
Jia-Hong Huang1†, C.-H. Huck Yang2†,3, Fangyu Liu4, Meng Tian5, Yi-Chieh Liu3,9, Ting-Wei Wu3,7, I-Hung Lin6, Kang Wang8, Hiromasa Morikawa2, Hershghua Chang9, Jesper Tegner2, Marcel Worring1

1University of Amsterdam, 2King Abdullah University of Science and Technology (KAUST), 3Georgia Institute of Technology, 4University of Cambridge, 5Department of Ophthalmology, Bern University Hospital, 6Tri-Service General Hospital, 7University of California, Berkeley, 8Beijing Friendship Hospital, and 9National Taiwan University

† corresponding author: j.huang@uva.nl
‡ work conducted during visiting KAUST.

Abstract

In this work, we propose an AI-based method that intends to improve the conventional retinal disease treatment procedure and help ophthalmologists increase diagnosis efficiency and accuracy. The proposed method is composed of a deep neural networks-based (DNN-based) module, including a retinal disease identifier and clinical description generator, and a DNN visual explanation module. To train and validate the effectiveness of our DNN-based module, we propose a large-scale retinal disease image dataset. Also, as ground truth, we provide a retinal image dataset manually labeled by ophthalmologists to qualitatively show the proposed AI-based method is effective. With our experimental results, we show that the proposed method is quantitatively and qualitatively effective. Our method is capable of creating meaningful retinal image descriptions and visual explanations that are clinically relevant.

1. Introduction

The World Health Organization (WHO) estimates that typical retinal diseases such as Age-related Macular Degeneration (AMD) and Diabetic Retinopathy (DR) are expected to affect over 500 million people worldwide shortly [49]. Besides, generally speaking, the traditional process of retinal disease diagnosis and creating a medical report for a patient takes time in practice. The above means that ophthalmologists will become busier and busier.

As we may know, the current state of the art in Artificial Intelligence (AI) involves deep learning research, and we claim deep learning is one of the promising ways to help ophthalmologists and improve the traditional retinal disease treatment procedure. Deep learning based models such as convolutional neural networks (CNN) or recurrent neural networks (RNN) for computer vision or natural language processing tasks, respectively, have achieved, and, in some cases, even exceeded human-level performance. There is no better time than now to propose an AI-based medical diagnosis method to aid ophthalmologists.

In this paper, we propose an AI-based method for automatic medical report generation based on an input retinal image, as illustrated in Figure 1. The proposed method intends to improve the traditional retinal disease diagnosis procedure, referring to Figure 2, and help ophthalmologists increase diagnosis efficiency and accuracy. The main idea of this method is to exploit the deep learning based models, including an effective retinal disease identifier (RDI) and an effective clinical description generator (CDG), to automate part of the traditional treatment procedure. Then, the proposed method will make the diagnosis more efficient.
To train our deep learning models and validate the effectiveness of our RDI and CDG, we introduce a new large-scale retinal disease image dataset, called DeepEyeNet (DEN). Besides, as ground truth, we provide a retinal image dataset manually labeled by ophthalmologists to qualitatively show that the proposed AI-based model is effective. The dataset helps us show the activation maps of our deep models are aligned with image features that are clinically recognized by ophthalmologists as linked with the identified disease. Our experimental results show that the proposed AI-based method is effective and successfully improves the traditional retinal disease treatment procedure. Our main contributions are summarized as follows:

Contributions.

- To improve the traditional retinal disease treatment procedure and help ophthalmologists increase diagnosis efficiency and accuracy, we propose an AI-based method to generate medical reports for retinal images. In this method, we exploit the deep learning based models including an RDI and a CDG to automate part of the conventional treatment procedure.

- We propose a large-scale retinal disease image dataset, called DeepEyeNet (DEN) dataset, with 15,709 images to train our deep models and validate the effectiveness of the proposed RDI and CDG quantitatively.

- We provide another dataset with 300 retinal images labeled by ophthalmologists to qualitatively show our method is effective by visually confirming the activation maps of our models are aligned with image features clinically recognized by ophthalmologists.

2. Related Work

In this section, we divide the related works into retinal disease classification, image captioning, neural networks visual explanation, and retinal dataset comparison.

2.1 Retinal Disease Classification

Optical Coherence Tomography (OCT), Fluorescein Angiography (FA), and Color Fundus Photography (CFP) are the three most commonly used and important imaging methods for the diagnosis of retinal diseases [61]. Optical Coherence Tomography (OCT) is a technology of emerging biomedical imaging, and it provides high-resolution and non-invasive real-time imaging of highly scattering tissues. That is, OCT images [11, 34, 18] usually are used to show the structure of the retina. [5] have proposed an algorithm to segment and detect six different retinal layers, including Nerve Fiber Layer (NFL), Ganglion Cell Layer (GCL) + Inner Plexiform Layer (IPL), Inner Nuclear Layer (INL), Outer Plexiform Layer (OPL), Outer Nuclear Layer (ONL) + Photoreceptor Inner Segments (PIS), and Photoreceptor Outer Segments (POS), in OCT retinal images. Fluorescein angiography (FA) has been used to realize the pathophysiology course of Retinopathy of Prematurity (ROP) following intravitreal anti-Vascular Endothelial Growth Factor (Anti-VEGF) [36]. Color Fundus photography (CFP) is a simple and cost-effective technology for trained medical professionals. Image preprocessing is one of the important issues in the automated analysis of CFP. The authors of [67] have proposed a method to reduce the vignetting effect caused by non-uniform illumination of a retinal image. In this work, we mainly exploit the DNN-based methods [19, 53, 65] to further investigate the retinal disease classification [62] toward a multi-label task combining language information.

2.2 Image Captioning

Recently, computer vision researchers have proposed a new task, image captioning, and [31, 59, 14] are early works. In [31], the proposed model can embed visual and language information into a common multimodal space. The authors of [14] exploit a natural language model to combine a set of possible words, which are related to several small parts of the image, and then generate the caption of the given image. The authors of [59] use CNN to extract the image feature and use it as the input at the first time step of the RNN to generate the caption of the input image. The authors of [17] propose a new deliberate residual attention network for image captioning. The layer of first-pass residual-based attention prepares the visual attention and hidden states for generating a preliminary version of the captions, while the layer of second-pass deliberate residual-based attention refines them. Since the second-pass is based on the global features captured by the hidden layer.
and visual attention in the first-pass, their method has the potentials to generate better captions. In [41], the authors mention that existing image captioning models are usually trained via maximum likelihood estimation. However, the log-likelihood score of some captions cannot correlate well with human assessments of quality. Standard syntactic text evaluation metrics, such as METEOR [6], BLEU [48], and ROUGE [39], are also not well correlated. The authors of [41] show how to use a policy gradient method to optimize a linear combination of CIDEr [58] and SPICE [3]. In [20], the authors propose a method that focuses on discriminating properties of the visible object, jointly predicts a class label, and explains why the predicted label is proper for a given image. Through a loss function based on reinforcement learning and sampling, their model learns to generate captions. According to [59, 31, 17], existing image captioning models are only able to generate the rough description for a given image. So, in this work, we exploit keywords to make our CDG have better reasoning ability.

2.3 Neural Networks Visual Explanation

There are some popular CNN visualization tools, [70, 52, 64]. The authors of [70] have proposed a technique, called Class Activation Mapping (CAM), for CNN. It makes classification-trained CNN learn how to perform the task of object localization, without using a bounding box. In our previous works [42, 63], we exploit class activation maps to visualize the predicted class scores on retinal images, highlighting the discriminative object parts which are detected by the CNN. In [52], the authors have proposed the other similar features visualization tool, called Gradient-weighted Class Activation Mapping (Grad-CAM), for making a CNN-based model transparent by producing visual explanations of features. The authors of [69] introduce a CNN visualization technique that gives insight into the operation of the classifier and the function of intermediate feature layers. These visualizations allow us to find architectures of CNN models. The authors of [8] propose a generalized method, Grad-CAM++, based on Grad-CAM. The Grad-CAM++ method provides better visual explanations of CNN model predictions than Grad-CAM, in terms of better object localization [66] and occurrences explanation of multiple object instances in a single image. In [38], the authors propose another method different from the above methods which are trying to explain the network. They build up an end-to-end model to provide supervision directly on the visual explanations. Furthermore, the authors validate that the supervision can guide the network to focus on some expected regions. The aforementioned is more related to image data only visualization. The authors of [68, 51, 29, 16] have proposed some methods for the multimedia data, such as text and images, visualization. In [68], the authors introduce five popular multimedia visualization concepts, including basic grid, similarity space, similarity-based, spreadsheet, and thread-based concepts. In this work, we exploit CAM to visually show that the activation maps of our deep models are aligned with image features that are clinically recognized by ophthalmologists as linked with the identified disease. In addition, we use a table-based concept, similar to the static spreadsheet concept, to visualize our medical report.

2.4 Retinal Dataset Comparison

Retinal disease research already has long history and many retinal datasets have been proposed, such as [55, 50, 7, 21, 54, 13, 32, 33, 44, 10, 2, 43, 12, 45, 1, 15, 22, 46, 57]. The DRIVE dataset [55] contains 40 retina images which are obtained from a diabetic retinopathy screening program in the Netherlands. These 40 images have been divided into a half training set and a half test set. For the training images, a single manual segmentation of the vasculature is available. For the test cases, two manual segmentations are available. The IDRiD dataset [50] is a dataset for retinal fundus image consisting of 516 images. The authors of IDRiD dataset provide ground truths associated with the signs of Diabetic Macular Edema (DME) and Diabetic Retinopathy (DR) and normal retinal structures given below and described as follows: (i) Pixel level labels of typical DR lesions and optic disc; (ii) Image level disease severity grading of DR, and DME; (iii) Optic disc and fovea center coordinates. The DRIONS-DB dataset [7] consists of 110 color digital retinal images, and it contains several visual characteristics, such as cataract (severe or moderate), light artifacts, some of the rim blurred or missing, moderate peripapillary atrophy, concentric peripapillary atrophy/artifacts, and strong pallor distractor. The FIRE dataset [21] consists of 129 retinal images forming 134 image pairs, and image pairs are split into three different categories depending on their characteristics. The Drishti-GS dataset [54] contains 101 images, and it is divided into 50 training and 51 testing images. The MESSIDOR dataset [13] has 1200 eye fundus color numerical images. Although the dataset contains a medical diagnosis for each image, there is no manual annotation, such as lesions contours or position, on the images. The DIARETDB0 dataset [32] consists of 130 color fundus images of which 20 are normal, and 110 contain signs of the DR. The DIARETDB1 dataset [33] consists of 89 color fundus images of which 84 contain at least mild non-proliferative signs of the DR, and five are considered as normal which do not contain any signs of the DR. The INSPIRE-AVR dataset [44] has 40 colorful images of the vessels and optic disc and an arterio-venous ratio reference standard. The ONHSDD dataset [10] has 99 retinal images and it is mainly used for the segmentation task. The REVIEW dataset [2] consists of 14 images, and it is also mainly used for the segmentation task. The ROC dataset [43] aims to help patients with diabetes through improving computer-aided detection and diagnosis of DR. The e-ophtha [12] is a dataset...
of color fundus images specially designed for scientific research in DR. The HRF dataset [45] contains at the moment 15 images of healthy patients, 15 images of patients with DR and 15 images of glaucomatous patients. Also, binary gold standard vessel segmentation images are available for each image. The RODREP dataset [1] contains repeated 4-field color fundus photos (1120 in total) of 70 patients in the DR screening program of the Rotterdam Eye Hospital. The CHASE-DB1 dataset [15] is mainly used for retinal vessel analysis, and it contains 14 images. The STARE dataset [22] has 397 images and it is used to develop an automatic system for diagnosing diseases of the human eye. The VARIA [46] is a dataset of retinal images used for authentication purposes, and it includes 233 images. The e-ophth [12] is a large-scale retinal images dataset, DeepEyeNet (DEN), to train our deep learning based models and validate our RDI and CDG. For convenience, we summarize the above retinal datasets in Table 1.

### Table 1: Summary of available retinal datasets

<table>
<thead>
<tr>
<th>Name of Dataset</th>
<th>Field of View</th>
<th>Resolution</th>
<th>Data Type</th>
<th>Number of Images</th>
</tr>
</thead>
<tbody>
<tr>
<td>VICAVR [57]</td>
<td>45°</td>
<td>768 * 584</td>
<td>Image</td>
<td>58</td>
</tr>
<tr>
<td>VARIA [46]</td>
<td>20°</td>
<td>768 * 584</td>
<td>Image</td>
<td>233</td>
</tr>
<tr>
<td>STARE [22]</td>
<td>≈ 30° - 45°</td>
<td>700 * 605</td>
<td>Image + Text</td>
<td>397</td>
</tr>
<tr>
<td>RODREP [1]</td>
<td>45°</td>
<td>2000 * 1312</td>
<td>Image</td>
<td>1,120</td>
</tr>
<tr>
<td>HRF [45]</td>
<td>45°</td>
<td>3504*2336</td>
<td>Image</td>
<td>45</td>
</tr>
<tr>
<td>e-ophth [12]</td>
<td>≈ 45°</td>
<td>2544 * 1696</td>
<td>Image</td>
<td>463</td>
</tr>
<tr>
<td>ROC [43]</td>
<td>≈ 30° - 45°</td>
<td>768 * 576</td>
<td>Image</td>
<td>100</td>
</tr>
<tr>
<td>REVIEW [2]</td>
<td>≈ 45°</td>
<td>1360<em>1024 - 3584</em>2438</td>
<td>Image</td>
<td>14</td>
</tr>
<tr>
<td>INSPIRE-AVR [44]</td>
<td>30°</td>
<td>2392 * 2048</td>
<td>Image</td>
<td>40</td>
</tr>
<tr>
<td>DIARETDB1 [33]</td>
<td>50°</td>
<td>1500 * 1152</td>
<td>Image + Text</td>
<td>89</td>
</tr>
<tr>
<td>DIARETDB0 [32]</td>
<td>50°</td>
<td>1500 * 1152</td>
<td>Image</td>
<td>130</td>
</tr>
<tr>
<td>MESSIDOR [13]</td>
<td>45°</td>
<td>1440<em>960 - 2304</em>1536</td>
<td>Image + Text</td>
<td>1,200</td>
</tr>
<tr>
<td>Drishti-GS [54]</td>
<td>≈ 25°</td>
<td>2045 * 1752</td>
<td>Image</td>
<td>101</td>
</tr>
<tr>
<td>FIRE [21]</td>
<td>45°</td>
<td>2912 * 2912</td>
<td>Image</td>
<td>129</td>
</tr>
<tr>
<td>DRIONS-DB [7]</td>
<td>≈ 30°</td>
<td>600 * 400</td>
<td>Image</td>
<td>110</td>
</tr>
<tr>
<td>IDRiD [50]</td>
<td>50°</td>
<td>4288 * 2848</td>
<td>Image</td>
<td>516</td>
</tr>
<tr>
<td>DRIVE [55]</td>
<td>45°</td>
<td>565 * 584</td>
<td>Image</td>
<td>40</td>
</tr>
<tr>
<td><strong>DeepEyeNet (DEN)</strong></td>
<td><strong>≈ 30° - 60°</strong></td>
<td><strong>various</strong></td>
<td><strong>Image + Text</strong>*</td>
<td><strong>15,709</strong></td>
</tr>
</tbody>
</table>

In this section, we start to describe our proposed DEN dataset in terms of types of retinal images and labels and some statistics of the dataset. Note that some of our group members are experienced ophthalmologists and they help us build the proposed DEN dataset sorted by 265 unique retinal symptoms from the clinical definition and their professional domain knowledge. In our proposed DEN dataset, there are two types of retinal images, grey scale FA and colorful CFP. The total amount of images is 15,709, including 1,811 FA and 13,898 CFP. As with most of the large-scale datasets for deep learning research, we create standard splits, separating the whole dataset into 60%/20%/20%, i.e., 9425/3142/3142, for training/validation/testing, respectively. Each retinal image has three corresponding labels including the name of the disease, keywords, and clinical description. For the total number of retinal diseases, the dataset contains 265 different retinal diseases including the common and non-common. For the keyword and clinical description, it contains 15,709 captions and 15,709 keywords labels. Keyword label denotes important information in the diagnosis process. Clinical description label represents the corresponding caption of a given retinal image. Note that all the labels are defined by retina specialists or ophthalmologists. To better understand our dataset, we show some data examples from the DEN dataset in Figure 3. Also, in Figure 4, we show the word length distribution of the keyword and clinical description labels. Based on Figure 4, we observe the longest word length in our
4. Methodology

In this section, we start to describe the proposed AI-based method for automatic medical report generation. The proposed method is mainly composed of the DNN-based module and DNN visual explanation module.

4.1 DNN-based Module

The DNN-based module contains two components, i.e., a retinal disease identifier (RDI) and a clinical description generator (CDG). We introduce them in the following subsections. Note that we hypothesize an effective RDI and effective CDG help improve the conventional retinal disease treatment procedure and help ophthalmologists increase diagnosis efficiency and accuracy.

Retinal Disease Identifier (RDI). To identify retinal diseases, in our RDI sub-module, we provide two types of deep learning models based on [19, 53], pre-trained on ImageNet, and then trained on the proposed DEN dataset. From the lower level feature perspective, such as color, most of the medical images, e.g., radiology images of the chest, are mainly grey-scale [35] but retinal images are mainly colorful in our dataset. Using the ImageNet pre-trained at least helps extract the better lower level features information. So,
Table 2: This table shows the quantitative results of different RDI models based on our DEN. The RDI model based on [53] with ImageNet pre-training has the best performance. “Pre-trained” indicates the model is initialized from the pre-trained weights of ImageNet. “Random init” means the model’s weights are initialized randomly. Prec@k indicates how often the ground truth label is within the top k ranked labels after the softmax layer. We investigate Prec@1 and Prec@5 due to the need to shortlist candidates of diseases in real-world scenarios. Note that since we have 265 retinal disease candidates and limited training data, it is hard to have good performance in the sense of Prec@1. The situation of limited data is common in medicine.

<table>
<thead>
<tr>
<th>Model</th>
<th>Precision</th>
<th>Random init</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Pre-trained</td>
<td>Prec@5</td>
</tr>
<tr>
<td>He, et al. [19]</td>
<td>37.09</td>
<td>63.36</td>
</tr>
<tr>
<td>Simonyan, et al. [53]</td>
<td>54.23</td>
<td>80.75</td>
</tr>
<tr>
<td>Jing, et al. [30]</td>
<td>32.72</td>
<td>63.75</td>
</tr>
</tbody>
</table>

in this case, we expect that pre-training on ImageNet can improve model performance.

**Clinical Description Generator (CDG).** To generate the clinical description for an input retinal image, we use a pre-trained CNN-based model, such as MobileNetV2, VGG16, VGG19, or InceptionV3, as our image feature encoder and a Long Short-term Memory (LSTM) as our decoder to generate text, referring to Figure 6. When we try to generate the clinical description by the LSTM unit, we incorporate the beam-search mechanism to get the better final output description. In ophthalmological practice, commonly existing keywords, with unordered nature, help ophthalmologists create medical reports. Inspired by this, we exploit keywords to reinforce our CDG sub-module. As shown in Figure 6, we use a keyword embedding module, such as bag of words, to encode our keyword information. Note that when keywords are used to reinforce CDG, it means we will have two types of input features, i.e., image and text features. In this work, we use the average method to fuse these two types of features, referring to Figure 6.

**4.2 DNN Visual Explanation Modules**

There are some existing DNN visual explanation methods, such as [70, 52, 23]. The authors of [70] have proposed a technique, called Class Activation Mapping (CAM), for CNN. It makes classification-trained CNN learn how to perform the task of object localization, without using a bounding box. Furthermore, they exploit class activation maps to visualize the predicted class scores on a given image, highlighting the discriminative object parts which are detected by the CNN. To improve the conventional retinal disease treatment procedure, we incorporate the DNN visual explanation module in our proposed AI-based method. Also, we exploit this module to help verify the effectiveness of the method, referring to our Experiments section.

**4.3 Medical Report Generation**

According to [68, 51], proper multimedia data visualization helps people get insight from the data efficiently. In some sense, we can say that multimedia visualization is a way to visually arrange multimedia data, and it sometimes even helps people get a deeper understanding and extra information from the visualized data. In this work, it contains five multimedia data, including the name of the disease, keyword, clinical description, retinal image, and CAM result image. So, we exploit the table-based concept, which is similar to the static spreadsheet-based concept [68], to visualize our medical report, referring to Figure 7. The medical report visualization intends to help ophthalmologists get the insight from the above image and text data efficiently and also increase the diagnostic accuracy.

**5. Experiments**

In this section, we compare our proposed method to baselines for verifying the effectiveness based on the assumption described in our Methodology section.

**5.1 Retinal Disease Identifier (RDI) Verification**

In our experiment, we try to show that the RDI model with ImageNet pre-training is better than the RDI model without ImageNet pre-training, i.e., our baseline. We exploit the ImageNet-pre-trained DNN-based deep model and non-ImageNet-pre-trained DNN-based deep model with different architectures to do fine-tuning on DEN. For empirical reasons, we use two recipes to train different models. For the RDI model based on [19], we start with a learning rate of 0.1 and decay it 5 times for every 50 epoch. For the RDI model based on [53], we start with a learning rate of 0.001 and decay it 5 times for every 50 epoch. According to the evaluation results in Table 2, we find that the RDI model based on [53] with ImageNet pre-training has better performance than others. We conjecture that RDI models based on [19, 30] may be too complicated for the proposed DEN dataset. Although DEN is a large-scale dataset from the retinal field perspective, the number of training images is still not enough for very deep models. Note that our proposed DEN dataset has 265 classes, including common and non-common retinal diseases or symptoms, and only 8512 training images, so it is not easy to achieve high Prec@1 accuracy for human doctors and AI machines. That is one of the reasons why we investigate both Prec@1 and Prec@5. Also, reporting Prec@5 accuracy is more appropriate from the real-world scenario perspective.

**5.2 Clinical Description Generator (CDG) Verification**

In [25, 28, 26], the authors mention that the evaluation of image description generators is very subjective and there is no such thing as the most proper metric to evaluate the text-to-text similarity. Different text-to-text similarity metrics have different properties, so we exploit six commonly used metrics, including BLEU-1, BLEU-2, BLEU-
Table 3 contains the evaluation results of our CDGs based on the above six different text-to-text similarity metrics. All CDG modules with the keyword-driven method have better performance than the non-keyword-driven CDGs, i.e., our baseline. It implies that using keywords to reinforce the CDGs is effective. Based on Table 3 and [30, 60], we find that the evaluation score of the medical image captioning, based on the above commonly used text evaluation metrics, is much lower than the evaluation score of the natural image captioning. One reason is that, typically, the length of the medical image caption is much longer than the natural image caption. Also, the medical image caption has more abstract words or concepts than the natural image caption. These abstract words/concepts will make algorithms difficult to generate correct captions. The other possible reason is that the innate property of the commonly used text-to-text similarity metrics [25, 28, 26] makes this happen. In addition, in Figure 8, we show some generated clinical description results. Based on Figure 8, we find that although our CDG module cannot always generate correct “age” or “gender”, the models are capable of generating correct descriptions to important characteristics for retinal images.

5.3 Evaluation by DNN Visual Explanation Module

The main idea of DNN visual explanation module evaluation is that if our DNN visual explanation results generated
Table 3: This table shows the evaluation results of our keyword-driven and non-keyword-driven clinical description generators (CDGs). Note that we highlight the best scores of keyword-driven and non-keyword-driven generators in each column, respectively. “w/o” denotes non-keyword-driven baseline generators, and “w/” denotes our proposed keyword-driven generators. “BLEU-avg” denotes the average score of BLEU-1, BLEU2, BLEU-3, and BLEU-4. Note that the model based on “Jing, et al. [30]” has the best performance among all the non-keyword-driven models, and the keyword-driven model based on “Jing, et al. [30]” also has the best performance among all the models. All the keyword-driven models, with the average feature fusion method, are superior to the non-keyword-driven models. So, using keywords to reinforce the CDGs is effective.

<table>
<thead>
<tr>
<th>Model</th>
<th>BLEU-1</th>
<th>BLEU-2</th>
<th>BLEU-3</th>
<th>BLEU-4</th>
<th>BLEU-avg</th>
<th>CIDEr</th>
<th>ROUGE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Karpathy, et al. [31]</td>
<td>0.081</td>
<td>0.031</td>
<td>0.009</td>
<td>0.004</td>
<td>0.031</td>
<td>0.117</td>
<td>0.134</td>
</tr>
<tr>
<td>w/</td>
<td>0.169</td>
<td>0.103</td>
<td>0.060</td>
<td>0.017</td>
<td>0.087</td>
<td>0.120</td>
<td>0.202</td>
</tr>
<tr>
<td>Vinyals, et al. [59]</td>
<td>0.054</td>
<td>0.018</td>
<td>0.002</td>
<td>0.001</td>
<td>0.019</td>
<td>0.056</td>
<td>0.083</td>
</tr>
<tr>
<td>w/</td>
<td>0.144</td>
<td>0.092</td>
<td>0.052</td>
<td>0.021</td>
<td>0.077</td>
<td>0.296</td>
<td>0.197</td>
</tr>
<tr>
<td>Jing, et al. [30]</td>
<td>0.130</td>
<td>0.083</td>
<td>0.044</td>
<td>0.012</td>
<td>0.067</td>
<td>0.167</td>
<td>0.149</td>
</tr>
<tr>
<td>w/</td>
<td>0.184</td>
<td>0.114</td>
<td>0.068</td>
<td>0.032</td>
<td>0.100</td>
<td>0.361</td>
<td>0.232</td>
</tr>
<tr>
<td>Li, et al. [37]</td>
<td>0.111</td>
<td>0.060</td>
<td>0.026</td>
<td>0.006</td>
<td>0.051</td>
<td>0.066</td>
<td>0.129</td>
</tr>
<tr>
<td>w/</td>
<td>0.181</td>
<td>0.107</td>
<td>0.062</td>
<td>0.032</td>
<td>0.096</td>
<td>0.453</td>
<td>0.230</td>
</tr>
</tbody>
</table>

by CAM [70] are accepted by ophthalmologists, it implies that the proposed method is qualitatively effective. To prove the claim, we build the other retinal image dataset with 300 retinal images labeled by ophthalmologists and exploit the CNN visualization tool, CAM, to visualize the learned feature and compare it to the ground truth retinal image. We show the qualitative results in Figure 9. In Figure 9, row (a) shows the four different kinds of raw images of retina diseases and each raw image has a yellow sketch labeled by the ophthalmologist to highlight the lesion areas on the retina. The numbers from (1) to (4) denote the four different diseases, including Optic Neuritis, Macular Dystrophy, Albinotic Spots in Macula, and Stargardt Cone-Rod Dystrophy, respectively. We exploit CAM to generate row (b) to demonstrate the visualization results of our DNN-based model. Then, row (c) is produced by the same method as the row (b). Note that both row (b) and row (c) use the same pre-trained weights of ImageNet but row (b) has fine-tuning on DEN dataset and row (c) has no fine-tuning on DEN. The comparison of row (b) and row (c) shows that the DNN-based model successfully learns the robust features of retinal images by training on our DEN dataset. Also, row (b) indicates that the features learned by DNN agree with the domain knowledge of ophthalmologists. That is to say, the activation maps of our deep models are aligned with image features that are clinically recognized by ophthalmologists as linked with the identified disease. The above experimental results show our proposed AI-based method is qualitatively effective.

6. Conclusion

To sum up, we propose an AI-based method to automatically generate medical reports for retinal images to improve the traditional retinal diseases treatment procedure. The proposed method is composed of a DNN-based module, including RDI and CDG sub-modules, and DNN visual explanation module. To train our deep models and validate the effectiveness of our RDI and CDG, we propose a large-scale retinal disease image dataset, DEN. Also, we provide another retinal image dataset manually labeled by ophthalmologists to qualitatively evaluate the proposed method. Our experimental results show the proposed AI-based method is effective and successfully improves the conventional treatment procedure of retinal diseases.

7. Acknowledgments

This work is supported by competitive research funding from King Abdullah University of Science and Technology (KAUST) and University of Amsterdam.
References


