

Semi-supervised Blindness Detection with Neural Network Ensemble

Jinghan Hu

Harvey Mudd College, CA, U.S.A

ahu@hmc.edu

Abstract. Diabetic retinopathy (DR), a common complication of diabetes mellitus, is a major cause of visual loss among the working-age population. Since DR vision loss is irreversible, early detection of DR is crucial for preventing vision loss in patients. However, manual detection of DR remains time costly and inefficient. In this paper, an ensemble of 6 pre-trained neural networks (including EfficientNets, ResNet, and Inception) are combined. The compatibility of different networks is tested by creating different combinations of networks and evaluating their relative performance. Pseudo-labelling is used to further increase accuracy. With a limited training data set of only 5592 images, the final neural network ensemble achieved an accuracy of 0.864.

Keywords: Machine learning, deep learning, neural network.

1. Introduction

Diabetes has become increasingly prevalent around the world. Cases of diabetes milieus are predicted to double in the upcoming decade, increasing to an astounding estimate of 592 million in 2035 [1]. The most common complication caused by diabetes milieus is Diabetic retinopathy (DR), which remains the leading cause of visual loss among working-age population [2]. Therefore, frequent retina screenings for diabetic patients and accurate early diagnosis are crucial to preventing vision loss [3].

The diagnosis of diabetic retinopathy depends on the detection of vascular abnormalities in the retina [4]. Common identifiers for DR diagnosis include microaneurysms, hemorrhages as well as soft and hard exudates. Microaneurysms (MA) are common symptoms in earlier developments of DR [5]. In this case, the swelling of blood vessels causes small bulges in retinal blood vessels, and it appears as a circular red dot on the blood vessels in retina imaging [5]. Hemorrhages are larger shaded spots on the retina. They appear as irregularly shaped spots in the retina caused by blood leaking from damaged blood vessels [5]. Exudates are caused by the leakage of plasma from damaged blood vessels for hard exudates or the swelling of nerve fibers for soft exudates [5]. They appear as small yellow (for hard exudates) or white (for soft exudates) spots in retina imaging [5].

Manual diagnosis of DR is often costly and inaccurate. Therefore, an accurate and cost-efficient automated diagnosis of DR is crucial to help prevent vision loss among diabetic patients. Current automated methods for detecting different stages of DR from retina images either suffer from low accuracy or are overly reliant on large datasets to train, which are generally expensive and difficult to obtain given that labeled medical images are generally not publicly available. Therefore, it is crucial to create automated methods that have a relatively high accuracy using

smaller training sets. In this paper, we use a stacked model of convolutional neural networks to diagnose DR severity from retina images using a small dataset of 5592 images. This experiment seeks to classify retina images into one of the five categories, marked by various severity levels of DR. Level 0 corresponds to the diagnosis of no DR, whereas levels 1 to 3 correspond to mild, middle, and severe non-proliferate DR, and level 4 corresponds to proliferate DR.

2. Related Work

Automated detection of DR is more cost and time efficient than manual detection. With the development of Deep Learning approaches, efforts have been made to automate the detection of DR.

Earlier efforts have focused on creating binary classifications on retina images to classify a fundus image into the category DR or non-DR. Xu et al. attempted to classify images into binary categories using only 1000 fundus images from the Kaggle data set and yielded an accuracy of 94.5% [6]. Other binary classification efforts utilized transfer learning. M. T. Esfahan et al. used the pre-trained ResNet34 model. The model was trained on 35,000 images from the Kaggle dataset and yielded an accuracy of 85% [7].

H. Pratt et al. purposed a CNN-based approach to classify DR into 5 stages and fine-tuned the model using the Kaggle dataset [8]. While their model had some issues with detecting lesions in images, they achieved an accuracy of 75% and a sensitivity of 30% [8]. Others explored transfer learning for multi-class classifications. X. Wang et al. studied the performance of three pre-trained CNN architectures: VGG16, AlexNet, and InceptionNet V3 to classify images into 5 categories, corresponding to 5 different stages of DR severity [9]. They used a limited data set of only 166 images and reported an accuracy of 37.43% for AlexNet, 50.03% for VGG16, and 63.23% for InceptionNet V3 [9]. With a more difficult classification task, the limited size of available data sets and labeled images becomes challenging. Current results with high accuracy have generally relied on large training data sets. The research by Li et al. used a data set of 13,673 self-collected fundus images and model transfer learning [10]. With the high volume of training data, they were able to achieve both high accuracy and specificity ($\geq 90\%$) [10]. Yet, it is still worth considering ways to train high accuracy networks with small training data sets, since medical images are hard to obtain and label.

3. Methodology

3.1. Data Set

The dataset used for this experiment is the Asia Pacific Tele-Ophthalmology Society (APTOS) 2019 public data set [11]. Out of these, 3662 images are publicly available with labels, and 1928 are public but without labels.

3.2. Preprocessing

As the images were taken in different batches with different lighting conditions and different sizes, preprocessing is necessary to remove the correlations between images. As images taken in the same batch tend to have similar lighting and noise conditions, it is crucial that those correlations be removed to prevent overfitting. The images are first turned to grayscale and then resized to 300×300 . A Gaussian blur filter is then applied to the images to remove noise and further eliminate correlations between images.

Firstly, the image is converted to grayscale so a mask could be created for cropping the image. A tolerance level of 7 was set so that mask black margins in the image could be removed and the image could focus on as much retina as possible without cropping out important portions of the retina. The images are then resized to 300×300 . Finally, a Gaussian blur filter was applied to remove noise and correlations between images. A standard deviation value of 10 was selected for both dimensions. In one dimension, the Gaussian function is given by Eq. (1):

$$G(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-x^2}{2\sigma^2}} \quad (1)$$

The blurred and original images are then combined with weights. The original image was given a weight of 4, the processed image was given a weight of -4, and a gamma value of 128 is selected. Therefore, Eq. (2) is used to obtain the final image:

$$\text{img} = 4 \times \text{img1} + (-4) \times \text{img2} + \gamma \quad (2)$$

To remove the amount of noise and speckles in the train images, a low pass filter using Gaussian blur is created and applied to the image as a final step of preprocessing. The Gaussian blur filter

removes high frequency noise that exceeds the filter threshold and creates images with cleaner edges and less noise that could interfere with training.

3.3. Data Augmentation

Given the limited size of trainable images, data augmentation was used to remove correlations in the training image data set as well as to augment the training data size. Data augmentation was achieved using the built-in image data generator in the Tensorflow Keras library. The augmentation techniques applied are outlined in the chart below:

Table 1. Settings for data augmentation

Rotation Range	360
Zoom Range	0.2
Width Shift Range	0.2
Height Shift Range	0.2
Horizontal Flip	TRUE
Vertical Flip	TRUE

3.4. Neural Network Architecture

An ensemble of transfer learned neural networks is assembled to predict the categories of images. We use transfer learning with weights for image net as a starting point for training. A blend of ResNet, Inception, and Efficient Net was selected for the ensemble. For the ensemble, we selected and trained EfficientNet B5, EfficientNetB6, EfficientNetB4 [12], EfficientNetV2S [13], ResNet50 [14], and InceptionV3 [15]. The training and validation data set is created with a 0.8 to 0.2 split of all labelled images in the data set.

Each model is loaded with pre-trained image net data and then trained for 20 epochs on the training data set with a custom cosine decay scheduler. In this scheduler, the learning rate grows linearly from an initial warm up learning rate of 0 to the base learning rate of 0.001. Afterwards, the model follows a cosine decay schedule to prevent overfitting. Moreover, the scheduler implements early stopping and restores weights with the highest validation accuracy to further optimize the training process. After all models are trained. Models are combined to create an ensemble and the accuracy of the ensemble is tested on the validation data set.

The following is a flow chart of

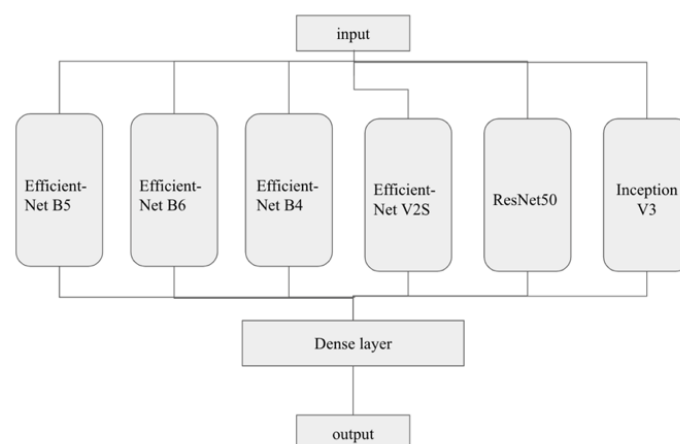


Figure 1. Ensemble Architecture

3.5. Pseudo Labeling

Using the methodology outlined in the above section. The accuracy of the model ensemble is tested and the best ensemble combination is selected. Due to the relatively small size of the training data set, pseudo labeling is used to add confident predicted data to the training data set to further increase the

accuracy of the trained model [16]. The model assembled in the previous section is used to predict labels for the unlabeled test set. These confident predicted test observations are then added to the training set. Each of the individual models is separately fine-tuned on this new data set that includes the previously unseen public test set. The models are trained with the same learning rate scheduler for 5 epochs.

4. Results

4.1. Ensemble Prediction

Various ensemble blends are tested. The most intuitive blend was having all 6 pre-trained models and taking the average of those predictions. This ensemble blend is named Blend 1 in the following charts. Blend 2 is the blend of all efficient net models excluding the ResNet 50 and the InceptionV3 network. Blend 3 is the blend of 4 networks, excluding two EfficientNet networks for computational efficiency.

Thus, Blend 3 contains ResNet50, InceptionV3, EfficientNetB5 and EfficientNetB6. The performance of the three blends is tested against the validation set and recorded in the charts below. All results are the running average of 3 trials.

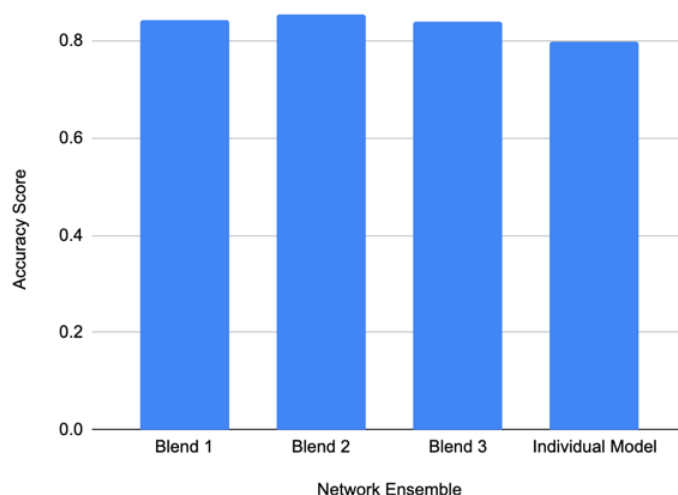


Figure 2. Accuracy score of ensembles and average

To get a more comprehensive evaluation of ensemble performance. The ensembles are also evaluated using the cohen kappa score, which qualitatively measures the level to which raters' scores agree. Using the metric, the performance of the ensembles is shown in figure 3.

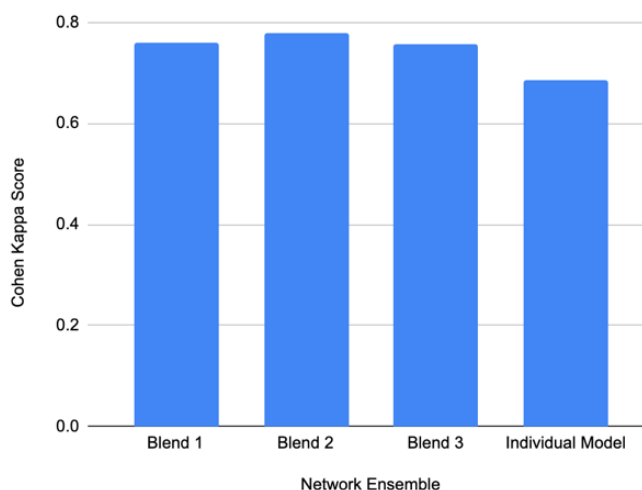


Figure 3. Cohen Kappa score of ensembles and individual model average individual model

All 3 network ensembles had higher accuracy and kappa scores than the average individual model. This difference is especially pronounced when comparing the kappa scores of an individual model and the overall model. While Blend 2 (with only EfficientNet networks) outperformed the other network ensembles, the difference is not especially pronounced.

4.2. Data Set Pseudo Labeled Prediction

While the ensemble blend with EfficientNets was the best performing blend of network ensembles, the slight advantage in model accuracy does not outweigh the significant computational cost of this ensemble. The Efficient Net networks are significantly deeper than ResNet50 and InceptionV3. Therefore, considering both accuracy and computational efficiency, blend 3 was selected for pseudo labeling. After pseudo labeling, the accuracy of each model increased slightly, generally showing a 0.01 increase. The accuracy of the ensemble increased to 0.864.

5. Discussions

The ensemble of networks performed better than averaged individual models. Though for certain well-trained models like EfficientNetB5, the ensemble generally failed to perform significantly better. This is due to the limited number of publicly available limited images and imbalances in categories. Since it is more likely that a patient would have no DR than having any DR, the data set is skewed towards category 0. Due to the small size of publicly available training data and the inherent imbalance in the data set, there is model insufficiency in all trained models. Therefore, after several epochs, all models begin showing similar biases and inadequacies. Due to model inadequacies, the ensemble of models sometimes agrees on incorrect predictions. It is also worth noting that the ensemble with only EfficientNet networks showed higher accuracy than the blend with EfficientNet, ResNet, and Inception. It is likely that although ResNet and Inception tend to show good compatibility, either or both networks do not blend well with EfficientNet.

Pseudo-labelling was effective in increasing the accuracy of the model. Though the accuracy of the model did not show a very pronounced increase, it was still effective in helping boost the accuracy of the model.

6. Conclusion

In this paper, 6 pre-trained neural networks were fine tuned on the public training set of the Kaggle APTOS 2019 data set. The models were trained for 20 epochs with a cosine learning rate scheduler. The models are then assembled to create a model ensemble. Various model combinations have been tested and one ensemble was selected for pseudo-labeling in order to further boost model accuracy.

Previous attempts at multi-class classification (classifying an image into one of the various stages of DR) have relied heavily on large training data sets to achieve high accuracy. In this paper, we restricted our data set to only 5590 images and achieved an accuracy of over 86%. Therefore, we have demonstrated the possibility of training accurate models without the use of large datasets, which are often difficult and expensive to obtain.

Future work could be to investigate the compatibility between various pre-trained models and the EfficientNet ensemble, which was shown to be the highest scoring ensemble. While ResNet50 and InceptionV3 blended well together, the two models did not blend well with other EfficientNet. Therefore, since EfficientNet was more effective in DR classification. It is worth considering whether other models could be added to the EfficientNet ensemble to boost accuracy.

Literature

- [1] Wejdan L. Alyoubi, Wafaa M. Shalash, and Maysoon F. Abulkhair. Diabetic retinopathy detection through deep learning techniques: A review. *Informatics in Medicine Unlocked*, 20: 100377, 2020.

- [2] Friedman D. S. Lietman T. Congdon, N. G. Important causes of visual impairment in the world today. JAMA, 290: 2057–2060, 2003.
- [3] Gregori N. Diabetic retinopathy: Causes, symptoms, treatment. American Academy of Ophthalmology, 2021.
- [4] Cheung C. Y. Sim S. S. Tan P. C. Tan G. S. Wong T. Y. Goh, J. K. Retinal imaging techniques for diabetic retinopathy screening. Journal of diabetes science and technology, 10: 282–294, 2016.
- [5] R. Taylor and D. Batey. Handbook of retinal screening in diabetes. WileyBlackwell, 2012.
- [6] Kele Xu, Dawei Feng, and Haibo Mi. Deep convolutional neural networkbased early automated detection of diabetic retinopathy using fundus image. Molecules, 22 (12), 2017.
- [7] Mehdi Torabian Esfahani, Mahsa Ghaderi, and RJLEJPT Kafiye. Classification of diabetic and normal fundus images using new deep learning method. Leonardo Electronic Journal of Practices and Technologies, 17 (32): 233–248, 2018.
- [8] Harry Pratt, Frans Coenen, Deborah M Broadbent, Simon P Harding, and Yalin Zheng. Convolutional neural networks for diabetic retinopathy. Procedia computer science, 90: 200–205, 2016.
- [9] Xiaoliang Wang, Yongjin Lu, Yajuan Wang, and Wei-Bang Chen. Diabetic retinopathy stage classification using convolutional neural networks. In 2018 IEEE International Conference on Information Reuse and Integration (IRI), pages 465–471. IEEE, 2018.
- [10] Tao Li, Yingqi Gao, Kai Wang, Song Guo, Hanruo Liu, and Hong Kang. Diagnostic assessment of deep learning algorithms for diabetic retinopathy screening. Information Sciences, 501: 511–522, 2019.
- [11] Kaggle Dataset [Online]. Available www.kaggle.com/competitions/aptos2019-blindness-detection
- [12] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In International conference on machine learning, pages 6105–6114. PMLR, 2019.
- [13] Mingxing Tan and Quoc Le. Efficientnetv2: Smaller models and faster training. In International Conference on Machine Learning, pages 10096–10106. PMLR, 2021.
- [14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015.
- [15] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision, 2015.
- [16] Dong hyun Lee. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks, 2013.