Deep learning-based research on the influence of training data size for breast cancer pathology detection

Chongyang Cui$^{1,2}$, Shangchun Fan$^{1,2}$, Han Lei$^{1,2}$, Xiaolei Qu$^{1,2}$, Dezhi Zheng$^{1,2}$

$^1$School of Instrumentation Science and Opto-electronics Engineering, Beihang University, Beijing, People's Republic of China
$^2$Beijing Advanced Innovation Center for Big Data-based Precision Medicine, Beihang University, Beijing, People's Republic of China

Abstract: In pathological diagnosis of breast cancer, there are problems such as shortage of pathologists, difficulties in sample labeling, and huge workload of manual diagnosis. Therefore, deep learning-based computer-assisted pathology analysis systems have been developed to diagnose breast cancer and have achieved impressive results. However, it is difficult to obtain a large number of training sets due to the scarcity of pathological images and the huge labeling costs. Therefore, the size of the training set should be planned before building the pathology computer-assisted breast cancer analysis system. Here, the authors present a study to determine the optimal size of the training data set needed to achieve high classification accuracy when developing a pathology computer-assisted breast cancer analysis system. The authors trained two kind of CNNs using six different sizes of training set and assessed the resulting system with a total of 10,000 images. All images were acquired from the Camelyon17 challenge. Here, the authors propose a scheme for determining the size of the training set and the size of the model in developing the pathology computer-assisted breast cancer analysis systems, which can be easily applied to develop systems for other different pathological images.

1 Introduction

Pathological diagnosis is the gold standard in the process of diagnosis of the entire disease, and it is also the core of the prognosis of the disease. It is irreplaceable in pathological diagnosis. However, because of the high work pressure and high risk, the number of pathologists worldwide is very low. The current answer to this problem is computer-assisted pathology analysis systems, but the limitation of these systems is that they still use the features extracted manually [1, 2]. In recent years, convolutional neural networks (CNN) are undertaking a more and more important role in image classification tasks. Along with the rise and rapid development of digital pathology, deep learning has been applied to the analysis of digital pathological images, which has garnered the interest of the medical image analysis community, resulting in increasing numbers of publications on histopathologic image analysis [3].

In the pathological examination of breast cancer, pathologists examine the slides of human tissues by microscopy. Tissue samples are usually collected during surgery or collected by biopsy. The ultra-high resolution digitisation of slides in recent years makes deep learning an ideal choice for pathological image analysis.

Constructing a computer-assisted pathology analysis system based on deep learning requires the classification model to reach a certain accuracy, which is crucial for patient diagnosis and treatment. However, due to the patients' privacy and security policies, it is difficult to guarantee the acquisition of a large number of pathological images. Furthermore, after obtaining the pathological image, high-cost manual labelling is needed before using it in the training of the CNN model. This requires the following questions to be solved before developing a pathology computer-assisted breast cancer analysis system: when training CNN models with target classification accuracy, how to plan the size of the training set, when obtaining a certain number of training sets, how high the classification accuracy can be achieved, and how to select the CNN model that can achieve better results.

Several approaches to the question that how many training data sets are needed have been introduced and explored in different applications [4–6]. Upon thorough evaluation and consideration of these approaches, we chose the learning curve method due to its shown promise and robustness within other applications; and we have considered a number of CNN models and selected two different models (Alexnet [7] and Vgg16 [8]) for research.

Here, we propose a scheme for selecting the training set size in developing a pathology computer-assisted breast cancer analysis system to achieve the desired high precision; at the same time, we studied the effect of the size of the model on the results to ensure the optimal use of the training set. In this planning scheme, some methods can be easily generalised in different pathological image analysis tasks, and even applied to other image analysis fields.

1.1 Detection of breast cancer metastases

Our research mainly focus on the detection of breast cancer metastases in lymph nodes. Lymph nodes are small glands that filter lymph, the fluid that circulates through the lymphatic system. The lymph nodes in the axilla are the first place breast cancer is likely to spread. Prognosis is poorer when cancer has spread to the lymph nodes. This is why lymph nodes are surgically removed and examined microscopically.

In clinical pathology, human tissue is examined through a microscope by a pathologist: a medical doctor specialised in detecting and characterising diseases on a cellular level. Tissue samples are mostly collected during surgery or via biopsy and need to be further processed in order to make glass slides, which hold histological sections of just a few micro-meters thick.

Tissue processing includes fixation, embedding, cutting, and staining. The haematoxylin and eosin (H&E) stain is most widely used.

The rise of digital pathology has enabled the high-resolution digitisation of pathological tissue slides, making pathological images more accessible and more manoeuvrable. However, the diagnostic procedure for pathologists is tedious and time-consuming. Most importantly, small metastases are very difficult to detect and sometimes they are missed.

2 Methodology

2.1 Data acquisition

All data sets we use in the research were acquired from the CAMELYON17 challenge which is the second grand challenge in
pathology organised by the Diagnostic Image Analysis Group (DIAG) and Department of Pathology of the Radboud University Medical Centre (Radboudumc) in Nijmegen, The Netherlands. The data in this challenge contain whole-slide images (WSI) of H&E stained lymph node sections, 500 TIFF images are provided for training and another 500 TIFF images for testing with five slides per patient. All the data set for CAMELYON17 is collected from five medical centres in the Netherlands. (Fig. 1)

Detailed transfer area annotations are provided at the lesion level and patient level. All ground truth annotations were carefully prepared under supervision of expert pathologists. Here, we will mainly focus on the lesion level to detect the lesion area in the images.

2.2 Image Pre-processing

We first identify tissue within the WSI and exclude background white space by adopting a threshold-based segmentation method to automatically detect the background region [9]. The optimal threshold values in each channel within RGB colour space are computed using the Otsu algorithm [10]. The detection results are visualised in Fig. 2, where the tissue region are highlighted using green curves. According to the detection results, the average percentage of background region per WSI is more than over 50%.

We extract millions of small positive and negative patches with size of 256 × 256 from the set of training WSIs. If the small patch is located in a tumour region, it is labelled with 1, otherwise, it is labelled with 0. Our research only focuses on the patch-based classification stage. We select positive and negative examples to train a supervised classification model to discriminate between these two classes of patches [11]. (Fig. 3)

2.3 Convolution neural network

The detection of metastasis of breast cancer requires the examination of the entire pathological image, and the missed determination of micrometastases or other insignificant metastatic regions is very prone to serious errors.

After extensive review of these algorithm, we chose to implement and use Alexnet [7] and Vgg16 [8] for our specific application because they are computationally efficient and provides high-quality classification compared to other CNN. Meanwhile, they have different parameter quantities, we studied the relationship between the size and precision of training sets in different models, the purpose is to select the more suitable CNN models for the pathology computer-assisted breast cancer analysis systems. (Table 1)

2.4 Learning curve

The learning curve approach of modelling classification performance as a function of the training sample size can predict the sample size needed to train a certain image classification system [4]. The classification accuracy (y) is expressed as a function of the training set size (x) where given unknown parameter \( b = b_1, b_2 \). The learning curve was modelled by the following equation.

\[
y = f(x; b) = 100 + b_1 \cdot x^{b_2}
\]

where \( x = [x_1, x_2, \ldots, x_k]^T \), \( y = [y_1, y_2, \ldots, y_k]^T \), and \( b = [b_1, b_2]^T \), \( b_1 \) and \( b_2 \) represent the learning rate and decay rate, respectively. The model fit assumes that the classification accuracy (y) grows asymptotically to 100%, or maximum achievable performance.

Using the observed classification accuracy at six different sizes of training sets (20, 50, 100, 200, 1000, and 2000), unknown parameters \( b = [b_1, b_2]^T \) were estimated using weighted non-linear regression. Since the variance of the classification accuracy in models of different training data sets sizes are different, we weight them accordingly.

Repeat the modelling process on the two CNN models to obtain the accuracy rate learning curve for the two models.

3 Results

In the actual pathological image, the tumour area is much smaller than the normal area, but in order to obtain more intuitive experimental results, we set the size of the negative and positive patches ratio to 1:1. We set up six sets of training sets of different sizes, using Alexnet and Vgg16 models based on tensorflow, and trained each set of data 10 times. After that, we use the prepared test set to measure the classification accuracy of the model, take the average of 10 times as the final accuracy, calculate its variance, and use it to fit the weights of the learning curve. Finally, we use a larger training set (5000), to test the trained models and compare the results with those predicted by the learning curve.

Fig. 4 shows the learning curves of the two models, respectively. During the experiment, we found that with the increase in the number of training samples, the variance of the 10 test accuracy gradually decreases, and we use the weighted least squares method to find the learning curve based on the standard

---

Table 1 Scale of the two models

|          | Layers | Kernel size |
|----------|--------|-------------|
| AlexNet  | 8      | 11, 5, 3    |
| Vgg16    | 19     | 3           |

---

They are computationally efficient and provides high-quality classification compared to other CNN. Meanwhile, they have different parameter quantities, we studied the relationship between the size and precision of training sets in different models, the purpose is to select the more suitable CNN models for the pathology computer-assisted breast cancer analysis systems. (Table 1)
deviation as the weight value. We can also see that in Fig. 4, the point with the largest number of training sets is closest to the learning curve, and the curve fits best at this sample size.

At the same time, Fig. 4 shows that at the same place of two models, when the number of training samples is less than or equal to 250, the training accuracy of the model increases rapidly, and the precision growth of the model after 250 starts to become slow.

Fig. 5 shows the learning curve for the two models. It can be seen that AlexNet's performance is better than Vgg16 when the training set size is very small. After training set size exceeds 100, Vgg16's performance exceeds that of AlexNet. The results show that when the data volume of the training set is large, Vgg16, which is deeper and more complex, will have higher classification accuracy.

Fig. 6 shows the accuracy of the prediction when the training set size is 5000 compared with the actual precision of the training set when it is 5000 sets. The training set used to test the prediction is completely different from the training set used to generate the model. Also, the ratio of positive to negative patches is 1:1. On AlexNet, the actual accuracy of the 5000 training set is 97.8%, which is basically consistent with the prediction accuracy of 97.2%. Similarly, on Vgg16, 98.2% of the 5000 training set had a prediction accuracy of 98.6%. The prediction ability of learning curve is demonstrated.

4 Conclusion

With the digitisation of pathological images, people's research focuses on the deep learning-based pathological detection of breast cancer. However, due to the difficulty in obtaining pathological images and labelling, the preparation of the training set and the selection of the model during the development of the computer-assisted pathology analysis systems have an important influence on the detection accuracy of the system. The important issues to be solved when developing a computer-assisted pathology analysis system include: how to plan the size of the training set when

![Fig. 4](image1.png) *The predicted learning curve and observed average value of Alexnet and Vgg16 (a) The predicted learning curve of Alexnet, (b) The predicted learning curve of Vgg16*

![Fig. 5](image2.png) *Predicted learning curve of Alexnet and Vgg16*

![Fig. 6](image3.png) *Tested result at large data set on Alexnet and Vgg16 (a) Tested result at large data set on Alexnet, (b) On Vgg16*
certain accuracy is desired, and how to select the deep learning model when the size of the labelled training set is already determined.

We study the pathological images of breast cancer and use the learning curve method to predict the classification effect of breast cancer in patch-based classification stage, or to select the training set size for the computer-assisted pathology analysis systems when the accuracy is determined. At the same time, we conducted studies on AlexNet and Vgg16, respectively, to give guidance to the developers of the computer-assisted pathology analysis systems through the performance of training sets of the same size on the deep learning models of different complexity. In this way, we can guarantee the efficiency of the training set and ensure the classification accuracy of the model.

We have conducted studies on pathological images of breast cancer. Due to the common characteristics of pathological images, we can easily use the results of breast cancer research as a guide for the development of computer-assisted pathology analysis systems, even promote the application of more other image analysis field.

5 Acknowledgments

This research was supported by The National Youth Talent Support Program held by Beihang University.

For this research, we would like to thank the innovation team of the ministry of education for advanced sensing technology in aerospace. More importantly, we would like to thank my supervisor Dr Qu and Dr Zheng for their trust and generous support. We also thank for classmates’ self-giving help in coding and valuable suggestions.

6 References

[1] Hua, K.L., Hsu, C.H., Hidayati, S.C., et al.: ‘Computer-aided classification of lung nodules on computed tomography images via deep learning technique’, Oncotargets Therapy, 2015, 8, p. 2015
[2] Shiraishi, J., Li, Q., Appelbaum, D., et al.: ‘Computer-aided diagnosis and artificial intelligence in clinical imaging’, Semin. Nucl. Med., 2011, 41, (9), p. 449
[3] Shen, D., Wu, G., Suk, H.I.: ‘Deep learning in medical image analysis’, Annu. Rev. Biomed. Eng., 2017, 19, (1), pp. 221–248
[4] Figueroa, R.L., Zemp-Treitler, Q., Randula, S., et al.: ‘Predicting sample size required for classification performance’, BMC Med. Inform. Decis. Mak., 2012, 12, (1), p. 8
[5] Roth, H.R., Lee, C.T., Shin, H.C., et al.: ‘Anatomy-specific classification of medical images using deep convolutional nets’, 2015, pp. 101–104
[6] Dobbin, K.K., Zhao, Y., Simon, R.M.: ‘How large a training set is needed to develop a classifier for microarray data?’, Clinical Cancer Research: An Official Journal of the American Association for Cancer Research, 2008, 14, (1), pp. 108–114
[7] Krizhevsky, A., Sutskever, I., Hinton, G. E.: ‘Imagenet classification with deep convolutional neural networks’. Int. Conf. on Neural Information Processing Systems, 2012, pp. 1097–1105
[8] Simonyan, K., Zisserman, A.: ‘Very deep convolutional networks for large-scale image recognition’, Comput. Sci., p. 2014, arXiv:1409.1556 [cs.CV]
[9] Wang, D., Khosla, A., Gargeya, R., et al.: ‘Deep learning for identifying metastatic breast cancer’, 2016
[10] Otsu, N.: ‘A threshold selection method from gray-level histograms’, IEEE Trans. Syst. Man Cybernet., 2007, 9, (1), pp. 62–66
[11] Liu, Y., Gadepalli, K., Norouzi, M., et al.: ‘Detecting cancer metastases on gigapixel pathology images’, 2017

8732

This is an open access article published by the IET under the Creative Commons Attribution License (http://creativecommons.org/licenses/by/3.0/)