# A Study on Interpretation of DR Deep Learning Classification Models

## Abstract

Diabetic retinopathy (DR) can be a series complication of diabetes that can lead to vision impairment and blindness if left unchecked. The early detection and classification of DR are crucial for preventing severe outcomes.

In this study, we compare the performance state-of-the-art deep learning models in the task of DR classification and demonstrate that DenseNet121 achieves superior accuracy and sensitivity. We also investigate popular interpretable method of Grad-CAM onto DenseNet121. Contrary to previous research, our results suggest that Grad—CAM is not suitable for interpreting DR classifiers based on pretrained deep learning models.

## 1. Introduction

Diabetic retinopathy (DR) is a complication of diabetes that, if left unchecked and untreated, can lead to impaired vision or even blindness, which will place a great burden on the patient, the patient's family, and society. This disease caused by poor glycemic control in patients who generally have some problems with glycemic control contributes to high incidence and life-span periodical tests, which surges the workload of ophthalmologists and the cost of medical care. With the exceptional performance of Deep Learning in image classification task, it is necessary to use this tool on the diagnosis burden.

Many researchers focus on this topic. Dekhil et al. achieved 77% diagnostic accuracy on Kaggle APTOS 2019 dataset using transfer learning and convolutional neural network [1]. Yuchen, Wu et al. classified DR diseases on pre-trained VGG19, ResNet50 network and showed that transfer learning achieved higher accuracy than no transfer learning [2]. In the works [3]–[6], different algorithms were used to realize the diagnosis of DR and all of them achieved highly accurate: reaching a diagnostic accuracy of over 97% or comparable to or exceeding that of ophthalmologists on their respective different datasets. Among them, in 2019, Jordi et al. proposed a multiple iteration approach based on convolutional neural networks to achieve expert-level accuracy. In 2020, Wang, Xiang-Ning et al. utilized reinforcement learning and multiple iterations on a dataset of more than 500 DR fundus images where the damage and features had been manually annotated as well as on multiple publicly available datasets convolutional neural network training for DR classification, achieving over 98% accuracy [6]. In 2021, Phong et al. utilized a particle swarm algorithm to classify DR, and Shankar et al. used the Inception model.

In 2019, Li T. et al. compared the performance of various commonly used models in DR classification and feature recognition and found that although most algorithms have high accuracy, they are not the same as doctors' judgment of DR condition (by observing and measuring the degree of fundus loss), and these methods perform poorly in lesion segmentation and detection [7] [8], which explains why those high-accuracy classification models have still not been used in medical practice.

To find which model classifies the level of DR in the right way, many model interpretation methods has been developed such as CAM [8], HaS, SPG, ADL, and CutMix as a weakly supervised localization of models. However, in 2020, Junsuk Choe et al. proposed a method to show focus of models, based on which several recent methods including CAM, HaS, SPG, ADL, and CutMix, were evaluated, and it was found that none of the other methods proposed later worked better than CAM [9]. This reflects the superiority of CAM in finding model focus. But the shortcomings of CAM (the need to change the network structure as well as the reduced classification accuracy) limited its use. And an advanced version of CAM, Gradient Weighted Category Activation Mapping (Grad-CAM) [10] was developed, and were used in many papers [11]–[19]. In this study, we analysis the usefulness of Grad-CAM in DR classification models.

In order to analyze the Grad-CAM, we need to build an effective and reliable classifier for DR diagnosis by investigating which state-of-the-art model is outstanding and then embed the interpretable heat map, Grad-CAM, onto the model and test its usefulness in DR classification task.

Contribution:

1. We found that DenseNet is superior to the other state-of-the-art models in DR classification, and also, we built a DR multi-classification model based on a pretrained DensNet121;

2. Contrary to previous researched, our experiment and analysis showed that Grad-CAM had quite limitation in interpretating a DR classification model.

## **2.** DR Classifier Contrast and Construction

### 2.1 Data source and processing

The data were obtained from the Kaggle public dataset Diabetic Retinopathy Detection public competition dataset[20], which was collected from several hospitals, where the images and labels contained noise, the images were taken by many imaging devices, some images contained defects such as artifacts, underexposure or overexposure, unfocused and different sizes, and for some of the images, it is not possible to determine which category they belonged to. The dataset includes 35126 fundus images, of which 33545 are fundus images that do not affect vision and the other 1581 are DR fundus images that cause visual impairment, and it is clear that there is an uneven distribution of the data.

### 2.1.1 Data pre-processing and data enhancement

Base on the characters of our dataset and above analysis, we pre-process and augment our data with the methods of random rotation, horizontal flip, vertical flip, translation, random grayscale, and random color change. Plus, oversampling is also used in our study and rate of sampling and weight of diverse DR levels as in (2 – 1).

$$d = [25810, 2443, 5292, 873, 708]\ w = [0.1, 1, 0.5, 3, 4] \qquad (2-1)$$

### 2.1.2 Framework selection.

The experiment environment is as follows:

OS: Windows 10; Programming language: Python 3.8.4; Deep learning framework: PyTorch 1.71; CPU：IntelCorei3-9100F@ 3.60GHz；31GB RAM; GPU：GeForce RTX 2080 Ti.

### 2.2 DR Classification Model

In order to carry out the DR image classification task, considering the complexity of diabetic retinopathy images, a relatively deep network is used to achieve better results, as deep networks are proven to be better at handling complex classification tasks. The use of gradient descent to optimize neural networks makes deep networks prone to the problem of gradient disappearance, and the common methods used to deal with this problem are the use of long and short-term memory (LSTM)[21], ReLU[22], Residual neural networks (ResNets)[23], Batch Normalization (BN)[24], and DenseNet[25]. ReLU and BN were of limited help for the gradient disappearance problem until the introduction of ResNets, which enable the massive use of deep network models. Based on ResNets, DenseNet achieves better result with fewer model parameters and shorter training time[25].

### 2.3 Contrast of DR Classification Models

To verify which network is best for DR classification, we chose state-of-the-art deep learning architectures for image classification task, specifically DenseNet121, Resnet101 and VGG16, all of which with similar amount parameters are trained on the same large image dataset ImageNet, and the last fully connected layers of the networks were changed with the same settings of the fully connected layer, and all selected the last network block containing convolution as trainable accordingly. The other shallow layers are set to be frozen, and the hyperparameters are selected identically. We compare those models from training process and accurate result in the following two secessions.

#### 2.3.1 Models Training Process

The loss curves of the models (DenseNet121, VGG16, and ResNet101) visualized with Tensorboard are shown in Figures 1, 2, and 3, where the thick solid line is the loss smoothed curve and the light blue line is the actual loss curve.

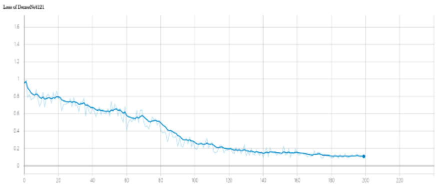

Figure 1 Loss curve of DenseNet121 network training process

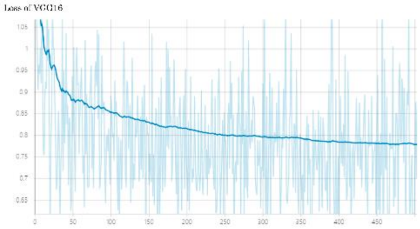

Figure 2 Loss curve of VGG16 network training process

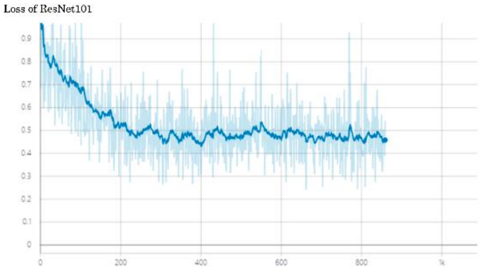

Figure 3 Loss curve of Resnet101 network training process

Comparing Figures 1, 2, and 3, it can be seen that the Densenet121 network converged after 100 rounds of training, faster than the VGG network and the Resnet network, and the training process was more stable. Apparently, the Densenet121 network is easier to train, and the loss decreases more smoothly.

### 2.3.2 For the overall accurate of the classification of the models

Further comparing the sensitivity, specificity, and accuracy between DenseNet121, ReseNet101, and VGG16 networks, the calculation formulas are shown in 2-2, 2-3, and 2-4:

$$\text{Sensitivity} = \frac{TP}{TP + TF} \qquad (2-2)$$

$$\text{Specificity} = \frac{TN}{TN + FP} \qquad (2-3)$$

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \qquad (2-4)$$

Where TP is the number of true samples, TN is the number of true negative samples, FP is the number of false positive samples, and FN is the number of false negative samples.

Table 1 Comparison of model classification results

| Models | Sensitivity | Specificity | Accuracy |
|---|---|---|---|
| Resnet101_based network | 29% | 89% | 76% |
| VGG16_based network | 6% | 96% | 74% |
| Densenet121_based network | 77% | 91.3% | 88.3% |

As seen in Table 1, the performance of the Densenet121 network based on Transfer Learning is better, only the specificity is slightly worse than the VGG network, and the sensitivity and accuracy are the highest among the three. Moreover, the model converges quickly, and the loss drops to only about 0.15 for more than 100 batches, showing the superiority of the DenseNet121 network relatively.

### 2.3.3 model selected for DR classification

Therefore, in this study, DenseNet121 network was chosen for DR classification interpretive study. The model uses a pre-trained DenseNet121 neural network with fine-tuned Transfer Learning, which means changing the last layer of the network (FC layer) so that the output channel is 5 for classes of DR levels. We trained that FC layer and the last three dense layers in the last dense block, and fixed the network parameters of the other layers. The detail of network used in this study is shown in Table 2.

Table 2 Model structure and parameters

| Name of Layers | Output Size | Parameters |
|---|---|---|
| Convolution | 112, 112 | 7×7 conv, stride 2 |
| Polling | 56, 56 | 3×3 max pool, stride 2 |
| Dense Block (1) | 56, 56 | [1×1 conv, 3×3 conv]×6 |
| Transition Layer (1) | 56, 56 | 1×1 conv, |
| | 28, 28 | 2×2 average pool, stride 2 |
| Dense Block (2) | 28, 28 | [1×1 conv, 3×3 conv]×12 |
| Transition Layer (2) | 28, 28 | 1×1 conv, |
| | 14, 14 | 2×2 average pool, stride 2 |
| Dense Block (3) | 14, 14 | [1×1 conv, 3×3 conv]×24 |
| Transition Layer (3) | 14, 14 | 1×1 conv, |
| | 7, 7 | 2×2 average pool, stride 2 |
| Dense Block (4) | 7, 7 | [1×1 conv, 3×3 conv]×16 |
| Classification Layer | 1, 1 | 7×7 global average pool |
| | | FC: 1024…5 |

Hyperparameter settings:

Learning rate: 0.0001, batch size: 256 (set larger to improve the training speed and training effect).

Data set: 80% of the data set is used for the training set and 20% for the test set.

Optimizer: Adam; Loss function: Cross entropy.

# 3. DR Classification Model with Grad-CAM

## 3.1 Analysis of Grad-CAM on DenseNet121 model

In medical field, truth-worthy is the first quality for deep learning technique since the wrong diagnosis will cause substantial consequences. Although DL has achieved amazing progress in classing images, the issue of 'black box' hinders its expand in medical imaging field.

To solve the issue, many research used Grad-CAM as a tool of interpretation in their models due to its plug-in-like feature. Researchers, as we do in this study, can embed it in the last layer of maps in a model (commonly pretrained in a public image data set and with deep neural network structure) to get a heat map for understand which part in the input image matters for the judgement of the model.

## 3.2 DenseNet121 feature maps and Grad-CAM map

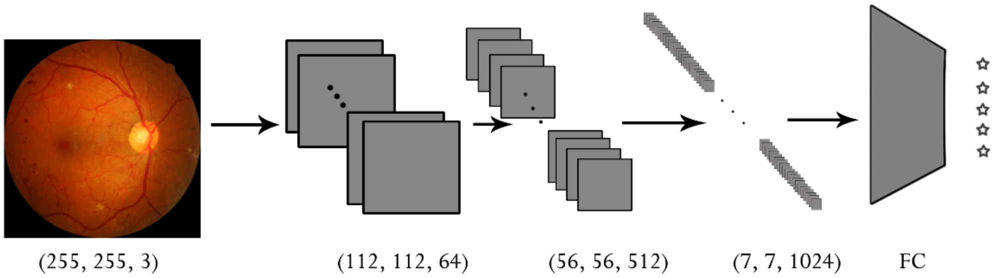

Figure (a) change of the maps of DenseNet121 model

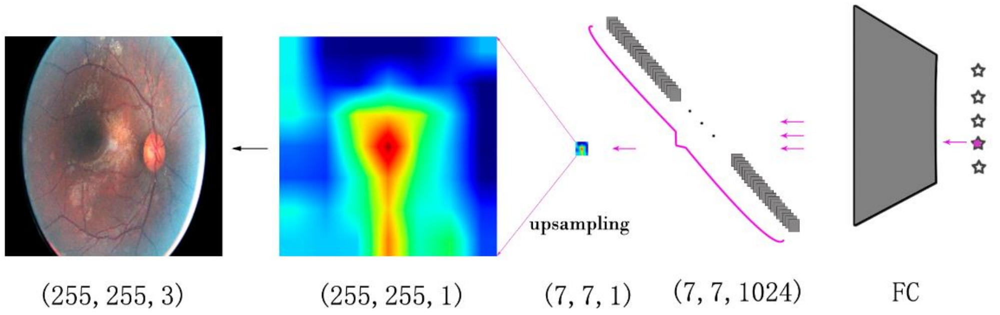

Figure (b) Grad-CAM steps

The specific steps are as the following.

First, Grad-CAM is added to the output of the last Dense Block of the DensNet121, and the size of the output map is (7, 7, 1024), which is a long and thin square as in the Figure b.

Next, to get the heat map of the model, based on the result of classifier, the loss function is calculated and back-propagated to calculate the gradient of those 1024 maps and averaged (in the last axis) to get a (7, 7) size map, which means the 1024 maps are weighted and compressed to only one of 7×7 size. Then, the map is up-sampled to the same size as the input picture of (255, 255) as the heat map for the model interpretation.

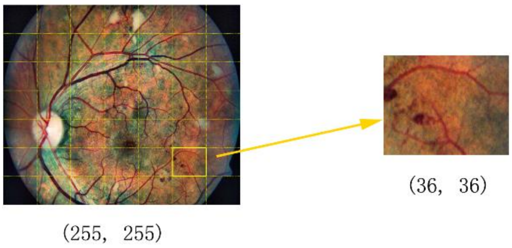

Figure (c) Proportion of lesion in Input Image

Fundus features such as microaneurysms make up only a small percentage of the fundus image or even less than 0.08% of the entire fundus image, however, Like in Figure (c), Grad-CAM divides the image into 7 * 7 = 49 segmentation blocks (each about 2% of the original image) to find that portion that is important. So, the feature points make up only 0.08% / (5%) = 1.6% of the segmented block. And the feature points are distributed in other features such as the vascular network, and it is impossible to distinguish whether the model is focusing on the vascular feature or the fundus lesions, because the fundus lesions neared or messed with vascular makes it difficult to identify.

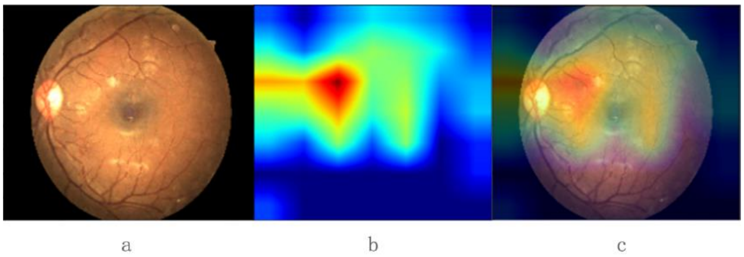

Figure 5 Heat map showing the results of diabetic retinopathy classification

As shown in Figure 5, Figure a shows the DR fundus image, Figure b shows the corresponding heat map, and Figure c shows the effect of the first two superimposed. As seen in Figure 5, the red area is the part of the DR picture that has a large impact on the model classification result, followed by yellow. As seen by the heat map, the network judgment is based on some features in the pictures including a large area of the picture, whereas lesion sites are small and scattered, and the red part even include part of the black background.

Therefore, Grad-CAM is difficult to be used as an interpretable method for DR classifiers.

# 4. Conclusion:

This paper compares the performance of different deep learning models in this DR classification task, and DenseNet121 performs the best. It also investigated the effect of Grad-CAM in DR classification model interpretation, and found that due to the last layer of feature maps of the pre-trained model is too small, resulting in the heat map produced by Grad-CAM focusing on too large a scope, it is difficult to accurately identify the lesion points in the DR fundus image as well as to distinguish between different features of the fundus image, showing limitation of the interpretation of Grad-CAM in DR classification task.

## Acknowledge

I appreciate my wife Shuo Cao who supports me both in everyday life and my research. I also thank Kaggle and all the data sharers making possible that individual researchers like me do research.

# Bibliography

[1] O. Dekhil, A. Naglah, M. Shaban, M. Ghazal, F. Taher, and A. Elbaz, "Deep Learning Based Method for Computer Aided Diagnosis of Diabetic Retinopathy," in *2019 IEEE International Conference on Imaging Systems and Techniques (IST)*, 2019, pp. 1–4, doi: 10.1109/IST48021.2019.9010333.

[2] Y. Wu and Z. Hu, "Recognition of Diabetic Retinopathy Basedon Transfer Learning," in *2019 IEEE 4th International Conference on Cloud Computing and Big Data Analysis (ICCCBDA)*, 2019, pp. 398–401, doi: 10.1109/ICCCBDA.2019.8725801.

[3] K. Shankar, E. Perumal, M. Elhoseny, and P. T. Nguyen, "An IoT-cloud based intelligent computer-aided diagnosis of diabetic retinopathy stage classification using deep learning approach," *Computers, Materials \& Continua*, vol. 66, no. 2, pp. 1665–1680, 2021.

[4] P. T. Nguyen, V. D. B. Huynh, K. D. Vo, P. T. Phan, E. Yang, and G. P. Joshi, "An optimal deep learning based computer-aided diagnosis system for diabetic retinopathy," *Comput. Mater. Contin*, vol. 66, no. 3, pp. 2815–2830, 2021.

[5] J. Wang, Y. Bai, and B. Xia, "Simultaneous diagnosis of severity and features of diabetic retinopathy in fundus photography using deep learning," *IEEE Journal*

*of Biomedical and Health Informatics*, vol. 24, no. 12, pp. 3397–3407, 2020.

[6] X.-N. Wang, L. Dai, S.-T. Li, H.-Y. Kong, B. Sheng, and Q. Wu, “Automatic grading system for diabetic retinopathy diagnosis using deep learning artificial intelligence software,” *Current Eye Research*, vol. 45, no. 12, pp. 1550–1555, 2020.

[7] T. Li, Y. Gao, K. Wang, S. Guo, H. Liu, and H. Kang, “Diagnostic assessment of deep learning algorithms for diabetic retinopathy screening,” *Information Sciences*, vol. 501, pp. 511–522, 2019.

[8] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, “Learning deep features for discriminative localization,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2921–2929.

[9] J. Choe, S. J. Oh, S. Lee, S. Chun, Z. Akata, and H. Shim, “Evaluating weakly supervised object localization methods right,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 3133–3142.

[10] H. Jiang *et al.*, “A Multi-Label Deep Learning Model with Interpretable Grad-CAM for Diabetic Retinopathy Classification,” in *2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, 2020, pp. 1560–1563, doi: 10.1109/EMBC44109.2020.9175884.

[11] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, “Grad-cam: Visual explanations from deep networks via gradient-based localization,” in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 618–626.

[12] H. Jiang *et al.*, “A multi-label deep learning model with interpretable Grad-CAM for diabetic retinopathy classification,” in *2020 42nd Annual*

*International Conference of the IEEE Engineering in Medicine \& Biology Society (EMBC)*, 2020, pp. 1560–1563.

[13] K. Duvvuri, S. Chethana, S. S. Charan, V. Srihitha, T. K. Ramesh, and K. S. Srikanth, “Grad-cam for visualizing diabetic retinopathy,” in *2022 3rd International Conference for Emerging Technology (INCET)*, 2022, pp. 1–4.

[14] O. Daanouni, B. Cherradi, and A. Tmiri, “Automatic detection of diabetic retinopathy using custom cnn and grad-cam,” in *Advances on Smart and Soft Computing: Proceedings of ICACIn 2020*, 2021, pp. 15–26.

[15] H. S. Alghamdi, “Towards explainable deep neural networks for the automatic detection of diabetic retinopathy,” *Applied Sciences*, vol. 12, no. 19, p. 9435, 2022.

[16] R. H. Paradisa, A. Bustamam, A. A. Victor, A. R. Yudantha, and D. Sarwinda, “Diabetic retinopathy detection using deep convolutional neural network with visualization of guided grad-CA,” in *2021 4th International Conference of Computer and Informatics Engineering (IC2IE)*, 2021, pp. 19–24.

[17] S. Sridhar and S. Sanagavarapu, “Detection and prognosis evaluation of diabetic retinopathy using ensemble deep convolutional neural networks,” in *2020 International Electronics Symposium (IES)*, 2020, pp. 78–85.

[18] T. Van Craenendonck, B. Elen, N. Gerrits, and P. De Boever, “Systematic comparison of heatmapping techniques in deep learning in the context of diabetic retinopathy lesion detection,” *Translational vision science \& technology*, vol. 9, no. 2, p. 64, 2020.

[19] Z. Hai *et al.*, “A novel approach for intelligent diagnosis and grading of diabetic retinopathy,” *Computers in Biology and Medicine*, vol. 172, p. 108246, 2024.

[20] “Diabetic Retinopathy Detection.” 2020, [Online]. Available: https://www.kaggle.com/c/diabetic-retinopathy-detection.

[21] S. Hochreiter and J. Schmidhuber, “Long short-term memory,” *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[22] V. Nair and G. E. Hinton, “Rectified linear units improve restricted boltzmann machines,” in *Proceedings of the 27th international conference on machine learning (ICML-10)*, 2010, pp. 807–814.

[23] K. He, X. Zhang, S. Ren, and J. Sun, “Deep Residual Learning for Image Recognition,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

[24] S. Ioffe and C. Szegedy, “Batch normalization: Accelerating deep network training by reducing internal covariate shift,” in *Proceedings of the 32nd International Conference on Machine Learning (ICML)*, 2015, pp. 448–456.

[25] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, “Densely connected convolutional networks,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 4700–4708.