Fundus Image Quality-Guided Diabetic Retinopathy Grading

Kang Zhou[1,2(✉)], Zaiwang Gu[2,3], Annan Li[4], Jun Cheng[2], Shenghua Gao[1], and Jiang Liu[2]

[1] School of Information Science and Technology, ShanghaiTech University, Shanghai, China
zhoukang@shanghaitech.edu.cn
[2] Cixi Institute of Biomedical Engineering, Ningbo Institute of Materials Technology and Engineering, Ningbo, China
[3] School of Mechatronic Engineering and Automation, Shanghai University, Shanghai, China
[4] School of Computer Science and Engineering, Beijing University of Aeronautics and Astronautics, Beijing, China

Abstract. With the increasing use of fundus cameras, we can get a large number of retinal images. However there are quite a number of images in poor quality because of uneven illumination, occlusion and so on. The quality of images significantly affects the performance of automated diabetic retinopathy (DR) screening systems. Unlike the previous methods that did not face the unbalanced distribution, we propose weighted softmax with center loss to solve the unbalanced data distribution in medical images. Furthermore, we propose Fundus Image Quality (FIQ)-guided DR grading method based on multi-task deep learning, which is the first work using fundus image quality to help grade DR. Experimental results on the Kaggle dataset show that fundus image quality greatly impact DR grading. By considering the influence of quality, the experimental results validate the effectiveness of our propose method. All codes and fundus image quality label on Kaggle DR dataset are released in https://github.com/ClancyZhou/kaggle_DR_image_quality_miccai2018_workshop.

Keywords: Fundus image quality classification · DR screening
Multi-task · Deep learning

1 Introduction

The fundus image quality has a significant effect on the performance of automated ocular disease screening, such as diabetic retinopathy (DR), glaucoma and age-related macular degeneration (AMD). The symptoms of the above diseases are well defined and visible in fundus images. Research communities have put great effort towards the automation of a computer screening system which is able to promptly detect DR in fundus images. The evaluation of fundus image

D. Stoyanov et al. (Eds.): COMPAY 2018/OMIA 2018, LNCS 11039, pp. 245–252, 2018.
https://doi.org/10.1007/978-3-030-00949-6_29

Table 1. In our Kaggle DR image quality dataset (Sect. 3.1), the number of good and poor quality images are shown as follows. The ratio is extremely unbalanced.

Data set	Total	Good	Poor	Ratio (poor/good)
Training	35126	33841	1285	0.038
Validation	10906	10680	226	0.021
Testing	42670	41797	873	0.021

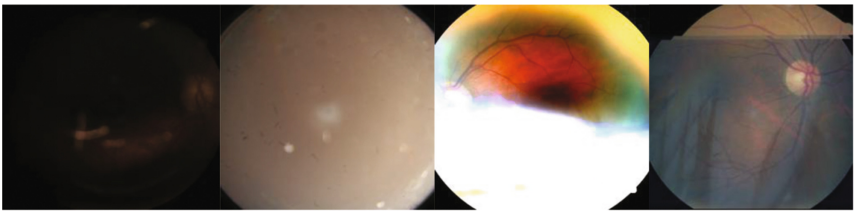

Fig. 1. Four instances of poor quality images in Kaggle DR dataset, and the quality of these images are too poor to identify the lesion.

quality involves a computer-aided retinal image analysis system that is designed to assist ophthalmologists to detect eye diseases. Consequently, automated evaluations of ophthalmopathy can be performed to support the diagnosis of doctors. However, the success of these automatic diagnostic systems heavily relies on the image quality. In reality, due to some inevitable disturbances in the image acquisition, e.g. the operator's expertise, the type of image acquisition equipment, the situation of different individuals, the images are often blurred, which affects the follow-up diagnosis. Therefore, the image quality plays an extremely important role in the computer-aided screening system (Fig. 1).

In the context of retinal image analysis, image quality classification is used to determine whether an image is useful or the quality of a retinal image is sufficient for the subsequent automated diagnosis. Many methods based on hand-crafted features have been proposed for fundus image quality assessment for disease screening. Lee *et al.* [6] use a quality index Q which is calculated by the convolution of a template intensity histogram to measure the retinal image quality. Lalonde *et al.* [5] adopt the features which are based on the edge amplitude distribution and the pixel gray value to automatically assess the quality of retinal images. Traditional feature extraction methods with low computational complexity only can obtain some characteristic that represents image quality rather than always acquiring diversity factors that affect image quality.

With the development of convolution neural network (CNN) in image and video processing [4], automatic feature learning algorithms using deep learning have emerged as feasible approaches and are applied to handle the medical image analysis. Recently, some methods based on deep learning have been proposed for fundus images [2,3]. Specially, methods to handle the fundus image quality

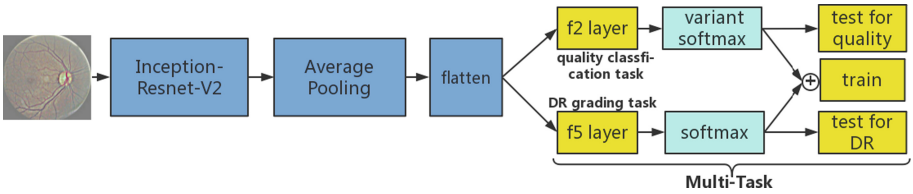

Fig. 2. The overall architecture of our method.

assessment problem also have been proposed. Yu *et al.* [9] first introduced CNN and treated it as a fixed high-level feature extractor, replacing low-level features such as hand-crafted geometric and structural features. Then, SVM algorithm was adopted to automatically classify high quality and poor quality retinal fundus images. Sun *et al.* [7] directly used four CNN architectures to assess fundus images quality. However, in these two papers the authors randomly selecting training set and testing set in Kaggle DR dataset [1], which make it difficult for other to reproduce and compare. In addition, in these two papers the amount of training set and testing set are equal, but it dose not reflect the real data distribution, in which the amount of good quality fundus images is much more than that of poor quality. For example, as Table 1 shown, in Kaggle DR dataset the amount of good quality fundus images and poor quality fundus images are extremely unbalanced. Both of the work avoided the unbalanced data distribution, which is a very common but complex problem in the field of medical image analysis. In this paper, we propose weighted softmax with center loss to handle the problem of unbalanced data distribution.

In the realistic process of computer-aided screening system, fundus image quality assessment is important for subsequent disease diagnosis, such as DR grading. To the best of our knowledge, there is no work using fundus image quality information to help grade DR. In this paper, we propose Fundus Image Quality (FIQ)-guided DR grading method based on multi-task deep learning.

The contributions of our work are summarized as follows:

1. We propose weighted softmax with center loss to solve the unbalanced data distribution in medical images.
2. We propose FIQ-guided DR grading method based on multi-task deep learning, which is the first work using fundus image quality information to help grade DR.
3. Experimental results on the Kaggle dataset show that fundus image quality greatly impact DR grading. By considering the influence of quality, the experimental results validate the effectiveness of our propose method.

The rest of the paper is organized as follows. In Sect. 2, we introduce our method in detail. Section 3 introduce kaggle image quality dataset, as well as the experimental results and quantitative analysis. In the last section, the conclusion is presented.

2 Method

The overall architecture of our FIQ-guided DR grading method is shown in Fig. 2.

2.1 Variant Softmax Loss for Unbalanced Problem

A commonly used loss function for classification in machine learning is softmax loss function, which is shown in Eq. (1):

$$L_{q0} = -\frac{1}{m}\left[\sum_{i=1}^{m}\sum_{j=1}^{k} 1\{y^{(i)} = j\}\log(\text{Prob}_{ij})\right] \tag{1}$$

where m denotes the number of input instances, k denotes the number of classes, $1\{\cdot\}$ denotes the indicator function, $y^{(i)}$ denotes the label of i-th instance and Prob_{ij} denotes the probabilities output by softmax activation. However, this loss function is not appropriate for unbalanced problem because the loss dosen't consider the unbalanced distribution.

The image quality data distribution of Kaggle DR dataset is shown in Table 1, which is extremely unbalanced. To solve the unbalanced problem, there are two popular variant softmax loss called weighted softmax loss (i.e. Eq. 2) and center loss (i.e. Eq. 4).

Weighted Softmax Loss. The weighted softmax loss is shown as follow, where each class is weighted inversely proportional to the number of its samples.

$$L_{q1} = -\frac{1}{\sum_{i=1}^{m} w_i}\left[\sum_{i=1}^{m} w_i \sum_{j=1}^{k} 1\{y^{(i)} = j\}\log(\text{Prob}_{ij})\right] \tag{2}$$

where

$$w_i = \begin{cases} \beta, & y^{(i)} = 0 \\ 1, & y^{(i)} = 1 \end{cases} \tag{3}$$

and scalar β is a hyperparameter.

Center Loss. In order to enhance the discriminative power of the deeply learned features, Wen *et al.* [8] proposed a new supervision signal, called center loss. Specifically, the center loss simultaneously learns a center for deep features of each class and penalizes the distances between the deep features and their corresponding class centers.

$$L_{q2} = -\frac{1}{\sum_{i=1}^{m} w_i}\left[\sum_{i=1}^{m}\sum_{j=1}^{k} 1\{y^{(i)} = j\}\log(\text{Prob}_j) + \lambda L_c\right] \tag{4}$$

where

$$L_c = \frac{1}{2}\sum_{i}^{m} \|x_i - c_{y_i}\|_2^2 \tag{5}$$

and scalar λ is a hyperparameter, which is used for balancing the two loss functions.

Weighted Softmax with Center Loss. In order to make full use of weighted softmax loss and center loss, we propose weighted softmax with center loss:

$$L_{q3} = -\frac{1}{\sum_{i=1}^{m} w_i}\left[\sum_{i=1}^{m}\sum_{j=1}^{k} 1\{y^{(i)} = j\}\log(\text{Prob}_j)w_i + \lambda L_c\right] \tag{6}$$

The conventional softmax loss can be considered as a special case of this joint supervision, if λ is set to 0 and β is set to 1.

2.2 Multi-task Learning

To use fundus image quality information for improving DR grading, we propose multi-task learning that train quality classification task and DR grading task at the same time. As shown in Fig. 2, the propose loss function in training stage is defined as follow:

$$L = L_{dr} + L_q + L_{reg} \tag{7}$$

where L_{dr} denotes the softmax loss of DR grading task, L_q denotes the loss of image quality classification task and L_{reg} denotes the regularization loss (weight decay term) used to avoid overfitting. In testing period, we can simultaneously predict image quality class and DR grade.

3 Experiment

3.1 Datasets

To validate the propose multi-task method and analysis the influence of image quality, we use two dataset as follows:

Kaggle DR Dataset. Kaggle organized a comprehensive competition in order to design an automated retinal image diagnosis system for DR screening in 2015 [1]. The retinal images were provided by EyePACS, which is a free platform for retinopathy screening. The dataset consists of 35126 training images, 10906 validate images and 42670 testing images. Each image is labeled as $\{0, 1, 2, 3, 4\}$ and the number represents the level of DR. We will use this dataset to evaluate the performance of DR grading.

Kaggle DR Image Quality Dataset. To verify the effectiveness of variant softmax loss methods for unbalanced medical images and analysis the influence of image quality qualitatively, we label Kaggle DR Dataset as Image Quality Dataset, which is shown in Table 1. All images are tagged by the professionals to identify the quality of the dataset, in which label 1 represents the image of good quality and label 0 stands for the poor quality images.

3.2 Evaluation Protocols

DR Grading. To evaluate the performance of DR grading, we use the quadratic weighted kappa (shown as Eq. 8) to evaluate our methods, which is used in Kaggle DR Challenge [1]. The quadratic weighted kappa not only measures the agreement between two ratings but also considers the distance between the prediction and the ground truth.

$$k = 1 - \frac{\sum_{i,j} w_{i,j} O_{i,j}}{\sum_{i,j} w_{i,j} E_{i,j}} \tag{8}$$

where $w_{i,j} = \frac{(i-j)^2}{(N-1)^2}$ and O, E are N-by-N histogram matrix.

Image Quality Classification. On the one hand, since this is a binary classification problem, we use the popular metrics: specificity, sensitivity, precision. On the other hand, this is an unbalanced binary classification problem and these negative samples are few but important, so we use mean_acc and specificity as the mainly metrics:

$$\text{mean\_acc} = \frac{\text{acc\_0} + \text{acc\_1}}{2} = \frac{\text{specificity} + \text{sensitivity}}{2} \tag{9}$$

where acc_0, acc_1 denoted the accuracy of class 0, class 1 respectively. Futhermore, specificity = acc_0, sensitivity = acc_1.

3.3 Hyper-parameters

During the training stage, the learning rate in our network is empirically set as 0.001, $\beta = 27$ in weighted softmax loss, $\lambda = 0.1$ in center loss.

3.4 Experiments

A. Image Quality Classification

To evaluate each softmax loss and its variant, we conduct ablation experiments and the results are shown in Tables 2 and 3. All of these results are evaluated on Kaggle Image Quality Dataset.

Performance on **validation set** is shown in Table 2. Results about mean_acc and specificity in row 1 (i.e. L_{q0} with Adadelta) and row 2 (i.e. L_{q1} with Adadelta) show that weighted softmax loss is more appropriate for unbalanced

Table 2. Performance on **validation set**. L_{q0} denotes naive softmax loss, L_{q1} denotes weighted softmax loss, L_{q3} denotes weighted softmax with center loss. For the unbalanced binary classification problem and the negative samples are few, mean_acc and specificity metrics are important.

Loss	Optimizer	*mean_acc*	*Specificity*	acc	Sensitivity	Precision
L_{q0}	Adadelta	0.845	0.704	0.980	0.986	0.994
L_{q1}	Adadelta	0.897	0.827	0.965	0.968	0.996
L_{q1}	Momentum	0.961	0.947	0.974	0.974	0.999
L_{q3}	Momentum	**0.962**	**0.969**	0.955	0.954	0.999

Table 3. Performance on **testing set**, on which is similar with validate set.

Loss	Optimizer	*mean_acc*	*Specificity*	acc	Sensitivity	Precision
L_{q0}	Adadelta	0.850	0.711	0.983	0.989	0.994
L_{q1}	Adadelta	0.905	0.838	0.969	0.971	0.997
L_{q1}	Momentum	0.966	0.954	0.977	0.978	0.980
L_{q3}	Momentum	**0.965**	**0.976**	0.955	0.954	0.999

quality dataset. Results in row 3 (i.e. L_{q1} with Momentum) and row 4 (i.e. L_{q3} with Momentum) show that our weighted softmax with center loss is effective. Performance on **testing set** is shown in Table 3, which is similar in Table 2.

B. DR Grading and Quantitative Analysis

The performance of our method and quantitative experimental results are shown in Table 4, and these results show: (i) $b > a > c$: Fundus image quality greatly impact DR grading; (ii) $d > a$: Our proposed FIQ-guided DR grading method is effective; (iii) $e > b, f < c$ and the raise of ratio: Explain why our proposed method is effective.

Table 4. Quantitative analysis on Kaggle DR dataset. **Single-task** denotes single naive DR grading task, **multi-task** denotes our FIQ-guided DR grading method, **good** denotes kappa on good quality images set while $\mathbf{poor}_k$ denotes kappa on the opposite set, **true** denotes the number of true prediction while $\mathbf{poor}_n$ denotes the number of poor quality image in true set.

Date set	Methods	Kappa			Num		
		Overall	Good	$poor_k$	True	$poor_n$	Ratio
Validation	Single-task	0.718_a	0.721_b	0.629_c	8854	164	18.52‰
	Multi-task	0.745_d	0.750_e	0.616_f	9095	167	18.36‰
Testing	Single-task	0.710	0.715	0.589	34298	633	18.46‰
	Multi-task	0.724	0.730	0.549	34908	623	17.85‰

4 Conclusion

In this paper we propose weighted softmax with center loss to solve the unbalanced data distribution in medical images. Futhermore, we propose FIQ-guided DR grading method based on multi-task deep learning, which is the first work using fundus image quality information to help grade DR. Experimental results on the Kaggle dataset show that fundus image quality greatly impact DR grading. By considering the influence of quality, the experimental results validate the effectiveness of our propose method.

References

1. EyePACS: Diabetic retinopathy detection. https://www.kaggle.com/c/diabetic-retinopathy-detection/data
2. Fu, H., Cheng, J., Xu, Y., Wong, D.W.K., Liu, J., Cao, X.: Joint optic disc and cup segmentation based on multi-label deep network and polar transformation. IEEE Trans. Med. Imaging (2018)
3. Fu, H., et al.: Disc-aware ensemble network for glaucoma screening from fundus image. IEEE Trans. Med. Imaging (2018)
4. Krizhevsky, A., Sutskever, I., Hinton, G.E.: ImageNet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems, pp. 1097–1105 (2012)
5. Lalonde, M., Gagnon, L., Boucher, M.C.: Automatic visual quality assessment in optical fundus images. In: Proceedings of Vision Interface, Ottawa, vol. 32, pp. 259–264 (2001)
6. Lee, S.C., Wang, Y.: Automatic retinal image quality assessment and enhancement. In: Medical Imaging 1999: Image Processing, vol. 3661, pp. 1581–1591. International Society for Optics and Photonics (1999)
7. Sun, J., Wan, C., Cheng, J., Yu, F., Liu, J.: Retinal image quality classification using fine-tuned CNN. In: Cardoso, M. (ed.) FIFI/OMIA-2017. LNCS, vol. 10554, pp. 126–133. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-67561-9_14
8. Wen, Y., Zhang, K., Li, Z., Qiao, Y.: A discriminative feature learning approach for deep face recognition. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9911, pp. 499–515. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46478-7_31
9. Yu, F., Sun, J., Li, A., Cheng, J., Wan, C., Liu, J.: Image quality classification for DR screening using deep learning. In: 2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), pp. 664–667. IEEE (2017)