MULTI-LEVEL BATCH NORMALIZATION IN DEEP NETWORKS FOR INVASIVE DUCTAL CARCINOMA CELL DISCRIMINATION IN HISTOPATHOLOGY IMAGES

Francisco Perdigón Romero[⋆]   *An Tang*[†]   *Samuel Kadoury*[⋆†]

[⋆]MedICAL Laboratory, Polytechnique Montreal, Montréal, Canada
[†]Centre de recherche du CHUM (CRCHUM), Montréal, Canada

ABSTRACT

Breast cancer is the most diagnosed cancer and the most predominant cause of death in women worldwide. Imaging techniques such as the breast cancer pathology helps in the diagnosis and monitoring of the disease. However identification of malignant cells can be challenging given the high heterogeneity in tissue absorbotion from staining agents. In this work, we present a novel approach for Invasive Ductal Carcinoma (IDC) cells discrimination in histopathology slides. We propose a model derived from the Inception architecture, proposing a multi-level batch normalization module between each convolutional steps. This module was used as a base block for the feature extraction in a CNN architecture. We used the open IDC dataset in which we obtained a balanced accuracy of 0.89 and an F1 score of 0.90, thus surpassing recent state of the art classification algorithms tested on this public dataset.

Index Terms— breast cancer, pathology, deep learning, batch normalization, convolutional neural networks

1. INTRODUCTION

According to a recent census from 140 countries worldwide [1], breast cancer is the most frequently diagnosed cancer in women, with an estimated incidence of 1.7 million new cases per year. It represents 25% of all types of cancers diagnosed in women. Furthermore, 15% of cancer deaths in women are due to breast cancer [2]. In recent years, a combination of improved detection techniques and earlier diagnosis, added to more effective treatments, has caused the mortality rate to decrease in developed or emerging countries [1].

This has motivated groups to pursue efforts in improving detection methods at early stages of the disease. Breast cancer detection through histopathology images allows to quantify the cancer cell ratio with respect to healthy cells. Oncologists use this information to determine cancer aggressiveness and select the appropriate treatment option. Previously, researchers have worked on methods for cancer/healthy cell classification [3, 4]. However, the scientific community is still working to improve these classification methods by identifying predominant features in cell images.

Convolutional neuronal networks (CNN) have become increasingly popular in the processing of histopathological images [5]. The main difference with respect to traditional machine learning approaches is that feature extraction is done intrinsically and hierarchically. This allows the network to choose in a an unsupervised fashion the features for optimal cell discrimination. In this work, we present a new CNN-based model for cancer/healthy cell classification on histopathology slides (HPS). We leverage these recent discoveries in deep learning, particularly using the Inception architecture [6] by combining regularization techniques such as batch normalization [7] in order to reduce the intrinsic covariate shift.

2. METHODS AND MATERIALS

2.1. Dataset

We used a publicly available dataset [1] that was first introduced by Cruz-Roa et al. [3] for Invasive Ductal Carcinoma (IDC) identification. The histopathological images of the dataset were not released in their original format but in RGB patches of 50x50 pixels. The zoom factor was 2.5x (4m/pixel). The dataset contained 277,525 patches from 279 subjects. The patches were comprised of 28.39% invasive ductal carcinoma cells, while the remaining patches consisted of healthy tissue or non-invasive ductal carcinoma (see Figure 1).
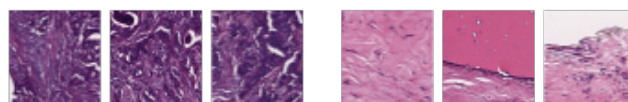

Fig. 1. Left: three IDC cell patches; Right: three normal tissue patches.

2.2. Proposed model

The Inception module [6] introduced by Szegedy C. et al. was initially proposed to deal with the size variability problem in

---

[1]Available at `http://andrewjanowczyk.com/wp-static/IDC_regular_ps50_idx5.zip`

hierarchical representations. The presented solution was simple and achieved excellent results by using parallel filters with different kernel sizes, and allowed the back-propagation process to select which filter would be used in each case.

Our model is composed of blocks derived from Inception that integrates batch normalization (BN) [7] after each convolution step. For each Inception module, the feature number per branch is determined during validation. Due to the concatenation that takes place in each module, the feature output number is four times greater than the feature number per branch value. In the branches that contain the filters with 5x5 and 3x3 kernels, dimensionality reduction is applied to minimize memory consumption through filters with 1x1 kernels. The number of dimensionality reduction filters is set by the (α) and (β) parameters, which vary depending on the feature number per branch (see Figure 2).
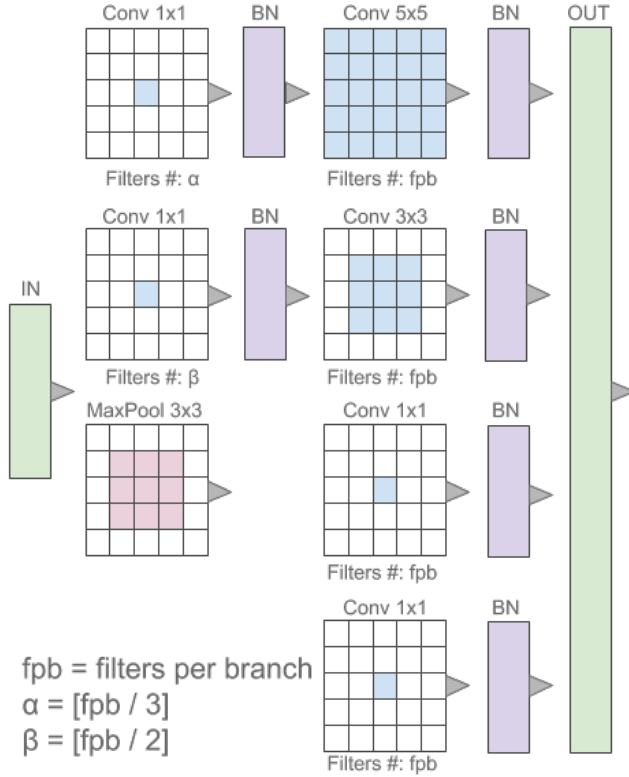

Fig. 2. Schematic illustration of the proposed Inception module.

The batch normalization steps reduce the internal covariate shift, improving network training stability. This aids in the training as the technique re-parametrizes the underlying optimization problem thereby making the gradients more predictable [8].

MaxPool layers are used to reduce the size of the internal network representations. These layers increase the number of extracted features without excessive memory consumption and provide a certain level of rotational invariance. In the fi-

nal stages where the classification is performed, we include two fully connected layers of 512 neurons with batch normalization and dropout of 0.4 each. The network output contains two neurons followed by a SoftMax layer. See Table 1 for the detailed architecture.

Table 1. Details of the framework architecture.

Block Type	Block Name	Feature #
Feature Extraction	Inception	64
Feature Extraction	Inception	64
Feature Extraction	MaxPool	-
Feature Extraction	Inception	128
Feature Extraction	Inception	126
Feature Extraction	MaxPool	-
Feature Extraction	Inception	256
Feature Extraction	Inception	256
Feature Extraction	MaxPool	-
Feature Extraction	Inception	512
Feature Extraction	Inception	512
Classification	Fully connected	256
Classification	Batch Normalization	-
Classification	Dropout (0.4)	-
Classification	Fully connected	256
Classification	Batch Normalization	-
Classification	Dropout (0.4)	-
Classification	Fully connected	2
Classification	Softmax	-

2.3. Training Protocol

In this section, we present the training protocol with the parameter selection strategies.

Padding: Most of the patch images were 50x50 pixels in size, but in some isolated cases, images which contained 1-2 fewer rows or columns were zero-padded.

Normalization: Mean centering and standard deviation normalization was applied to each image. This decreases the variability of the input data, thus improving the training stability.

Train/Validation/Test split: We train the network with a balanced training set of 94,543 images. The validation set contained 31,514 images, also balanced. For the final evaluation, our test set is unbalanced, the total number of images were 151,465 images, of these 15,757 images contain cancer cells.

Batch size: A batch size of 32 images was used for training. This makes the training faster as it takes advantage of the speed-up of matrix-matrix products [9].

Optimization and Learning Rate Scheduling: Adam optimization algorithm [10] was used with an initial learning rate of 10^{-3}. This learning rate was reduced by a factor of 2 after 10 epochs without improvements in the validation set accuracy. The allowed minimum learning rate was 10^{-10}.

Iterations: The initial number of epochs was 10^3, but due to the early stopping criterion that was implemented, the final number of epochs was 55. For early stopping, the validation set accuracy was monitored, and stopped after 50 epochs if no improvements was detected.

Training time on an NVIDIA Titan Xp GPU was 170 minutes, while the inference time per single patch was 3 ms (89 ms on an I7 7th generation CPU). The above timings were calculated for the proposed model, include the Inception architecture.

3. RESULTS AND DISCUSSION

Table 2 shows a comparison between the balanced accuracy and F1-score obtained with our method and previous methods using the same dataset. Both metrics are designed to deal with unbalanced datasets because in these cases, metrics such as accuracy can lead to misinterpretations in the classifier efficiency. Our model showed an improvement in the results obtained by Cruz-Roa et al. [3], that presented a custom deep neural network, and the results obtained by Janowczyk and Madabhushi [4] who used AlexNet architecture [11] for the classification task in the same dataset.

Table 2. Ensemble evaluation results comparing our model with others approaches in the same dataset.

Model	Balanced Accuracy	F1 score
2D-CNN [3]	0.842	0.718
AlexNet with BN [4]	0.847	0.765
Original InceptionNet [6]	0.868	0.883
Proposed method	**0.890**	**0.897**

Figures 3 and 4 shows the efficiency of our classifier. The area under the curve (AUC) of the receiver operating characteristic (ROC) was 0.956 for Inception+BN, while the original Inception module was 0.949. The improvement in accuracy was 2.2% compared to the original model. Figure 4 shows the confusion matrix of the obtained results for the proposed architecture.

These results demonstrate the effectiveness of our Inception module variant. Due to network's multi-scale capability of processing information through filters with different kernel sizes, the network is able to extract highly discriminative features. Furthermore, the use of batch normalization aids the training by reducing internal covariate shift, thus increasing the stability of the deep neural network. Finally, the training strategy allowed us to obtain the maximum benefit of each learning rate step, improving the convergence of the network.

Because the original patches in the database contain patch positions, this enables to reconstruct the entire original slides. The trained models were then applied at the patch level to
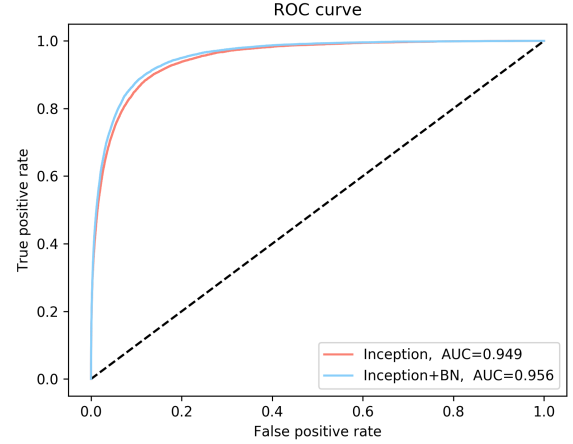

Fig. 3. Models performance represented by the ROC curves.
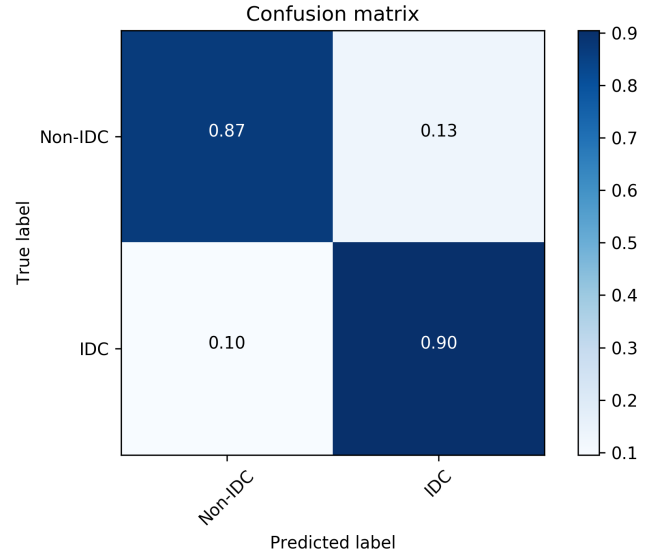

Fig. 4. Model performance evaluated by the confusion matrix.

generate entire processed slides. Because our model processes patches of 50x50 pixels, the generated heatmap appeared "pixelated". To improve the final rendering, we applied a Gaussian filter with a 25x25 kernel. Figure 5 (a) shows the slide number 14154, while Figure 5 (b) shows an overlap of the predictions heat map generated by our model. Our framework can be applied to HPS of other datasets, provided they have a similar zoom factor of x2.5.

4. CONCLUSION

We presented a novel approach for Invasive Ductal Carcinoma cell discrimination in histopathology slides. The model is composed of Inception variant blocks with combined batch normalization steps to reduce internal covariances. Results
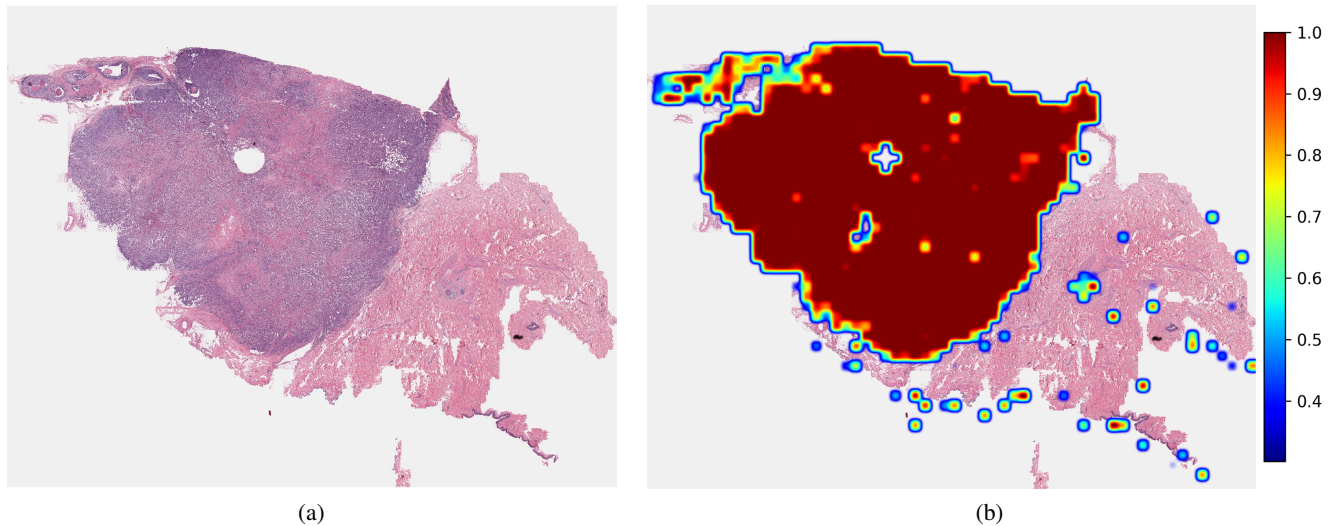
(a)

(b)

Fig. 5. IDC cell prediction results. (a) Original reconstructed pathology slide. (b) Prediction heat map overlapped to the original slide.

surpass state of the art classification methods which were trained and tested on the same dataset. The quantitative evaluation of the results was carried out through balanced accuracy and F1-score, metrics used for unbalanced datasets. Future work will be focused on the inclusion of residual blocks [12] to deal with the vanishing gradients during the training, which in theory will lead to improved accuracy.

5. REFERENCES

[1] W.B. Stewart and C.P. Wild, "World cancer report 2014. lyon, france: International agency for research on cancer," *World Health Organization*, p. 630, 2014.

[2] J. Ferlay, I. Soerjomataram, R. Dikshit, S. Eser, C. Mathers, M. Rebelo, M.D. Parkin, D. Forman, and F. Bray, "Cancer incidence and mortality worldwide: sources, methods and major patterns in globocan 2012," *International journal of cancer*, vol. 136, no. 5, pp. E359–E386, 2015.

[3] A. Cruz-Roa, A. Basavanhally, F. Gonz'alez, H. Gilmore, M. Feldman, S. Ganesan, N. Shih, J. Tomaszewski, and A. Madabhushi, "Automatic detection of invasive ductal carcinoma in whole slide images with convolutional neural networks," in *Medical Imaging 2014: Digital Pathology*. International Society for Optics and Photonics, 2014, vol. 9041, p. 904103.

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

[5] G. Litjens, T. Kooi, E.B. Bejnordi, A.A. Setio, F. Ciompi, M. Ghafoorian, A.J. van der Laak, B. Van Ginneken, and I.C. S'anchez, "A survey on deep learning in medical image analysis," *Medical image analysis*, vol. 42, pp. 60–88, 2017.

[6] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1–9.

[7] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *arXiv preprint arXiv:1502.03167*, 2015.

[8] S. Santurkar, D. Tsipras, A. Ilyas, and A. Madry, "How does batch normalization help optimization?(no, it is not about internal covariate shift)," *arXiv preprint arXiv:1805.11604*, 2018.

[9] Y. Bengio, "Practical recommendations for gradient-based training of deep architectures," in *Neural networks: Tricks of the trade*, pp. 437–478. Springer, 2012.

[10] P.D. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[11] A. Krizhevsky, I. Sutskever, and E.G. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[12] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.