Frontiers in Signal Processing, Vol. 3, No. 4, October 2019
https://dx.doi.org/10.22606/fsp.2019.34003

Diabetic Retinopathy Detection Based on Deep Learning

Qiongyao Liang, Xiangkui Li and Yansong Deng*

Key Laboratory of Electronic and Information Engineering (Southwest Minzu University), China
State Ethnic Affairs Commission, China
Email: 1945528511@qq.com

Abstract Recent years, deep learning in the image identification has made great progress, showing good application prospects in medical image reading. Diabetic Retinopathy (DR) is an eye disease due to diabetes, which is the most ordinary cause of blindness. Traditional diabetic retinopathy detection is a manual and time-consuming and labor-intensive process, which requires a highly experienced clinician to examine and evaluate the digital color fundus photos of the retina. Therefore, it is crucial to use the deep learning technique to achieve automatic detection of diabetic retinopathy. In this paper, we proposed a diabetic retinopathy detection method based on deep learning and proposed a network structure named multi-self-attention. At first, the image features were extracted through the InceptionV3 model, and then the feature maps was directly generated. Secondly, the feature maps, which can reflect condition of retina, will be input into multi-self-attention network structure, to calculate multi-self-attention feature. Finally, by convolutional layer and fully connected layer, the stage results about diabetic retinopathy will be obtained. With the experiments in TensorFlow framework , the effectiveness of multi-self-attention network structure for feature extraction and classification is proved.

Keywords: Deep learning, multi-self-attention mechanism, Diabetic Retinopathy (DR), InceptionV3 model.

1 Introduction

The concept of deep learning [1] was proposed by Hinton et al. in 2006 and it is one of the technologies and research areas of machine learning. Deep learning realizes artificial intelligence in computer systems by constructing an artificial neural network with hierarchical structure. At present, deep learning technology has been successfully applied in the fields of image identification, voice recognition [2,3] , natural language processing [4,5,6,7], etc., and is gradually applied to the medical field. The image identification technology in deep learning has a good application prospect in the completion of medical image reading. Many technology companies are trying to use deep learning techniques to reduce the doctor's work intensity, improve work efficiency, and make up for the lack of medical resources.

In modern people's life, diabetes as a high-risk disease has been paid more and more attention. Diabetic retinopathy is a common eye disease in diabetic patients and is the main cause of blindness in the population. Early detection of diabetic retinopathy protects patients from losing their vision [8]. At present, traditional diabetic retinopathy detection is a manual and time-consuming and labor-intensive process, which requires a highly experienced clinician to examine and evaluate the digital color fundus photos of the retina [9]. The expertise and equipment required for the detection of diabetic retinopathy is extremely scarce in places with high prevalence of diabetes. As the number of people with diabetes continues to grow, the infrastructure to prevent retinopathy will become even more inadequate. Therefore, the automatic detection of diabetic retinopathy is particularly urgent. Therefore, the development of an effective retinopathy detection system has great practical significance.

Recently, automated systems for detecting diabetic retinopathy stages have widely explored and gained a lot of acceptances [10,11,12,13,14,15,16]. Azzopardi et. al [17] proposed a blood vessels detection method based on Bar-Combination of Shifted Filter Responses BCOSFIRE approach. Prasad et.al [18] proposed a method to detect blood vessels, exudates and microaneurysms using Haar Wavelet transform and Principal Component Analysis techniques. Enrique V.Carrera [8] proposed a computerassisted diagnosis based on the digital processing of retinal images to automatically classify the grade of non-proliferative diabetic retinopathy at any retinal image. These methods are performed in two steps, first detecting the amount of bleeding and permeate and then giving the results of the lesion after synthesis. The method proposed in this paper input the images, it automatically performs feature extraction and feature processing to directly obtain the classification result. The processing is simpler and the recognition effect is similar.

In this paper, we proposed a diabetic retinopathy detection method based on deep learning. Through experiments, the effectiveness of multi-self-attention network model for feature extraction and classification is proved.

The paper is organized as follow: In Section 2, the main idea of our proposed approach is presented. Section 3 is based on experimental analysis and results. In Section 4, we concluded our proposed DR detection method.

2 Methodology

Diabetic retinopathy is a complication of diabetes. In long-term high glucose environment, retinal vessels will produce a series of pathological changes, such as micro aneurysms, hard exudate and soft exudate. According to the severity of lesions, DR can be divided into No DR, Mild DR, Moderate DR, Severe DR, Proliferative DR in five stages [8]. About 50% patients with diabetes have some stages of the disease, resulting in visual impairment or blindness. How to distinguish between these stages accurately is a complex problem to be solved urgently.

Diabetes mainly affects the entire retina by affecting the blood vessels in the retina. According to the changes in various characteristics of the blood vessels in the lesion image, the texture features of the retinal image and the color characteristics of the retinal image are taken as the main features of the detection of diabetic retinopathy.

In this paper, drawing on the thoughts of the attention mechanism in deep learning and the self-attention generative adversarial network [19] proposed by Han Zhang and Ian Goodfellow in 2018, we proposed a network structure named multi-self-attention. The retinal image features extracted by the InceptionV3 network model were processed by a network model of multiple-self-attention mechanisms, and finally the results of diabetic retinopathy detection are obtained. The processing of the method proposed in this paper is shown in Figure 1. The details for the methods used in each stage are explained on the following sections.

2.1 Feature Extraction

The traditional feature extraction method manually preprocesses the images, which reduces the accuracy of image identification. The feature extraction algorithm based on deep learning builds a multi-layer network, on which the computer automatically learns and obtains the implicit relationship of the data, extracts higher-dimensional and more abstract relationship, which makes the learned image features more expressive power [20]. In deep learning, image feature extraction is very important for image classification, which largely determines the quality of image classification results.

The deep learning models used for feature extraction in image identification mainly include deep belief network (DBN), convolutional neural network (CNN), recurrent neural network (RNN), and capsule network (CapsNet). The DBN can reflect the similarity of the similar data, and its classification accuracy is not high. The input data has a translation that is not deformed.The CNN models are more generalized and have fewer training parameters. The pooling operation reduces the spatial dimension of the network, and the translation of the input data is not required to be deformed. Gradient dissipation problems are prone to occur. The spatial relationship is poorly recognized. The recognition ability is low after the object is rotated a lot. The RNN can model the sequence content. There are many parameters that need to be trained, and gradient dispersal or gradient explosion problems are prone to occur. Does not have feature learning ability. CapsNet solves the difference in the spatial relationship recognition of the CNN model and the low recognition ability after the object is rotated a lot. The network structure of the model is shallow, and the accuracy in image recognition classification is still far from the current popular CNN model.

According to the characteristics of each network model, the InceptionV3 model with good performance in CNN is selected for feature extraction. InceptionV3 uses a smaller convolution kernel to reduce

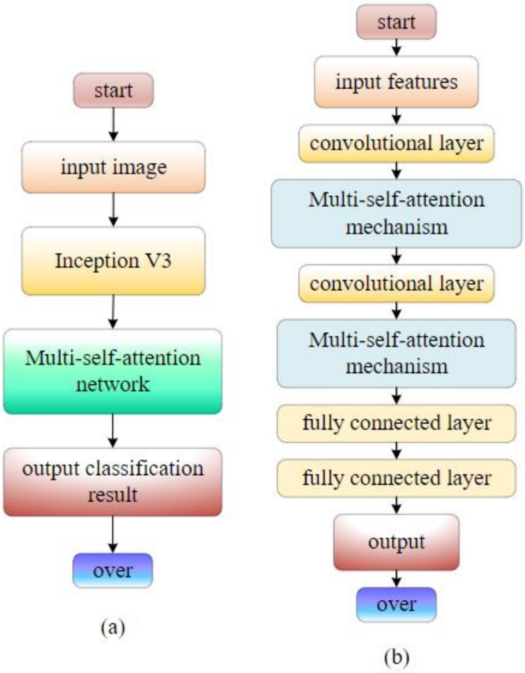


Figure 1. Flow charts. (a) The processing of the method proposed in this paper. (b) The design flow of the multi-self-attention network

parameters while maintaining performance. Meanwhile, InceptionV3 also converts full connection or even general convolution into sparse connections, reducing computational complexity and saving computing resources. This is also an important reason why we decided to use the InceptionV3 model to extract image features.

In order to get the better multi-self-attention features, we modify the InceptionV3 model and use InceptionV3 transfer learning process the extracted features to make the output different. The input images in InceptionV3 model is processed by using the bottleneck layer in transfer learning, and the extracted feature graph is represented as $8 \times 8 \times 2048$.

2.2 Multi-self-attention Network Model

The main purpose of designing the multi-self-attention network model is to process the convolutional feature map extracted by the improved InceptionV3 model, then update the weights, and finally obtain the classification results. The design flow of the multi-self-attention network model is shown in Figure 1.

Previous models rely heavily on convolution operations to simulate the dependencies between different regions of the image. Each convolution operation has a local receptive field, and the long-range correlation between features often goes through several convolutional layers. It will show up at the expense of efficiency and computation. In contrast, the self-attention model can take into account the three aspects of simulating long-distance dependence, efficiency and computational complexity, which is a more appropriate choice [19].

The self-attention mechanism is adopted in the multi-self-attention network model proposed in this paper, which is more advantageous than the traditional algorithm in dealing with the tiny details of the image. The traditional network model discards important but small details in the image after multiple convolutions, which has a serious impact on the classification results. Traditional convolution is a function calculation for local points on low-level feature maps, but the self-attention mechanism can use information from all feature locations to process details.

The internal structure of the multi-self-attention network model is shown in Figure 2. We use InceptionV3 model to extract the image features, getting the convolutional feature map $\boldsymbol{x}$, and after

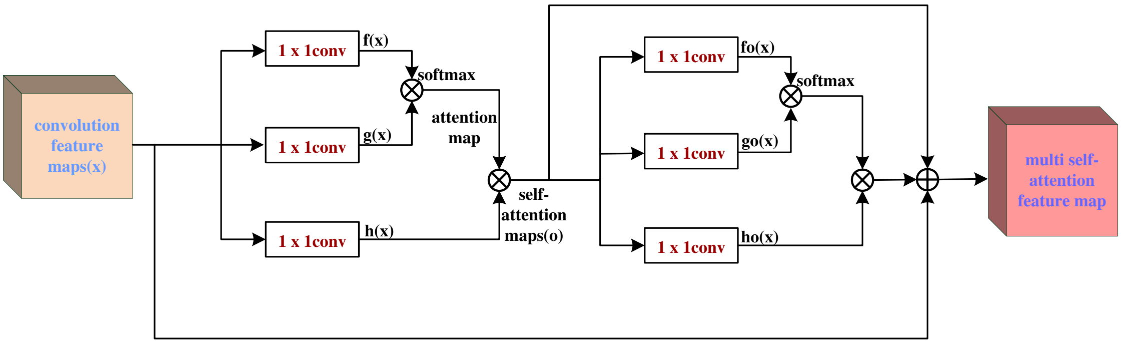


Figure 2. The proposed multi-self-attention mechanism.

1×1 convolution, we get $\boldsymbol{f}(\boldsymbol{x})$, $\boldsymbol{g}(\boldsymbol{x})$, $\boldsymbol{h}(\boldsymbol{x})$, after $\boldsymbol{f}(\boldsymbol{x})$ and $\boldsymbol{g}(\boldsymbol{x})$ were matrix multiplied, the attention feature was obtained by the softmax function. The attention feature and $\boldsymbol{h}(\boldsymbol{x})$ were matrix multiplied to obtain the self-attention feature $\boldsymbol{o}$. The self-attention calculation process can be freely selected. It has been proved by experiments that the calculation of multi-self-attentions network has a positive effect on the update of weights.

The image features from the previous hidden layer $\boldsymbol{x} \in \mathbb{R}^{C\times N}$ are first transformed into two feature spaces f, g to calculate the attention, where $\boldsymbol{f}(\boldsymbol{x}) = \boldsymbol{W_f x}, \boldsymbol{g}(\boldsymbol{x}) = \boldsymbol{W_g x}$

$$\beta_{j,i} = \frac{\exp(s_{ij})}{\sum_{i=1}^{W} \exp(s_{ij})}, \quad s_{ij} = \boldsymbol{f}(\boldsymbol{x}_i)^T \boldsymbol{g}(\boldsymbol{x}_j) \tag{1}$$

and $\beta_{j,i}$ indicates the extent to which the model attends to the i-th location when synthesizing the j-th region. Then the output of the attention layer is $\boldsymbol{o} = (\boldsymbol{o}_1, \boldsymbol{o}_2, \ldots, \boldsymbol{o}_j, \ldots, \boldsymbol{o}_N) \in \mathbb{R}^{C\times N}$, where,

$$\boldsymbol{o}_j = \sum_{i=1}^{N} \beta_{j,i} \boldsymbol{h}(\boldsymbol{x}_i), \quad \boldsymbol{h}(\boldsymbol{x}_i) = \boldsymbol{W_h x_i} \tag{2}$$

In the above formula, $\boldsymbol{W_g} \in \mathbb{R}^{\overline{C}\times C}$, $\boldsymbol{W_f} \in \mathbb{R}^{\overline{C}\times C}$, $\boldsymbol{W_h} \in \mathbb{R}^{C\times C}$ are the learned weight matrices, which are implemented as 1×1 convolutions. We use $\overline{C} = C/8$ in all our experiments.

After getting the self-attention feature o, like x, the same operation is performed by o, and oo is obtained. Therefore, the final output is given by,

$$\boldsymbol{Y} = \boldsymbol{o} + \boldsymbol{oo} + \boldsymbol{x} \tag{3}$$

This allows the network to first rely on the cues in the local neighborhood - since this is easier - and then gradually learn to assign more weight to the non-local evidence. The intuition for why we do this is straightforward: we want to learn the easy task first and then progressively increase the complexity of the task.

In this network, the attention mechanism can use the information from all feature locations to process the details, saving important features in the image while retaining more subtle features in the image.

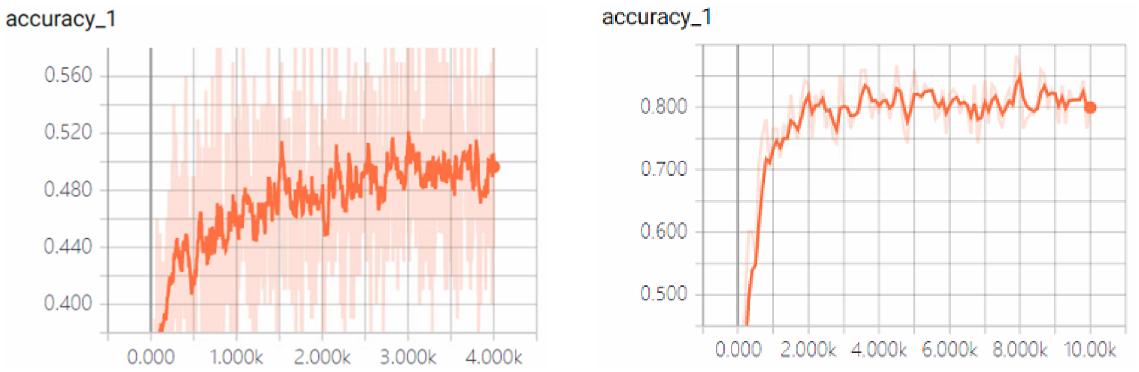
3 Result and Discussion

In this experiment, DR can be divided into No DR, Mild DR, Moderate DR, Severe DR, Proliferative DR in five stages[8]. In order to balance the number of training data set for each stage, the training data set for each stage is 5000 images, the validation set is 500, and the test set is 500.

In the multi-self-attention network model, we can set the number of times the self-attention network model was used. When $n = 0$, it means that the extracted diabetic retinal image features have not been processed by the self-attention network, just use the improved InceptionV3 model for classification. When

Table 1. Accuracy of each model in the classification task of retinopathy.

Algorithm name	Accueacy
InceptionV3 model	50.3%
Improved InceptionV3 model ($n = 0$)	84.9%
Multi-self-attention model	86.7%



(a) The InceptionV3 model accuracy.

(b) The accuracy when $n = 2$.

Figure 3. The experimental accuracy of $n = 2$ and original InceptionV3 model.

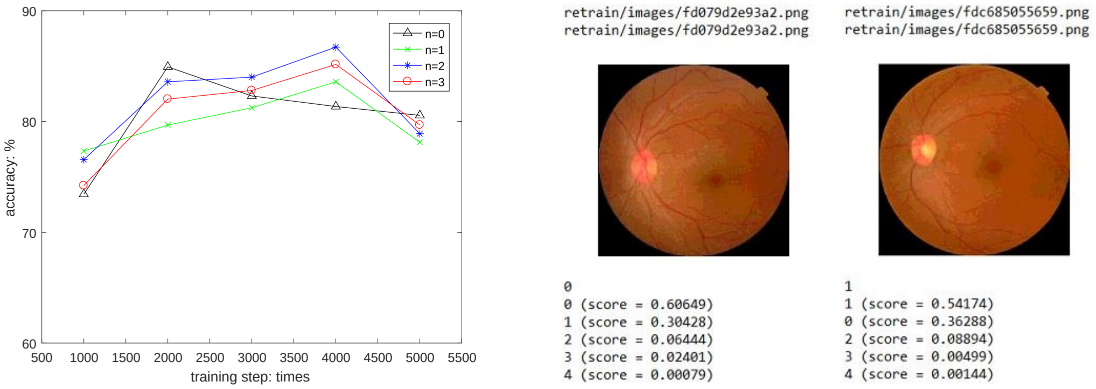
$n = 1$, it means that the process of the self-attention network is performed. When $n = 2$, it means that the process was performed twice.

Table 1 shows the accuracy of each model in the classification task of retinopathy. We can see that using the original InceptionV3 model proposed by Google for training, using the pre-training method, only the last layer of classification training, the accuracy is only 50.3%. The improved InceptionV3 model and the multi-self-attention model proposed in this paper also use the pre-training method, and the results are significantly higher than the original InceptionV3 model. Similarly, in the comparison of the InceptionV3 model of Figure 3 and when $n = 2$, it can be seen that the classification accuracy of $n = 2$ is much higher than that of the original InceptionV3 model.

Table 2. Modifying the training steps, the experimental accuracy of different n.

train step	n	Training acc	Valid acc	Test acc
1000	0	97.6%	72.4%	73.4%
	1	89.8%	78.2%	77.3%
	2	96.8%	75.8%	76.5%
3000	0	100%	81.5%	82.3%
	1	98.3%	80.3%	81.2%
	2	98.4%	83.6%	84.1%
4000	0	100%	80.1%	81.3%
	1	99.0%	80.2%	81.2%
	2	99.8%	84.6%	86.7%

Modifying the training steps, the experimental accuracy of different n is shown in Table 2. When $n = 2$, the classification results are the best, making the accuracy to 86.7%, exceeded the highest 82.3% when $n = 0$. The reason is that when updating the weights in the multi-attention network model, not only the close-range features, but also the long-distance features can be included, and more details can be paid attention to. When n is larger, the more complex the network, the more parameters are trained, and the longer the training time is, the more likely it is to overfit, the lower the classification accuracy.



```
retrain/images/fd079d2e93a2.png
retrain/images/fd079d2e93a2.png
0
0 (score = 0.60649)
1 (score = 0.30428)
2 (score = 0.06444)
3 (score = 0.02401)
4 (score = 0.00079)
```

```
retrain/images/fdc685055659.png
retrain/images/fdc685055659.png
1
1 (score = 0.54174)
0 (score = 0.36288)
2 (score = 0.08894)
3 (score = 0.00499)
4 (score = 0.00144)
```

(a) The test accuracy of different n.

(b) Test result

Figure 4. Experimental accuracy comparison and test results.

In (b) of the Figure 3 shows the change of the accuracy of the validation set when $n = 2$. According to the results of the validation set, modifying parameters setting, the test accuracy is shown in (a) of the Figure 4. It can be seen from the curve in (a) of figure 4 that when the training step is 2000 times, $n = 0$ with higher accuracy. As the training step increasing, the accuracy begins to decrease due to over-fitting. When the training step reaches 3000 times, the self-attention network gradually shows the advantage. When the training step reaches 4000 times, the self-attention network model with $n = 2$ has the highest accuracy. The self-attention model uses more training times to achieve better classification results. The self-attention model uses more training times to achieve better classification results.

In the diabetic retinopathy detection system, the user upload the retina images that meet the requirements and can get the test results. In (b) of Figure 4 shows an example of detecting diabetic retinopathy. The left side is an image without DR, the test result shows that the probability of 60.6% is No DR. The right side is a Mild DR, and the test result shows that the probability of 54.2% is Mild DR. The test results are correct.

4 Conclusion

This paper proposes a method for the effective detection of diabetic retinopathy. This method uses InceptionV3 model to extract features, and inputs the obtained features into a network model based on multi-self-attention mechanism for processing, and then inputs to the convolution layer and fully connected layer to realizes the feature extraction and classification of the retina picture. Experiments show that the multi-self-attention network model proposed in this paper has an advantage in the identification of diabetic retinal images.

The diabetic retinopathy detection system designed in this paper can be used as a useful screening tool for diabetic retinopathy. It can be used to deploy the website to give flexibility to the system. It is more convenient for users, and can also be used as a reference for medical judgment.

For the problem of the quality and quantity of the data set, it is also necessary to further study by transforming the data set and then transforming the existing data set to generate a new data set.

Acknowledgments. This work was supported by the Southwest Minzu University Graduate Innovative Research Project (Master Program CX2018SP290). A special acknowledgement should give to Southwest Minzu University for its experimental conditions and technical support. Thanks to the careful guidance of the teacher and the help of the classmates.

References

1. Y. LeCun, Y. Bengio, and G. Hinton, “Deep learning,” *nature*, vol. 521, no. 7553, p. 436, 2015.
2. S.-H. Fang, Y. Tsao, M.-J. Hsiao, J.-Y. Chen, Y.-H. Lai, F.-C. Lin, and C.-T. Wang, “Detection of pathological voice using cepstrum vectors: A deep learning approach,” *Journal of Voice*, 2018.
3. H. S. Bae, H. J. Lee, and S. G. Lee, “Voice recognition-based on adaptive mfcc and deep learning for embedded systems,” *Journal of Institute of Control, Robotics and Systems*, vol. 22, no. 10, pp. 797–802, 2016.
4. T. Young, D. Hazarika, S. Poria, and E. Cambria, “Recent trends in deep learning based natural language processing,” *ieee Computational intelligenCe magazine*, vol. 13, no. 3, pp. 55–75, 2018.
5. A. Siddhant and Z. C. Lipton, “Deep bayesian active learning for natural language processing: Results of a large-scale empirical study,” *arXiv preprint arXiv:1808.05697*, 2018.
6. S. Revay and M. Teschke, “Multiclass language identification using deep learning on spectral images of audio signals,” *arXiv preprint arXiv:1905.04348*, 2019.
7. B. Marinelli, M. Kang, M. Martini, J. R. Zech, J. Titano, S. Cho, A. B. Costa, and E. K. Oermann, “Combination of active transfer learning and natural language processing to improve liver volumetry using surrogate metrics with deep learning,” *Radiology: Artificial Intelligence*, vol. 1, no. 1, p. e180019, 2019.
8. E. V. Carrera, A. González, and R. Carrera, “Automated detection of diabetic retinopathy using svm,” in *2017 IEEE XXIV International Conference on Electronics, Electrical Engineering and Computing (INTERCON)*. IEEE, 2017, pp. 1–4.
9. Z. Omar, M. Hanafi, S. Mashohor, N. Mahfudz, and M. Muna’im, “Automatic diabetic retinopathy detection and classification system,” in *2017 7th IEEE International Conference on System Engineering and Technology (ICSET)*. IEEE, 2017, pp. 162–166.
10. S. Yu, D. Xiao, and Y. Kanagasingam, “Exudate detection for diabetic retinopathy with convolutional neural networks,” in *2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 2017, pp. 1744–1747.
11. N. Yalçin, S. Alver, and N. Uluhatun, “Classification of retinal images with deep learning for early detection of diabetic retinopathy disease,” in *2018 26th Signal Processing and Communications Applications Conference (SIU)*. IEEE, 2018, pp. 1–4.
12. J. Yadav, M. Sharma, and V. Saxena, “Diabetic retinopathy detection using feedforward neural network,” in *2017 Tenth International Conference on Contemporary Computing (IC3)*. IEEE, 2017, pp. 1–3.
13. S. Suriyal, C. Druzgalski, and K. Gautam, “Mobile assisted diabetic retinopathy detection using deep neural network,” in *2018 Global Medical Engineering Physics Exchanges/Pan American Health Care Exchanges (GMEPE/PAHCE)*. IEEE, 2018, pp. 1–4.
14. R. Shalini and S. Sasikala, “A survey on detection of diabetic retinopathy,” in *2018 2nd International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC)I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC), 2018 2nd International Conference on*, Aug 2018, pp. 626–630.
15. K. K. Palavalasa and B. Sambaturu, “Automatic diabetic retinopathy detection using digital image processing,” in *2018 International Conference on Communication and Signal Processing (ICCSP)*. IEEE, 2018, pp. 0072–0076.
16. L. Li and M. Celenk, “Detection and identification of hemorrhages in fundus images of diabetic retinopathy,” in *BIBE 2018; International Conference on Biological Information and Biomedical Engineering*. VDE, 2018, pp. 1–5.
17. G. Azzopardi, N. Strisciuglio, M. Vento, and N. Petkov, “Trainable cosfire filters for vessel delineation with application to retinal images,” *Medical image analysis*, vol. 19, no. 1, pp. 46–57, 2015.
18. D. K. Prasad, L. Vibha, and K. Venugopal, “Early detection of diabetic retinopathy from digital retinal fundus images,” in *2015 IEEE Recent Advances in Intelligent Computational Systems (RAICS)*. IEEE, 2015, pp. 240–245.
19. H. Zhang, I. Goodfellow, D. Metaxas, and A. Odena, “Self-attention generative adversarial networks,” *arXiv preprint arXiv:1805.08318*, 2018.
20. F. Chollet, *Deep Learning mit Python und Keras: Das Praxis-Handbuch vom Entwickler der Keras-Bibliothek*. MITP-Verlags GmbH & Co. KG, 2018.