Improving weakly-supervised lesion localization with iterative saliency map refinement

Cristina González-Gonzalo, Bart Liefers, Bram van Ginneken, Clara I. Sánchez
Diagnostic Image Analysis Group, RadboudUMC, Nijmegen, the Netherlands
Cristina.GonzalezGonzalo@radboudumc.nl

Abstract

Interpretability of deep neural networks in medical imaging is becoming an important technique to understand network classification decisions and increase doctors' trust. Available methods for visual interpretation, though, tend to highlight only the most discriminant areas, which is suboptimal for clinical output. We propose a novel deep visualization framework for improving weakly-supervised lesion localization. The framework applies an iterative approach where, in each step, the interpretation maps focus on different, less discriminative areas of the images, but still important for the final classification, reaching a more refined localization of abnormalities. We evaluate the performance of the method for the localization of diabetic retinopathy lesions in color fundus images. The results show the obtained visualization maps are able to detect more lesions after the iterative procedure in the case of more severely affected retinas.

1 Introduction

Interpretation of classification networks is gaining attention in medical imaging in order to increase the expert trust on the obtained prediction. Interpretation algorithms based on visual attribution allow the identification of regions discriminant for the final decision, and, consequently, the weakly-supervised localization and/or segmentation of lesions. This can help assess disease status and grading as well as provide a better understanding of the disease mechanisms without requiring specific lesion-level annotations. Nevertheless, visual attribution based directly on neural network classifiers have been shown to localize only the most significant regions, ignoring lesions that have less influence on the classification result. These lesions could, however, still be important for disease understanding and grading [1]. We propose a novel deep visualization framework that iteratively increases the attention to less discriminative areas through selective inpainting in order to obtain a more accurate localization of abnormal regions. Specifically, we assess the localization performance for the detection of diabetic retinopathy (DR) lesions in retinal color fundus (CF) images.

2 Methods

The proposed framework consists of a baseline deep convolutional neural network to classify CF images by DR stage, from 0 to 4. After training and validating the network, we are interested in finding evidence for the predicted stage. By calculating the saliency map [2] with guided back-propagation [3] we can find which regions in the image could contribute the most to changes in the predicted DR stage if their pixel intensities were modified. Next, we inpaint these regions in the image using [4] to remove the detected abnormalities and increase the "normality" of the images. We iteratively apply the mentioned procedure with the goal of generating a refined saliency map that also includes less discriminative lesions that were not detected at first.

1st Conference on Medical Imaging with Deep Learning (MIDL 2018), Amsterdam, The Netherlands.

2.1 Baseline image-level classification

The baseline network that was used in this study is based on the VGG-16 network architecture [5], pre-trained on ImageNet, and adapted to work on input images of size 512×512 by applying a stride of 2 in the first layer of the first convolutional block, and using a valid instead of padded convolution for the first layer of the last convolutional block. Dropout layers (p=0.5) were added in between the fully connected layers. The network was trained for 500 epochs using stochastic gradient descent, with Nesterov momentum of 0.9. Data augmentation and class balancing were applied during the training phase to reduce overfitting. We follow a regression approach in which the output of the network consists of a single node, representing a continuous value which corresponds to the predicted DR stage. During training, the loss was defined as the mean squared error between the prediction and the ground-truth label.

2.2 Lesion-level detection with iterative saliency map refinement

Those images classified with a category higher than 1, i.e. referable DR, are considered for the localization task. For each image I, we follow the following procedure:

1. Calculate the saliency map s_i, i.e. the derivative of the classification score with respect to the input image, with guided back-propagation. We consider negative gradients in order to observe regions that would lead to a decrease in output if modified, that is, towards a normal prediction.
2. Binarize s_i to create a mask, then inpaint the masked regions to obtain a modified image I_i.
3. Calculate the classification prediction of the baseline network on the modified image I_i.
4. Repeat steps 1-3 until the classification prediction is 1 or lower (non-referable DR) or a number of maximum iterations is reached.
5. The final saliency map S is obtained by an exponentially decaying weighted sum of the iteratively generated maps s_i.

3 Experiments and results

For this study images from the Kaggle DR database were used. The images were graded in different categories, from 0 to 4, where categories 0 and 1 are considered non-referable DR and categories 2 to 4 referable DR. The network was trained on the 80% of the Kaggle training set (28,098 images) and validated on the remaining 20% (7,028 images). All images were rescaled to 512×512 pixels. The baseline network obtained on the validation set an area under the Receiver Operating Characteristic (ROC) curve of 0.93 and a quadratic weighted kappa (κ) of 0.72.

To evaluate the performance of the proposed framework for weakly-supervised lesion localization, 50 images from the validation set of category 2 and 3 were selected. Two experienced graders annotated in consensus DR-related lesions (red lesions, cotton-wool spots and hard exudates) on these images. Figure 1 shows a comparison between saliency map s_0 obtained without the iteration approach and final saliency map S obtained with the proposed approach. The Free-response ROC (FROC) curves can be found in Figure 2 for images from category 2 and 3.

4 Discussion and conclusions

With this study we demonstrate that by means of saliency maps we can achieve lesion-level detections in CF images of DR patients using just image-level labels. Furthermore, we show how an iterative inpainting process can be applied to refine these saliency maps and detect less discriminative lesions that were not present in the original classification, increasing the final performance of the lesion localization task. In addition, the proposed method follows a faster, more accurate and simpler approach than the ones in [1] and [6]. We observe that the original saliency maps focus mainly on red lesions. Regarding the grading protocol, these are the main indicative features of DR and the basis of the ground-truth labels, which guide the learning process of the network. After applying the iterative inpainting method, other lesions related to severe DR stages such as cotton-wool spots and hard exudates are also highlighted, while previously detected abnormalities become emphasized and

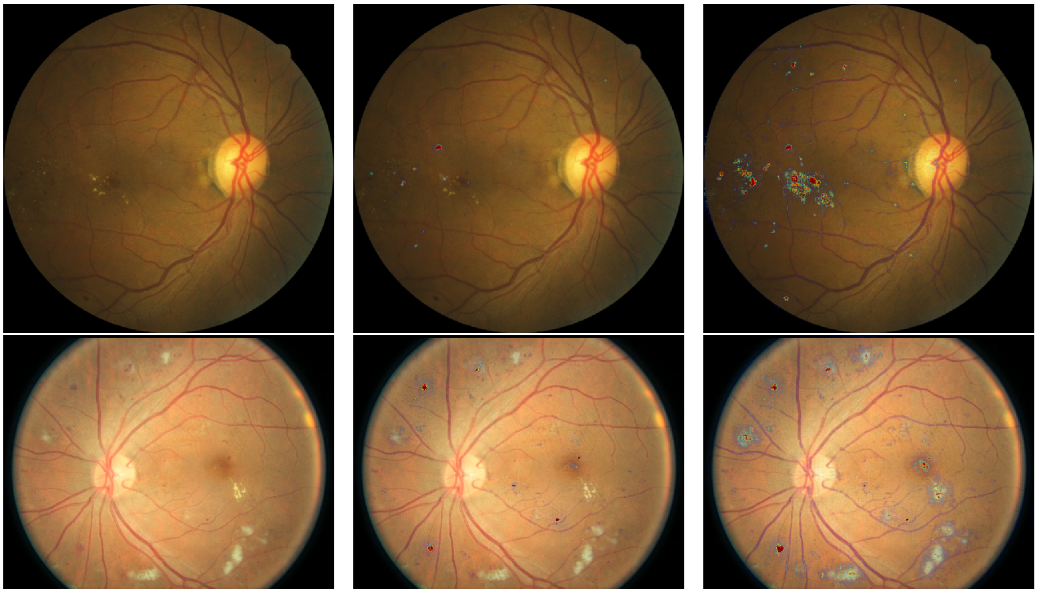

Figure 1: Example heatmaps for two images from DR stage 3. Original image (left), saliency map before (center) and after (right) iterative refinement.

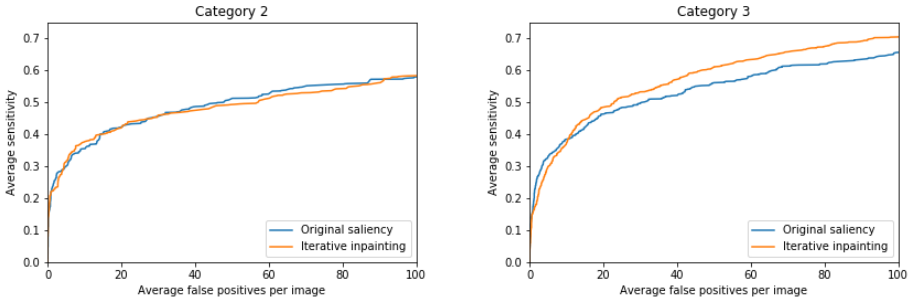

Figure 2: FROC curves for the selected images from category 2 (left) and 3 (right).

better delineated. This can especially be observed in higher DR stages, where the final map differs more from the original one. Future work includes separate evaluation of detection for different types of lesions. We also aim to study variations in classification's interpretability for different network models.

References

[1] Baumgartner, C. F., et al. "Visual Feature Attribution using Wasserstein GANs." arXiv preprint arXiv:1711.08998 (2017).

[2] Simonyan, K., et al. "Deep inside convolutional networks: Visualising image classification models and saliency maps." arXiv preprint arXiv:1312.6034 (2013).

[3] Springenberg, J. T., et al. "Striving for simplicity: The all convolutional net." arXiv preprint arXiv:1412.6806 (2014).

[4] Bertalmio, M., et al. "Navier-Stokes, fluid dynamics, and image and video inpainting." CVPR. Proceedings of the 2001 IEEE Computer Society Conference on. Vol. 1. IEEE, 2001.

[5] Simonyan, K., and Zisserman, A. "Very deep convolutional networks for large-scale image recognition." arXiv preprint arXiv:1409.1556 (2014).

[6] Quellec, G., et al. "Deep image mining for diabetic retinopathy screening." Medical image analysis 39 (2017): 178-193.