Diabetic Retinopathy Grading System

Ji Yang
Dept. Electrical and Computer Engineering
University of Alberta
Edmonton, Canada
jyang7@ualberta.ca

Shuo Jiang
Dept. Electrical and Computer Engineering
University of Alberta
Edmonton, Canada
sjiang4@ualberta.ca

***Abstract*—Diabetic retinopathy (DR) is a prevalent vision impairment, and shows no symptoms or mild vision problems at the early stage. Medical and preventive diagnosis and treatment are mainly introduced to avoid DR. In this paper, we propose a hybrid pipeline to address this image grading task. The pipeline combines different latent representations learned from a pretrained ResNet-50 which performs the grading task as a regression task with a supervised setting, as well as a conditional residual-generation network and an auxiliary classifier discriminator that generates DR images and classify its severity. Finally, we fuse the representation from the ResNet-50 and the auxiliary classifier discriminator to train a final classifier. Quantitative results show that the pipeline successfully encoded the ordinal property of the label and therefore improved the final performance from 83.28% to 84.03%. The automated grading system can significantly improve the speed of diagnosis for DR image for downstream process, and the proposed pipeline can be applied to other image analysis problem with ordinal label involved.**

***Index Terms*—Diabetic Retinopathy, Generative Adversiral Network, Ordinal Classifcation**

I. Introduction

Diabetic retinopahty (DR) is one leading cause of blindness among working-age adults, and it is a prevalent vision impairment when blood vessels of the light sensitive tissues are damaged at the retina. DR does not have any symptoms and mild vision problems at the early stage, but it is avoidable by early diagnosis and treatment to prevent blindness in the late stage. In order to develop and test digital grading system, large and diverse retinal image datasets is required [6]. Conventionally, a human expert has to manually annotate DR images for further analysis. This process is costly in terms of both financially and labor time efficiency. Thanks to recent advances in deep learning algorithm and the fast evolution in computational hardware specifications, a reliable deep learning model can provide comparable results on images with varying conditions within a second. It will be able to offer patients with more time on early diagnosis and treatment.

Along with the gradual and abnormal changes in vasculature structure, DR can be classified into non-proliferative DR (NPDR) and proliferative DR (PDR). As figure 1 shows, five progressive levels is defined based on the severity of cases.

Existing works usually fall into the following categories: in [8], authors leverage a CNN architecture trained with a manually annotated dataset under the fully supervised setting. The deployed model can then perform the grading task for DR images. On the other hand, many experiments [9], [11] tried to modify existing CNN architecture or grading pipeline to improve the performance where popular architecture includes VGGNet, ResNet and GoogleNet are heavily investigated. In addition, conventional methods use different feature extraction techniques such as dimensional reduction on the image space or image feature descriptors [9]. It is also a known issue that the lack of data is common in medical image analysis problems. Therefore, generative modelling methods are leveraged under different mindsets. In [4], a conditional GAN can generate a massive amount of high resolution MR images for training deep models. [5] is more related to our work where it is a modified Pix2Pix that generates eye fundus image by using vessel tree segmentation mask. These work mainly focused on using CNN to perform automatic grading of DR images, modifying existing CNN architecture to improve the performance of previous work, and using generative neural network or ts derivative to generate training images to perform data augmentation.

Disease Severity Level	Findings
Grade – 0: No apparent retinopathy	No visible sign of abnormalities
Grade – 1: Mild – NPDR	Only presence of Microaneurysms
Grade – 2: Moderate – NPDR	More than just microaneurysms but less than severe NPDR
Grade – 3: Severe – NPDR	Moderate NPDR and any of the following: • > 20 intraretinal hemorrhages • Venous beading • Intraretinal microvascular abnormalities • No signs of PDR
Grade – 4: PDR	Severe NPDR and one or both of the following: • Neovascularization • Vitreous/preretinal hemorrhage

Fig. 1. The figure shows the international clinical DR severity scale [6]

Our work, instead, we not only use a standard CNN to perform the image grading task, but we also included an adversarial training framework that helps the system to include diversified latent representations. Specifically, we use a ResNet-50 to perform the grading task as an ordinal regression task. The learned features are then forced to encode the ordinal property of the images in order to achieve better results for the grading task. Then, residual-based conditional GAN is used to (1) generate novel training images by conditioning on disease severity level (2) train an auxiliary classifier that performs the "fake or real" classification and image grading classification

task simultaneously. This model builds an advancing architecture of convolutional neural networks and adversarial training frameworks, and using a novel and effective way of latent representation learned by different convolutional architecture and varied training methodologies. Quantitative results show this simple yet effective pipeline has potentials to be extended to other ordinal image grading task.

II. Related Works

Existing automated grading system usually employs convolutional neural network for automated detection of vision-threatening referable DR. It uses 106,244 nonstereoscopic retinal images to generate an ophthalmologists graded DR severity panel with development and internal validation datasets, and a reference standard grading is assigned after obtained three consistent grading outcomes [8]. The computer-assisted diagnostic (CAD) system is introduced to analyze the grade of NPDR from optical coherence tomography (OCT) images. CAD segments the retina into 12 different layers. Reflectivity, curvature, and thickness of each layer will be quantified and implement into the neural network in the training model. Both normal and NPDR images can be classified and then further grades the level of DR [11]. Another automatic method is to classify a given set of fundus images. It implements convolutional neural network (CNN) for classification, segmentation and detection. Transfer learning and hyper-parameter with AlexNet, visual geometry group network (VGGNet), GoogleNet, and residual neural network (ResNet) are implemented to obtain the high accuracy of CNN and transfer learning on DR classification [12]. Another fundus image classification, however, uses VGG-19 architecture, which is a symmetrically optimized solution through the combination of a Gaussian mixture model (GMM) for region segmentation, VGGNet for high dimensional feature extraction, single value decomposition (SVD) and principle component analysis (PCA) for feature selection, and softmax for fundus image classification [9]. GANs are used for related problems as well. The improvement of training model uses auxiliary classifier GAN (ACGAN) by adding more structure to GAN latent space [10]. The residual and illumination with GAN for shadow removing (RISGAN) is a helpful neural network for image enhancement by combining with the coarse shadow-removal image, the estimated negative residual images and inverse illumination maps in order to construct indirect shadow-removal images to improve the coarse shadow-removal result to the refined shadow-free image in a coarse-to-fine fashion [2].

III. Dataset

Two datasets are used in this work. First, a dataset that includes retinal images from a publicly available platform for retinopathy screening. There are 35127 images of a 1024×1024 resolution in total and all images are used for training purpose. The second dataset contain images from the Asia Pacific tele-ophthalmology society (APTOS) 2019 challenge. The training dataset includes about 3800 images in total, and the test set has around 2000 images [3]. To make the different dataset being feasible to train with a CNN-based system, we preprocesse the images with cropping where the cropped version has been identified the center and radius of the circle of the fundus images, and black space is cropped out as much as possible [1]. Finally, all processed images are resized to 256×256 with the consideration of our limited amount of available computational resources. All the images are labeled by human experts with the same severity index ranged from 0 to 4 (5 levels). Note that the labels are discrete.

IV. Method

We formulate our entire pipeline as a 2-stage method: during the first training stage, the ResNet-50 for regression, denoted as G_{reg} is trained with steps for supervised tasks, given an image I, there is always a corresponding ground truth label, y. We denote the latent representation used for the final regression as f_{reg}. For the residual generation ACGAN (R-ACGAN), given an image with healthy condition(i.e., severity index 0), I_{n_0}, we use an conditional encoder-decoder generator, G_{res}, where a conditional embedding, c_n concatenated right after the encoder, and also add an extra auxiliary classifier head to the discriminator, G_{dis} where the discriminator can output both the fake/real prediction as well as which class the input image belongs to. With this modified discriminator, we obtain another latent representation f_{cls} from the discriminator. At the second training stage, weight from both the regression ResNet-50 and the R-ACGAN is fixed. The two latent representations, f_{reg} and f_{cls} are extracted from each image and used to train the fusion network G_{fusion} and make the final prediction.

In the rest of this section, we introduce each of the components in our proposed pipeline in detail, namely, we introduce a ResNet-50 for ordinal regression, then we discuss our R-ACGAN architecture. Finally, we demonstrate how we fuse the latent representations learned from the two components.

A. ResNet-50 for Original Regression

We use ResNet-50 as the baseline model. ResNet-50 is the smallest model that leverages bottleneck residual block and it's well fitted to our computational resources. Compared with conventional sequential architectures such as VGGNet and ResNet-18/34, it provides a significant more amount of non-linearity with even lower number of parameters due to the extensive usage of 1×1 convolutional layers in the residual bottleneck block, as shown in Figure 2.

Existing work tried to treat this problem as a classification task due to the discrete fact of the label. In our case, however, we consider this setup presents several drawbacks. First, under the classification setting, the model is trained purely for the classification task. The ordinal property of the label cannot be covered well. Also, classification prediction is always a hard prediction, i.e., the predicted results fall into one exact category where in the real world cases we may consider a patient's case is in transition from one level to the next.

With the consideration above, we modify the a standard ResNet-50 to tackle the ordinal regression task as shown in

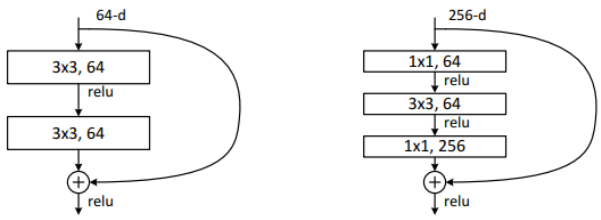

Fig. 2. Left: the naive residual block used in ResNet-18/34. Right: the bottleneck residual block used in ResNet-50.

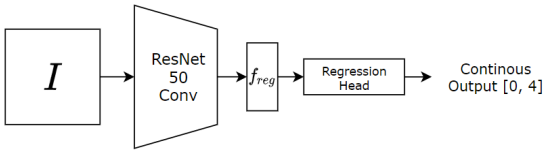

Fig. 3. The ResNet-50 architecture used for regression task. We keep all the layers from the pre-trained model except for replacing the last output layer where previously used for classification.

Figure 3. In particular, under the classification setting, the output layer of the ResNet has N outputs where N is the number of classes, the output is now changed to a single scalar value to accommodate the regression task. To make the model more robust, we leverage pre-trained weight by ImageNet for training.

B. Residual-based Auxiliary Classifier GAN

The proposed Residual-based Auxiliary Classifier GAN (R-ACGAN) is illustrated in Figure 4. The entire network is trained from scrach by using our DR image dataset. Given an image I_{g_0} without any diabetic symptom, the encoder-decoder structure is applied to generate a residual image $I_{R_{g_n}}$ with the concatenated condition embedding C_n. The condition embedding is generated with an intuitive method, we define a constant factor c, and multiply it by the target label index to obtain $C_n = c \times y$. At the time that we need to concatenate this scalar value to a 3-dimensional output ($H \times W \times C$) from the encoder, c is broadcasted to C_n. After an element-wise addition with the input image I_{g_0} and the generated residual image $I_{R_{g_n}}$, we obtained a synthetic image $\hat{I}_{g_n}$.

In principle, any fully convolutional encoder-decoder architecture can be used for our R-ACGAN. However, we do not want to design any particular architecture for our framework, therefore a commonly used framework called DenseUNet is employed in our framework [2]. The DenseUNet consist of a contracting path to capture context at varied feature scale and a symmetric expanding path for upsampling. Compared with the original UNet [7], DenseUNet replaces the sequential convolutional layers with dense blocks in the network, which concatenate all layers' output with its input. This enhancement helps the information passing and gradient flow during training.

The discriminator used here is just like the common discriminator used in most of GANs, but we follow the design from [10] where the discriminator not only predict whether the image is real or fake, but also predict which class the given image belongs to. The discriminator is a simple 5-layer convolutional neural network, where each of the convolutional layer is followed by batch normalization and leaky ReLU layers. All the kernel sizes used is 3×3 with a stride size of 2 on both height and width dimensions.

Two classification heads are employed after the convolutional layers. They share the same structure, which is fully-connected layers with output sizes as 1 (fake/real, sigmoid activation) and 5 (severity classification, softmax activation), respectively.

C. Representation Fusion Network

We use a simple multi-layer perceptron with 2 fully connected layers as the representation fusion network. We demonstrate it in Figure 5. Note this network is only trained once the regression ResNet-50 and the R-ACGAN are done training. For each of the image, we obtain two fixed latent embeddings, f_{reg} and f_{cls}. We consider these as the input to the fusion network and concatenate them together after applying a global adaptive average pooling operation on each of them. Therefore, both embeddings are reshaped by the global pooling to a 1-d vector. The intuition behind this design is straightforward.

Compared with f_{cls}, f_{reg} focuses more on the features that encode information for a regression grading task, i.e., the image features that represent the progressively increased severity in the image. Also, f_{cls} contains knowledge of the pretrained model. The f_cls, instead, focuses more on discrimination features that can distinguish different classes of DR images, and is fully trained only on DR images. By fusing these two different latent representation, we implicitly cover the ordinal property of the image grading labels and also include varied feature representation in the entire framework.

D. Loss Function

Consider our proposed method as a 2-stage method, each of the components is actually trained separately. During training, the regression ResNet-50 is trained to try to minimize the mean square error (MSE). We noticed that in previous competitions [3], many attempts fall into using the mean absolute error (MAE). However, in our case, we consider MSE is a better fit. The quantitative results prove our intuition that in this disease image grading task, for example, predicting the image as 2 levels away from the ground truth is much worse than 2 times when we predict it just 1 level difference, i.e., medical use case is sensitive and need high recall. Therefore, the loss functions for regression ResNet-50 is

$$L_{reg} = \frac{\sum_{n-1}^{n}(y_n - \hat{y}_n)^2}{n}$$

where n is the number of samples, y_n is the ground truth label and $\hat{y_n}$ is the predicted value.

The R-ACGAN employ the same loss function as its base architecture, ACGAN [10]. The loss function has two parts: the

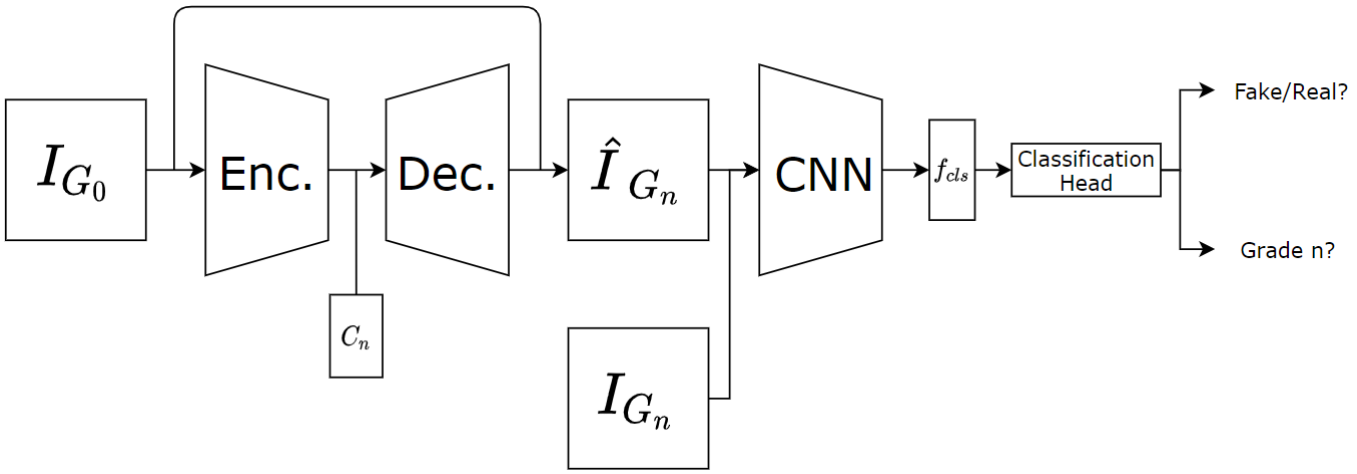

Fig. 4. The proposed R-ACGAN structure. It is composed of a residual-based encoder-decoder structure and an auxiliary classifier discriminator. It takes a healthy DR image and a conditional embedding after the encoder as the input and outputs a residual image that includes diabetic features. The auxiliary classifier discriminator accepts synthetic inputs from the generator or real images from the datasets, it is trained to predict both whether the image is a synthetic one or a real one as well as the disease severity index.

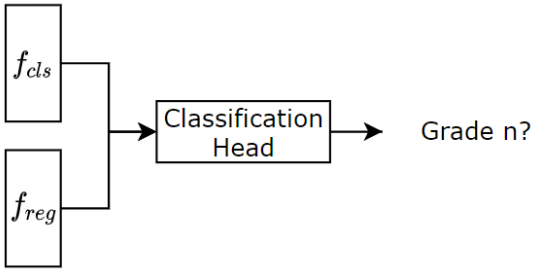

Fig. 5. The representation fusion network, which is a 2-layer feedforward fully-connected neural network. It accept two different latent representations from the ResNet-50 and R-ACGAN and output a final prediction of the severity index.

log-likelihood of the correct source, L_s, and the log-likelihood of the correct class, L_c.

$$L_s = E[logP(S = \text{real}|X_{real})] + E[logP(S = \text{fake}|X_{fake})]$$

$$L_c = E[logP(C = c|X_{real})] + E[logP(C = c|X_{fake})]$$

The residual-based generator is trained to maximize $L_c - L_s$ and the discriminator is trained to maximize $L_s + L_c$.

Finally, the representation fusion network is trained with the standard supervised classification setting therefore it tries to minimize negative log-likelihood.

E. Implementation

All the code base is developed with a popular auto-differentiation framework, Pytorch. We use a workstation with 3 GTX 2080Ti. In our experiments, the input size of image is 256×256. The learning rate for training all the models is initialized to 0.001 and is reduced by a factor of 10 at each plateau epoch. The batch sizes for the regression ResNet-50, R-ACGAN and the fusion network are 96, 12, and 128, respectively. The generator in R-ACGAN is optimized by stochastic gradient descent optimizer with momentum factor of 0.9 where all other components are trained by using Adam optimizer.

V. Experiments

To verify our contribution and the effectiveness of the proposed pipeline, we conduct various experiments based on the two available DR dataset [1], [3]. To evaluate the performance, we employ two metrics: accuracy and quadratic kappa score.

A. Regression ResNet-50

We first examine whether treat the classification with ordinal labels as regression is feasible in this problem setting. It is worth to notice that both the accuracy and quadratic kappa score are classification metrics, when we evaluate the result from a regression model, we need to use a threshold to round the prediction. Therefore, we consider the quantitative evaluation as a reference only but a definite standard of which model is outperforming others.

To obtain a hard prediction from the regression, we use simple cutoff thresholds as the boundaries of intervals with the same size. For example, $[-\infty, 0.5]$ means severity level 0, $(0.5, 1.5]$ is level 1, and so on so forth. It is possible to find better thresholds for cutting these intervals but we leave it as future engineering work.

The quantitative results in Table I shows that classification setting is better than regression by 1.32% in terms of accuracy and 0.007 according to the quadratic Kappa score. However, we consider this is an acceptable result as we mentioned the regression result may suffer from our naive selection of cutoff thresholds. In following section, we also demonstrate that regression ResNet-50 helps more compared with a ResNet-50 under the classification setting.

TABLE I
A COMPARISON BETWEEN PERFORMING SUPERVISED TRAINING FOR REGRESSION AND CLASSIFICATION SETTING.

	Accuracy	Kappa Score
Regression	82.56	0.811
Classification	83.18	0.818

TABLE II
THE CLASSIFICATION RESULT OBTAINED FROM BOTH DISCRIMINATOR AND THE RESNET-50.

	Accuracy	Kappa Score
Classification by Discriminator	81.34	0.805
Classification by ResNet-50	83.18	0.818

B. R-ACGAN

With the ResNet-50 we discussed in the previous section, pretrained model weight from ImageNet is leveraged. However, to ensure that the training more stable, our R-ACGAN is trained from scratch. Therefore, a completely fair comparison of the classification performance between the ResNet-50 and the discriminator in R-ACGAN is not feasible in this case.

As shown in Table II, surprisingly, although the discriminator we used has only 5 convolutional layers, the performance is still comparable with the ResNet-50 classification result.

C. Fusion Network

We finally use the two trained network, the regression ResNet-50 and R-ACGAN, to extract features from the images. Then we consider the two latent embedding, f_{cls} and f_{reg}, as input to train the representation fusion network. We compared different combinations of all the components we have experimented so far and provide a summary of results in Table III.

We noticed that our intuition that fusing the regression related feature vector and the classification/discrimination feature vector is better than fusing two classification models. In particular, by fusing the representation from R-ACGAN and regression ResNet-50, we are able to achieve 0.830 Kappe score where fusing R-ACGAN and classification ResNet-50 can only give 0.822, though it's still better than a single classification model.

In addition, to complete an ablative study for the fusion network, we also tried to combine 3 latent representations to train the model, however, we observed a slight performance drop of the Kappa score when we combine all of them. We argue this may caused by the co-adaptation and/or co-variance in the two classification representation.

TABLE III
AN ABLATIVE FASHION EXPERIMENT FOR EVALUATING THE PERFORMANCE OF THE REPRESENTATION FUSION NETWORK.

Reg. ResNet-50	Cls. ResNet-50	R-ACGAN	Kappa Score
	✓	✓	0.822
	✓		0.818
✓		✓	0.830
✓	✓	✓	0.828

VI. CONCLUSION

In this work, we propose a 2-stage pipeline that leverages recent advances in GAN as well as combining varied types of representation to achieve better performance for ordinal label classification problem. Based on the result, although the performance improvement should be able to achieve with including more modalities such as ultrawide-view version of DR images, this proposed method can significantly improve the speed of diagnosis for DR image for downstream process. The applied adversarial training frame work is currently a basic form of such advancing strategies, where many recently proposed works has shown significant advances. Furthermore, the automated grading system can be modified and apply to other medical image analysis.

REFERENCES

[1] Diabetic retinopathy (resized)-resized version of the diabetic retinopathy kaggle competition dataset.

[2] Ris-gan: Explore residual and illumination with generative adversarial networks for shadow removal. 34.

[3] Asia Pacific Tele-Ophthalmology Society (APTOS). Aptos 2019 blindness detection-detect diabetic retinopathy to stop blindness before it's too late.

[4] Agisilaos Chartsias, Thomas Joyce, Mario Valerio Giuffrida, and Sotirios A Tsaftaris. Multimodal mr synthesis via modality-invariant latent representation. *IEEE transactions on medical imaging*, 37(3):803–814, 2017.

[5] Pedro Costa, Adrian Galdran, Maria Inês Meyer, Michael David Abràmoff, Meindert Niemeijer, Ana Maria Mendonça, and Aurélio Campilho. Towards adversarial retinal image synthesis. *arXiv preprint arXiv:1701.08974*, 2017.

[6] DeepDRiD. The 2nd diabetic retinopathy - grading and image quality estimation challenge.

[7] Thorsten Falk, Dominic Mai, Robert Bensch, Özgün Çiçek, Ahmed Abdulkadir, Yassine Marrakchi, Anton Böhm, Jan Deubner, Zoe Jäckel, Katharina Seiwald, et al. U-net: deep learning for cell counting, detection, and morphometry. *Nature methods*, 16(1):67–70, 2019.

[8] Zhixi Li, Stuart Keel, Chi Liu, Yifan He, Wei Meng, Jane Scheetz, Pei Ying Lee, Jonathan Shaw, Daniel Ting, Tien Yin Wong, et al. An automated grading system for detection of vision-threatening referable diabetic retinopathy on the basis of color fundus photographs. *Diabetes care*, 41(12):2509–2516, 2018.

[9] Muhammad Mateen, Junhao Wen, Sun Song, Zhouping Huang, et al. Fundus image classification using vgg-19 architecture with pca and svd. *Symmetry*, 11(1):1, 2019.

[10] Augustus Odena, Christopher Olah, and Jonathon Shlens. Conditional image synthesis with auxiliary classifier gans. In *International conference on machine learning*, pages 2642–2651. PMLR, 2017.

[11] Harpal Singh Sandhu, Ahmed Eltanboly, Ahmed Shalaby, Robert S Keynton, Schlomit Schaal, and Ayman El-Baz. Automated diagnosis and grading of diabetic retinopathy using optical coherence tomography. *Investigative ophthalmology & visual science*, 59(7):3155–3160, 2018.

[12] Shaohua Wan, Yan Liang, and Yin Zhang. Deep convolutional neural networks for diabetic retinopathy detection by image classification. *Computers & Electrical Engineering*, 72:274–282, 2018.