Classification of Findings with Localized Lesions in Fundoscopic Images Using a Regionally Guided CNN

Jaemin Son[1], Woong Bae[1], Sangkeun Kim[1], Sang Jun Park[2],
and Kyu-Hwan Jung[1](✉)

[1] VUNO Inc., Seoul, Korea
{woalsdnd,iorism,sisobus,khwan.jung}@vuno.co

[2] Department of Ophthalmology, Seoul National University Bundang Hospital
and Seoul National University College of Medicine, Seongnam, Korea
sangjunpark@snu.ac.kr

Abstract. Fundoscopic images are often investigated by ophthalmologists to spot abnormal lesions to make diagnoses. Recent successes of convolutional neural networks are confined to diagnoses of few diseases without proper localization of lesion. In this paper, we propose an efficient annotation method for localizing lesions and a CNN architecture that can classify an individual finding and localize the lesions at the same time. Also, we introduce a new loss function to guide the network to learn meaningful patterns with the guidance of the regional annotations. In experiments, we demonstrate that our network performed better than the widely used network and the guidance loss helps achieve higher AUROC up to 4.1% and superior localization capability.

1 Introduction

Fundoscopic images provide comprehensive visual clues about the condition of the eyes. For the analysis, ophthalmologists search for abnormal visual features called *findings* from the images and make decisions on *diagnoses* based on the findings discovered. For instance, the severity of diabetic retinopathy (DR) is clinically judged by the existence and the extent of relevant findings (microaneurysm, hemorrhage, hard exudate and cotton wool patch, etc) [14].

In recent years, Convolutional Neural Networks (CNN) have achieved the level of professional ophthalmologists in diagnosing DR and diabetic macular edema (DME) [5,13]. However, CNNs in the literature are trained to make decisions on the diagnoses directly without localizing lesions. There exist several studies that visualize the lesions that contribute to the decision on diagnoses [3,12], though, types of findings were not identified for the lesions.

D. Stoyanov et al. (Eds.): COMPAY 2018/OMIA 2018, LNCS 11039, pp. 176–184, 2018.
https://doi.org/10.1007/978-3-030-00949-6_21

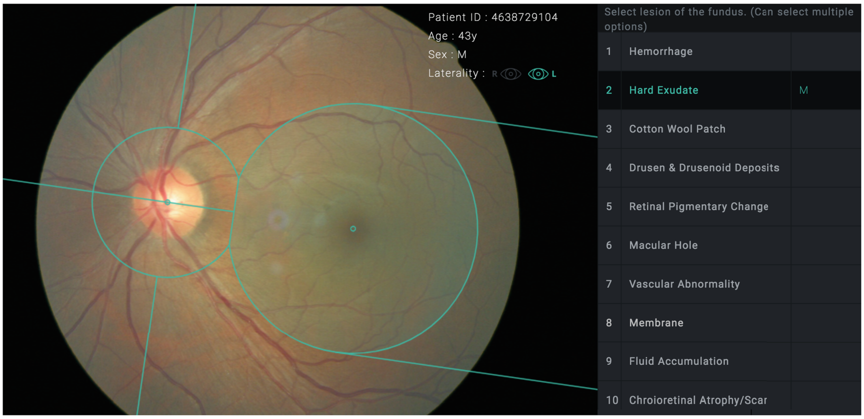
In the past, segmentation methods with hand-crafted feature-extractors had been proposed for the detection of hemorrhage [1], hard Exudate [11], drusen deposits [9] and cotton wool patch [8]. However, since the heuristic feature-extractors embed biases of the human designer regarding visual properties of the target findings, unexpected patterns are not well detected severely constraining the performance in real world applications. CNN for segmentation [2] or detection [10] would improve the performance, however, manual annotation of lesions is labor-intensive especially when they are spread in the images, thus, renders the process of data collection highly expensive.

In this paper, we demonstrate an inexpensive and efficient approach to collecting regional annotation of findings and propose a CNN architecture that classifies the existence of a target finding and localizes the lesions. We show that training with the guidance of regional cues not only helps localize the lesions of findings more precisely but also improves classification performance in some cases. This is possible because the regional guidance encourages the network to learn right patterns of findings instead of biases in the images.

2 Proposed Methods

2.1 Data Collection

We collected regional annotations of findings for macular-centered images obtained at health screening centers and outpatient clinics in Seoul National University Bundang Hospital using a data collection system (Fig. 1). Annotators chose a type of finding on the right panel and selected the corresponding regions on the left panel. When the eye was normal, no finding was annotated to the image.



Patient ID : 4638729104
Age : 43y
Sex : M
Laterality : R L

Select lesion of the fundus. (Can select multiple options)
1 Hemorrhage
2 Hard Exudate M
3 Cotton Wool Patch
4 Drusen & Drusenoid Deposits
5 Retinal Pigmentary Change
6 Macular Hole
7 Vascular Abnormality
8 Membrane
9 Fluid Accumulation
10 Chroioretinal Atrophy/Scar

Fig. 1. Data collection system developed to retrieve regional annotations of findings in fundoscopic images.

We divided an image into 8 regions in a way that each region reflects the anatomical structure of the eyes and the regional characteristics of findings. When the distance between the optic disc and fovea is D, circles are drawn at the centers of the optic disc and fovea with the radius of $\frac{2}{5}D$ and $\frac{2}{3}D$ and the intersections of the two circles are connected with a line segment. Then, a half-line passing through the optic disc and fovea (L) cuts the circle of the optic disc in half and two half-lines parallel to L and tangent to the circle of fovea are drawn in a direction away from the optic disc. Finally, a line perpendicular to L is drawn to pass through the center of the optic disc.

We also separated annotations into training and test sets based on the expertise of the annotators. Training set was annotated by 27 board-certified ophthalmologists and the test set was annotated by 16 certified retina specialists and 9 certified glaucoma specialist. Each fundoscopic image was annotated by 3 ophthalmologists in total. Training and test set amount to 66,473 and 15,451 images respectively. This study was approved by the institutional review board at Seoul National Bundang Hospital (IRB No. B-1508-312-107) and conducted in accordance with the tenets of the Declaration of Helsinki.

2.2 Network Architecture

As shown in Fig. 2, our network architecture consists of residual layer (feature maps after residual unit [6]), reduction layer (feature maps after 3 × 3 conv with stride 2, batch-norm, ReLU), average pooling layer, atrous pyramid pooling layer [2] and 1 × 1 conv (depth = 1) layer.

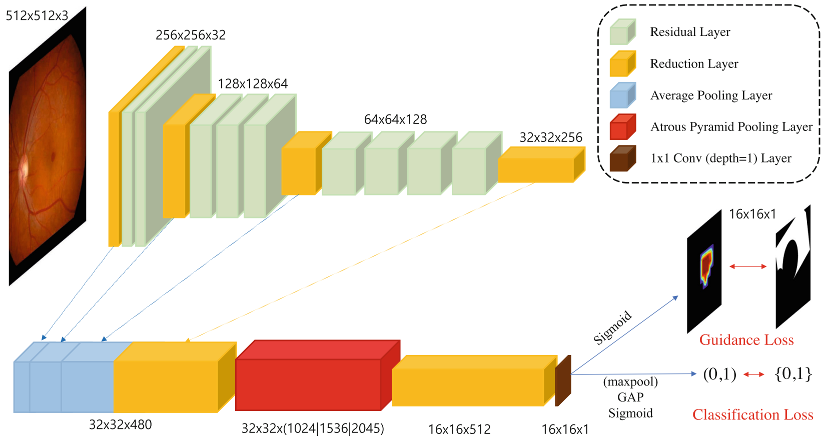


Fig. 2. Proposed network architecture for localization and classification of a specific finding.

The depth of layers doubles when the height and width halve. First four reduction layers with different sizes are concatenated after average pooling to exploit both low level and high level features. The concatenated feature maps are atrous-pyramid-pooled with the dilation rates of $1, 2, 4, 8$ (findings with large size), or the rates of $1, 2, 4$ (findings with medium size) or the rates of $1, 2$ (findings with small size). We employed atrous pyramid pooling to aggregate features with various size of receptive fields [2].

Note that the 1×1 conv layer is a linear combination of the previous feature maps as in class activation map(CAM) [15]. With a sigmoid function, values in the layer are normalized to $(0, 1)$, thus, the consequent layer can be considered as normalized activation map. Our activation map differs from CAM in that additional loss function guides the activation to appear only in desirable areas. Also, the 1×1 conv layer is globally average-pooled and normalized with sigmoid function to yield the prediction regarding the existence. Therefore, the activation map is directly related to the prediction in our network architecture and does not require external operations for visualization. In case of findings with small lesions (hemorrhage, hard exudate, cotton wool patch, drusen and retinal pigmentary change), max-pooling layer is inserted before the GAP layer to compensate for the low GAP values to compensate for low GAP values.

2.3 Objective Function

For a fundoscopic image ($I \in R^{W_I \times H_I}$), the existence of a target finding in the image I is encoded to $y_{true} \in \{0, 1\}$ and the probability of the existence $y_{pred} \in (0, 1)$ is the output from the network. When k images are given as a mini-batch, binary cross entropy for classification loss in Fig. 2 is given by

$$L_{class}(\mathbf{y}_{true}, \mathbf{y}_{pred}) = \frac{1}{k} \sum_{i=1}^{k} \big[-y^i_{true} \log y^i_{pred} - (1 - y^i_{true}) \log(1 - y^i_{pred}) \big] \quad (1)$$

where $\mathbf{y}_{true} = \{y^1_{true}, \ldots, y^k_{true}\}$ and $\mathbf{y}_{pred} = \{y^1_{pred}, \ldots, y^k_{pred}\}$.

When the last feature maps are size of $W_F \times H_F$, a region mask for a target finding ($M \in \{0, 1\}^{W_F \times H_F}$) is given as label and the activation map ($A \in (0, 1)^{W_F \times H_F}$) is generated from the network. With a mini-batch of size k, guidance loss in Fig. 2 is given by

$$L_{guide}(\mathbf{A}, \mathbf{M}) = \frac{1}{kW_F H_F} \sum_{i=1}^{k} \sum_{l=1}^{W_F H_F} (1 - m^i_l) \log(\max(a^i_l, \epsilon)) \quad (2)$$

where $\mathbf{A} = \{A^1, \ldots, A^k\}$ and $\mathbf{M} = \{M^1, \ldots, M^k\}$ and m^i_l and a^i_l are values at lth pixel in M^i and A^i for $l = 1, \ldots, W_F H_F$. Note that $\epsilon > 0$ is added inside the logarithm for numerical stability when $a^i_l \approx 0$. In a nutshell, the guidance loss suppresses any activation (a^i_l) in regions where the value of the mask is 0 ($m^i_l = 0$) and has no effect for activation inside the mask ($m^i_l = 1$).

Then, total loss is given by combining the classification loss and the guidance loss,

$$L_{total} = L_{class}(\mathbf{y}_{true}, \mathbf{y}_{pred}) + \lambda L_{guide}(\mathbf{A}, \mathbf{M}) \tag{3}$$

where λ balances two objective functions.

3 Experiments

3.1 Experimental Setup

We selectively show results of clinically important findings - findings associated with DR and DME [14] (hemorrhage, hard exudate, drusen, cotton wool patch (CWP)), macular hole, membrane and retinal nerve fiber layer defect (RNFL defect). The training set is split into derivation set (90%) and validation set (10%). Model was optimized with the derivation set until the validation loss stagnates and exacerbates. The model with the lowest validation loss is tested on the test set which we regard as gold standards. We defined that a target finding is absent when no ophthalmologists annotated and present when more than 2 out of 3 ophthalmologists annotated. The union of annotated regions is provided as regional cues during training.

We aim to measure the effectiveness of the guidance loss by experimenting with our CNN architecture (Fig. 2) with/without the regional guidance and comparing the results in terms of Area Under Receiver Operating Characteristic curve (AUROC), specificity, sensitivity and activations in the regional cues (AIR). AIR is defined as the summation of activations inside the regional cues divided by the summation of all activations. AIR is measured for both true positive and false negative in classification where the regional cues are available. We used the same network architecture (Fig. 2) to implement the networks with or without the regional guidance by changing λ in Eq. 3 ($\lambda = 0$ without the guidance).

Original color fundoscopic images are cropped to remove black background and resized to 512×512 for the network input. The resized images are randomly augmented by affine transformation (flip, scaling, rotation, translation, shear) and random re-scaling of the intensity. An image is normalized to $[0, 1]$ by dividing by 255. Weights and biases are initialized with Xavier initialization [4]. As an optimizer, we used SGD with *Nesterov* momentum 0.9 and decaying learning rate. Batch size is set to 32 following the recommendation that small batch size leads to better generalization [7]. We set $\epsilon = 10^{-3}$ in Eq. 2 to obtain numerical stability and $\lambda = 1$ in Eq. 3 to treat classification loss and guidance loss equally.

3.2 Experimental Results

Comparison of the performance between inception-v3 [5] and our two models (with/without guidance loss) is summarized in Table 1. We can observe positive effects of the guidance loss to AIR for TP and FN throughout all findings. This is desirable, since it means that the network attends inside the regional cues for

Table 1. Comparison of inception-v3 and our two models (with/without the guidance) on the test set with respect to AUROC. Activations in the regional cues (AIR) on true positive (TP) and false negative (FN) in classification. Among multiple operating points, specificity and sensitivity that yield the best harmonic mean are chosen.

Findings	AUROC			AIR (TP)		AIR (FN)	
	Inception-v3	With	Without	With	Without	With	Without
Macular hole	0.9592	**0.9870**	0.9676	**0.9999**	0.3156	**0.9999**	0.2611
Hard exudate	0.9889	**0.9938**	0.9910	**0.8089**	0.5999	**0.5750**	0.4614
Hemorrhage	0.9760	0.9862	**0.9895**	**0.8890**	0.6388	**0.4899**	0.3857
Membrane	0.9654	**0.9831**	0.9795	**0.9699**	0.3696	**0.8446**	0.2499
Drusen	0.9746	**0.9811**	0.9786	**0.8292**	0.5611	**0.5931**	0.4012
Cotton wool patch	0.9633	**0.9792**	0.9741	**0.8058**	0.5450	**0.4672**	0.4538
RNFL defect	0.9037	**0.9263**	0.8870	**0.7233**	0.4024	**0.4801**	0.2838

classification, thus the network is less likely to learn biases of datasets. Also, difference in AIR is larger in the cases of TP between the two models than those of FN. This is reasonable since FN consists of hard cases for the networks to classify, while TP is relatively easy to be classified with high confidence.

When it comes to AUROC, only macular hole and RNFL defect showed significant improvements. It is interesting to notice that these findings are observed in specific regions. This can be explained by the fact that learning becomes easier as the network is guided to attend to important regions for classification. On the other hand, findings that spread over the extensive areas such as hemorrhage, hard exudate and drusen took less or no advantage of regional cues for classification. We suspect that this happens because these findings would have wide regional cues that the guidance is marginal and the lesions are small that guidance would be more difficult. It is observed that when AUROC is higher, sensitivity is also higher and specificity is lower. However, significant difference is seen only for macular hole and RNFL defect.

In Fig. 3, we qualitatively compare activation maps of networks with/without the guidance loss. Before superimposed onto the original image, activation maps are upscaled through bilinear interpolation and blurred with 32×32 Gaussian filter for natural visualization and normalized to $[0, 1]$. As obvious in the figure, the network generates much more precise activation maps when trained with the regional cues. Though it is not as salient as is segmented pixelwise and includes few false positives in some cases, our activation maps provide meaningful information about the location of the findings which would be beneficial to clinicians. Without the guidance loss, activation maps span far more than the surroundings of lesions and sometimes highlight irrelevant areas.

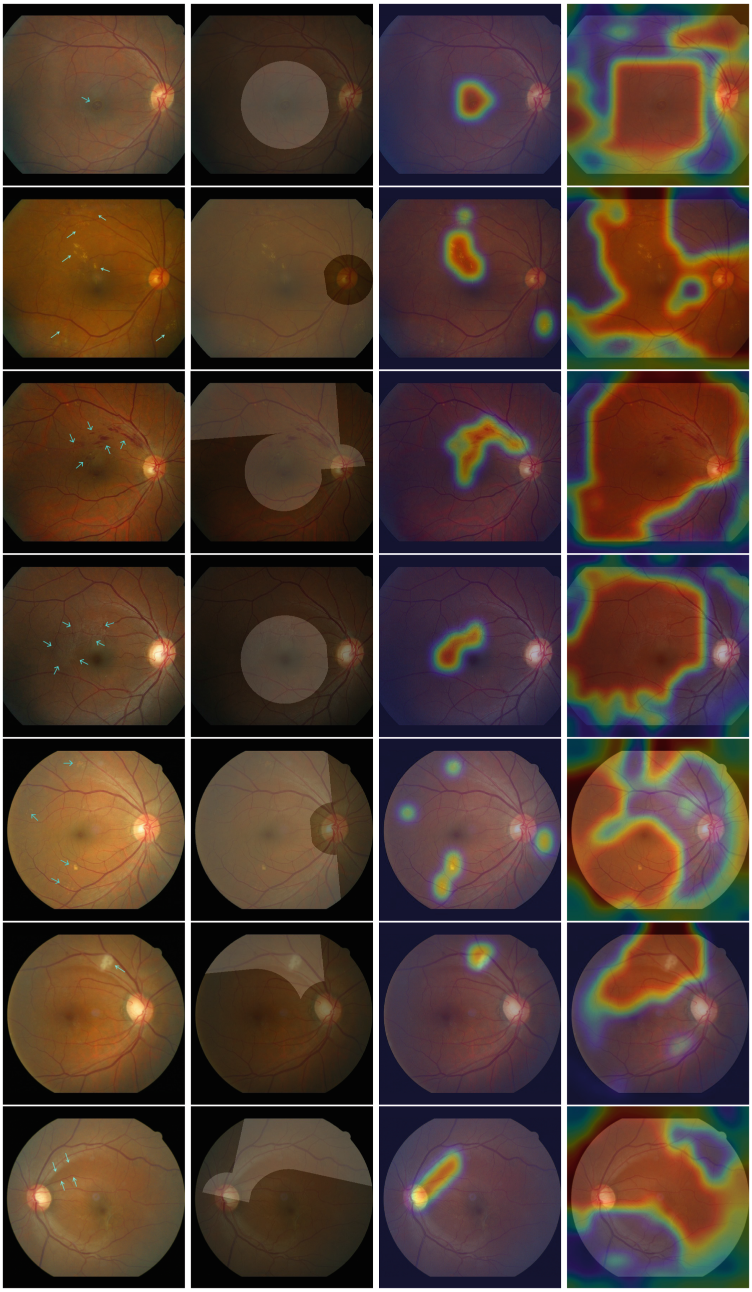


Fig. 3. (From left to right) original fundus image, mask, activation map with and without guidance loss. (From top to bottom) Macular Hole, Hard Exudate, Hemorrhage, Membrane, Drusen, Cotton Wool Patch, RNFL Defect.

4 Conclusion and Discussion

In this paper, we introduced an approach to exploiting regional information of findings in fundoscopic images for localization and classification. We developed an efficient labeling tool to collect regional annotations of findings and proposed a network architecture that classifies findings with localization of the lesions. When trained with the guidance loss that makes use of the regional cues, our network generates more precise activation maps with better attention to the relevant areas for classification. Also, the proposed regional guide also improves the classification performance of findings that occur only at specific regions.

References

1. Bae, J.P., Kim, K.G., Kang, H.C., Jeong, C.B., Park, K.H., Hwang, J.M.: A study on hemorrhage detection using hybrid method in fundus images. J. Dig. Imaging **24**(3), 394–404 (2011)
2. Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: DeepLab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. arXiv preprint arXiv:1606.00915 (2016)
3. Gargeya, R., Leng, T.: Automated identification of diabetic retinopathy using deep learning. Ophthalmology **124**(7), 962–969 (2017)
4. Glorot, X., Bengio, Y.: Understanding the difficulty of training deep feedforward neural networks. In: Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics, pp. 249–256 (2010)
5. Gulshan, V., et al.: Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. JAMA **316**(22), 2402–2410 (2016)
6. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
7. Keskar, N.S., Mudigere, D., Nocedal, J., Smelyanskiy, M., Tang, P.T.P.: On large-batch training for deep learning: generalization gap and sharp minima. arXiv preprint arXiv:1609.04836 (2016)
8. Köse, C., ŞEvik, U., İKibaş, C., Erdöl, H.: Simple methods for segmentation and measurement of diabetic retinopathy lesions in retinal fundus images. Comput. Methods Programs Biomed. **107**(2), 274–293 (2012)
9. Rapantzikos, K., Zervakis, M., Balas, K.: Detection and segmentation of drusen deposits on human retina: potential in the diagnosis of age-related macular degeneration. Med. Image Anal. **7**(1), 95–108 (2003)
10. Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: towards real-time object detection with region proposal networks. In: Advances in Neural Information Processing Systems, pp. 91–99 (2015)
11. Sasaki, M., Kawasaki, R., Noonan, J.E., Wong, T.Y., Lamoureux, E., Wang, J.J.: Quantitative measurement of hard exudates in patients with diabetes and their associations with serum lipid levels. Invest. Ophthalmol. Vis. Sci. **54**(8), 5544–5550 (2013)
12. Takahashi, H., Tampo, H., Arai, Y., Inoue, Y., Kawashima, H.: Applying artificial intelligence to disease staging: deep learning for improved staging of diabetic retinopathy. PLoS One **12**(6), e0179790 (2017)

13. Ting, D.S.W., et al.: Development and validation of a deep learning system for diabetic retinopathy and related eye diseases using retinal images from multiethnic populations with diabetes. JAMA **318**(22), 2211–2223 (2017)
14. Wilkinson, C., et al.: Proposed international clinical diabetic retinopathy and diabetic macular edema disease severity scales. Ophthalmology **110**(9), 1677–1682 (2003)
15. Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A.: Learning deep features for discriminative localization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2921–2929 (2016)