Diabetic Retinopathy: Segmentation, Detection and Grading.

Khalil Ouardini[1,2], Ramzi Charradi[1], and Benjamin Maheu[1]

[1] Ecole Normale supérieure Paris-Saclay, 61 Avenue du Président Wilson, 94230 Cachan
ramzi.charradi@ens-paris-saclay.fr
[2] CentraleSupélec, 3 Rue Joliot Curie, 91190 Gif-sur-Yvette
{khalil.ouardini,benjamin.maheu}@student.ecp.fr

Abstract. In this work, we tackle the IDRiD automatic grading, segmentation and detection challenge. For each task, we provide solutions and an analysis of our results.

1 Introduction

Diabetic retinopathy is the leading cause of blindness in the working-age population of the developed world. Currently, detecting DR is a time-consuming and manual process that requires a trained clinician to examine and evaluate digital color fundus photographs of the retina. Clinicians can identify DR by the presence of lesions associated with the vascular abnormalities caused by the disease. While this approach is effective, its resource demands are high. The need for a comprehensive and automated method of DR screening has long been recognized, and previous efforts have made good progress using image classification, pattern recognition, and machine learning. The aim of this work is to **evaluate algorithms for automated detection, grading and segmentation** of Diabetic Retinopathy using the IDRiD [13] (Indian Diabetic Retinopathy Image Dataset) dataset of retinal fundus images.

2 Datasets and tasks

The problem is divided in three sub-tasks:

- **Disease Grading**: For this task, the medical experts graded a full set of **516** images with a variety of pathological conditions of DR ranging from 0 if the patient is sane (No apparent retinopathy) to 4 in case of proliferative diabetic retinopathy (PDR). For simplicity we refer to each class in range 0-4 as "Sane", "Mild", "Moderate", "Severe", "Severe+".
- **Lesion Segmentation**: The dataset for the segmentation task provides expert markups of typical diabetic retinopathy lesions and normal retinal structures.

This dataset consists of **81** color fundus images with signs of DR (Diabetic Retinopathy). Precise pixel level annotation of hard exudates (**EX**) is provided as a binary mask. It includes color fundus images (.jpg files) and binary masks made of lesions (.tif files). The EX dataset thus contains 81 images (54 training images and 27 testing images).

– **Automatic localization of optic disc and fovea center coordinates**: The training dataset contains **413** fundus images provided along with all optic disc and fovea center coordinates.

3 Related work

For the first disease grading task, a majority of the proposed solutions use a mix of the IDRiD [13] and the larger Kaggle DR [12] datasets. For instance Son et. al [11] pretrain a custom CNN on the Kaggle DR dataset before fine-tuning the network on the IDRiD dataset. On the other hand, few other teams don't use additional data and leverage models pretrained on large datasets such as ImageNet (e.g The AVASAVA team Varghese Alex et al., used ensembles of pre-trained CNNs for the task of disease grading). In our work we mainly focus on transfer learning techniques and explore effective strategies to achieve competitive results with limited ressources and data.

For the second task (segmentation of rethinopatic lesions) most of the approaches use a UNet architecture [1]. The green channel and CLAHE preprocessing we used for this task was inspired by SAIHST's work [2] which was ranked **5th** in the official challenge's leaderboard. However in our work we do not make use of the padding (They pad each image into a 4352×3072 before cropping them into 204 512×512 images).

For the third task of localizing the retinal landmarks. VRT (Jaemin Son et al.) [3] used a GAN-based network to extract the vessel segmentations and used them as inputs to a CNN in order to find the centers.
Dai et al. (team Deepdr) [4] who was ranekd first in the onsite challenge proposed a two-step method. They first designed a global CNN encoder to localize the OD and fovea centers as a whole by solving a regression task. Second, they use a prior knowledge about the size of the optic disc and Fovea in order to propose some regions randomly but centered on the previously computed centers. Finally, they use a VGG16 [6] architecure to compute precisely the centers from the croped images.
Our proposed method is inspired by their work. However, we make use of the prior before defining our model. Thus, we construct some boxes around the centers and we transform the problem to a regression problem over the boxes. This way, we can use some state-of-the-art object detection methods. To the best of our knowledge, No one of the participants used object detection networks to solve the problem. The majority focused on a regression over the centers. So we found it interesting to explore the power of these networks to solve the problem.

4 Task 1: Disease grading

The objective is to design an image classifier to output a severity grade (0-4) of a given image.

4.1 Approach

To tackle the disease grading challenge, we used transfer learning techniques and finetune ImageNet pre-trained CNNs on the IDRiD dataset. In particular, we used Resnets [5] (precisely the lightweight ResNet50 to deal with ressources constraints). We add three dense layers and freeze all the earlier layers of the network. Deep learning approaches offer state-of-the-art performance for a variety of medical image classification tasks. However, a major challenge in practical translation of these methods is that model training and/or fine-tuning requires thousands of images labelled by domain experts or clinical specialists. Aware of these limitations, we focus most of our efforts on data augmentation techniques to increase the size of our training set ($\approx$ 500 samples initially). To do so we considered multiple strategies.

Data Augmentation : A first method is to randomly introduce small perturbations in the training set (horizontal/vertical flip, rotation, color jitter etc...) to prevent from over-fitting. Few samples are shown in Figure 1. A second method we explored is to incorporate similar labelled images from another (larger dataset). We use 2000 images from the Kaggle Diabetic Retinopathy Detection dataset (Kaggle-DR [12]). This large dataset of 35k samples also contains high-resolution retina images rated by clinicians (on a scale of 0-4), but classification here is far more challenging due to huge variety of imaging conditions

Preprocessing : it is clear that image preprocessing was a crucial part of the Kaggle DR challenge [12] in order to reduce variations in the dataset. To tackle these issues, we squared cropped to cut the neutral background and resized the images to 256 pixels, before subtracting the local average color to reduce differences in lighting. In Figure 2, we show samples of the preprocessed images.

Training and Evaluation: models initialized with the pre-trained ImageNet weights and the parameters of the fully connected layers were optimized by reducing cross-entropy loss with ADAM as an optimizer. The learning rate was initialized to 10^{-4}. Each network was trained for 20 epochs and the model parameters that yielded the highest validation accuracy were used for inference. To facilitate comparisons with the challenge's leaderboard, we report accuracy metrics. After identifying the best model, we will also provide a confusion matrix and the AUC score.

4.2 Results

The main classification experiments we ran are summarized in Table 4.2. In our case, although it demonstrated significant improvements for the Kaggle DR challenge, the preprocessing does not improve the test accuracy. We also find that

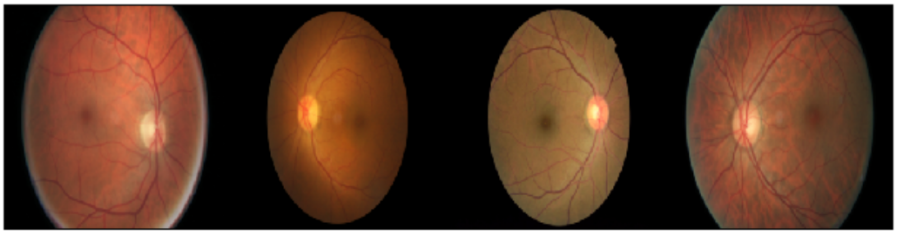

Fig. 1. Samples of the disease grading dataset after data augmentation

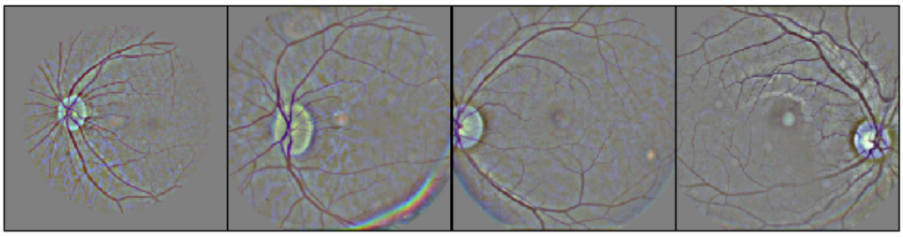

Fig. 2. Samples of the disease grading dataset after preprocessing

data augmentation is not effective. Not surprisingly, the most effective strategy here was to directly add samples form kaggle DR dataset (without data augmentation, and without preprocessing). We reach a test accuracy of 56.3% which places us at the 3rd and 7th spots respectively of the on-site challenge and online challenge leaderboards. Furthermore, for our best model, the AUC score is 0.75 and the confusion matrix is provided in figure 3 of the appendix. A quick analysis of the confusion matrix shows that most misclassification errors lie either in a confusion between classes 0-2 (from sane to mild/moderate), and classes 2-4 (From mild to severe +).

Dataset	**preprocessing**	**Data augmentation**	**accuracy**
IDRiD	No	No	54.5%
IDRiD	No	Yes	51.3%
IDRiD	Yes	Yes	53.4%
IDRiD + Kaggle-DR	Yes	No	48.5%
IDRiD + Kaggle-DR	No	No	**56.3%**

Table 1. Test accuracy (averaged over 2 runs) for various experiments (IDRiD + Kaggle-DR refers to the use of additional Kaggle DR images in the IDRiD training set)

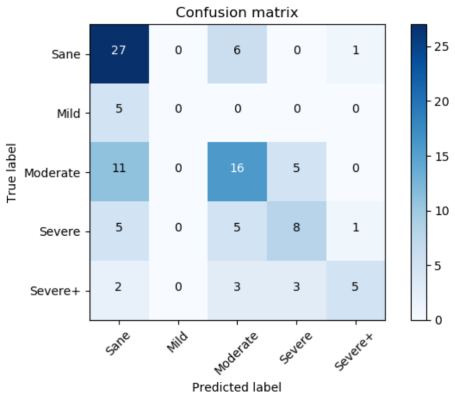

Fig. 3. confusion matrix for the best performing model of the disease grading task

4.3 Discussion

This application is once again an evidence that transfer learning methods can be a very good starting point when dealing with medical imaging classification task, even when the target domain is significantly different. However our experiments also highlight that it is crucial to rely on large datasets when performing disease grading for DR, as we witnessed an improvement of 2 accuracy points, when introducing a small subset of the Kaggle DR images in the IDRiD dataset. The error analysis (3) highlights the difficulties encoutered by the network when discriminating between "similar" classes (e.g "Normal" and "Mild"/"Moderate" or "Severe"/"Severe +"). A possible path of improvement is to pair the multi-label classifier with binary classifier specialized on the harder sub-tasks.

5 Task 2: Segmentation of Rethinopathic Lesions

In this part we focus on Hard Exudates (EX) segmentation but . Our model can of course be generalized to other segmentation tasks on retinal images.

5.1 Approach

Preprocessing - Data Augmentation We resized the input (2848×4288) to a size of (520×782) while conserving the aspect ratio of the original image. We then used multiple preprocessing techniques based on enlightment, color, contrast and HUE enhancement which led us to use CLAHE (Contrast Limited Adaptative Histogram Equalization) from the Albumentations library. Such a preprocessing has proved its efficiency for retinal images [2]. We only take into account the green channel for the EX segmentation as most of the information is stored in this channel.

We have access to a limited number of high resolution images wich led us to use Data Augmentation techniques. To augment our training set we randomly crop the (520×782) inputs into (512×512) images before using a random rotation of $n * 90$ degrees. We then tried different approaches, the first one consisted in arbitrarly enhance the brightness and contrast of the images (respectively of factors 1.3 and 1.2). Another approach was to randomly enhance the brightness, contrast and HUE of the image (with respectively factors 0.4, 0.2, 0.05 using the ColorJitter implementation from Pytorch). Finally, we apply CLAHE and keep the green channel only . Using multiple data augmentation allowed us to choose the one that fits our problem. We also considered using all the RGB channels of our images. In the 5.2 we will discuss and evaluate these different methods.

Training and Evaluation To address this problem we used a UNet architecture with Softmax activation. For the training we used an Adam optimizer with a ReduceLROnPlateau callback (set with patience of 7) and a learning rate of 10^{-4}. We trained our network during 20 epochs on batches of size 8 and experiment with different losses (Mean Square Error, Binary Cross-Entropy, Dice Loss and Mean Dice Loss) to study their influences on the Area Under Precision-Recall Curve (AUPR - metric used in the challenge's leaderboard). We also tried another approach using Global Convolutional Network without success.

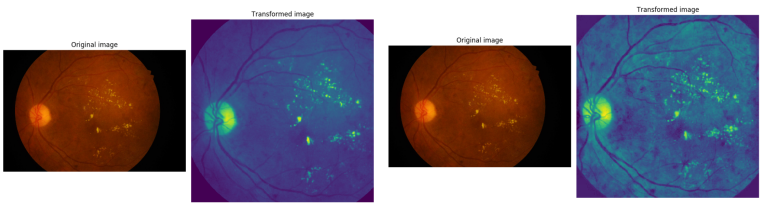

Fig. 4. Illustration of the preprocessing used for the segmentation task. On the left an arbitrary enhanced image (green channel only). On the right the green channel of the original image enhanced by CLAHE

5.2 Results

We report the Areas Under Precision-Recall curve (AUPR) on the test set for the different preprocessing/data-augmentation setups. The results are summarized in Table 2.

5.2.1 Discussion If the overall AUPR for our best model is competitive (ranked **15th** on the official leaderboard [10]) with 56% AUPR, we noticed that the sensitivity is rather low when compared to the precision. Thus by looking at the predicted masks for the Hard Exudate task (EX) we can easily shed light on a few failure cases when the image is not well contrasted or presents insufficient

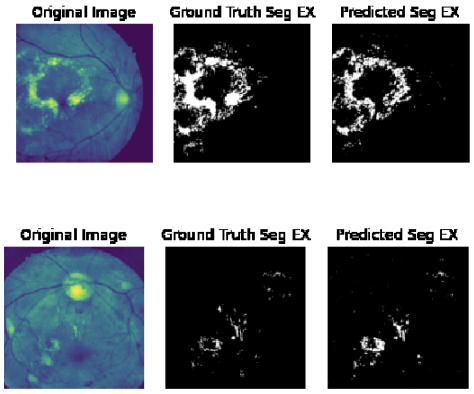

Fig. 5. Some predicted masks from our algorithm (using MSE loss and Color Jitter + CLAHE). (Top) An example of accurate segmentation. (Bottom) On the right a failure case.

brightness after processing. We tried to solve this issue with the preprocessing explained earlier but some cases still lead to a low sensitivity.

In fact we can observe on 5 that when the lesions (EX) are stacked together the model tends to have a better overall precision/recall (by studying the area under ROC curve). However our pipeline seems to struggle with cases where the lesions are isolated ones from the other (few pixels scattered here and there), as the bottom image of 5 shows. Moreover on this latter image a lot of false negative are present in the predicted mask. On most of the cases, in medical imaging we want the False Negative rate to be as low as possible even if this is at the expense of the sensitivity of our model.

The following table 2 shows that our model seems to perform better when we only take the green channel of the pre-processed image. By training with different losses we figured the one that gave us best results on the validation set was the Mean Square Error loss with an overall AUPR of **55%**. The best approach is thus to use either a combination of Color Jitter and CLAHE before taking the green channel or using CLAHE only. For this latter method, we believe the model would be less robust and will most likely be sensitive to changes of enlightment (as no random enlightment would be used during training).

Further improvements To prevent from the highlighted behaviour in presence of scattered lesions we could have considered a patch based approach . The idea is to extract hundreds of patches (520×520) from each original image without resizing, to let the network focus on tiny details (approach explored in [2]).

To decrease our False Positive rate we could also have thought of a watershed technique or an Otsu thresholding: finding a good threshold and rejecting every lesion that have an output below this threshold.

	Arbitrary Image Enhancer	Color Jitter	Only CLAHE	Color Jitter + CLAHE
MSE (RGB)	42%	36%	46%	34%
BCE (RGB)	39%	28%	32%	32%
Dice Loss (RGB)	43%	36%	50%	39%
Mean Dice Loss (RGB)	41%	36%	51%	39%
MSE (Green)	**56%**	**53%**	**55%**	**56%**
BCE (Green)	35%	32%	50%	42%
Dice Loss (Green)	54%	42%	51%	45%
Mean Dice Loss (Green)	51%	45%	49%	43%

Table 2. Area under Precision-Recall Curve for various setups of preprocessing/data-augmentation

6 Task 3: Optic Disc and Fovea Detection

The goal of this task is to detect two points, namely the center of the optic disc and the center of the Fovea.

6.1 Approach

The detection does not have the same degree of difficulty for the two points. More precisely, detecting the optic disc seems to be easier thanks to the brightness of the disc distinguishing it from the rest of the fundus parts. However, some images are not taken in good conditions and we almost have a saturated image.
On the other hand, Fovea detection is a really challenging task, due to the simlilarities of the Fovea with multiple objects in the image and low contrast in some cases.

Data preprocessing : First, we resize all the images to (800,800) RGB images. Since we are working with coordinates, we need to keep track of the resizing factor in order to update the centers' coordinates and the evaluation metrics. Second, we enhance the contrast of the image using the CLAHE transformation and some brightness enhancement using PIL and albumentations libraries.

Hypothesis : We suppose that all the discs have the same size, similarly, we suppose that all the Foveas have the same size. Based on this hypothesis, we construct, for each image, a bounding box of size (108,148) for the optic disc and a bounding box of size (120,120) for the Fovea. Figure 6 illustrates the transformed image.

Training : Now, that we have the bounding boxes we train our object detection model. We chose to test two object detection networks, namely, a two stages network FasterRCNN [7] and a one stage object detection network RetinaNet [8]. For FasterRCNN, We used Pytorch's implementation with a ResNet-50 backbone. The loss function is a weighted sum of regression loss of bounding boxes for the region proposal network and a classification loss over the two classes (OD, Fovea).
For RetinaNet, we used an implementation from github [9] that we modified and

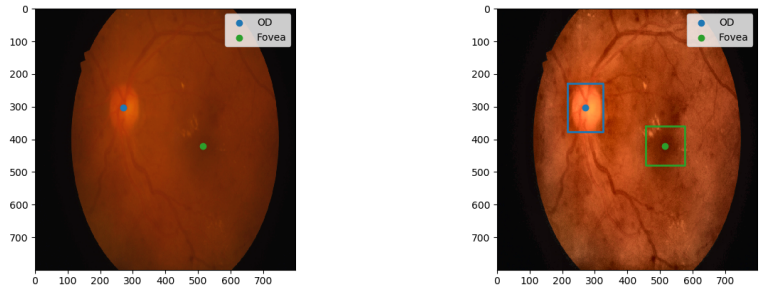

Fig. 6. Illustration of the contrast enhancement and bounding boxes' construction.

adapted to match our dataset and our evaluation's configuration. The loss used is the focal loss. For both networks we used Adam optimizer with a learning rate starting from 10^{-5} and decreasing by half each two epochs. We augment the data by adding some random horizontal flips and we train the network on a Nvidia GTX 1650 GPU.
In order to choose the right box for fovea, we first thought of taking the box that presents a distance from the detected optic disc center, that is the closest to the mean distance between optic disc and fovea in the training set. However, this did not work as expected and the results were unsatisying.
Finally, we decided to use the non maximum supression technique and take the box with the highest score. Once we have the box, we consider its center as the predicted center of the optic disc or fovea.

6.2 Results

We evaluate our methods on the data provided for submission to the IDRiD challenge. We compute the mean euclidean distance between predicted opic disc center and ground truth disc center (resp. Fovea center). We also compute the Intersection Over Union separately for OD boxes and Fovea boxes. Table 3 summarises our results compared to the first and the second ranked participants in the onsite leaderboard. We tested FasterRCNN and RetinaNet with different depths of the backbone network (Resnet-50, Resnet-101, Resnet-152)

Our proposed method with a RetinaNet of depth 101 ranks **2nd** in optic disc localization and **1st** in Fovea localization in the onsite leaderboard (web page [10]). In the online leaderboard, we rank **1st** in optic disc localization and **7th** in Fovea localization (same web page).

6.3 Discussion

All the networks performed well on the localization task but RetinaNet with depth 101 yielded the best results. We expected the deeper network (depth

	OD distances	Fovea distances	IoU for OD	IoU for Fovea
Deepdr (1st)	**21.072**	64.492	-	-
VRT (2nd)	33.538	68.466	-	-
FasterRCNN	42.818	76.817	0.891	0.796
RetinaNet (50)	44.626	114.884	0.892	0.766
RetinaNet (101)	21.398	**54.309**	**0.91**	**0.836**
RetinaNet (152)	43.531	112.076	0.897	0.757

Table 3. IoU, Euclidean distance of OD and fovea location from the ground truth. Unit of the distance is pixel in the original image of 4288×2848.

152) to perform better. However, since we have a limited number of images, the network seems to overfit quickly on the training set.
Previously, we fixed our hypothesis that all the sizes of the bounding boxes are the same which is obvioulsy not true. However, we see that the network was able to detect the objects and propose boxes with a different size.
Figure 7 show some failure cases. The major reasons are :
- The optic disc box is smaller than what it should be, in this case, the network detects the optic disc but, inside of the box we have probably a homogeneous part, so we could lack some precision. However, it's not drastically bad.(image 1)
- Network misses the Fovea because of the similarities with other parts in the image. (image 2 and 4)
- The image is not well contrasted (image 3)

The choice of the box size when assigning the boxes for the Fovea is crucial, a small box will yield a higher variance in the prediction because of the similarities in the image. On the other hand, a big box wil help localizing the Fovea but it will decrease the precision.

Further improvements :
A possible way to improve the model is to refine the hypothesis and find a better way to assign the bounding boxes. This could be done manually since the dataset is not very large. However, in a more general case with much more data it would be problematic.
Another way of improvement, is to add a second stage like in the work of Dai et al [4]. In our work, once we predict the bounding box we simply compute its center and consider it as the center of OD (resp.Fovea). A second CNN stage would aim to find the true OD or Fovea center given that bounding box.
Furthermore, We could try with a better and more robust network architecture like YOLOv3.

7 Conclusion

In this work we proposed algorithms to solve automatic grading, segmentation and detection of the IDRiD dataset, and competed with the top performing solutions of the leaderboard. For the disease grading task, we reach an accuracy of

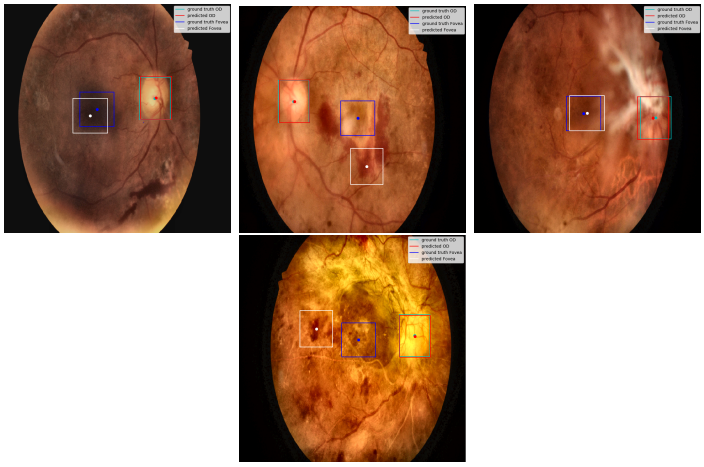

Fig. 7. Examples of failure cases for the detection task

56.3% which places us at the 3rd and 7th spots respectively of the on-site challenge and online challenge leaderboards. We also demonstrated the potential of transfer learning for retinal images classification while highlighting the need for additional data in such supervised tasks. We proposed an encouraging U-Net architecture for hard exudates segmentation and highlighted the feasibility of such a task without extra post-processing techniques. For the detection of OD and Fovea, we reformulate the prblem to a box regression problem and we use state of the art object detection networks to find the coordinates. Our approach was ranked first for OD in the online leaderboard and second in onsite leaderboard, Similarly for Fovea, our proposed method after advanced preprocessing reaches high ranks in the leaderboard.

References

1. Ronneberger et al., 2015: U-Net: Convolutional Networks for Biomedical Image Segmentation
2. Choi et al.: DENSELY CONNECTED U-NET FOR HARD EXAUDATE SEGMENTATION
3. Jaemin Son, Woong Bae (team VRT).: ISBI 2018 SUB-CHALLENGE 3: OPTIC DISC AND FOVEA DETECTION.
4. Dai et al. (team Deepdr).: DEEP OPTIC DISC AND FOVEA LOCALIZATION USING GLOBAL AND LOCAL FEATURE ENCODERS.
5. Kaiming He, Xiangyu Zhang et al : Deep Residual Learning for Image Recognition, https://arxiv.org/pdf/1512.03385.pdf
6. Karen Simonyan, Andrew Zisserman : Very Deep Convolutional Networks for Large-Scale Image Recognition, https://arxiv.org/pdf/1409.1556.pdf
7. Shaoqing Ren, Kaiming He et al. : Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks

8. Tsung-Yi Lin, Priya Goyal et al. : Focal Loss for Dense Object Detection
9. RetinaNet implementation : https://github.com/yhenon/pytorch-retinanet
10. https://idrid.grand-challenge.org/Leaderboard/
11. J. Son, W. Bae, S. Kim, S.J. Park, K.-H. Jung Classification of findings with localized lesions in fundoscopic images using a regionally guided CNN Computational Pathology and Ophthalmic Medical Image Analysis, Springer (2018), pp. 176-184
12. Diabetic Retinopathy Detection: https://www.kaggle.com/c/diabetic-retinopathy-detection
13. Prasanna Porwal, Samiksha Pachade, Ravi Kamble, Manesh Kokare, Girish Deshmukh, Vivek Sahasrabuddhe, Fabrice Meriaudeau. (2018). "Indian Diabetic Retinopathy Image Dataset (IDRiD)." Web.