DRAN: Densely Reversed Attention based Convolutional Network for Diabetic Retinopathy Detection

Cam-Hao Hua[1], Thien Huynh-The[2], and Sungyoung Lee[1]

Abstract— **Diabetic Retinopathy (DR), the complication leading to vision loss, is generally graded according to the amalgamation of various structural factors in fundus photography such as number of microaneurysms, hemorrhages, vascular abnormalities, etc. To this end, Convolution Neural Network (CNN) with impressively representational power has been exhaustively utilized to address this problem. However, while existing multi-stream networks are costly, the conventional CNNs do not consider multiple levels of semantic context, which suffers from the loss of spatial correlations between the aforementioned DR-related signs. Therefore, this paper proposes a Densely Reversed Attention based CNN (DRAN) to leverage the learnable integration of channel-wise attention at multi-level features in a pretrained network for unambiguously involving spatial representations of important DR-oriented factors. Consequently, the proposed approach gains a quadratic weighted kappa of 85.6% on Kaggle DR detection dataset, which is competitive with the state-of-the-arts.**

I. INTRODUCTION

Diabetic Retinopathy (DR), the complication developed by long-term suffering from diabetes mellitus, is one of the leading causes of vision impairment and blindness [1]. Therefore, accurately recognizing the severity of DR (no, mild, moderate, severe, and proliferative DR) from fundus photography enables the ophthalmologists to diagnose and design treatment plan efficiently. Besides that, Convolutional Neural Network (CNN) [2]–[6], a popular deep learning architecture, has recently nailed impressive achievements in various recognition-based topics such as human activity recognition [7], semantic image segmentation [8], DR risk progression identification [1], etc. Accordingly, grading DR scale automatically by CNN becomes an active research domain due to its importance for diagnosis and treatment.

Basically, the DR severity scales are determined by the amalgamation of various structural factors inside the color fundus images such as number of microaneurysms, hemorrhages, neural degeneration, vascular abnormalities. Hence, there are two major CNN-based approaches proposed in the literature for expressing potentially DR-oriented features: (i) utilization of conventional CNN architecture built for classification problem [9]–[13] and (ii) integration of additional network streams [14]–[16]. As for the first group, the authors of [9], [10], [11], and [12] employed 11-, 17-, 18-, and 20-layer CNNs (which are commonly designed by sequential pipeline of convolution, rectified linear unit (*ReLU*) activation, max/average pooling, Fully Connected (*FC*) layers), respectively, for classifying corresponding DR grades. Meanwhile, the light-weight Inception-v3 architecture [3] is utilized for transfer learning in such DR detection domain [13]. Regarding the second group, a triplet of sub-CNNs [14], i.e., Main, Attention, and Crop Networks, was introduced for exhaustively examining multiple clinical details existing in the fundus photography. On the other hand, the methodology in [15] introduced two stages of patch-level learning to comprehensively acquire fine-to-coarse details at multiple scales from the raw fundus photography for DR grade prediction. Besides that, the authors in [16] slightly modified the original ResNet-18 [4] by involving an additional attention stream to enhance inter-class discrimination.

It can be realized that the existing multi-stream networks are subject to costly computation while the remaining models do not involve multiple levels of semantic context in the constructed CNN. Particularly, the fact that multiple down-sampling stages during feedforward process of the CNN leads to the loss of certain spatial correlations between the aforementioned DR-related signs, which are hardly encoded along depth dimension. Thus, we hypothesize that only taking into account highest-level features for the classifier is insufficient in terms of predicting DR severity level.

Based on those observations, a Densely Reversed Attention based CNN (DRAN) is proposed to leverage the learnable integration of channel-wise attention at multi-level features in a pretrained CNN, which allows accomplishing superior recognition performance in a cost-effective way. In concrete, given that informative features are channel-wisely encoded from shallow to deep layers, we opt for densely embedding such semantically-rich details into the finer-grained patterns by the attention extractors (which are inspired from [6]) in a reversed manner. As a consequence, the attachment of the densely reversed attention (DRA) stream into the backbone CNN enables spatial representations of important DR-oriented factors to be comprehensively involved in the final prediction of the severity grade. Finally, we evaluate the proposed DRAN using Kaggle DR detection dataset [17], of which the experimental results in terms of quadratic weighted kappa (QWK) are competitive with the state-of-the-arts.

*This research was supported by the MSIT (Ministry of Science and ICT), Korea, under the ITRC (Information Technology Research Center) support program (IITP-2017-0-01629) supervised by the IITP (Institute for Information & communications Technology Promotion). This work was supported by Institute for Information & communications Technology Promotion (IITP) grant funded by the Korea government (MSIT) (No.2017-0-00655); NRF-2016K1A3A7A03951968; NRF-2019R1A2C2090504.

[1] The Department of Computer Science and Engineering, Kyung Hee University, Gyeonggi-do, 446-701, South Korea hao.hua@oslab.khu.ac.kr, sylee@oslab.khu.ac.kr

[2] ICT Convergence Research Center, Kumoh National Institute of Technology, Gumi, South Korea thienht@kumoh.ac.kr

978-1-7281-1990-8/20/$31.00 ©2020 IEEE

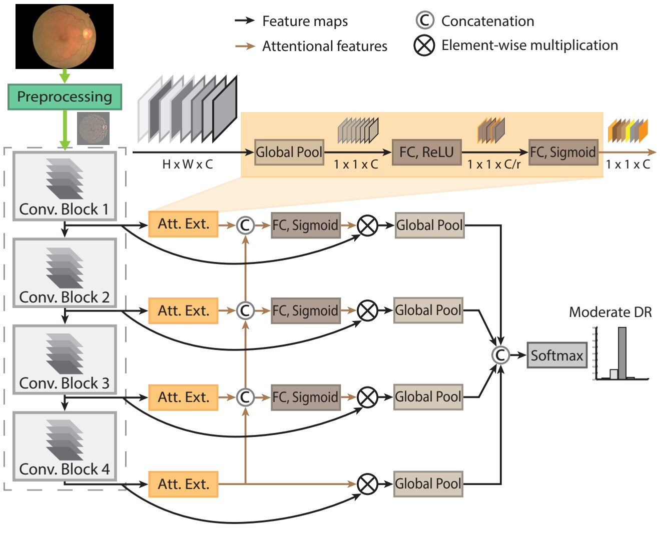

Fig. 1. Architecture of the proposed DRAN for DR severity classification. Note that 'Global Pool', 'Conv. Block', and 'Att. Ext.' represent *Global Pooling* layer, block of multiple convolutional layers, and Attentional Features Extractor, respectively.

II. METHODOLOGY

In general, the proposed architecture is constructed by a backbone CNN associated with the stream of DRA. As demonstrated in Fig. 1, convolution blocks in the dashed box represent the fundamental components of the backbone CNN while the remaining portion manifests the attention-based stream of aggregating multi-level features for the classification of DR severity grade.

A. Backbone CNN

Initially, we preprocess the raw fundus images by re-scaling to a predefined radius and then subtracting local average color for suppressing the diverse difference of illumination and resolution in the dataset [18]. Then, different backbone networks, i.e., VGG [2], ResNet [4], and DenseNet [5], are applied to demonstrate the flexibility of the proposed DRA stream regarding various representational capacities. In these classification networks, layers at the convolution block acquire and represent features at a particular scale with respect to a semantic degree. For instance, both ResNet and DenseNet possess four fundamental convolution blocks (see Fig. 1), of which the strides of final outputs are 4, 8, 16, and 32 with respect to the input's spatial resolution, respectively. Accordingly, for the purpose of securing the increment of computational burden reasonably, only the ultimate outputs of the above-described convolution blocks are involved to the learnable DRA stream. Meanwhile, as for the VGG architecture, outputs of *ReLU* layers followed by the last four max-pooling layers, which also have same strides as mentioned before, are taken into account for next stages.

B. Attentional Features Extractor (Att. Ext.)

This section further characterizes the channel-wise attentional features extractor inspired from Hu *et al.* [6]. In general, the primary objective of the attentional mechanism is to facilitate informative features to be attended more intensively in the learning procedure. Originally, the feature responses are self-recalibrated through the global context acquired across corresponding channels [6]. However, in our model, the cross-channel details of the selected feature maps are employed not only to enrich their own representational ability but also to integrate semantically-richer information into lower-level counterparts, which will be described at next sub-section.

As manifested by the 'Att. Ext.' blocks in Fig. 1, each channel of the considered feature map of size $H \times W \times C$ is spatially averaged to generate a C-length vector at first. Let $\mathbf{F}_i$ be the feature maps inferred by the above-mentioned convolution blocks, where $i = 1, 2, 3, 4$ corresponds to the feedforward route in the core CNN. That means, $\mathbf{F}_1$ and $\mathbf{F}_4$ represent the lowest- and highest-level feature maps of interest, respectively. Then, the computed vectors, defined as $\boldsymbol{g}_i \in \mathrm{R}^{\mathrm{C}}$, manage details of cross-channel dependencies as follows

$$\boldsymbol{g}_i = [g_1(\mathbf{F}_i), \ldots, g_c(\mathbf{F}_i), \ldots, g_C(\mathbf{F}_i)]^T \qquad (1)$$

where $g_c(\mathbf{F}_i)$ refers to as the *Global Pool* operator that handles the c^{th} channel of the feature maps $\mathbf{F}_i$ through following formulation

$$g_c(\mathbf{F}_i) = \frac{1}{H \times W} \sum_{h=1}^{H} \sum_{w=1}^{W} \mathbf{F}_{i h,w,c} \qquad (2)$$

where $h = 1, \ldots, H$ and $w = 1, \ldots, W$ are pixel coordinates in the feature maps $\mathbf{F}_i$. Consequently, channel-wise semantic details are encoded into the vectors $\boldsymbol{g}_i \in \mathrm{R}^{\mathrm{C}}$.

Afterwards, to exploit the correspondingly underlying cross-channel dependencies, the vectors $\boldsymbol{g}_i$ are fed into two *FC* layers with a *ReLU* layer in the middle. Notably, the capacity of these learnable layers is equivalent to C/r and C, respectively, where r is the compression ratio fixed at 16 in this model for the reduction of computational cost. These learning processes can be described by following expression

$$\boldsymbol{g}_{int_i} = \mathbf{W}^T_{fc_2}\left(ReLU(\mathbf{W}^T_{fc_1}\boldsymbol{g}_i + \mathbf{B}_{fc_1})\right) + \mathbf{B}_{fc_2} \qquad (3)$$

where $\{\mathbf{W}_{fc_1} \in \mathrm{R}^{\mathrm{C}\times\frac{\mathrm{C}}{\mathrm{r}}}, \mathbf{B}_{\mathrm{fc}_1} \in \mathrm{R}^{\frac{\mathrm{C}}{\mathrm{r}}}\}$ and $\{\mathbf{W}_{fc_2} \in \mathrm{R}^{\frac{\mathrm{C}}{\mathrm{r}}\times\mathrm{C}}, \mathbf{B}_{\mathrm{fc}_2} \in \mathrm{R}^{\mathrm{C}}\}$ are learnable parameters of the first and second *FC* layers, respectively, and $\boldsymbol{g}_{int_i} \in \mathrm{R}^{\mathrm{C}}$ are the yielded channel-wise attention feature vectors. Next, we adopt *Sigmoid* activation function $\sigma(.)$ to rescale the vectors $\boldsymbol{g}_{int_i}$'s entries to the range of $(0, 1)$. As a result, channel-wisely attentional feature vectors, namely $\boldsymbol{g}_{att_i} \in \mathrm{R}^{\mathrm{C}}$, are acquired for enriching informative representations of the feature maps typically involved from the backbone network.

C. Stream of Densely Reversed Attention

To this end, the attentional feature vectors $\boldsymbol{g}_{att1}$, $\boldsymbol{g}_{att2}$, $\boldsymbol{g}_{att3}$, and $\boldsymbol{g}_{att4}$ (illustrated by outbound brown arrows of Att. Ext. modules in Fig. 1) corresponding to the chosen feature maps $\mathbf{F}_1$, $\mathbf{F}_2$, $\mathbf{F}_3$, and $\mathbf{F}_4$, respectively, are obtained from the Attentional Features Extractor. Different backbone CNNs make the vectors $\boldsymbol{g}_{att_i}$ have dissimilar lengths of

TABLE I

LENGTHS C OF EXTRACTED ATTENTIONAL FEATURE VECTORS $\boldsymbol{g}_{att_i}$ ($i = 1, 2, 3, 4$) WITH REGARD TO VARIOUS BACKBONE CNNS.

Backbone CNN	$\boldsymbol{g}_{att1}$	$\boldsymbol{g}_{att2}$	$\boldsymbol{g}_{att3}$	$\boldsymbol{g}_{att4}$
VGG-16 [2]	128	256	512	512
ResNet-101 [4]	256	512	1024	2048
DenseNet-161 [5]	384	768	2112	2208

C as shown in Table I. Obviously, later $\boldsymbol{g}_{att_i}$ with larger dimension contain more informative context, which can be applied to recalibrate higher-resolution features extracted at preceding convolution blocks. Thus, in order to utilize fine-grained features without contextual ambiguities extensively to the final classifier, it is essential to integrate the depth-wisely semantic details from higher-level features of the backbone network in a reversed style. Moreover, because each attentional feature vector $\boldsymbol{g}_{att_i}$ possesses a particular degree of descriptive statistics, we opt for an exhaustive combination manner to embed those multi-level semantic features comprehensively, by which spatial appearances of coarser-level features are refined more flexibly.

As demonstrated in Fig. 1, following operations are performed after extracting attentional features

$$\begin{aligned}
\mathbf{F}_{att4} &= \mathbf{F}_4 \otimes \boldsymbol{g}_{att4} \\
\mathbf{F}_{att3} &= \mathbf{F}_3 \otimes \sigma(\mathbf{W}^T_{fc_{33}}\left(\mathcal{C}[\boldsymbol{g}_{att3}, \boldsymbol{g}_{att4}]\right)) \\
\mathbf{F}_{att2} &= \mathbf{F}_2 \otimes \sigma(\mathbf{W}^T_{fc_{32}}\left(\mathcal{C}[\boldsymbol{g}_{att2}, \boldsymbol{g}_{att3}, \boldsymbol{g}_{att4}]\right)) \\
\mathbf{F}_{att1} &= \mathbf{F}_1 \otimes \sigma(\mathbf{W}^T_{fc_{31}}\left(\mathcal{C}[\boldsymbol{g}_{att1}, \boldsymbol{g}_{att2}, \boldsymbol{g}_{att3}, \boldsymbol{g}_{att4}]\right))
\end{aligned} \tag{4}$$

where $\otimes$ denotes the element-wise multiplication operator at corresponding channels, $\mathcal{C}[.]$ means the vector concatenation procedure, and $\{\mathbf{W}_{fc_{33}}, \mathbf{W}_{fc_{32}}, \mathbf{W}_{fc_{31}} \in \mathrm{R}^{\mathrm{D}\times\mathrm{C}}\}$ are the learnable parameters of *FC* layers. Clearly, C is the channel size of the considered feature map $\mathbf{F}_i$ while the value of D depends on dimension of the output concatenated by attentional vectors $\boldsymbol{g}_{att_i}, \boldsymbol{g}_{att_{i+1}}, \dots, \boldsymbol{g}_{att_4}$. For example, in the case of employing ResNet-101 as the backbone CNN, values D of $\mathbf{W}_{fc_{33}}$, $\mathbf{W}_{fc_{32}}$, and $\mathbf{W}_{fc_{31}}$ are 3072, 3584, and 3840, respectively (following the related lengths of C in Table I). Furthermore, with such kind of densely reversed mechanism, features acquired at shallower layers can still gain semantically-richer details for a more efficient contribution to the detection performance.

Then, the re-weighted feature maps $\mathbf{F}_{att_i}$ are continuously processed by the *Global Pool* modules in (2) and a subsequent concatenation operator for another dense aggregation of globally-essential information at multiple representational scales as follows

$$\mathbf{F}_{dr} = \mathcal{C}\left[g_1(\mathbf{F}_{att1}), g_2(\mathbf{F}_{att2}), g_3(\mathbf{F}_{att3}), g_4(\mathbf{F}_{att4})\right] \tag{5}$$

where $\mathbf{F}_{dr}$ stands for the finalized DR-oriented features of our DRAN model. Finally, a Softmax classifier is adopted to classify the corresponding DR grade based on a predefined number of severity categories.

TABLE II

QWK ON DR KAGGLE [17] VALIDATION SET WITH DIFFERENT TYPES OF BACKBONE CNN AND ATTENTION-EMBEDDED SCHEME.

Backbone CNN	Strategy			QWK (%)
	Baseline	AN	DRAN	
VGG-16 [2]	✓			84.9
		✓		85.4
			✓	86.3
ResNet-101 [4]	✓			85.4
		✓		86.1
			✓	86.7
DenseNet-161 [5]	✓			85.5
		✓		86.5
			✓	86.9

III. EXPERIMENTS

Notably, the Institution's Ethical Review Board approved all experimental procedures involving human subjects.

A. Benchmark Dataset

Kaggle DR detection dataset [17] is used to evaluate the proposed methodology. It contains approximately 35,000 training, 11,000 validation (public test), and 43,000 private test images, which are categorized into five severity scales as aforementioned. Remarkably, all the color fundus images are supplied by EyePACS, a retinopathy screening platform.

B. Implementation Details

The proposed model and relevant evaluations are implemented using Pytorch [19]. Similar to existing work, augmentation techniques such as randomly cropping to size of 448 × 448, horizontal and vertical flipping, and arbitrary rotation are also applied to the training batches. In addition, weight decay of 5e-4 is employed to generalize the proposed model intensively. As for the optimization phase, we use stochastic gradient descent with initial learning rate of 0.005 and momentum of 0.9. During training, the learning rate decreases by half after every 20 epochs. In total, we execute 80 training epochs on one NVIDIA GTX 1080TI GPU.

C. Ablation Study

To this end, we conduct three experimental strategies, i.e., Baseline, AN, and DRAN with respect to each backbone CNN, to benchmark the effectiveness of the proposed architecture. Notably, the Baseline means that the pretrained backbone CNN is fine-tuned end-to-end. Meanwhile, the AN refers to as the additional engagement of attentions at the end of basis convolution blocks but without the densely reversed stream. Finally, QWK measures of these strategies on the validation set are presented in Table II.

Apparently, AN and DRAN with respect to different backbone CNNs show superior performance over the corresponding Baselines with higher QWK of 0.5-1.0% and 1.3-1.4%, respectively. It is argued that the utmost reason comes from the benefit of utilizing attention strategy at different scales followed by a concatenation operator, which is able to maintain features of different DR-oriented factors throughout various representational levels in the backbone networks.

TABLE III

COMPARISON OF QWK ON KAGGLE DR [17] TEST SET.

Approach	QWK (%)
11-layer CNN [9]	76.7
SI2DRNet-v1 [12]	80.4
14-layer CNN [11]	85.1
Zoom-in-Net [14]	85.7
DRAN (VGG-16)	**84.9**
DRAN (ResNet-101)	**85.4**
DRAN (DenseNet-161)	**85.6**

Moreover, compared to AN, DRAN is capable of boosting the QWK more 0.4% (in the case that VGG-16 is the backbone CNN), 0.6% (ResNet-101), and 0.9% (DenseNet-161). Accordingly, these improvements imply the advantage of exhaustively embedding deeper attentional feature vectors to recalibrate shallower features. In concrete, as previously mentioned in Section II-C, the operator of dense concatenation in reversed manner further enables low-level features (which contain spatially informative details of DR-related factors) to be extensively involved in the final classifier.

Also, it can be observed that higher capacity of backbone CNN is able to produce better prediction performance. Particularly, DRAN equipped with backbones of ResNet-101 and DenseNet-161 increases the QWK by 0.4% and 0.6%, respectively, compared to that using VGG-16.

D. Comparisons with State-of-the-arts

For the comparison with other methods, the proposed DRAN with three different backbone CNNs is evaluated by the Kaggle DR test set. The benchmark results reported in Table III show that of the proposed approach is competitive with the state-of-the-arts. Specifically, although Zoom-in-Net [14] achieves highest QWK, the superiority over the proposed DRAN (which utilizes DenseNet-161 as backbone network) is insignificant (0.1%). It should be noted that their results are obtained from the expensive triplet of sub-CNNs and ensemble learning. In a nutshell, thanks to the dense attention of higher-level depth-wise features to spatially-rich details at lower levels in reversed scheme, which allows early involvement of finely-patterned features, the proposed architecture can achieve an impressive performance on such challenging dataset.

IV. CONCLUSIONS

In this paper, a CNN with densely reversed attention, i.e., DRAN, has been introduced to effectively address the DR detection problem. Concretely, the proposed architecture enables finely-patterned (high-resolution) feature maps, which possess well-organized representation of DR-oriented factors, to be smoothly combined with the semantically-rich (low-resolution) counterparts for a better recognition performance. The key for such utilization is the dense embedding of a channel-wise attention mechanism into a pretrained CNN in reversed manner. As a consequence, experimental results have demonstrated consistent improvements of the proposed model, which is constructed from a baseline network to the involvement of multi-scale attentional features extractor and a further stream of densely reversed attention. In the future, DRAN can be potentially extended for tackling other disease recognition problems besides detecting DR severity scale.

REFERENCES

[1] C.-H. Hua, T. Huynh-The, K. Kim, S.-Y. Yu, T. Le-Tien, G.H. Park, J. Bang, W.A. Khan, S.-H. Bae, and S. Lee, "Bimodal learning via trilogy of skip-connection deep networks for diabetic retinopathy risk progression identification," *International Journal of Medical Informatics*, vol. 132, pp. 103926, 2019.

[2] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *CoRR*, vol. abs/1409.1556, 2014.

[3] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016, pp. 2818–2826.

[4] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016, pp. 770–778.

[5] G. Huang, Z. Liu, L. v. d. Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017, pp. 2261–2269.

[6] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[7] T. Huynh-The, C.-H. Hua, and D. Kim, "Encoding pose features to images with data augmentation for 3d action recognition," *IEEE Transactions on Industrial Informatics*, pp. 1–1, 2019.

[8] C.-H. Hua, T. Huynh-The, and S. Lee, "Retinal vessel segmentation using round-wise features aggregation on bracket-shaped convolutional neural networks," in *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, July 2019, pp. 36–39.

[9] M. C. A. Trivino, J. Despraz, J. A. L. Sotelo, and C. A. Peña, "Deep learning on retina images as screening tool for diagnostic decision support," *CoRR*, vol. abs/1807.09232, 2018.

[10] J. Torre, A. Valls, and D. Puig, "A deep learning interpretable classifier for diabetic retinopathy disease grading," *Neurocomputing*, 2019.

[11] S. M. S. Islam, Md M. Hasan, and S. Abdullah, "Deep learning based early detection and grading of diabetic retinopathy using retinal fundus images," *CoRR*, vol. abs/1812.10595, 2018.

[12] Y. Chen, T. Wu, W. Wong, and C. Lee, "Diabetic retinopathy detection based on deep convolutional neural networks," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, April 2018, pp. 1030–1034.

[13] J. Sahlsten, J. Jaskari, J. Kivinen, L. Turunen, E. Jaanio, K. Hietala, and K. Kaski, "Deep learning fundus image analysis for diabetic retinopathy and macular edema grading," *Scientific Reports*, vol. 9, no. 1, pp. 10750, 2019.

[14] Z. Wang, Y. Yin, J. Shi, W. Fang, H. Li, and X. Wang, "Zoom-in-net: Deep mining lesions for diabetic retinopathy detection," in *Medical Image Computing and Computer Assisted Intervention, MICCAI 2017*, 2017, pp. 267–275.

[15] L. Zhou, Y. Zhao, J. Yang, Q. Yu, and X. Xu, "Deep multiple instance learning for automatic detection of diabetic retinopathy in retinal images," *IET Image Processing*, vol. 12, no. 4, pp. 563–571, 2018.

[16] P. Junjun, Y. Zhifan, S. Dong, and Q. Hong, "Diabetic retinopathy detection based on deep convolutional neural networks for localization of discriminative regions," in *2018 International Conference on Virtual Reality and Visualization (ICVRV)*, Oct 2018, pp. 46–52.

[17] "Kaggle: Diabetic retinopathy detection," https://www.kaggle.com/c/diabetic-retinopathy-detection.

[18] B. Graham, "Kaggle diabetic retinopathy detection competition report," 2015.

[19] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, "Automatic differentiation in PyTorch," in *NIPS Autodiff Workshop*, 2017.