SEA-NET: SQUEEZE-AND-EXCITATION ATTENTION NET FOR DIABETIC RETINOPATHY GRADING

Ziyuan Zhao[*†], *Kartik Chopra*[*‡], *Zeng Zeng*[♯†], *Xiaoli Li*[†]

[†]Institute for Infocomm Research, Agency for Science, Technology and Research, Singapore
[‡]Institute of Systems Science, National University of Singapore, Singapore

ABSTRACT

Diabetes is one of the most common disease in individuals. *Diabetic retinopathy* (DR) is a complication of diabetes, which could lead to blindness. Automatic DR grading based on retinal images provides a great diagnostic and prognostic value for treatment planning. However, the subtle differences among severity levels make it difficult to capture important features using conventional methods. To alleviate the problems, a new deep learning architecture for robust DR grading is proposed, referred to as SEA-Net, in which, spatial attention and channel attention are alternatively carried out and boosted with each other, improving the classification performance. In addition, a hybrid loss function is proposed to further maximize the inter-class distance and reduce the intra-class variability. Experimental results have shown the effectiveness of the proposed architecture.

***Index Terms*—** Convolutional neural network; Squeeze-and-Excitation net; Diabetic retinopathy grading; Attention mechanism

1. INTRODUCTION

Diabetic retinopathy is the most prevalent microvascular complication among patients with diabetes mellitus [1], as well as one of the most frequent cause of blindness of humans [2]. High blood glucose levels can damage the tiny blood vessels at the back of the eyes, even in the prediabetes stage. In Singapore, around 1 out of 12 people aged from 19 to 69 years are affected by diabetes, and 43.5% among them suffer from different severity of DR [3]. Moreover, there are no early warning symptoms for DR, which lead to difficulties in timely diagnosis and early treatment. Conventionally, DR grading relies on a manual process performed by experts based on fundus photography, which is tedious, costly and time-consuming. Because of human intervention, DR grading also suffers from high intra- and inter-observer variability.

Automatic methods based on computer vision have shown the promising performance in DR detection. Early work in the literature rely on handcrafted features, in which, retinal features such as vessel enhancement, optic disk detection and lesion segmentation are extracted by using image processing techniques and followed by a binary classifier [4, 5]. These methods can not be well embedded into an end-to-end framework, and may suffer from the sensitiveness of conditions, like noises and artifacts. Recently, deep learning, especially *convolutional neural networks* have achieved great success in this scenario [6–9]. Most of them have focused on binary severity-level classification, while multiple severity-level classification can better help patients and doctors in clinical diagnosis and treatments. Bravo *et al.* [10] adopted pre-trained CNNs with different image processing techniques to get an average class accuracy of 50.5% on DR grading. BiRA-Net proposed in [11] utilized the bilinear learning strategy with attention models for fine-grained classification [12], which achieves a higher accuracy. However, even using low-rank bilinear pooling [13], high-dimensional bilinear features also easily lead to some issues, such as overfitting and heavy computation.

In this work, we propose a deep CNN architecture comprising of feature attention and channel recalibration. More specifically, a residual neural network (ResNet) [14] is first implemented to extract features from retinal images, followed by the proposed attention model, which combines a series of 1×1 convolution layers for feature refinement in the spatial dimension. To explore the interdependencies between the channels of convolutional features, the *Squeeze-and-Excitation* (SE) block [15] is implemented in the proposed architecture for higher quality of representations. Moreover, a combination of the weighted cross entropy loss and center loss is proposed to prevent overtraining, enhancing the discriminative ability of the architecture. Experimental results have demonstrated the effectiveness of the proposed method compared to other methods.

The remainder of this paper is structured in the following way. We present an overview of the related work in DR grading in Section 2. The proposed method is shown in Section 3. Section 4 shows experimental results and evaluates the proposed method with respect to different metrics. Finally, we present our conclusions in Section 5.

* Both authors contribute equally to this work.

♯ Corresponding author. The work was supported by Singapore-China NRF-NSFC Grant (Grant No. NRF2016NRF-NSFC001-111).

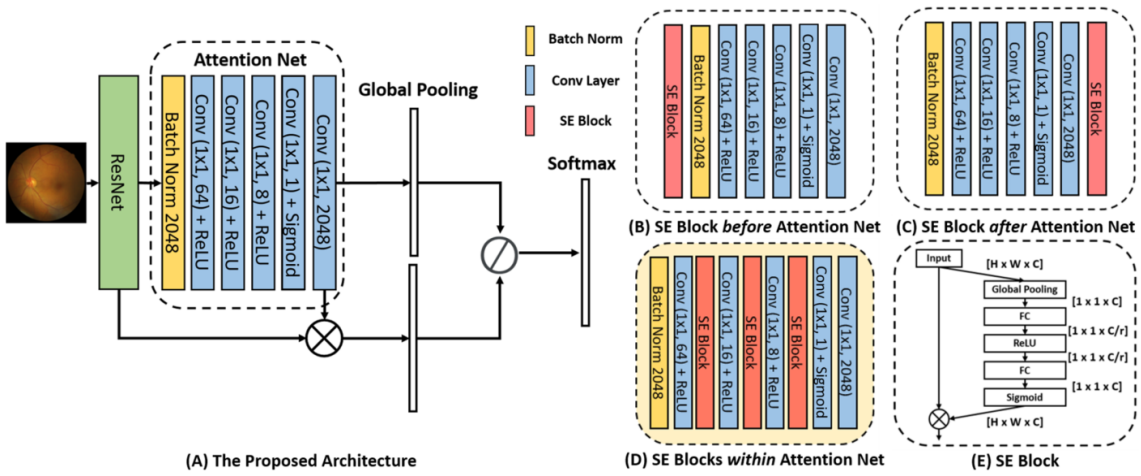

Fig. 1. The overview of the proposed framework is shown in (A). The different defined positions of SE block placed in the architecture are illustrated in (B), (C) and (D). The optimal position (D) is highlighted, which is termed as "SEA-Net". The details of SE block are shown in (E).

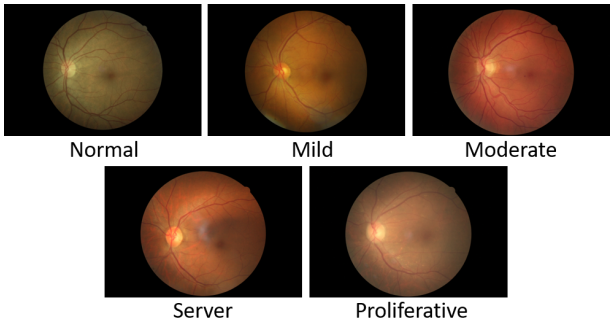

Fig. 2. Examples of retinal images provided by EyePACS. The appearance of images from different classes is similar.

2. RELATED WORK

Most automatic DR classification frameworks heavily rely on handcrafted features or extracted features using image processing. In recent years, many work on binary classification of DR have demonstrated the effectiveness of deep learning based methods, while few work focused on multiclass classification of DR, which is more effective for diagnosis and treatments. Bravo *et al.* integrated different image processing techniques along with VGG16-based architecture for DR grading on a balanced dataset [10]. Zeng *et al.* proposed a siamese neural network for DR grading using left and right retina images [16]. However, only subtle differences among classes are observed in retina images as Fig. 2, which poses a challenge for DR grading.

Retina images contain some irrelevant information, and only some features like microaneurysms are critical for doctors. Therefore, to mimic the clinician diagnosis process, Wang *et al.* proposed a deep learning network using attention mechanism, termed as Zoom-in Net [17]. In Zoom-in Net, the suspicious areas generated by attention maps are zoomed in for details, and more local information is considered for classifying diabetic retinopathy. In [11], a bilinear learning strategy with attention network is proposed to classify DR images at a fine-grained level, boosting meaningful features while suppressing weak ones. These methods capture spatial correlations, achieving competitive accuracy by incorporating spatial attention.

It is well noted that *Squeeze-and-Excitation Networks* (SE Nets) give us a different aspect of attention mechanism, in which, the channel relationship is modeled by introducing SE blocks. The SE Nets perform feature recalibration, utilizing global information for channel attention.

3. METHODOLOGY

The proposed framework is shown in Fig. 1, in which, a ResNet is firstly trained on the processed images to extract features, followed by Attention Net with a sequence of 1×1 convolution layers and pooling layers for dimensionality reduction and spatial attention. Considering the channel relationship of learned features, SE blocks are introduced to recalibrate channel-wise feature maps for fine-grained classification.

3.1. Residual Neural Network

Residual Neural Network (ResNet) [14] is implemented first for deep feature extraction, in which, the shortcut connections skip some layers and perform identity mapping, thus avoiding the gradient vanishing problem and increasing the training speed and effects. Therefore, the ResNet-50 pre-trained

on ImageNet [18] is applied as the backbone of the proposed architecture, in which, the weights of all layers are freezed.

3.2. Attention Net

Let us assume an input image $\mathbf{I} \in \mathbb{R}^{H\times W\times C'}$ which first passes through the ResNet-50 to generate output feature map $\mathbf{U} \in \mathbb{R}^{H\times W\times C}$, $\mathbf{F}_{res} : \mathbf{I} \rightarrow \mathbf{U}$. Here H and W are the spatial height and width, with C' and C being the input and output channels, respectively. We define Attention Net as $\mathbf{F}_{atten} = [\mathbf{c}_1, \mathbf{c}_2, \cdots, \mathbf{c}_n]$, where $\mathbf{c}_n$ is n-th convolution layer in Attention Net. Through a series of convolutional layers and non-linearities defined by $\mathbf{F}_{atten}(\cdot)$, the refined feature map $\mathbf{A} \in \mathbb{R}^{H\times W\times C}$ is generated. We consider the feature map $\mathbf{A} = [\mathbf{a}_1, \mathbf{a}_2, \cdots, \mathbf{a}_C]$ as a combination of channels $\mathbf{a}_i \in \mathbb{R}^{H\times W}$. Spatial squeeze is performed by a global average pooling (GAP) layer, which produces vector $\mathbf{x} \in \mathbb{R}^{1\times 1\times C}$ with its k-th element

$$x_k = \frac{1}{H \times W} \sum_{i}^{H} \sum_{j}^{W} \mathbf{a}_k(i, j). \tag{1}$$

The GAP layer provides a receptive field of whole spatial extent and embeds the global spatial information in vector x. The same operations are performed on $\mathbf{U} \in \mathbb{R}^{H\times W\times C}$, resulting in vector $\mathbf{y} \in \mathbb{R}^{1\times 1\times C}$. Finally, to filter unrelated information, an element-wise division is used followed by a softmax layer.

3.3. Squeeze-and-Excitation Block

To exploit channel dependencies and contextual information, we propose to incorporate the SE block [15] into the proposed architecture. The details of Squeeze-and-Excitation block is described in Fig. 1 (E). The input $\mathbf{G} \in \mathbb{R}^{H\times W\times C}$ is first embedded in to vector $\mathbf{z} \in \mathbb{R}^{1\times 1\times C}$ using a GAP layer, and then transformed to $\hat{\mathbf{z}} = \mathbf{W}_1 (\delta (\mathbf{W}_2 \mathbf{z}))$, where $\mathbf{W}_1 \in \mathbb{R}^{C\times \frac{C}{r}}$ and $\mathbf{W}_2 \in \mathbb{R}^{\frac{C}{r}\times C}$ are the weights of two fully-connected (FC) layers, and $\delta(\cdot)$ is the ReLU activation layer. The parameter r is the reduction ratio for dimensionality reduction, indicating the bottleneck in the channel excitation. In our experiments, we set $r = 4$. After passing $\hat{\mathbf{z}}$ through a sigmoid layer $\sigma(\hat{\mathbf{z}})$, the activations of $\hat{\mathbf{z}}$ are limited into the interval $[0, 1]$, which is used to recalibrate the input $\mathbf{G} = [\mathbf{g}_1, \mathbf{g}_2, \cdots, \mathbf{g}_C]$, $\mathbf{g}_i \in \mathbb{R}^{H\times W}$. The output feature map $\mathbf{G}_{se} \in \mathbb{R}^{H\times W\times C}$, $\mathbf{F}_{se} : \mathbf{G} \rightarrow \mathbf{G}_{se}$ is computed as

$$\mathbf{G}_{se} = [\sigma (\hat{z}_1) \mathbf{g}_1, \sigma(\hat{z_2})\mathbf{g}_2, \cdots, \sigma(\hat{z_C})\mathbf{g}_C] . \tag{2}$$

When integrating the SE blocks in the proposed architecture, the positions of SE blocks in the network influence the performance of DR grading. To find the optimal position, we explore three different positions of SE blocks. Fig. 1 (B), (C) and (D) show different defined positions of SE blocks.

3.4. The Proposed Hybrid Loss

Considering the distance between different classes, we propose to implement center loss [19] to reduce the loss-accuracy discrepancy and get an improved convergence. The center loss function is formulated as

$$\mathfrak{L}_{ct} = \frac{1}{2} \sum_{i=1}^{m} \| \boldsymbol{x}_i - \boldsymbol{c}_{y_i} \|_2^2 , \tag{3}$$

where x_i is the i^{th} training sample and c_{yi} is the y^{th} class center of deep features.

In addition, the DR dataset is heavily imbalanced, and most images are labeled as 0. The weighted cross entropy loss is used to alleviate the problem, which is defined by

$$\mathfrak{L}_{ce} = \text{weight}_y \left(- \log \left(\frac{\exp(x[y])}{\sum_j \exp(x[j])} \right) \right), \tag{4}$$

where x is the training sample, $y \in [0, C-1]$, C is the number of classes, weight_y is a manual rescaling weight given to each class. The weight_y is calculated by dividing the total number of training samples by samples in each class. This ensures that minority classes get higher weights.

Finally, the hybrid loss function is denoted as

$$\mathfrak{L} = \mathfrak{L}_{ce} + \lambda \mathfrak{L}_{ct}, \tag{5}$$

where λ is a scalar to control the strength of loss functions. It is noted that when $\lambda = 0$, only the weighted cross entropy loss is used for parameter estimation.

4. EXPERIMENTS

4.1. Dataset and implementation

The dataset used in this work is from a Kaggle competition provided by EyePACS [20]. The dataset is comprised of 33566 retinal images with 5 classes, where 0 represents no disease and 4 represents the highest severity of disease. Some examples are shown in Fig. 2. Due to the imbalanced dataset, following the data distribution adopted in [10], a balanced testing dataset of 1560 images was applied to our experiments for testing, and the rest were used for training.

To reduce the noises from background, the original images are cropped and resized to 610×610 so that only the retinal region is visible. Then the images are standardized across RGB channels by subtracting the mean and divided by the standard deviation of each channel in the training data. The histogram equalization is applied to improve contrast of the images. To reduce overfitting on models, some augmentation techniques are carried on the dataset, more specifically, the images are randomly rotated ± 10 degrees, flipped vertically or horizontally.

Table 1. Experimental results of various approaches on DR grading, which are described in Section 4.3. The last one is implemented with the proposed hybrid loss function, and the results are obtained when $\lambda = 0.1$.

Method	ACA	Marco-F1	AUC
Bravo et al. [10]	0.5051	0.5081	-
BiRA-Net [11]	0.5431	0.5725	-
AT-Net	0.5442	0.4951	0.8699
SE-AT-Net	0.5776	0.5505	0.8734
AT-SE-Net	0.5830	0.5892	0.8721
SEA-Net	0.5859	0.5872	0.8738
SEA-Net ($\lambda = 0.1$)	**0.5994**	**0.6047**	**0.8760**

The proposed framework is trained on a single NVIDIA Quadro P5000 GPU with batch size of 20, using the stochastic gradient descent (SGD) optimizer with a momentum of 0.9. The initial learning rate is 0.002 with weight decay factor of $1 \times e^{-8}$.

4.2. Performance metrics

To evaluate the performance on the multi-class classification task, we generate the confusion matrix, in which, the number of predictions in each class are presented. Based on the confusion matrix, the *average of classification accuracy* (ACA) can be calculated by taking the percentage of the diagonal elements, which represent the number of points for which the predicted label is equal to the true label. By adopting one-against-all strategy [21], F1 score of each class can be calculated, and Marco-F1 score is obtained by taking the average of the F1 scores for all 5 classes.

To evaluate the diagnostic performance of the proposed method, Receiver operating characteristics (ROC) [22] are calculated, which describe the true positive rate (TPR) and the false positive rate (FPR) at various threshold settings. For comparison, AUC (Area Under Curve) is calculated based on ROC.

4.3. Baseline methods

Previous deep learning approaches are described for comparison. In [10], a series of VGG-based classifiers were trained using different image processing techniques, such as circular RGB and color centered sets. In [11], a bilinear learning strategy was proposed to improve the classification performance on fine-grained images. In addition, the proposed architecture with different positions of SE block is implemented as depicted in Fig. 1, which are summarized as follows.

- **AT-Net:** The proposed Attention Net.
- **SE-AT-Net:** Net with SE block *before* attention block.
- **AT-SE-Net:** Net with SE block *after* attention block.
- **SEA-Net:** Net with SE blocks *within* attention block.

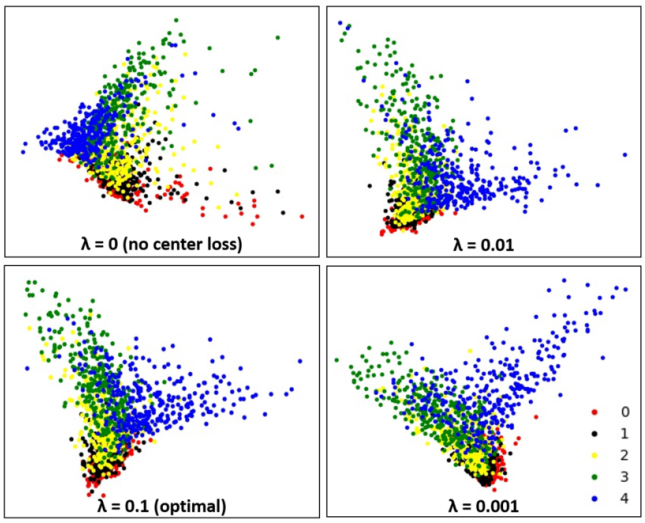

Fig. 3. The feature distribution learned with the proposed hybrid loss of different λ values. The points with different colors denote features from different classes.

4.4. Results and discussion

The experimental results of all methods on the testing data are shown in Table 1. We can see that the proposed framework outperforms other methods in all metrics, even without SE block. SE blocks are proved to be effective in the proposed methods, in which, the best results are obtained from SEA-Net, which is the optimal position in the network. In SEA-Net, the SE blocks are placed alternatively with convolution layers, recalibrating the learned feature maps in an adaptive manner.

It is well noted that the performance of SEA-Net is further improved with the proposed hybrid loss function. Following [19], the distribution of deeply learned features is visualized in Fig. 3, which proves that the proposed hybrid loss can learn better discriminative features, especially for confusing classes, *i.e.*, class 0 and class 1. These two classes are usually regarded as one class in binary classification [10]. In addition, the features in the same class are more close under the supervision of the hybrid loss, than only using the weighted cross entropy loss.

5. CONCLUSIONS

In this work, we proposed a novel deep learning architecture for DR grading, called "SEA-Net", in which, spatial attention and channel attention are implemented to boost each other, recalibrating the attention maps adaptively. In addition, to obtain better learned features, a hybrid loss function based on weighted cross entropy loss and center loss is implemented in SEA-Net. Extensive experiments were carried out using different methods, which demonstrate the effectiveness of the proposed architecture.

6. REFERENCES

[1] Martin M Nentwich and Michael W Ulbig, "Diabetic retinopathy-ocular complications of diabetes mellitus," *World journal of diabetes*, vol. 6, no. 3, pp. 489, 2015.

[2] R Priya and P Aruna, "Diagnosis of diabetic retinopathy using machine learning techniques," *ICTACT Journal on soft computing*, vol. 3, no. 4, pp. 563–575, 2013.

[3] "Updates in detection and treatment of diabetic retinopathy in Singapore," https://www.singhealth.com.sg/news/medical-news-singhealth/updates-in-detection-and-treatment-of-diabetic-retinopathy, [Online; accessed 20-May-2019].

[4] Muhammad Sharif and Jamal Hussain Shah, "Automatic screening of retinal lesions for grading diabetic retinopathy," *INTERNATIONAL ARAB JOURNAL OF INFORMATION TECHNOLOGY*, vol. 16, no. 4, pp. 766–774, 2019.

[5] Axel Pinz, Stefan Bernogger, Peter Datlinger, and Andreas Kruger, "Mapping the human retina," *IEEE Transactions on medical imaging*, vol. 17, no. 4, pp. 606–619, 1998.

[6] Gilbert Lim, Mong Li Lee, Wynne Hsu, and Tien Yin Wong, "Transformed representations for convolutional neural networks in diabetic retinopathy screening," in *Workshops at the Twenty-Eighth AAAI Conference on Artificial Intelligence*, 2014.

[7] Shuangling Wang, Yilong Yin, Guibao Cao, Benzheng Wei, Yuanjie Zheng, and Gongping Yang, "Hierarchical retinal blood vessel segmentation based on feature and ensemble learning," *Neurocomputing*, vol. 149, pp. 708–717, 2015.

[8] Varun Gulshan, Lily Peng, Marc Coram, Martin C Stumpe, Derek Wu, Arunachalam Narayanaswamy, Subhashini Venugopalan, Kasumi Widner, Tom Madams, Jorge Cuadros, et al., "Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs," *Jama*, vol. 316, no. 22, pp. 2402–2410, 2016.

[9] Zeng Zeng, Xulei Yang, Yu Qiyun, Yao Meng, and Zhang Le, "Sese-net: Self-supervised deep learning for segmentation," *Pattern Recognition Letters*, vol. 128, 08 2019.

[10] María A Bravo and Pablo A Arbeláez, "Automatic diabetic retinopathy classification," in *13th International Conference on Medical Information Processing and Analysis*. International Society for Optics and Photonics, 2017, vol. 10572, p. 105721E.

[11] Ziyuan Zhao, Kerui Zhang, Xuejie Hao, Jing Tian, Matthew Chin Heng Chua, Li Chen, and Xin Xu, "Bira-net: Bilinear attention net for diabetic retinopathy grading," in *2019 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2019, pp. 1385–1389.

[12] Tsung-Yu Lin, Aruni RoyChowdhury, and Subhransu Maji, "Bilinear cnn models for fine-grained visual recognition," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1449–1457.

[13] Shu Kong and Charless Fowlkes, "Low-rank bilinear pooling for fine-grained classification," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 365–374.

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[15] Jie Hu, Li Shen, and Gang Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7132–7141.

[16] Xianglong Zeng, Haiquan Chen, Yuan Luo, and Wenbin Ye, "Automated diabetic retinopathy detection based on binocular siamese-like convolutional neural network," *IEEE Access*, vol. 7, pp. 30744–30753, 2019.

[17] Zhe Wang, Yanxin Yin, Jianping Shi, Wei Fang, Hongsheng Li, and Xiaogang Wang, "Zoom-in-net: Deep mining lesions for diabetic retinopathy detection," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2017, pp. 267–275.

[18] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.

[19] Yandong Wen, Kaipeng Zhang, Zhifeng Li, and Yu Qiao, "A discriminative feature learning approach for deep face recognition," in *European conference on computer vision*. Springer, 2016, pp. 499–515.

[20] "Kaggle: Diabetic Retinopathy Detection," https://www.kaggle.com/c/diabetic-retinopathy-detection, [Online; accessed 20-May-2020].

[21] Christopher M Bishop, *Pattern recognition and machine learning*, springer, 2006.

[22] Karimollah Hajian-Tilaki, "Receiver operating characteristic (roc) curve analysis for medical diagnostic test evaluation," *Caspian journal of internal medicine*, vol. 4, no. 2, pp. 627, 2013.