CANCER METASTASES ANALYSIS WITH DEEP NEURAL NETWORKS

Mrinal Mohit, Ramesh Raskar
Media Lab, Massachusetts Institute of Technology

ABSTRACT

CAMELYON17 [1] was a competition organized by the International Symposium of Biomedical Imaging (ISBI), with the task of analyzing lymph node slides for metastatic breast cancer, classifying each patient into one of 5 stages of severity. We achieved a Cohen-Kappa score of 0.8684 on a 20% validation set. We describe here the multi-stage analysis framework, which includes a convolutional neural networks (CNN), a random forest and a rule-based classifier.

1. INTRODUCTION

The second grand challenge in pathology, CAMELYON17 aimed to evaluate algorithms for automated detection and classification of breast cancer metastases in whole-slide images of histological lymph node sections. The lymph nodes are the first place breast cancer is likely to spread. Metastatic involvement of lymph nodes is one of the most important prognostic factors in breast cancer. However, the diagnostic procedure requires extensive microscopic assessment by highly trained pathologists with subjective opinions. Small metastases are very difficult to detect and sometimes they are missed. The challenge built upon the one held a year before [2] by moving from slide level analysis to patient level analysis (i.e. combining the assessment of multiple lymph node slides into one outcome: the pN stage), bringing the efforts closer to direct usefulness in a clinical setting. The competition ran from November 2016 to April 2017.

In this paper, we present a multi-stage deep-learning based approach to determine the clinical stage of the breast cancer patient. We split each gigapixel-resolution image into millions of patches, use a convolutional neural network to make predictions about each patch’s probability of representing a tumor section, and then aggregate these predictions into a likelihood map. We train a random forest classifier on the geometric properties of the largest region / connected component in the map to separate out different sizes of tumors. These slide-level classification results from multiple lymph node samples of the patient are then combined with an expert-provided rule-based system to obtain the clinical stage.

2. DATASET AND EVALUATION

In this section, we describe the dataset provided, and the evaluation criterion used to rank submissions.

2.1. Dataset

The dataset consisted of whole-slide images (WSI) of hematoxylin and eosin (H&E) stained lymph node sections, collected from 5 medical centers. They were stored in a multi-resolution pyramid structure with each image in the pyramid stored as a series of tiles, to facilitate rapid retrieval of sub-regions of the image. A typical whole-slide image is approximately 200,000 x 100,000 pixels on the highest resolution level with 3 byte RGB pixel format. (i.e. 55.88GB of uncompressed pixel data from a single level)

2.1.1. Training Data

Lesion-level data has two sources -

- CAMELYON16 train and test data – 400 WSI with annotated binary masks
- CAMELYON17 train and test data – 50 WSI with annotated binary masks Patient-level data – 100 patients (5 WSI each) with an expert diagnosis for each slide (among negative, isolated tumor cells, micrometastasis, macro-metastasis) and for each patient (pN stage)
2.1.2. Test Data

The test dataset consisted of unannotated WSI from 100 patients across 5 centers, with 5 slides each.

2.2. Evaluation Criterion

There are four categories a lymph node WSI can belong to –

1. Macro-metastases: metastases greater than 2.0 mm.
2. Micro-metastases: metastases greater than 0.2 mm or more than 200 cells, but smaller than 2.0 mm.
3. Isolated tumor cells (ITC): single tumor cells or a cluster of tumor cells smaller than 0.2 mm or less than 200 cells.
4. Negative: None of the above

The pathologic lymph node classification (pN-stage) is defined as –

1. pN0: No micro-metastases or macro-metastases or ITCs found.
2. pN0(i+): Only ITCs found.
3. pN1mi: Micro-metastases found, but no macro-metastases found.
4. pN1: Metastases found in 13 lymph nodes, of which at least one is a macro-metastasis.
5. pN2: Metastases found in 49 lymph nodes, of which at least one is a macro-metastasis.

A five class quadratic weighted Cohen’s kappa score is used for evaluation, where the classes are the pN-stages.

\[
\kappa = 1 - \frac{\sum_{i=1}^{k} \sum_{j=1}^{k} w_{ij} x_{ij}}{\sum_{i=1}^{k} \sum_{j=1}^{k} w_{ij} m_{ij}}
\]

where \( k \) is the number of classes and \( w_{ij}, x_{ij}, \) and \( m_{ij} \) are elements in the weight, observed, and expected matrices, respectively. The weights in the quadratic set are defined as:

\[
w_i = 1 - \frac{i^2}{(k-1)^2}
\]

In case of a tie, the submissions and ranked by the kappa score calculated on the individual pN stages in reversed order (starting with pN2).

3. METHODS

In this section we describe how the data is processed to build a machine learning model, through a 3-stage framework.

3.1. Pre-processing

The slides were corrected for variations in staining color, as described in [3]. The algorithm utilizes color and spatial information to classify the image pixels into different stain components, thus aligning the chromatic and density distributions for each of the stain components in the hue-saturation-density color model from a template whole-slide image (WSI). It yields the smallest standard deviation and coefficient of variation of the normalized median intensity measure.

The WSI were pre-processed to remove areas not containing the biopsy sample. This was achieved by Otsu’s foreground segmentation technique [4] on the hue and saturation channels. Due to the massive size of the images, the resulting 80% reduction in image data was critical to performing the analysis in a reasonable timeframe.

![Non-tumorous patches](image1)

![Tumorous patches](image2)

Fig. 2. Sample 256 x 256 patches from the two classes at highest magnification.

3.2. Patch-wise classification

We perform all analyses at the highest available resolution (40x), as lower magnifications do not significantly improve performance, as investigated earlier in [5] and [6].

Patches of size 256x256 (Figure 2) were extracted from the highest resolution image available. 500 normal patches and 500 tumorous patches (when applicable) from each WSI formed a set of 200,000 and 120,000 samples respectively to train a binary classifier on.

We trained a 101-layer Residual Network [7] model on the data, which had been pre-trained on the Imagenet dataset [8], to a top-5 error rate of 6.21%. A number of augmentations were applied to make the model generalize better, including:

- Scale and aspect ratio augmentation (random crop 8%–100% and aspect ratio 3/4 to 4/3), as described in [9], instead of the one in the original ResNet paper.
Fig. 3. Example of a detected tumor. The original slide had been masked by the organizers to show only the actual slide.

- Color augmentation (photometric distortions) from [10] in addition to AlexNet style augmentation from the ResNet paper.

- Random rotations in multiples of 90 degrees, and horizontal and vertical flips. The testing was done on each orientation and flip, and the resulting confidences averaged.

Each WSI at the highest resolution was split into patches of size 256x256, which were tested through the binary classifier, as shown in Figure 3.

3.3. Whole slide classification

The classification results of the constituent patches (performed with a stride of 64) were combined into a new likelihood map image as input to the next stage.

We performed connected component analysis to find the largest tumor-classified region. The region properties (regionprops from the python skimage module) were extracted. These included, among others, the area, perimeter, eccentricity, centroid location, convex hull, axis of orientation and solidity of the region.

The features were fed into a random forest classifier, with 4 possible outputs - macrometasis, micrometasis, itc or none. Each slide was thus assigned a label.

3.4. Patient stage classification

The 5 slide labels for each patient were then combined with the ruleset described in Section 2.2. It must be noted that the ruleset is a heuristic simplification of the actual procedure used by expert pathologists, where the lymph node’s location of origin, molecular techniques and >10 slides per patient are used to inform the stage assignment.

4. RESULTS

We randomly shuffled and divided the patient-level training set into 80% training and 20% validation. Performing the techniques of Section 3 resulted in one pN stage label per patient. The quadratic weighted Cohen’s kappa evaluation metric gave a best score of 0.8684 on the validation set after a hyperparameter search.

5. DISCUSSION

A number of engineering tricks were required to work with such a large dataset (1.4 TB training, 1.4 TB test). In order to maintain very high throughput while extracting patches from the tiff and running them through the 8 GPUs, we wrote a custom data loader which ran multi-threaded in the background to match the extraction rate with the GPU processing rate. This led to 700% speedup over a naive extract-save-load-classify pipeline.

We used Torch on an Ubuntu Server machine with 2x8-core Intel Xeon 2.1 GHz processor, 128 GB DDR4 RAM, 2 TB SSD, and 8x GTX 1080 8GB graphics cards. To minimize memory usage, we shared gradient tensors between ResNet modules of the same type.

6. ACKNOWLEDGEMENTS

We thanks all the organizers of the CAMELYON17 grand challenge.

7. REFERENCES


