Automated Grading of Diabetic Retinopathy in Retinal Fundus Images using Deep Learning
Sagar B Hathwar∗ and Gowri Srinivasa†
PES Center for Pattern Recognition,
Dept. of Computer Science and Engineering, PES University,
Bengaluru, India
Email: ∗sagarbhathwar@gmail.com, †gsrinivasa@pes.edu

Abstract—The importance of early detection of Diabetic Retinopathy (DR) and the time consuming and error prone nature of manual screening of the disease necessitates an automated system that can identify and grade the disease reliably. In this paper, we propose a deep learning solution for automated grading of DR from retinal fundus images. Specifically, we employ Convolutional Neural Network (CNN) architectures which have achieved state-of-the-art performance in image recognition tasks. The experimental results demonstrate that the best performing deep learning model achieves a high sensitivity of 94.3% and specificity of 95.5% (for DR/no-DR classification) and a quadratic weighted kappa score ($\kappa^2$) of 0.88 (for grading from 0 through 4), indicating an improved $\kappa^2$ score compared to the state-of-the-art method ($\kappa^2$ score of 0.81).

Index Terms—Diabetic Retinopathy, Deep Learning, Convolutional Neural Network, Transfer Learning

I. INTRODUCTION

Diabetic Retinopathy is a complication in diabetic patients where the retina is damaged. It is the most prevalent cause of avoidable vision impairment in working age population. Clinical diagnosis of Retinopathy requires presence of one or more retinal lesions like Microaneurysms, Hemorrhages and Exudates. Other signs include intra-retinal micro-vascular abnormalities, neovascularization and vitreous hemorrhage. The most widely accepted protocol for grading the severity of DR from 0 through 4 is based on the presence and quantification of such lesions [1]. But, manual screening of DR is both time consuming and error prone. Computer aided diagnosis of the disease has the advantage of being more reliable than manual screening.

An image processing pipeline with m-mediods classifier for grading DR has been proposed [2]. A similar approach using SVM classifier has also met with some success [3]. In these predecessor works, much effort was invested in designing numerical descriptors of pathologies. The performance of these methods is dependent on the numerical descriptors and the classifier used. However, CNN classifiers have been shown to outperform traditional feature engineering in several image recognition tasks [4]. This can be attributed to the ability of such networks to automate the process of feature extraction.

Fig. 1. Images depicting grades of Diabetic Retinopathy severity

The best performing models in the Diabetic Retinopathy Challenge organized by Kaggle have been CNNs. A CNN architecture for grading DR severity has been proposed in [5]. When applied on our test set, it achieved a $\kappa^2$ score of 0.38. The architecture proposed by the winner of Diabetic Retinopathy Challenge achieved a $\kappa^2$ score of 0.81. Deep Learning methods for grading DR severity, without specifying the $\kappa^2$ score, have also been proposed in [6, 7].

In this paper, we propose a transfer learning approach employing CNN architectures that have achieved state-of-the-art performance for Image Recognition tasks. As shown in the experimental results, transfer learning using models pre-trained on ImageNet dataset not only achieves high prediction performance but the model converges much faster compared to random initialized model. The best performing model achieved a very high sensitivity of 94.3% and specificity of 95.5% and quadratic weighted kappa score ($\kappa^2$) of 0.88 for grading DR severity, which is an improvement compared to the state-of-the-art method($\kappa^2$ score of 0.81).
II. METHODS

Images for each grade of DR depicting the typical visual cues used by medical practitioners to grade the severity are shown in Fig. 1. Traditional feature engineering to distinguish grades of DR severity involves designing statistical descriptors of these visual cues. The performance of this approach is dependent on the set of selected features and the classifier used. When there is a large, pre-annotated dataset available, deep learning provides a significant advantage by automating the process of feature extraction, eliciting descriptors that are not necessarily amenable to verbal description or visually apparent.

A. Deep Learning classifiers

Deep learning classifiers are neural networks (functions with a large number of parameters) composed of multiple layers of neurons between the input layer and the output layer. A CNN is similar to ordinary neural networks with the explicit assumption that inputs are images. The constituent neurons are arranged in 3 dimensions and convolution layers in the network compute the output of these neurons (Fig. 2). CNNs enable learning highly representative, data-driven, hierarchical image features from sufficiently large data. However, obtaining a dataset that is as large and comprehensively annotated as the ImageNet dataset for DR severity grading is challenging. There are two techniques to employ CNNs for DR severity grading. 1) Training from scratch 2) Transfer learning. We used two CNN architectures which have achieved state-of-the-art performance in image recognition tasks - Inception-ResNet-V2 [8, 9], Xception [10] and compared their performance for the task of grading DR severity.

When training from scratch, the network weights are initialized to random values sampled from a Gaussian distribution [11]. This technique is termed Glorot Random Initialization. During the training process, these parameters of the CNN are modified slightly after every iteration by comparing the output of the network with known severity grade of the input image. This process is repeated several times over the entire dataset until the cross-entropy loss on a separate validation dataset starts to plateau. Given the right training set with adequate variation, the network is general enough to accurately grade DR severity on a new image.¹

The algorithm to recompute the network weights after every iteration was Adaptive moment estimation (Adam) optimizer [12]. The learning rate $\alpha$ was initialized to 0.0003. The exponential decay rates for moment estimates, $\beta_1$ and $\beta_2$ were initialized to 0.9 and 0.999 respectively. If cross-entropy loss on validation set did not improve after few epochs, the learning rate was decreased by a factor of 0.8, with a lower bound of 0.00005 set on the learning rate.

The training was stopped when validation loss plateaued for several epochs, indicating convergence of the model. To prevent overfitting and to improve the generalization of the CNN models, we employed early stopping, a regularization technique which stops the training process when training accuracy exceeds 90%.

¹Source code and models available at https://github.com/sagarbhathwar/IDRiD
B. Transfer Learning

As an alternative to random initialization, the network weights could be initialized from models pre-trained to classify objects in the ImageNet dataset. This method, known as transfer learning by fine tuning, is a technique where a model trained for one task is used as a starting point to train a model on a different task. The application and relevance of fine-tuning an ImageNet pre-trained CNN for Computer aided diagnosis has been substantiated in [13].

Fig. 3. shows block diagram of a typical transfer learning approach, where a model is pre-trained on a sufficiently large, well annotated ImageNet and further fine-tuned on a smaller dataset relevant to the domain of application. There are significant challenges in creating a sufficiently large, well annotated medical image dataset whereas the conventional deep learning models contains millions of trainable parameters, requiring sufficiently large, pre-annotated medical datasets. Thus, fine-tuning CNNs pre-trained on ImageNet dataset can make medical image recognition, specifically DR severity grading in our case, more effective.

For transfer learning, all layers of the CNN except the last, fully-connected layer are fine-tuned as suggested by [14]. The last layer(with 1000 outputs) is retrained by replacing with a random-initialized, fully connected layer(with 5 outputs) to accommodate the five severity grades.

C. Data preprocessing

The EyePACS dataset contains a large number of fundus images with varying resolutions. To mitigate this problem and to avoid training overhead, all images were resized to 512 x 512 pixels with the FOV(visible region of the fundus image) centered in each image. EyePACS and IDRiD datasets also exhibit a large variation in lighting. We employed local average magnitude subtraction to overcome variation in lighting between images. The local average color gets mapped to 50% gray. To remove boundary effect, the FOV was clipped to 85% diameter of the original FOV. An example for an overexposed and an underexposed image is shown in Fig. 4.

D. Data Augmentation

Due to limited number of available training examples (35,124) relative to the complexity of CNNs, we utilized real-time data augmentation, where the original data is synthetically modified to create new data for training. Data augmentation also avoids overfitting by using different, randomly transformed images (Fig. 5.) for training in every epoch, as opposed to the original images.

We employed the following random transformations to generate synthetic data in real-time:

- Rotation - Retinopathy lesions can appear in any orientation. Random rotation can synthesize such variations.
- Zooming - Aids in learning scale invariant features.
- Shearing - Synthetically generates lesions of different shapes.
- Flipping - Ensures positional invariance of Optic Disc and other retinal features in determining DR severity.
III. DATASET

For the experiments, we use two datasets which follow identical protocol for grading DR severity [1] - the EyePACS dataset and the IDRiD (Indian Diabetic Retinopathy image Dataset).

The EyePACS dataset includes 35,124 images with sufficient variability like varying resolution of images, artifacts in the images, out of focus shots, overexposed and underexposed images. A deep learning model trained using this dataset is robust to such variation. This dataset is heavily skewed since 73.5%(25808 images) are annotated as grade 0 with remaining 9316 images distributed among the 4 severity grades. A separate subset of EyePACS with 1000 images was used as validation set.

The IDRiD contains 413 annotated images. It is the first dataset representative of Indian population. The images were captured and graded by retinal specialists at an Eye clinic in Nanded, Maharashtra, India. Training on one dataset and testing on another shows the invariance of proposed models to the specific datasets. To the best of our knowledge, there are no published results on the IDRiD dataset.

IV. EXPERIMENTS AND RESULTS

We used the 35,124 images from EyePACS for training each of the models and the 413 images from IDRiD for evaluation. We also utilized a different, smaller subset(1000 images) of EyePACS dataset for validation after each epoch of training. This validation set was used to tune learning rate if the validation loss did not drop at the end of few epochs and also to determine model convergence. We stopped training when cross-entropy loss on this smaller dataset plateaued.

Performance of the various CNN architectures is shown in Table 1. We rank how well a model performs by the quadratic weighted kappa score (κ²) [15]. Although the severity grades are distinct classes, they form an order of grading from 0 (No apparent retinopathy) to 4 (Proliferative retinopathy). In this case, quadratic weighted kappa score is a better performance indicator in contrast to accuracy since misclassification can be weighted differently when calculating κ². The other two metrics widely used in assessing the performance of diagnostic tests are sensitivity (true positive rate) and specificity (true negative rate). We report both the metrics for binary DR/no-DR classification.

During training, we noted that CNN architecture proposed by [5] is susceptible to overfitting due to class imbalance, favoring class 0 due to higher label concentration for grade 0. Hence, we discontinued training the mode after 79 epochs when training accuracy exceeded 90% without any improvement in validation loss or validation accuracy.

We used EyePACS for training and validation and tested the models on an entirely different(IDRiD) dataset. The performance metrics proves that the proposed models are invariant to specific dataset and generalize across datasets.

V. CONCLUSION

We take advantage of transfer learning where we utilize pre-trained models, specifically models trained to classify objects in the ImageNet dataset, and fine tune them for predicting DR severity. This drastically reduces training time, which is very critical in most deep learning tasks. We found that the fine tuned Xception model has the best performance.

Since EyePACS and IDRiD adhere to similar imaging and grading protocols, we leverage the larger EyePACS dataset to train CNN networks that could be used in grading an entirely different set of images with reasonable accuracy. Since no images from the IDRiD dataset was used in training, we can assume that the models are invariant to specific dataset.

The best performing model achieved a very high quadratic weighted kappa score (κ²) of 0.88, sensitivity of 94.3% and specificity of 95.5%. Our approach achieves a higher κ² score of 0.88 compared to the baseline method with κ² score of 0.81. The improved prediction performance can be attributed to better CNN architecture selection and fine-tuning ImageNet pre-trained models.

We note that sensitivity and specificity are greater than 90% for all our proposed models, with κ² greater than 0.80. These results indicate that the proposed CNN architectures, specifically the Xception model trained using transfer learning technique, is appropriate for the task of grading Diabetic Retinopathy severity. The results also substantiate our claim that transfer learning offers a better initialization of network weights, not only resulting in slightly better performance but also aiding much faster convergence of models thereby reducing the computational power required to train such models from scratch.

We used EyePACS for training and validation and tested the models on an entirely different(IDRiD) dataset. The performance metrics proves that the proposed models are invariant to specific dataset and generalize across datasets.

We used EyePACS for training and validation and tested the models on an entirely different(IDRiD) dataset. The performance metrics proves that the proposed models are invariant to specific dataset and generalize across datasets.

We used EyePACS for training and validation and tested the models on an entirely different(IDRiD) dataset. The performance metrics proves that the proposed models are invariant to specific dataset and generalize across datasets.

We used EyePACS for training and validation and tested the models on an entirely different(IDRiD) dataset. The performance metrics proves that the proposed models are invariant to specific dataset and generalize across datasets.

We used EyePACS for training and validation and tested the models on an entirely different(IDRiD) dataset. The performance metrics proves that the proposed models are invariant to specific dataset and generalize across datasets.

We used EyePACS for training and validation and tested the models on an entirely different(IDRiD) dataset. The performance metrics proves that the proposed models are invariant to specific dataset and generalize across datasets.

We used EyePACS for training and validation and tested the models on an entirely different(IDRiD) dataset. The performance metrics proves that the proposed models are invariant to specific dataset and generalize across datasets.

We used EyePACS for training and validation and tested the models on an entirely different(IDRiD) dataset. The performance metrics proves that the proposed models are invariant to specific dataset and generalize across datasets.

We used EyePACS for training and validation and tested the models on an entirely different(IDRiD) dataset. The performance metrics proves that the proposed models are invariant to specific dataset and generalize across datasets.

We used EyePACS for training and validation and tested the models on an entirely different(IDRiD) dataset. The performance metrics proves that the proposed models are invariant to specific dataset and generalize across datasets.

We used EyePACS for training and validation and tested the models on an entirely different(IDRiD) dataset. The performance metrics proves that the proposed models are invariant to specific dataset and generalize across datasets.

We used EyePACS for training and validation and tested the models on an entirely different(IDRiD) dataset. The performance metrics proves that the proposed models are invariant to specific dataset and generalize across datasets.

We used EyePACS for training and validation and tested the models on an entirely different(IDRiD) dataset. The performance metrics proves that the proposed models are invariant to specific dataset and generalize across datasets.

We used EyePACS for training and validation and tested the models on an entirely different(IDRiD) dataset. The performance metrics proves that the proposed models are invariant to specific dataset and generalize across datasets.

We used EyePACS for training and validation and tested the models on an entirely different(IDRiD) dataset. The performance metrics proves that the proposed models are invariant to specific dataset and generalize across datasets.

We used EyePACS for training and validation and tested the models on an entirely different(IDRiD) dataset. The performance metrics proves that the proposed models are invariant to specific dataset and generalize across datasets.

We used EyePACS for training and validation and tested the models on an entirely different(IDRiD) dataset. The performance metrics proves that the proposed models are invariant to specific dataset and generalize across datasets.

We used EyePACS for training and validation and tested the models on an entirely different(IDRiD) dataset. The performance metrics proves that the proposed models are invariant to specific dataset and generalize across datasets.

We used EyePACS for training and validation and tested the models on an entirely different(IDRiD) dataset. The performance metrics proves that the proposed models are invariant to specific dataset and generalize across datasets.

We used EyePACS for training and validation and tested the models on an entirely different(IDRiD) dataset. The performance metrics proves that the proposed models are invariant to specific dataset and generalize across datasets.

We used EyePACS for training and validation and tested the models on an entirely different(IDRiD) dataset. The performance metrics proves that the proposed models are invariant to specific dataset and generalize across datasets.

We used EyePACS for training and validation and tested the models on an entirely different(IDRiD) dataset. The performance metrics proves that the proposed models are invariant to specific dataset and generalize across datasets.

We used EyePACS for training and validation and tested the models on an entirely different(IDRiD) dataset. The performance metrics proves that the proposed models are invariant to specific dataset and generalize across datasets.

We used EyePACS for training and validation and tested the models on an entirely different(IDRiD) dataset. The performance metrics proves that the proposed models are invariant to specific dataset and generalize across datasets.

We used EyePACS for training and validation and tested the models on an entirely different(IDRiD) dataset. The performance metrics proves that the proposed models are invariant to specific dataset and generalize across datasets.

We used EyePACS for training and validation and tested the models on an entirely different(IDRiD) dataset. The performance metrics proves that the proposed models are invariant to specific dataset and generalize across datasets.

We used EyePACS for training and validation and tested the models on an entirely different(IDRiD) dataset. The performance metrics proves that the proposed models are invariant to specific dataset and generalize across datasets.

We used EyePACS for training and validation and tested the models on an entirely different(IDRiD) dataset. The performance metrics proves that the proposed models are invariant to specific dataset and generalize across datasets.

We used EyePACS for training and validation and tested the models on an entirely different(IDRiD) dataset. The performance metrics proves that the proposed models are invariant to specific dataset and generalize across datasets.

We used EyePACS for training and validation and tested the models on an entirely different(IDRiD) dataset. The performance metrics proves that the proposed models are invariant to specific dataset and generalize across datasets.

We used EyePACS for training and validation and tested the models on an entirely different(IDRiD) dataset. The performance metrics proves that the proposed models are invariant to specific dataset and generalize across datasets.

We used EyePACS for training and validation and tested the models on an entirely different(IDRiD) dataset. The performance metrics proves that the proposed models are invariant to specific dataset and generalize across datasets.

We used EyePACS for training and validation and tested the models on an entirely different(IDRiD) dataset. The performance metrics proves that the proposed models are invariant to specific dataset and generalize across datasets.

We used EyePACS for training and validation and tested the models on an entirely different(IDRiD) dataset. The performance metrics proves that the proposed models are invariant to specific dataset and generalize across datasets.

We used EyePACS for training and validation and tested the models on an entirely different(IDRiD) dataset. The performance metrics proves that the proposed models are invariant to specific dataset and generalize across datasets.

We used EyePACS for training and validation and tested the models on an entirely different(IDRiD) dataset. The performance metrics proves that the proposed models are invariant to specific dataset and generalize across datasets.

We used EyePACS for training and validation and tested the models on an entirely different(IDRiD) dataset. The performance metrics proves that the proposed models are invariant to specific dataset and generalize across datasets.

We used EyePACS for training and validation and tested the models on an entirely different(IDRiD) dataset. The performance metrics proves that the proposed models are invariant to specific dataset and generalize across datasets.

We used EyePACS for training and validation and tested the models on an entirely different(IDRiD) dataset. The performance metrics proves that the proposed models are invariant to specific dataset and generalize across datasets.

We used EyePACS for training and validation and tested the models on an entirely different(IDRiD) dataset. The performance metrics proves that the proposed models are invariant to specific dataset and generalize across datasets.

We used EyePACS for training and validation and tested the models on an entirely different(IDRiD) dataset. The performance metrics proves that the proposed models are invariant to specific dataset and generalize across datasets.

We used EyePACS for training and validation and tested the models on an entirely different(IDRiD) dataset. The performance metrics proves that the proposed models are invariant to specific dataset and generalize across datasets.

We used EyePACS for training and validation and tested the models on an entirely different(IDRiD) dataset. The performance metrics proves that the proposed models are invariant to specific dataset and generalize across datasets.

We used EyePACS for training and validation and tested the models on an entire
REFERENCES


