Fundus Image Quality-guided Diabetic Retinopathy Grading

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

[1] ShanghaiTech University, Shanghai, China
[2] Chinese Academy of Sciences, Ningbo, China
[3] Shanghai University, Shanghai, China
[4]Beijing Univeity of Aeronautics and Astronautics, Beijing, China

Introduction

The fundus image quality has a significant effect on the performance of automated ocular disease screening, such as diabetic retinopathy (DR).

- **Problem:**
  Quality Label: 0, 1 (poor / good quality)
  DR Label: 0, 1, 2, 3, 4 (Larger number means the severity of the disease becomes more significant)
- **Task:**
  Input: image / ***Output***: it's quality and grade
- **Challenge :**
  - The poor quality images are little, which means the dataset is unbalanced;
  - There is no work using image quality to help grading diabetic retinopathy.

Contribution

- We propose weighted softmax with center loss to solve the unbalanced data distribution in medical images.
- 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.

Proposed Method

- **Loss for Unbalanced Problem.** Softmax loss is not appropriate for unbalanced problem because the loss dosen't consider the unbalanced distribution:

$$L_1 = -\frac{1}{m}\left[\sum_{i=1}^{m}\sum_{j=1}^{k} 1\{y^{(i)} = j\} \log(\text{Prob}_{ij})\right]$$

In order to make full use of weighted softmax loss and center loss, we propose weighted softmax with center loss:

$$L_q = -\frac{1}{\sum_{i=1}^{m} w_i}\left[\sum_{i=1}^{m}\sum_{j=1}^{k} 1\{y^{(i)} = j\} \log(\text{Prob}_{ij}) w_i + \lambda L_c\right]$$

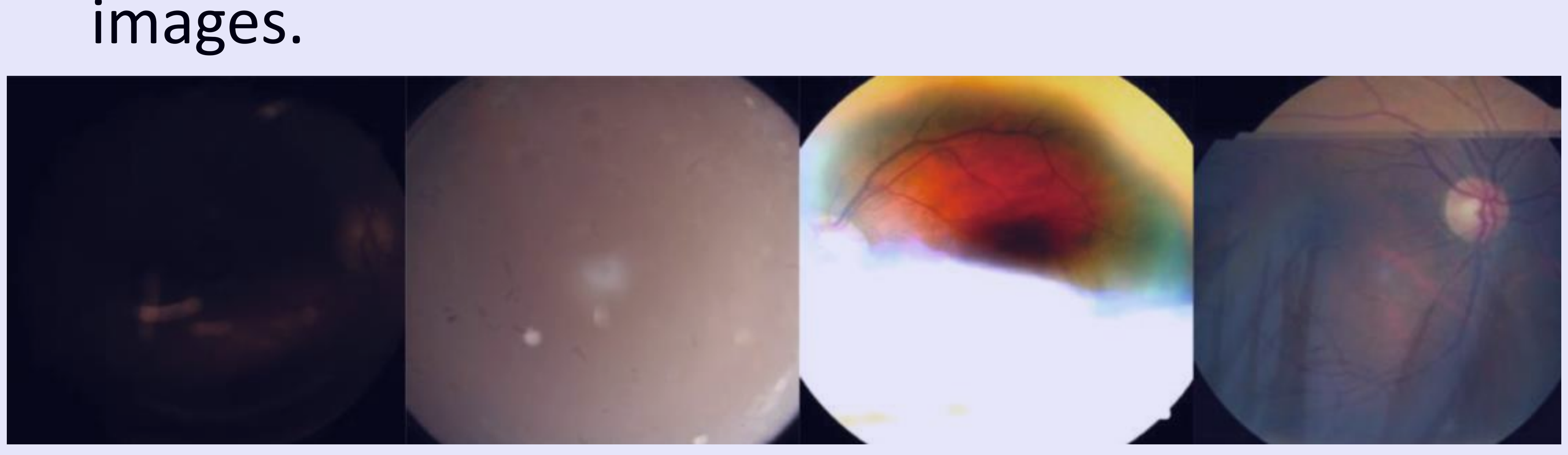
- **Multi-Task Learning.** To use fundus image quality information for improving DR grading, we propose multi-task learning that train quality classification task and DR grading task at the same time. As shown in the bottom figure, the propose loss function in training stage is defined as follow:

$$L = L_{dr} + L_q + L_{reg}$$

where Ldr denotes the softmax loss of DR grading task, Lreg denotes the regularization loss.

Dataset and Experiment

- **Dataset :**
  - **Kaggle DR Dataset.** The retinal images were provided by EyePACS, and Kaggle organized a comprehensive competition based on these images.
  - **Kaggle DR Image Quality Dataset.** To verify the effectiveness of our proposed loss and analysis the influence of image quality qualitatively, we label Kaggle DR Dataset as Image Quality Dataset. The follow figure shows four instances of poor quality images.



- **Experiment :**
  - **Image Quality Classification.** To evaluate each softmax loss and its variant, we conduct ablation experiments.

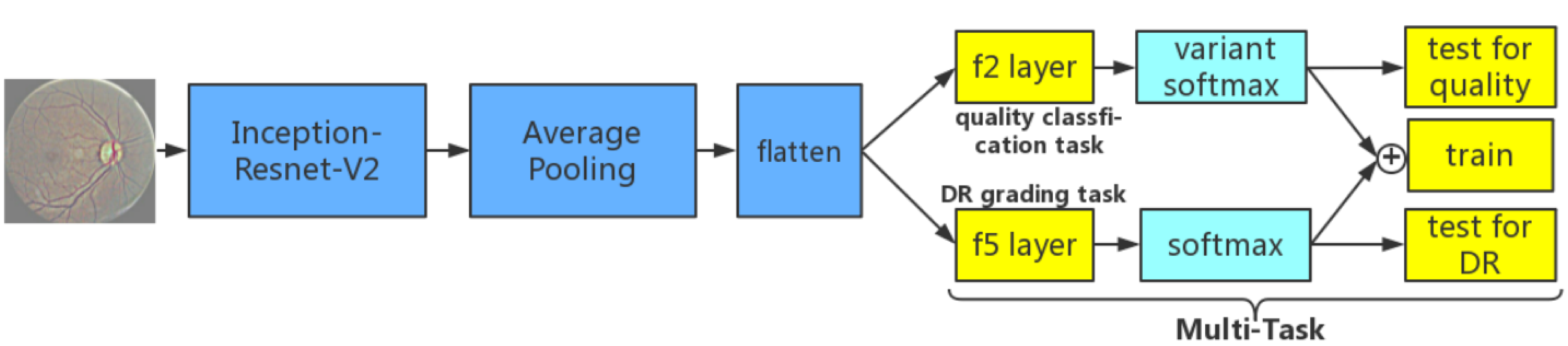
Table 2. Performance on **validation set**. L_{q0} denotes naive softmax loss, L_{q1} denotes weighted softmax loss, L_{q3} denotes weighted softmax with center loss. For the unbalanced binary classification problem and the negtive samples are few, mean_acc and specificity metrics are important.

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

- **DR Grading and Quantitative Analysis.** 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‰



The label we released in github