REFUGE Challenge Submission: Using Dense U-Nets to Detect Glaucoma and Segment Optic Disc and Cup

Apoorva Sikka*, Sai Samarth R Phaye*, and Deepti R. Bathula

Indian Institute of Technology Ropar

Abstract. In this paper, we describe our contribution to the REFUGE Challenge. We propose two Deep Learning models, CoarseNet (C-Nets) and FineNet (F-Nets), which are used to develop an efficient system for the Glaucoma Classification and Optic disc/cup segmentation. C-Net is used to localize the region where the optic disc is present, and then F-Net does the job of segmenting the cup and disc at a finer level. The complete framework is supported by various methods of preprocessing like using histogram matching to normalize the sample-space, and we also introduce how exponential transform of images helps to significantly improve the cup segmentation by getting clear boundaries between the cup and disc. Additionally, we introduce the modified version of pooling for segmentation architecture. Finally, we use the encoded information from F-Net to perform the classification. For Glaucoma Classification, our approach obtained an average accuracy of 0.877% on the ten-fold cross-validation. In the optic disc and cup segmentation, we achieved a dice value of 0.94 and 0.89, compared to the baseline results provided by the challenge of 0.91 and 0.87.

Keywords: Glaucoma Classification · Optic Disc and Cup Segmentation · Image Processing · Deep Learning

1 Introduction

One of the main reasons for irreversible blindness is Glaucoma. It is mainly characterized by damaging of optic nerves that carry signals from eyes to brain. The disease shows no symptoms in its initial stages and by the time it is noticed it is too late to treat it and prevent loss of eye sight [2]. The main cause is increase in intra-ocular pressure (IOP) which further leads to straining optic nerves. It is mainly characterized by increase in cup to disc ratio (CDR), disc hemorrhage, pale disc. Optic disc is the bright region where all vessels merge and optic cup is the central cup like area inside the disc. An increase in cup-to-disc ratio may be an indicator of glaucoma or other diseases [3]. Figure 2 shows an example of a scan from two different cameras where optic disc can be easily spotted as the bright region inside the disc. The goal is to detect and segment optic disc and

* states that authors have contributed equally.

cup which can be used as an indicator to detect glaucoma and hence classify glaucoma patients from non-glaucoma ones. It has been shown that CDR can be a good indicator of a diseased eye or a healthy eye.

2 Dataset

The dataset has been provided by REFUGE containing 1200 color Fundus images. For training, 400 color images are provided with an annotation mask of cup and disc and a label. Further, 400 unlabeled images are provided which are used as validation dataset. Each image in training has a size of 2124×2056 (captured by Zeiss Visucam 500) whereas each image in validation set is of size 1634×1634 (captured by Canon CR-2). All images are centered at the posterior pole with both macula and optic disc.

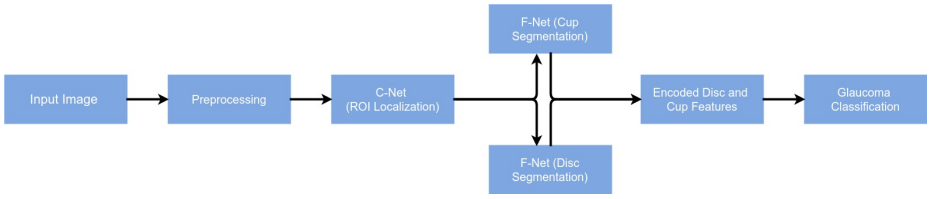

Fig. 1. Brief pipeline of proposed methodology

3 Methodology

Several groups have attempted to segment optic disc and cup using shape based and active contour based approaches. Different machine learning algorithms have also been used for Glaucoma classification. Recently, deep learning methods have shown great potential on a number of medical imaging challenges including Optic cup and disc segmentation. Convolutional Neural Networks (CNNs) have been employed on extracting complex features from images. A modified version of U-Net has been used for cup and disc segmentation. Recently, a network named M-Net has been proposed which works on polar transformed images. Our approach is mainly comprised of two deep learning models used to work at different resolutions of images. Both of them have similar architecture that is based on a combination of Densely Connected Convolutional Networks (DenseNets) and U-net, called DenseUNet. Figure 1 shows a brief pipeline of the proposed approach. The first network, CoarseNet or C-Net, localizes the region of fundus which is then cropped and passed to the finer network (F-Net). Both F-Net and C-Net have the same architecture but are two independent models that are trained to segment cup and disc independently. Finally, for glaucoma classification, the encoded features of F-Net are used to train a Deep Neural Network. For enhanced learning, several preprocessing steps are performed to normalize all the samples to the same reference. Following sections discuss these steps in detail.

3.1 Preprocessing

The image data is prepared for training with different preprocessing steps. Unlike some previous efforts that convert color Fundus images to grayscale, We used all three color channels as they provided discriminative features - while one channel provided clear boundary between the disc and the background, others helped delineate the cup and the disc. Our preprocessing steps mainly comprise of histogram matching and a color-channel normalization. Furthermore, image augmentation was employed as total number of images is not very high and data is skewed.

Histogram Matching: Since, the images in the training and validation sets were captured using different cameras, intensity distributions of the images varied significantly between them. Based on this observation, we applied histogram matching in which we took the average histogram of all the training samples, thus creating a reference histogram for color mapping. Further, each input sample is normalized by channel-wise mapping of its histogram to the reference histogram. This procedure was applied for both training and validations samples and the results showed significant improvement. Figure 2 depicts the difference between the original training/validation samples and the histogram matched samples. Clearly, such regularization can help the model converge faster and produce better results. After equalizing the color distributions, images were normalized to zero mean and unit variance.

Image augmentation: After normalization, image augmentation was performed on the training samples for both classes in a ratio (27 augmented images for one Glaucoma sample and 3 augmented images for 1 non-glaucoma sample) to make total images from each category equal. Image was only rotated in random angles, and as the images are spherical with black padding, the rotations didn't cause any information loss.

3.2 Architecture

Due to large sized training images and a very small size of region of interest (ROI), we adapt a two-stage process. First, C-Net is trained to work on coarse or down sampled versions of original images. Downsampling (to 112×112) was performed to reduce the network parameters by a huge amount since the task of C-Net is extract only the ROI and not be efficient in finding disc in the images. We adapt a fully convolutional DenseNet [4] which is an encoder-decoder architecture having dense blocks instead of the convolutional layers. DenseNets were mainly proposed for efficient gradient propagation and better feature learning in deeper networks [5]. On the other hand, U-Net was proposed for image segmentation where skip connections were added from encoder to decoder for enhanced image segmentation [6].

C-Net: Input to C-Net is a 112×112 size down sampled input image and output is binary mask trying to localize disc. The network consist of nine dense

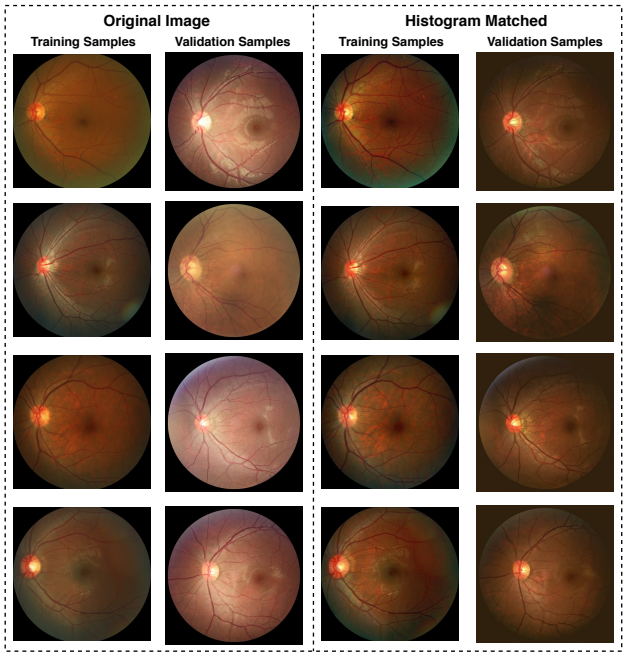

Fig. 2. Samples of Histogram matching when applied on training and validation samples. In the right part (after histogram matching), it's not possible to differentiate between the images from different camera.

blocks, where each dense block is a series of operations (Batch Normalization, ReLU, 3×3 convolution) with same padding and $2, 4, 8, 8, 8, 8, 8, 4, 2$ layers in each dense block with a growth rate and initial convolutional filters of $48, 32, 32, 64, 128, 64, 32, 32, 48$.

In the encoding sub-network, each dense block is connected by a transition layer where the average of average and max-pooling is applied. The reason for combining these pooling methods is to reduce the loss of information (accomplished by average pooling) while ensuring that the largest value still dominates (achived by max-pooling). In the decoding sub-network, each up block consists of up sampling with a size of 2×2 where features are concatenated from same sized encoder dense-block. We trained the network with dice coefficient as the loss function as it was giving better results, as compared to entropy loss or mean squared error loss. The network was trained using Adam optimizer with a learning rate of 10^{-3}.

Based on the C-Net generated activation mask that localizes the disc region, the goal is to crop the disc region. However, due to some erroneous activation regions in the predicted mask, cropping was not straight forward. To overcome this, we tried various methods but the best method was to use a kernel convolu-

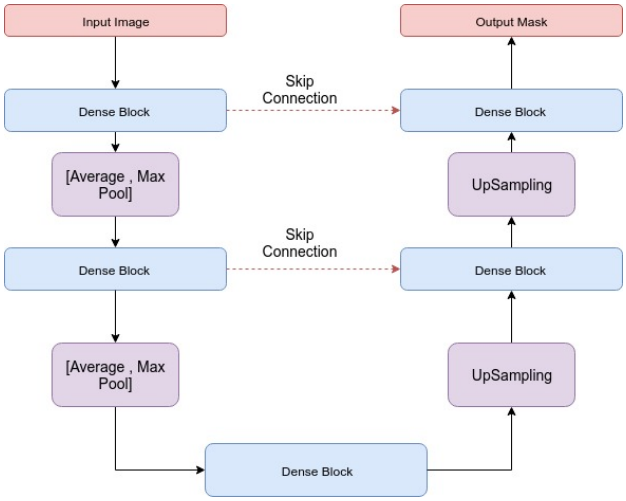

Fig. 3. Block Diagram of C-Net/ F-Net. This is just an illustration, and the original model has nine dense blocks in which the fifth block is the smallest feature map size (encoding).

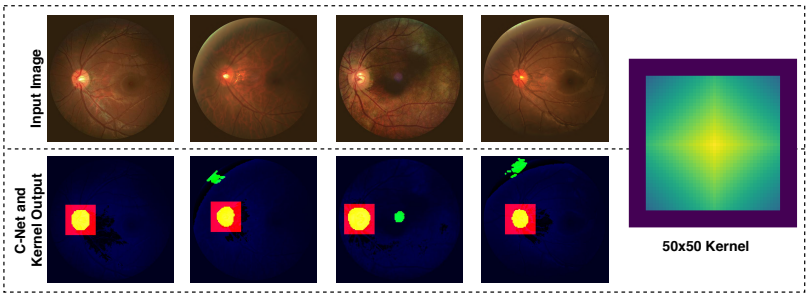

Fig. 4. The figure illustrates the output of C-Net model. Top row has the input samples, the bottom row has the output of C-Net model (Green Channel), the input image (Blue Channel) and the final cropped area after the kernel-convolution (Red Channel). Yellow region highlights the actual disc area in the image.

tion. First we resized the predicted mask to 163×163 and replaced zeros with -1 in the mask. We created a 50×50 kernel which has gradually increasing values from the 6th pixel from boundary till the center (where it is at the highest value of 10) and the first 5 pixels from each boundary has a negative value of -5. Now when we convolved this kernel onto the mask image and the center of this disc activated component reacted the most. Choosing this point as the center (and also by handling edge cases), we crop the 40×40 area and map this area onto the original image to obtain the localized region and localized mask label. This was working perfectly on both the training samples and the validation samples without any error. The padding helped to remove the binary activated components

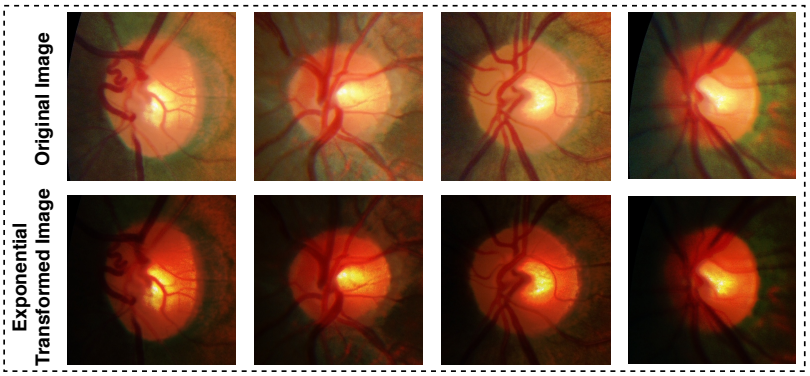

Fig. 5. Example of exponential transformed images (the disc region is cropped from C-Net and this is used for training F-Net for cup-segmentation). Here original images are images after preprocessing.

in the image which have high pixel values and are elongated in some direction, as shown in Figure 4. These image samples and labels are then used to train a finer network, F-Net.

F-Net: From the the output of C-Net, we obtain a 515×489 image region for the training sample (it is different for the test sample as the resolution of image is different) which is again resized to 112×112 for developing data for training. The corresponding maps are separated into two binary masks for cup and for disc. We separate the learning of disc and cup into two separate tasks (hence two different models) as learning was independent for both objects and preliminary experiments show that separate training leads to better learning of both tasks. The network architecture is same as C-Net and we perform similar augmentations for this network.

We observed that boundaries of disc are visible but for cup this is not the case with all the images. For some images, the boundaries of cup are highly fuzzy. This might be due to a low gradient between cup and disc region. Consequently, we used exponential transformation (a dual of logorithmic transformation) followed by scaling of intensity values to $[0 - 255]$ range. This was done to enhance the boundary between the cup and disc regions and it is passed as an input to the F-Net trained for cup as shown in figure 5. After obtaining the activations, we resize it to the original cropped-image size and first thresholded, then smoothened the output to get a smooth boundary and also performed morphological operations (opening and closing) to remove the minute grains in the output. Lastly, we merge the outputs of disc and cup, and place this mask on a image of zeros at the point given from C-Net. This is the final output of our segmentation model.

Classification: Since CDR ratio is an important indicator of glaucoma, we used the intermediate encodings from previously generated F-Net for disc and cup, which contain the most discriminative features of the cup and disc. As the number of kernels were 2048 for each of them, we first perform a simple

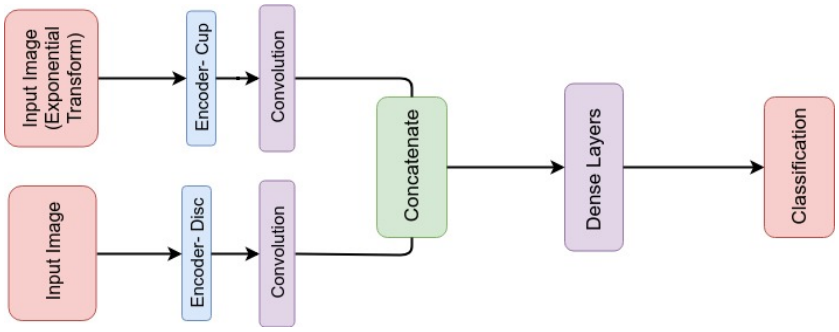

Fig. 6. Illustration of Glaucoma classification model

convolution of 64 kernels for both encodings (disc and cup) and then concatenate the flatten output. Further, we trained a simple classifier with 4 fully connected layers of size $1024, 512, 256, 64$ with ReLU as activation function and softmax function at the final layer. We freeze the weights of disc and cup F-Net model leading to very less final trainable parameters of the classifier. We trained the network using categorical cross-entropy loss function and Adam optimizer with a learning rate $1E-6$. We trained the models on 10-fold cross validation by keeping a validation set (from training data) aside from the training data for each fold, resulting in ten models.

We chose the best seven models from these ten models by analyzing the confusion matrices. We figured out that these seven models were very strong in predicting glaucoma, that is, whenever the models predicted Glaucoma, it was correct 99.25% times (not vice versa). So, we took the advantage of this and for final classification output, whenever any single model predicts Glaucoma, we tagged the sample as Glaucoma, otherwise it is labeled non-glaucoma (if all seven models agree). The average AUC achieved was 0.8319 on the validation samples.

4 Experiments

We trained all the models with NVIDIA GeForce K1080 with 8GB GPU RAM. For C-net we trained a model from training data by keeping a validation set aside(not the validation samples provided in challenge). The accuracy achieved was > 99.50 and further we encapsulated disc by padded rectangle to ensure there is no pixel missed by the C-Net. The F-Net was also trained on > 2000 augmented training images where the dice achieved for kept aside validation set was 0.94 and 0.89. The training for F-Net was done for 100 epochs for both the models. The figure shows the results of F-Net on both training and validation data provided. It is clear that discerning disc is lot more easier than segmenting cup. Some outputs on the validation samples are shown in Figure 7.

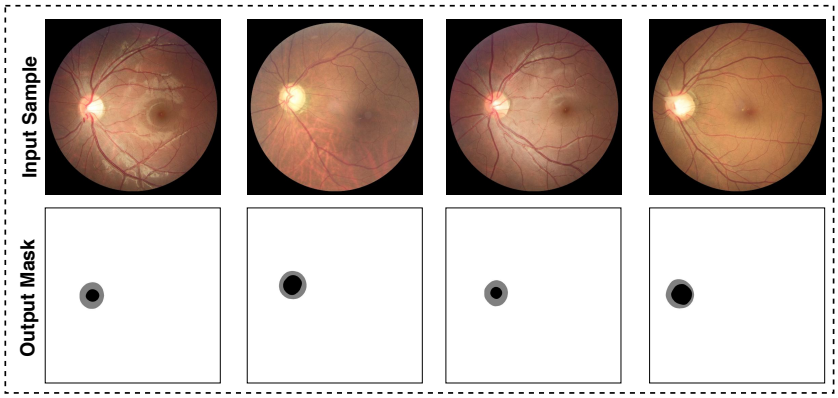

Fig. 7. Some examples of final output on the validation samples.

5 Conclusion

In this paper, we introduce the usage of DenseUNets for optic disc and cup segmentation, which is further used for classification of Glaucoma samples. We use a two-level system, in which the first sub-network (C-Net) localizes the disc region in the original sample which is further used by finer network (F-Net) for segmentation of cup and disc. We empower our methodology by using histogram matching for creating a reference for test samples and exponential transformed images to enhance the quality of image for cup segmentation. Further, we also introduce a new pooling layer for better feature accumulation. We use the dataset provided by REFUGE Challenge and produce our results. We aim to perform a deeper research on creating more efficient methods for cup and disc segmentation.

References

1. Sevastopolsky, A.: Optic disc and cup segmentation methods for glaucoma detection with modification of U-Net convolutional neural network., Pattern Recognition and Image Analysis(2017).https://doi.org/10.1134/S1054661817030269
2. All About Vision, https://www.allaboutvision.com/conditions/glaucoma.htm.
3. REFUGE, https://refuge.grand-challenge.org/background/
4. Jegou, Simon and Drozdzal, Michal and Vazquez, David and Romero, Adriana and Bengio, Yoshua: The One Hundred Layers Tiramisu: Fully Convolutional DenseNets for Semantic Segmentation, Computer Science - Computer Vision and Pattern Recognition(2016).https://doi.org/10.1134/S1054661817030269
5. Huang, G., Liu, Z., Van Der Maaten, L., and Weinberger, K. Q. (2017, July). Densely Connected Convolutional Networks. In CVPR (Vol. 1, No. 2, p. 3).
6. Ronneberger, O., Fischer, P., and Brox, T. (2015, October). U-net: Convolutional networks for biomedical image segmentation. In International Conference on Medical image computing and computer-assisted intervention (pp. 234-241). Springer, Cham.