ON DATA BALANCING FOR EFFICIENT DEEP LEARNING WITH PATHOLOGY IMAGES

Jisoo Kim [a], Hyungjoon Jang [b], Kanggeun Lee [b], Gyuhyun Lee [b], Se Young Chun [a], Won-Ki Jeong [b]

Department of EE [a], CSE [b], UNIST, Ulsan, Republic of Korea

ABSTRACT

We investigated a data balancing technique for effective deep learning with pathology images that often contain severely imbalanced data. Our proposed data balancing scheme with ResNet-101 yielded substantially improved performance for detecting cancer metastasis area, especially small area. Then, our random forest classifier yielded the final classification results of cancer staging for the submission of Camelyon'17.

1. INTRODUCTION

In digital pathology, the classification of cancer stage (pN-stage) through whole slide images (WSIs) is important. As the automation of pathology imaging is accelerated, it becomes easier to obtain a lot of WSIs to analyze. Unfortunately, it takes a lot of time and effort for pathologists to inspect several ultra-high-resolution WSI images for one diagnosis. Thus, algorithms that automatically detect and classify cancer metastases could help pathologists to deal with massive amount of pathology images efficiently and effectively.

Recent advances in deep learning-based classification (e.g. [1]) and the availability of massive amount of public dataset (e.g., Camelyon'17 challenge for breast cancer [2]) enabled significant improvements of localization / classification accuracies for cancer metastasis through inspecting ultra-high-resolution WSIs. However, several technically challenges still remain to improve performances for clinical usage, to apply to real working environment for pathologists, and to handle different metastasis features of various cancers in different organs. Here, we investigated a data balancing technique for effective deep learning with pathology images that often contain severely imbalanced data. Our proposed data balancing scheme with ResNet-101 yielded improved performance for detecting cancer metastasis area, especially small areas. Then, our random forest classifier yielded the final classification results of cancer staging for the final evaluation in Camelyon'17 challenge.

2. METHODS

Our proposed methods consist of four steps: 1) heatmap generation by detecting / localizing metastasis in WSIs, 2) feature extraction to yield a total of 11 feature vectors, 3) WSI classification using a random forest detector for each WSI into Negative, Micro-metastasis, Macro-metastasis, and isolated tumor cells (itc), and 4) pN-stage determination of each lymph node using the classification results of each WSI to be one of pN0, pN0(i+), pN1mi, pN1, pN2.

2.1. Heatmap generation for cancer metastasis

A deep learning network was used to detect breast cancer metastasis in lymph nodes. An entire slide image was divided into 256 by 256 patches and those image patches were fed into ResNet-101 model [1] for training so that the network can yield the probability of cancer metastasis in each patch. ResNet is a deep learning model that has been widely used for image classification [1] and it is also used for medical imaging and multiple image analysis. While the original ResNet-101 for image classification has 1000 outputs, we modified it to have only one output corresponding to the probability of having cancer metastasis. Once trained, the ResNet-101 can be applied to a WSI to yield a probability heatmap for each slide image. This heatmap shows a probability of having cancer metastasis for each patch so that spatial distribution and intensity of cancer metastasis can be clearly visualized.

2.2. Proposed data balancing

From generated heatmaps and ground truth label images for cancer metastasis, we were able to observe that the size of cancer metastasis area in each WSI greatly differs. Due to these differences, the number of cancer metastasis image patches and the number of normal tissue image patches are significantly different or imbalanced. Figure 1 shows one of the ground truth masks from the Camelyon17 dataset [2]. It is easy to observe that WSI01 and WSI03 have relatively large metastasis area while WSI02 and WSI04 have relatively small metastasis area. Note that the absolute ratio of the size of cancer lesion to the size of normal tissue is very small. Thus, significantly different numbers of cancer metastasis image patches are obtained as shown in Table 1.

Table. 1. The number of cancer metastasis patches per WSI before and after data balancing process.

	WSI01	WSI02	WSI03	WSI04
Before	1123	209	1823	84
After	811	810	811	810

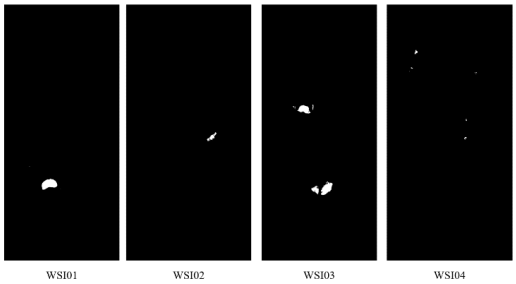

Figure 1. Binary masks to indicate metastasis areas (white) over normal tissue areas (black) from 4 different WSIs.

We propose a data balancing process as follows:

1) count the number of cancer metastasis image patches in a WSI for all WSIs and then obtain the average number over all WSIs.
2) discard cancer metastasis patches randomly if the number of patches is larger than the average OR augment cancer metastasis patches randomly if the number of patches is smaller than the average.

Thus, the final number of cancer patches for each WSI is the same as the average. For example, from 4 WSIs, about 810 (810 or 811) image patches for cancer metastasis were able to be obtained as shown in Table 1.

2.3. Classification of cancer metastasis status

A random forest classifier was used to identify cancer progression [3]. Firstly, a 11-feature vector was extracted from a probabilistic heat map of the entire slide image. All 11 features are described in Table 2 [3]. Then, a random forest classifier uses these 11 features to classify tumors in each full slide image among 4 categories such as negative, micro metastasis, macro metastasis, and isolated tumor cells (itc). Lastly, the final classification of tumor-node-metastasis (TNM) progression is determined based on classes from WSIs per patient as shown in Table 3 [2].

Table. 2. All 11 features in a feature vector for cancer metastasis classification [3].

No	Feature description
1	largest region's major axis length
2	largest region's maximum confidence probability
3	largest region's average confidence probability
4	largest region's area
5	average of all region's averaged confidence probability
6	sum of all region's area
7	maximum confidence probability in WSI
8	average of all confidence probability in WSI
9	number of regions in WSI
10	sum of all foreground area in WSI
11	foreground and background area ratio in WSI

Table. 3. TNM classification criteria [2]

pN0	No micro-metastases or macro-metastases or ITCs found.
pN0(i+)	Only ITCs found.
pN1mi	Micro-metastases found, but no macro-metastases found.
pN1	Metastases found in 1-3 lymph nodes, of which at least one is a macro-metastasis.
pN2	Metastases found in 4-9 lymph nodes, of which at least one is a macro-metastasis.

3. EXPERIMENTAL RESULTS

3.1. Experiment setup

First of all, we briefly validated our data balancing approach by training our modified ResNet-101 deep neural network with 4 WSIs. A WSI was divided into 256x256 patches for training deep learning model for lymph node breast cancer detection. In this training phase, a cross entropy function was used for a loss function and Adam optimizer was used with learning rate 0.0001, batch size 4, and 500 epochs. For the evaluation, Jaccard Index (Intersection over Union or IOU) was used for measuring the performance of the trained model.

Then, for the final classification results, all 500 WSIs were used to train the network again. We also trained a random forest classifier with a set of 500 WSIs where 400 of them were used as training data, and the remaining 100 were used as validation data.

3.2. Results

Table 4 shows the changes in the Jaccard Index scores before and after our proposed data balancing for 4 WSIs. For all WSIs, there were performance improvements by 139-1759%. Note that for WSI1 and WSI3, their performances were improved with smaller number of patches after data balancing. Figure 2 illustrates that data balancing efficiently trained the network to detect small cancer metastasis area (red in the rightmost sub-figure) while the network with no balancing was not able to detect as shown in the second rightmost sub-figure.

Table. 4. Jaccard Index scores without and with our proposed data balancing of cancer metastasis.

	WSI1	WSI2	WSI3	WSI4
Without proposed method	0.2869	0.0100	0.2664	0.0227
With proposed method	0.3978	0.1759	0.5627	0.0380
Performance improvement	139%	1759%	211%	167%

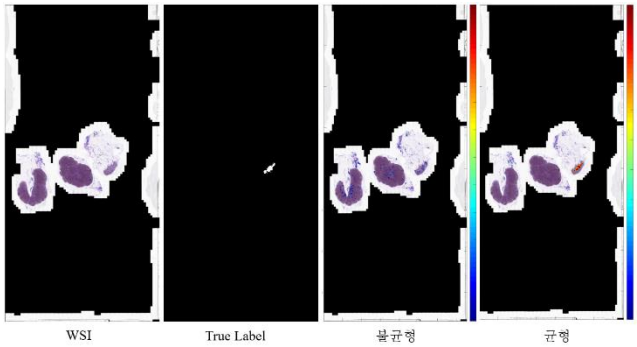

Figure 2. Results before and after data balancing of tumor patch counts (third and fourth sub-figures). The red region is the detected cancer metastasis area.

We extracted a total of 11 feature vectors per WSI as shown in Table 2 and used them for our random forest classifier for classification of lymph nodes. Table 5 shows the results of our random forest classifier using feature vectors extracted from a total of 500 slide images (400 training and 100 validation). The overall accuracy was 0.67.

Table. 5. Tumor classification results.

	Negative	Macro	Micro	itc
Sensitivity	0.9508	0.7176	0.6966	0.6842
Specificity	0.7941	0.3429	0.3214	0.3158

There are a total of 5 TNM classification results for a patient, determined by the type of metastasis of each lymph node, which are categorized by the criteria provided by Camelyon17 (Table 3) [2]. Table 6 shows the results of the final TNM classification. The experiment was conducted in 100 used tumor categories previously and a total of 20 lymph node data were used. Note that these results are from ground truth heatmaps instead of generated heatmaps to evaluate trained random forest classifier only. By submitting to Camelyon17, we will get results for the whole proposed process.

Table. 6. TNM classification results.

	pN0	pN0(i+)	pN1mi	pN1	pN2
Sensitivity	0.8571	0.8571	0.9000	0.7333	0.5789
Specificity	0.5833	0.4839	0.4893	0.4590	0.4211

REFERENCES

[1] He, Kaiming, et al. "Deep residual learning for image r ecognition." *Proceedings of the IEEE conference on co mputer vision and pattern recognition*. 2016.

[2] https://camelyon17.grand-challenge.org/

[3] Lee, Byungjae, and Kyunghyun Paeng. "A Robust and Effective Approach Towards Accurate Metastasis Dete ction and pN-stage Classification in Breast Cancer." *ar Xiv preprint arXiv:1805.12067*(2018).