DIABETIC RETINOPATHY DETECTION

A PREPRINT

Sun Bohan
Elektromobilität
University of Stuttgart
Stuttgart, 70569
st181715@stud.uni-stuttgart.de

Guo Xinru
Autonome Systeme
University of Stuttgart
Stuttgart, 70569
st181953@stud.uni-stuttgart.de

February 10, 2025

ABSTRACT

In this study. We explored various deep neural network archtectures for the classification of diabetic retinopathy. The goal was to determine wheter a model could predict if a patient has non-referable diabetic retinopathy (NRDR) or referable diabetic retinopythy (RDR) based on retinal images. We adopted two primary experimental approaches: (1) training base model such as VGG like structure convolution neural network, ResNet like stucture convolution neural network and Vision Transformer, from scratch, and (2) applying transfer learning using pre-trained models like MobileNetV3large with ImageNet weights. The highest test accuracy achieved was 86.41% for binary classification and 60.19% for five-classes classification. To better unterstand the training process, we implemented deep visualization techniques, such as Grad-CAM, to highlight critical regions in the images. Furthermore, we investigated Vision Transformers (ViTs) [1] in detail and conducted experiments to assess their performance on the same classification tasks. The ViTs are already implemented in the code, but modifications are still needed, so the results were not presented in the paper.

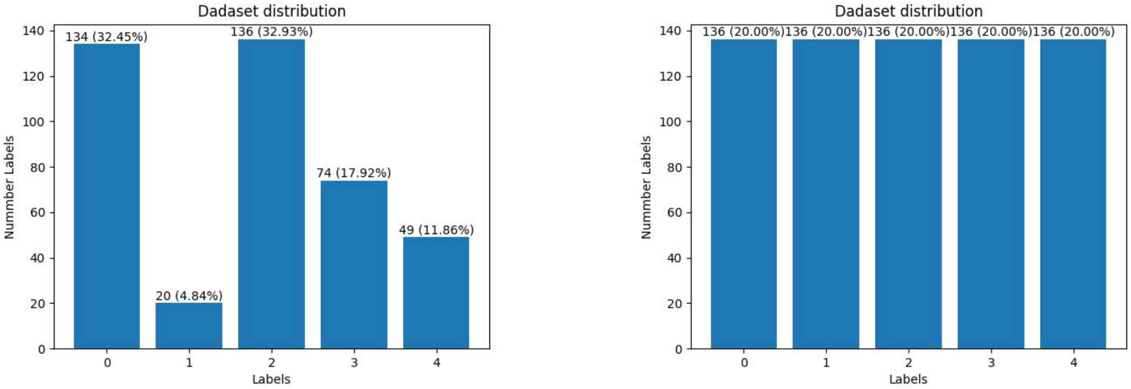
1 Introduction

Deep Learning techniques are increasingly being applied in the medical field, showcasing significant advantages across various methodologies. In this study, routine screening for diabetic retinopathy is both costly and time-consuming. However, this classfication task remains highly challenging due to the limited availability of datasets and the subtle differences between various stages of retinopathy. These deep learning approaches can also be extended to similar tasks in medical image recognition, emphasizing their versatility and potential in healthcare applications. In section 2, we will introduce the dataset analysis and preprocessing steps. Section 3 will present the results of the trained models, including deep visualization with Grad-CAM and the confusion metrics. Finally, Section 4 will summarize the conclusions derived from this project.

2 Dataset analyse and preprocessing inputpipeline

2.1 Dataset analyse (Sun Bohan)

In this Study, we utilized the IDRiD dataset. Figure A illustrates the distribution of the original dataset, while Figure B shows the balanced dataset. It is data analyse of dataset and it proves the original dataset is highly imbalanced. To address this issue, we applied oversampling to increase the data associated with label 1. Since we aimed to preserve all available information, we opted not to perform undersampling. Additionally, data augmentation was employed to enhance the diversity of the dataset, thereby improving the generalization and robustness of the models.



(a) original dataset

(b) balanced dataset

Figure 1: Dataset Distribution

2.2 Data preprocessing (Guo Xinru)

Before training the model, we perform data preprocessing and augmentation. The preprocessing pipeline includes cropping, masking, filtering, color channel conversion, letterbox adjustment, and resizing images to 256×256 pixels without distortion, ensuring aspect ratio is preserved. To optimize preprocessing, we compared grayscale images extracted from the green channel with full-color images and evaluated different enhancement techniques, including histogram equalization for global contrast enhancement, CLAHE for localized contrast enhancement to prevent noise amplification, and the Graham method (Graham) inspired by Benjamin Graham's Kaggle approach, which combines contrast enhancement with Gaussian filtering for noise reduction. These comparisons help identify the best preprocessing strategy for improving model performance.

2.3 Data Augmentation (Guo Xinru)

After data preprocessing and before the input pipeline, data augmentation is applied to enhance the robustness of the model. We observed that contrast and brightness significantly impact training performance. To address this, we applied data augmentation techniques such as flipping and random cropping of the images. Although we attempted to implement TFRecord as the input pipeline, we decided to read the dataset directly from the local disk due to the size of the IDRiD dataset. This approach is more practical for the given dataset size and ensures efficient data loading for model training.

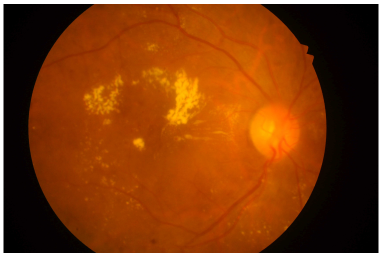


Figure 2: original image

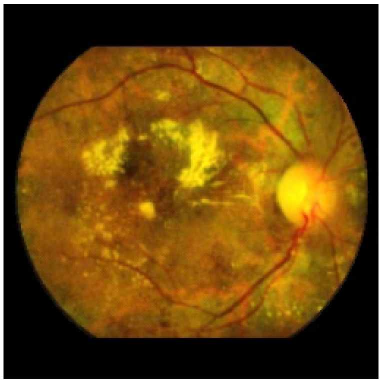


Figure 3: preprocessed image

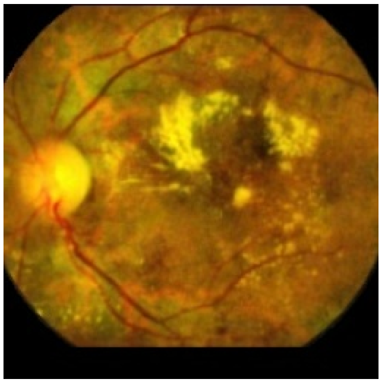


Figure 4: augmented image

3 Model and results

3.1 Model and Training (Sun Bohan and Guo Xinru)

In Section 3, we will present the models we trained and their corresponding results. For the baseline model, we initially its performance. During the training process, we encountered a gradient vanishing problem with the classic CNN model. To address this issue, we replaced the CNN blocks with ResBlocks and DenseBlocks. enabling deeper layers to learn more effectively and mitigating the gradient vanishing problem. The table 1 presents the results of the five-classes classification task across different models, while the table 2 shows the results for binary classification.

Method	Precision	Recall	F1-Score	Accuracy
CNN with resblock	63.88%	38.69%	35.68%	46.60%
CNN with resblock (balanced)	50.41%	43.68%	44.23%	59.22%
CNN with denseblock	38.87%	41.77%	38.32%	49.51%
CNN with denseblock (balanced)	37.87%	38.73%	37.45%	51.46%
MobilenetV3Large*	54.43%	45.55%	44.87%	60.19%
MobilenetV3Large (balanced)*	51.35%	49.95%	50.14%	60.19%

Table 1: Results of 5 classes classification

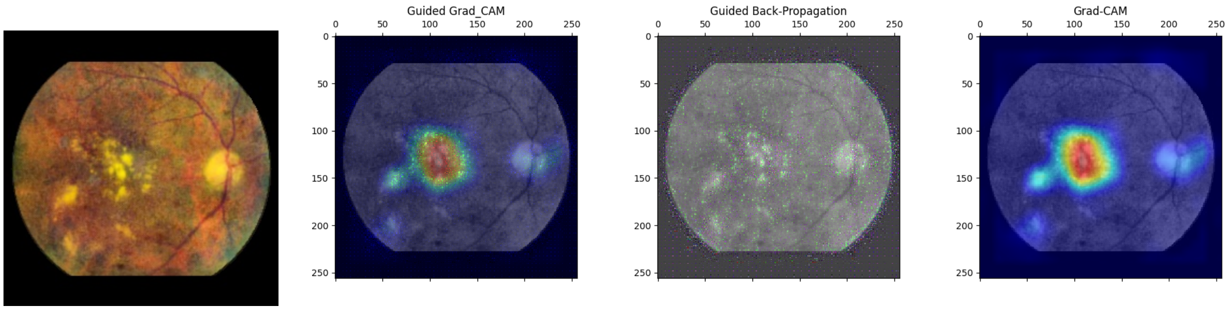
Method	Precision	Sensitivity	Recall	F1-Score	Accuracy
CNN with resblock	75.84%	70.31%	77.46%	75.39%	75.73%
CNN with resblock (balanced)	51.35%	49.95%	81.37%	80.03%	80.58%
CNN with denseblock	81.55%	78.12%	72.39%	80.43%	80.58%
CNN with denseblock (balanced)	72.18%	78.12%	72.39%	72.28%	73.39%
MobilenetV3Large*	84.56%	92.18%	81.99%	82.91%	84.47%
MobilenetV3Large (balanced)*	84.36%	85.93%	85.27%	84.74%	85.44%
MobilenetV3Small*	85.61%	82.81%	87.56%	86.02%	86.41%

Table 2: Results of binary classification

* Transfer Learning with pretrained model; (balanced) means the dataset upsample
all other Transfer Learning will be trained with the balanced dataset

3.2 Deep Visualization (Sun Bohan)

In this study, we implemented the Grad-CAM [2] method based on the algorithmus provided in the related research paper and generated the corresponding output results. The four image below illustrate the output results. We extracted features from the final convolutional layer and applied backpropagation to visualize the areas of interest on the images. From the results, we can observe that the ResNet like CNN model correctly focused on the diseased regions of the retina. However, in subsequent analyses, we also noticed that during overfitting, the model's attention became dispersed, highlighting areas unrelated to the disease.



(a) original image

(b) Grad-CAM

(c) guided back-propagation

(d) guided Grad-CAM

3.3 Metrics (Sun Bohan)

This picture is the metrics of confusion matrix. We can observe that on the label 1 there are not good result because the lack of the label 1. Even we balanced the dataset but we still have too few images with label 1.

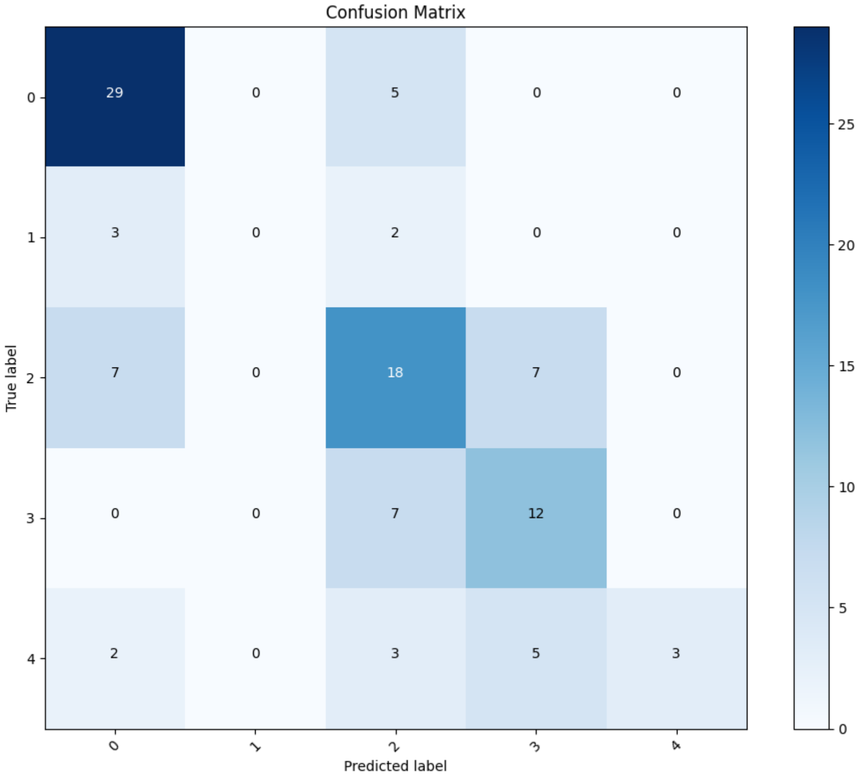


Figure 6: Confusion Matrix

4 Summary

In this study, we first preprocessed the data and then trained a baseline CNN as well as an improved CNN with modification to the ResBlock Structure. Additionally, we applied transfer learning by leveraging various large-scale models with pre-trained ImageNet weights. To better understand the training process, we implemented deep visualization Grad-CAM to observe the neural network's behavior and attention. Moreover, we explored the use of Vision Transformers to investigate the effectiveness of attention mechanisms in image recognition tasks.

References

[1] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *ArXiv*, abs/2010.11929, 2020.

[2] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. *International Journal of Computer Vision*, 2019.