Using convolutional neural network for diabetes mellitus diagnosis based on tongue images

Lintai Wu¹, Xiaoling Luo¹, Yong Xu¹

¹School of Computer Science and Technology, Harbin Institute of Technology (Shenzhen), Shenzhen, People's Republic of China
E-mail: laterfall@hit.edu.cn

Abstract: Tongue diagnosis plays a great role in traditional Chinese medicine. Diabetes mellitus (DM) diagnosis is a significant branch of tongue diagnosis. In recent years, many algorithms have been proposed to aid DM diagnosis based on tongue images. However, most of the previous studies are based on the traditional machine learning and extract only low-level features, such as colour and texture. Here, the authors used a convolutional neural network for tongue image classification by extracting and using the high-level features of tongue images. They conducted an experiment on a set of 422 DM images and 422 healthy images, which were captured by the specialised device. In order to solve the problem of a small dataset, the authors used a pre-trained model to fine-tune parameters of the network, which is a kind of transfer learning way to accelerate the training speed and improve the accuracy. Finally, the authors compared their experiment with the other state-of-the-art algorithms of DM diagnosis, and the results show that their method has the best performance in terms of many assessment criteria.

1 Introduction

Diabetes mellitus (DM) has become one of the major health problems in the 21st century. Medical institutions and governments are paying more and more attention to the examination and treatment of DM.

The diagnostic methods used now by hospitals are considered to be time consuming and invasive. For example, fasting plasma glucose is the standard method to diagnose DM. It requires the patient to keep fasting for at least 8 h, and take a sample of the patient's blood. Although this method is reliable, it can be painful. Therefore, it is very much necessary to develop a more simple and non-invasive method to diagnose DM.

In the past few decades, a number of diagnosis methods based on tongue images have been proposed. Chiu [1] built a computerised tongue examination system for quantising the properties of the tongue. He developed two algorithms to identify the colour of the tongue and the texture of its coating. Zhang et al. [2] detected DM by extracting 13 features of tongue images, including colour, texture and geometry features. The average accuracy of their method could achieve 80.52%. Cao et al. [3] introduced a way to compute statistical features and optimise them by a method based on doublet.

In addition to feature extraction, some researchers have devoted to optimising the classifiers. Jianfeng et al. [4] proposed a genetic algorithm-support vector machine (GA-SVM) model to classify DM tongue images. They used the GA to optimise parameters of SVM to achieve a better performance.

Even though the analysis of tongue images has been developed for decades, most of the previous studies are only based on low-level features, such as colour, texture, geometry etc. These features cannot present a holistic view of tongue images. Moreover, it costs a lot of time to extract each feature.

In recent years, the convolutional neural network (CNN), which can extract high-level features, has been proven to be highly successful in image classification, image denoising [5], object detection [6] and other tasks. As a result, it is widely applied to healthcare. Gulshan et al. [7] used the inception-v3 [8] architecture to detect the diabetic retinopathy based on over 120,000 retinal fundus photographs. The results almost reached the level of professional doctors. Kermany et al. [9] used the inception-v3 architecture combined with the transfer learning technique to diagnose age-related macular degeneration and diabetic macular oedema with over 200,000 optical coherence tomography (OCT) images of the retina.

Although CNN can be highly effective in most cases, it is well known that it relies on large amounts of data. However, different from retina images, the number of tongue images is too insufficient for CNN to achieve a good result. Accordingly, some researchers have attempted to design networks that are suitable for a small amount of data. Inspired by PCANet [10], Dan et al. [11] introduced a new network named CHDNet to diagnose gastritis. The CHDNet achieved a high level of accuracy in a small number of unbalanced tongue samples. Nevertheless, it costs too much time and memory to train a model.

Fortunately, the transfer learning technique, one effective method to address a lack of data, has been developed quickly. The main idea of transfer learning is to extract knowledge from the source domain and then apply it to the target domain [12]. One of the most commonly used methods of transfer learning is fine tuning, which makes use of the parameters of the source model to initialise the target model and then adjust them with the target data. It can not only make the loss converged faster so as to save much more time, but also improve the accuracy.

In this study, we used the SqueezeNet [13] architecture to detect DM tongue images. As the model size of SqueezeNet is less than 3M, it can easily be embedded into portable diagnosis devices. In order to address a lack of data, a pre-trained model with ImageNet weights was employed to fine tune our network, and the data augmentation technique was used for increasing the amount of data. Moreover, we slightly adjusted the architecture of the network to achieve better results.

2 Methodology

2.1 Image capture device

Tongue images were captured by an especially designed device. As shown in Fig. 1 [2], the designed device is composed of a three-chip CCD camera with 8-bit resolution and two D65 fluorescent tubes. These fluorescent tubes are installed symmetrically around the camera in order to produce uniform illumination. The angle between the incident light and emergent light is 45° in accordance with Commission Internationale de l'Eclairage (CIE).

Owing to the variation of illumination and device properties, the colour of each image may be slightly different. Therefore, the method of Wang and Zhang [14] was used to correct colour to...
eliminate the variation. The idea of this method is to map a matrix generated from the RGB vector of images to an objective sRGB vector, thereby obtaining a transformation model. Then we can use this model to calibrate the colour vector of the captured images.

After capturing all the tongue images, we applied the bi-elliptical deformable contour (BEDC) technique [15] to separate the tongue body from the entire image. The BEDC is based on a bi-elliptical deformable template (BEDT) and an active contour model. The most significant point of BEDC is that it replaces the traditional internal force term in the energy function of BEDT with the template force. Thus, it can deform to fit both global and local details. The tongue image after the aforementioned processing is shown in Fig. 2.

2.2 Data preprocessing

To conduct this experiment, we collected 1228 tongue images of healthy people and 411 ones of DM patients from Guangdong Provincial Hospital of Traditional Chinese Medicine, Guangdong, China and Hong Kong Foundation for Research and Development in Diabetes, Prince of Wales Hospital, Hong Kong SAR.

2.2.1 Undersample: As the dataset is unbalanced, in which the amount of the healthy samples is approximately three times as large as that of the DM samples, we undersampled 411 images randomly from the healthy samples to avoid this unbalance. Thus, the dataset in use is composed of 822 tongue images, in which each class accounts for 50%.

2.2.2 Three-fold cross-validation: In order to estimate the performance of the model, the three-fold cross-validation method was employed. We partitioned the dataset into three subsets evenly and then trained the model on two subsets and validated the model on the other subset. In total, as shown in Fig. 3, we performed validation with three rounds, and in each round the validation set and training set are different. Finally, the validation results were averaged over three rounds.

2.2.3 Data augmentation: For increasing the amount of dataset and improving the generalisation ability of the model, we conducted data augmentation to each training set. To be specific, we rotated the original images with the angle of 30°, flipped the original and rotated images, and added Gaussian noise to the original images with a mean of 0 and a standard deviation of 10.

Thus, we obtained 3836 images as a training set and 274 images as a test set in every cross-validation round.
The AUC value is the area under the receiver operating characteristic (ROC) curve, which is a graphical plot that illustrates the diagnostic ability of a binary classifier system.

In order to evaluate the performance of our method, four representative algorithms, k-nearest neighbour (k-NN), random forest (RF), SVM and GA-SVM were employed to compare with our method. For the first three algorithms, the most commonly used features of colour, texture and red spot were employed as the inputs. The GA-SVM is an improved algorithm applied to DM tongue image classification, which was proposed by Jianfeng et al. [4]. As Table 1 shows, our method exceeds the others in accuracy, sensitivity, specificity and AUC. In addition, as shown in Fig. 4, the ROC curves indicate that the classifier performs well both on the positive samples and the negative samples.

### 3.2 Comparison of networks

We have used other CNN architectures to compare with SqueezeNet, such as Caffenet [17] and Inception-v3 [8]. They were both run on Nvidia Tesla K80 GPUs with batch size 28 for 40 epochs and with base learning rate of 0.0004. For Caffenet, we used the learning policy of ‘sigmoid’ with gamma value of 0.02 and step size of 500. For Inception-v3, we used the learning policy of ‘sigmoid’ with gamma value of 0.004 and step size of 800. As Table 2 shows, not only did the SqueezeNet achieve the best result in terms of the corresponding assessment criteria – accuracy, sensitivity, specificity and AUC, but also has the smallest size which is helpful to embed the model into tongue diagnosis instruments.

### 3.3 Comparison of data augmentation and non-augmentation

To verify the importance of data augmentation, we conducted a contrast experiment without augmentation. As Table 3 shows, the accuracy, sensitivity, specificity and the AUC decreased by 6.94%, 5.35%, 9% and 0.046, respectively, without augmentation. These large differences indicate that data augmentation has a great effect on the improvement of results.

### 4 Conclusion

In this paper, we used a small-size CNN-SqueezeNet to identify DM based on tongue images. For the reliability of the experiment, tongue images were captured by the specially designed device and specialised preprocessed, such as calibrating the colour. In order to address the insufficiency of data, the transfer learning and the data augmentation technique were employed. In addition, we slightly modified the architecture of the network and used a more appropriate learning strategy to train the data to prevent overfitting. The contrast experiments indicated the effectiveness of these adjustments. The experimental results show that our method has the best performance in terms of several assessment criteria – accuracy, sensitivity, specificity and AUC. This lays the groundwork for a new method to diagnose DM more simply.

### 5 Acknowledgments

This work was supported by the Economic, Trade and Information Commission of Shenzhen Municipality (grant no. J. Eng., 2020, Vol. 2020 Iss. 13, pp. 635-638

This is an open access article published by the IET under the Creative Commons Attribution License (http://creativecommons.org/licenses/by/3.0/)
[8] Szegedy, C., Vanhoucke, V., Ioffe, S., et al.: ‘Rethinking the inception architecture for computer vision’, IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 2016, pp. 2818–2826

Kermen, D.S., Goldbaum, M., Cai, W., et al.: ‘Identifying medical diagnoses and treatable diseases by image-based deep learning’, Cell, 2018, 172, (5), pp. 1122–1131

Chen, T.H., Jia, K., Gao, S., et al.: ‘A simple deep learning baseline for image classification’, IEEE Trans. Image Process., 2014, 24, (12), pp. 5017–5032

Dan, M., Guo, T., Ye, D., et al.: ‘Tongue images classification based on constrained high dispension network’, Evidence-Based Complementary Altern. Med., 2017, 2017, (4), pp. 1–12

Pan, S.J., Yang, Q.: ‘A survey on transfer learning’, IEEE Trans. Knowl. Data Eng., 2010, 22, (10), pp. 1345–1359

Iandola, F.N., Han, S., Moskewicz, M.W., et al.: ‘Squeezenet: AlexNet-level accuracy with 50x fewer parameters and <0.5MB model size’, CoRR, abs/1602.07360, 2016

Wang, X., Zhang, D.: ‘An optimized tongue image color correction scheme’, IEEE Trans. Inf Technol. Biomed., 2010, 14, (6), p. 1355

Pang, B., Zhang, D., Wang, K., et al.: ‘The bi-elliptical deformable contour and its application to automated tongue segmentation in Chinese medicine’, IEEE Trans. Med. Imaging, 2005, 24, (8), p. 946

Glort, X., Bengio, Y.: ‘Understanding the difficulty of training deep feedforward neural networks’, J. Mach. Learn. Res., 2010, 9, pp. 249–256

Krizhevsky, A., Sutskever, I., Hinton, G.: ‘Imagenet classification with deep convolutional neural networks’, Conf. and Workshop on Neural Information Processing Systems (NIPS), Lake Tahoe, Nevada, USA, 2012, pp. 1097–1105