5] Henning Müller, Jayashree Kalpathy-Cramer, Dina Demner-Fushman, and Sameer Antani, "Creating a classification of image types in the medical literature for visual categorization," in *SPIE Medical Imaging*, 2012.

[6] Sven Koitka and Christoph M. Friedrich, "Traditional feature engineering and deep learning approaches at medical classification task of ImageCLEF 2016," in *CLEF2016 Working Notes*, Évora, Portugal, September 5-8 2016, CEUR Workshop Proceedings, CEUR-WS.org.

[7] Alba García Seco de Herrera, Roger Schaer, Stefano Bromuri, and Henning Müller, "Overview of the ImageCLEF 2016 medical task," in *Working Notes of CLEF 2016 (Cross Language Evaluation Forum)*, September 2016.

[8] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," in *Advances in Neural Information Processing Systems 25*, F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, Eds., pp. 1097–1105. Curran Associates, Inc., 2012.

[9] Tiffany L Sellaro, Robert Filkins, Chelsea Hoffman, Jeffrey L Fine, Jon Ho, Anil V Parwani, Liron Pantanowitz, and Michael Montalto, "Relationship between magnification and resolution in digital pathology systems," *Journal of pathology informatics*, vol. 4, 2013.

[10] Patrick Leo, George Lee, Natalie N. C. Shih, Robin Elliott, Michael D. Feldman, and Anant Madabhushi, "Evaluating stability of histomorphometric features across scanner and staining variations: prostate cancer diagnosis from whole slide images," *Journal of Medical Imaging*, vol. 3, no. 4, pp. 047502–047502, 2016.

[11] Gao Huang, Zhuang Liu, Kilian Q. Weinberger, and Laurens van der Maaten, "Densely connected convolutional networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017.