DeepOpht: Medical Report Generation for Retinal Images via Image Captioning and Visual Explanation

ABSTRACT

In this work, we propose a new multimedia-based method to automatically generate a medical report for retinal images to improve the efficiency of the conventional retinal diseases treatment procedure. This method is composed of two main modules, a deep neural networks-based (DNN-based) module for image caption generation and a DNN visual explanation module. To train our DNN-based module and, at the same time, foster retinal image research, we propose a large-scale retinal disease image dataset, called DeepEyeNet. Also, as ground truth, we provide a retinal image dataset manually labeled by ophthalmologists to qualitatively show that the proposed multimedia-based method is effective. With our experimental results, we show that the proposed method is capable of creating meaningful image captions and visual explanations which are clinically relevant.

KEYWORDS

Image captioning, Explainable AI, Multimedia report, Medical image dataset, Medical image classification

ACM Reference Format:

. 2019. DeepOpht: Medical Report Generation for Retinal Images via Image Captioning and Visual Explanation. In *Proceedings of ACM Conference (MM19).* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

The World Health Organization (WHO) estimates that the typical retinal diseases such as Age-related Macular Degeneration (AMD) and Diabetic Retinopathy (DR) are expected to affect over 500 million people worldwide by 2020 [1] so that the ophthalmologists will become busier and busier in the near future. To the best of our knowledge, the traditional process of retinal disease diagnosis and generating a medical report for a patient takes time in practice. As we know, the current state of the art in Artificial Intelligence (AI) involves deep learning research, and we claim deep learning is one of the best ways to help ophthalmologists increase the diagnosis efficiency and accuracy. 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 or multimedia-based medical diagnosis method to aid ophthalmologists.

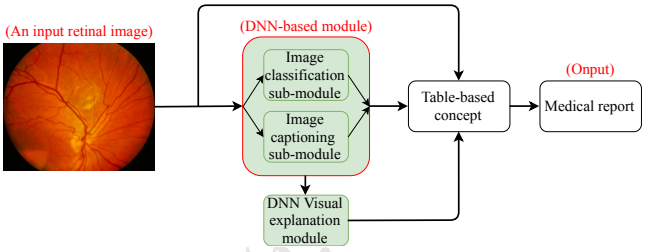


Figure 1: This figure shows the AI-based or multimedia-based medical diagnosis method in ophthalmology expert domain. It contains DNN-based and DNN Visual explanation modules. The DNN-based module is composed of two sub-modules, including image classification and image captioning sub-modules. In this method, the input is a retinal image, and the output is a table-based medical report. This medical report can be some diagnosis reference for ophthalmologists and help them to decide on further treatment. In Figure 2, it shows how to exploit this multimedia-based method to improve the traditional retinal diseases treatment procedure. Note that DNN indicates Deep Neural Networks.

In this paper, we propose a new multimedia-based method to automatically generate a medical report for retinal images, as illustrated in Figure 1, to improve the traditional retinal diseases diagnosis procedure, referring to Figure 2. This method not only helps doctors increase diagnosis efficiency but also improve the diagnosis accuracy. The main idea of this method is exploiting the deep learning based models, including image classification and image captioning models, to automate part of the traditional treatment procedure. Then, the proposed method will make the diagnosis more efficient. Also, we introduce a new large-scale retinal disease image dataset, called DeepEyeNet (DEN), to train our deep learning models and foster retinal image research. Besides, as ground truth, we provide a retinal image dataset manually labeled by ophthalmologists to qualitatively show that the proposed multimedia-based model is effective and thinks like ophthalmologists. Based on our experimental results, we show that our proposed method is capable of creating meaningful retinal image captions and visual explanations which are clinically relevant. The above implies that the proposed multimedia-based method is effective and successfully improves the traditional retinal diseases treatment procedure. In this work, our main contributions are summarized as follows:

Contributions.

- We propose a novel AI-based or multimedia-based method to automatically generate a medical report for retinal images to

MM19, Nice, France

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

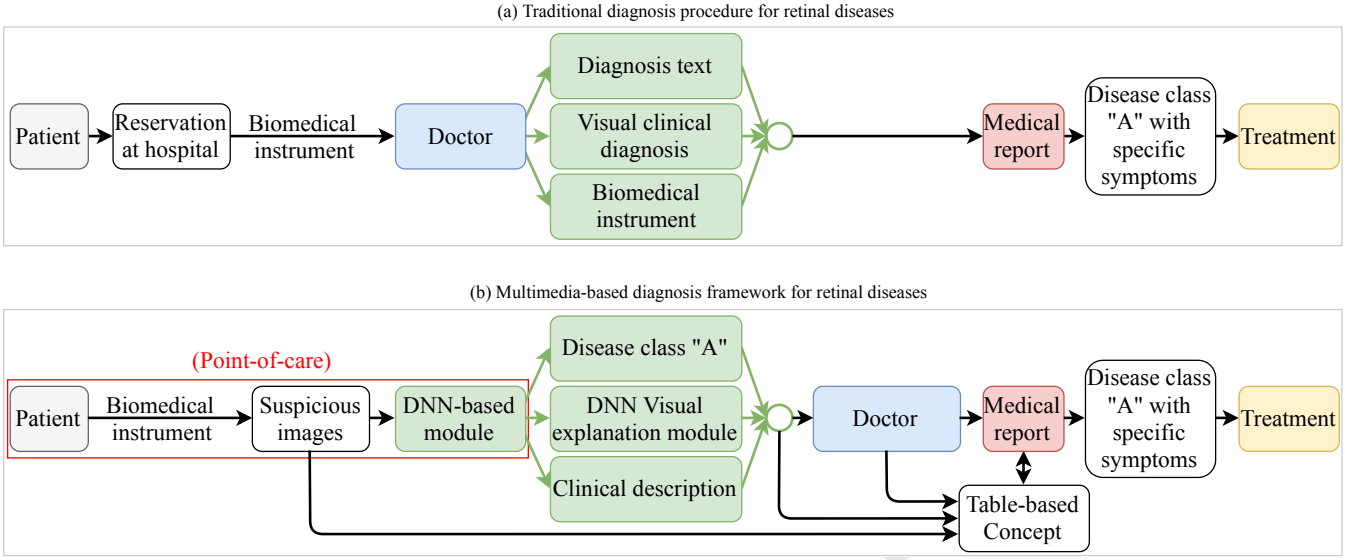


Figure 2: (a) is an existing traditional medical treatment procedure for retinal diseases [2]. Typically, doctors have to handle most of the jobs in the traditional procedure. In (b), we incorporate the AI-based or multimedia-based medical diagnosis method, referring to Figure 1, in the traditional treatment procedure to improve the efficiency of (a), based on the point-of-care (POC) [3] concept. In the proposed method, it mainly contains two modules, including DNN-based and DNN visual explanation modules. The outputs of DNN-based module are "Disease class "A"" and "Clinical description". The DNN visual explanation module will visualize the information from the DNN-based module. Please refer to the METHOD section for a more detailed explanation. Note that DNN indicates Deep Neural Networks in this figure.

improve the traditional retinal diseases treatment procedure. In this method, we exploit the deep learning based models including image classification and image captioning models to automate part of the conventional treatment procedure.

- We propose a large-scale retinal disease image dataset, called DeepEyeNet (DEN) dataset, with 14,185 images to train our image classification and image captioning deep models and foster the retinal disease research.
- We provide another dataset with 300 retinal images labeled by ophthalmologists to qualitatively show that the proposed multimedia-based method is effective and visually confirm that our deep models think like ophthalmologists.

The rest of our paper is organized as follows: In the beginning, we review the related work in Section 2. We introduce and analyze the proposed DEN dataset in Section 3. Then, we start to explain the proposed method in Section 4. Finally, in Section 5, we demonstrate how to evaluate the effectiveness of the proposed method.

2 RELATED WORK

In this section, we divide the related works into four parts including 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. 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 [23–25] usually are used to show the structure of the retina. [26] 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 pathophysiologic course of Retinopathy of Prematurity (ROP) following intravitreal anti-Vascular Endothelial Growth Factor (Anti-VEGF) [27]. 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 [28] 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, including ResNet [29] and VGG [30], to do the retinal disease classification.

2.2 Image Captioning

Recently, computer vision researchers have proposed a new task, image captioning, and [31–33] are early works. In [31], the proposed model can embed visual and language information into a common multimodal space. The authors of [33] 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 [32] use CNN to extract the image

Table 1: Summary of available retinal datasets. Based on this table, we find that our proposed DEN is much larger than the other retinal image datasets, and it contains three types of labels including the name of the disease, keywords, and clinical description. Most of the retinal dataset only contains image data, and the dataset size is not that large. Note that " Text*" denotes clinical description and keywords, referring to subsection 3.2. " Text" denotes only clinical description. So, our DEN dataset is unique.

Name of Dataset	Field of View	Resolution	Data Type	Number of Images
VICAVR [4]	45°	$768 * 584$	Image	58
VARIA [5]	20°	$768 * 584$	Image	233
STARE [6]	$\approx 30^\circ - 45^\circ$	$700 * 605$	Image + Text	397
CHASE-DB1 [7]	$\approx 25^\circ$	$999 * 960$	Image	14
RODREP [8]	45°	$2000 * 1312$	Image	1,120
HRF [9]	45°	3504*2336	Image	45
e-ophtha [10]	$\approx 45^\circ$	$2544 * 1696$	Image	463
ROC [11]	$\approx 30^\circ - 45^\circ$	$768 * 576 - 1386 * 1391$	Image	100
REVIEW [12]	$\approx 45^\circ$	$1360 * 1024 - 3584 * 2438$	Image	14
ONHSD [13]	45°	$640 * 480$	Image	99
INSPIRE-AVR [14]	30°	$2392 * 2048$	Image	40
DIARETDB1 [15]	50°	$1500 * 1152$	Image + Text	89
DIARETDB0 [16]	50°	$1500 * 1152$	Image	130
MESSIDOR [17]	45°	$1440 * 960 - 2304 * 1536$	Image + Text	1,200
Drishti-GS [18]	$\approx 25^\circ$	$2045 * 1752$	Image	101
FIRE [19]	45°	$2912 * 2912$	Image	129
DRIONS-DB [20]	$\approx 30^\circ$	$600 * 400$	Image	110
IDRiD [21]	50°	$4288 * 2848$	Image	516
DRIVE [22]	45°	$565 * 584$	Image	40
DeepEyeNet (DEN)	**$\approx 30^\circ - 60^\circ$**	**various**	**Image + Text***	**14,185**

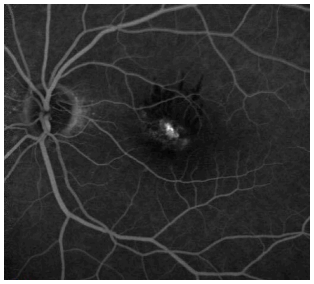
feature and use it as the input at the first time step of the RNN to generate the caption of the input image. The authors of [34] propose a new deliberate residual attention network for image captioning task. The layer of first-pass residual-based attention prepares the visual attention and hidden states 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 potentials to generate better captions. In [35], the authors mention that existing image captioning models are usually trained via maximum likelihood estimation. However, the log-likelihood score of some caption cannot correlate well with human assessments of quality. Standard syntactic text evaluation metrics, such as METEOR [36], BLEU [37], and ROUGE [38], are also not well correlated. The authors of [35] show how to use a policy gradient method to optimize a linear combination of CIDEr [39] and SPICE [40]. In [41], the authors propose a method which 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. To the best of our knowledge, most of the existing state-of-the-art image captioning models are only able to generate the rough description for a given image.

2.3 Neural Networks Visual Explanation

There are some popular CNN visualization tools, [42, 43]. The authors of [42] 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. In [43], 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 [44] 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 [45] 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 and occurrences explanation of multiple object instances in a single image. In [46], 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 [47, 48] have proposed some methods for the multimedia data, such as text and images, visualization. In [47], 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 our deep models think like ophthalmologists and use a table-based concept, similar to the static spreadsheet concept, to visualize our medical report.

2.4 Retinal Dataset Comparison

Name of disease: Geomorphologic Atrophy secondary to AMD
Keywords: Geomorphologic Atrophy; AMD
Clinical description: CFP of the right eye of a 76-year-old man with vision loss for two years shows a hypopigmented macular lesion. OCT reveals RPE atrophy in the macular area.



Name of disease: Central Serous Chorioretinopathy
Keywords: Central Serous Chorioretinopathy
Clinical description: FA of the left eye of a 23-year-old lady with vision loss for 3 weeks. FA shows dot hyperfluorescence in the macula fovea, and blocked fluorescence can be seen around the hyperfluorescence lesion.

Figure 3: Examples from our DEN dataset. Each image has three labels including the name of the disease, keywords, and clinical description. Note that ophthalmologists define all the labels.

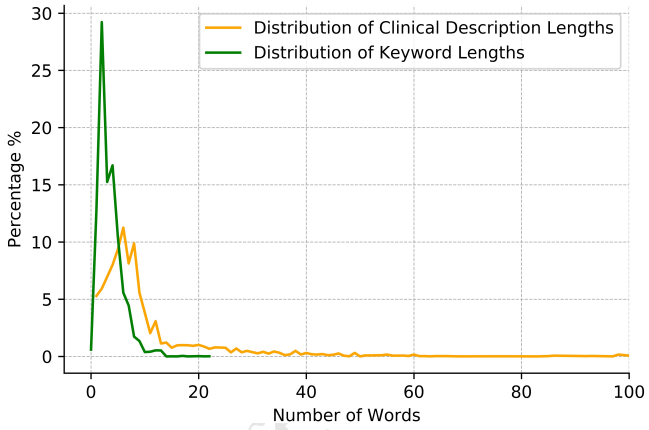


Figure 4: This figure shows the word length distribution of the keyword and clinical description labels. Based on the figure, the word length in our DEN dataset is mainly between 5 and 10 words.

Retinal disease research already has long history and many retinal datasets have been proposed, such as [4–22]. The DRIVE dataset [22] 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 [21] 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 [20] 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 [19] 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 [18] contains 101 images, and it is divided into 50 training and 51 testing images. The MESSIDOR dataset [17] 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 [16] consists of 130 color fundus images of which 20 are normal, and 110 contain signs of the DR. The DIARETDB1 dataset [15] 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 [14] has 40 colorful images of the vessels and optic disc and an arterio-venous ratio reference standard. The ONHSD dataset [13] has 99 retinal images and it is mainly used for the segmentation task. The REVIEW dataset [12] consists of 14 images, and it is also mainly used for the segmentation task. The ROC dataset [11] aims to help patients with diabetes through improving computer-aided detection and diagnosis of DR. The e-ophtha [10] is a dataset of color fundus images specially designed for scientific research in DR. The HRF dataset [9] 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 [8] 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 [7] is mainly used for retinal vessel analysis, and it contains 14 images. The STARE dataset [6] has 397 images and it is used to develop an automatic system for diagnosing diseases of the human eye. The VARIA [5] is a dataset of retinal images used for authentication purposes, and it includes 233 images from 139 different individuals. The VICAVR dataset [4] includes 58 images, and it is used for the computation of the ratio of A/V, (Artery/Vein). In this work, we propose a large-scale retinal images dataset, DeepEyeNet (DEN), to train our deep learning based models and foster the retinal disease research community. In DEN, it has 14,185 images and 265 classes of diseases. For convenience, we summarize the above existing retinal dataset in Table 1.

3 DATASET INTRODUCTION AND ANALYSIS

In this section, we start to describe our proposed DEN dataset in terms of types of retinal images, image 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.

3.1 Retinal Images

In our proposed DEN dataset, there are two types of retinal images, Fluorescein Angiography (FA) and Color Fundus Photography (CFP). FA is monochrome or grey scale, and CFP is colorful. The total amount of images is 14,185, including 1,618 FA and 12,567

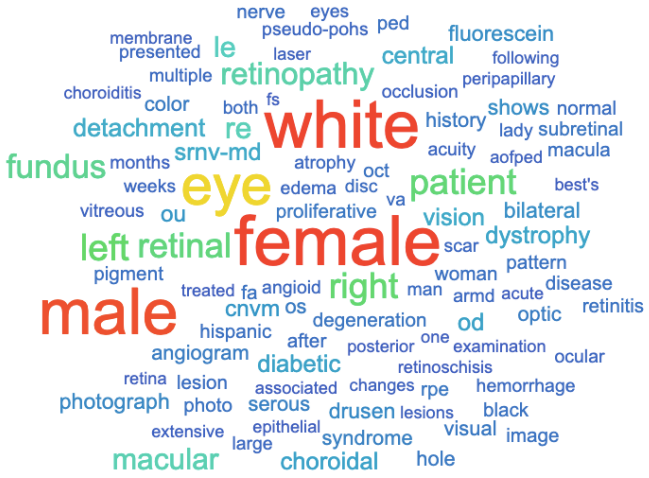


Figure 5: The figure represents Venn-style word cloud for clinical description labels. Note that the word size indicates the normalized counts.

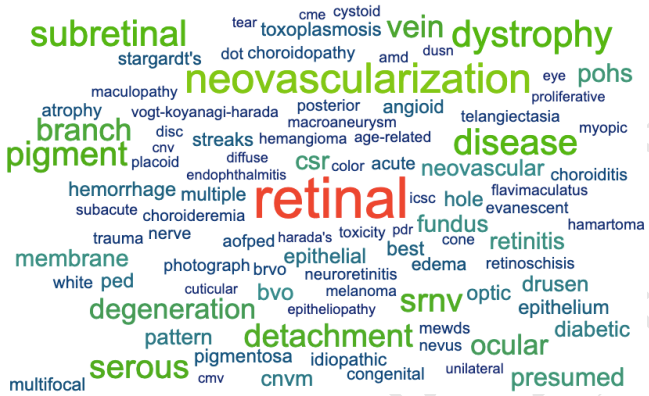


Figure 6: The figure represents Venn-style word cloud for keyword labels. Note that the word size indicates the normalized counts.

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., 8512/2837/2836, for training/validation/testing, respectively.

3.2 Image Labels

In our dataset, 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 14,185 captions and 14,185 keywords labels. Keyword label represents the important words in the label of the name of the disease and clinical description. Clinical description label represents the corresponding caption of a given retinal image. Note that all the labels are defined by retina specialists or ophthalmologists.

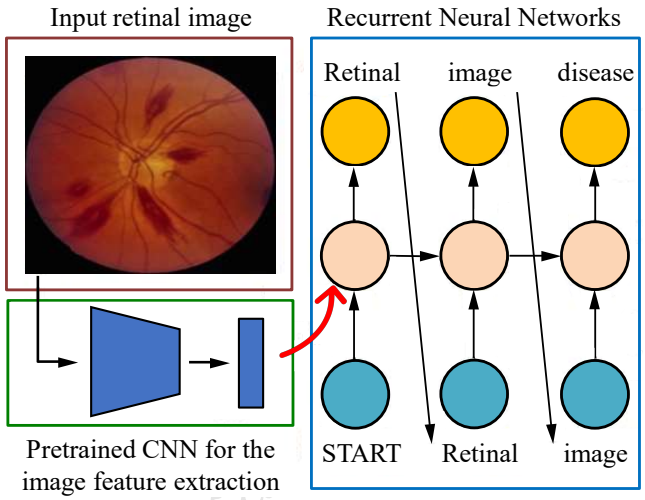


Figure 7: This figure conceptually depicts our image captioning model. In our retinal image captioning model, we exploit a pretrained CNN model to extract the retinal image feature. So, the CNN model is a so-called image encoder. Then, we use an LSTM model, i.e., recurrent neural networks (RNN), as a decoder to generate a word at each time step.

3.3 Statistics of DEN

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. Additionally, we provide the Venn-style word cloud visualization results for keywords and clinical description labels, referring to Figure 5 and Figure 6.

4 METHOD

In this section, we start to describe the proposed multimedia-based method. The proposed method is mainly composed of two modules, including DNN-based and DNN visual explanation modules, and we introduce them in the following subsections.

4.1 DNN-based Module

The main module of our proposed method contains two components, image classification and image captioning. We introduce each of the two sub-modules in the following subsections.

DNN-based Image Classification. In the DNN-based image classification sub-module, we provide two types of deep learning models, including ResNet and VGG models, trained on the proposed DEN dataset. We exploit the existing non-ImageNet-pretrained and ImageNet-pretrained DNN-based models with different architectures to do fine-tuning on our DEN dataset. For empirical reasons, we use two recipes to train different models. For ResNet model, we start with a learning rate of 0.1 and decay it 5 times for every 50 epochs. For VGG model, we start with a learning rate of 0.001 and decay it 5 times for every 50 epochs.

DNN-based Image Captioning. To generate the clinical description for a given retinal image, we use a pretrained CNN model, such as VGG16, VGG19, or InceptionV3, as our image feature encoders and

a Long Short-term Memory (LSTM) as our decoder to generate text, referring to Figure 7. When we try to generate the clinical description by the LSTM unit, we incorporate the beam-search mechanism to get the final output caption. In addition, we re-write the original LSTM formulas and propose a modified LSTM unit figure for better understanding, referring to Figure 8 and equations (1) to (7). Looking carefully at Figure 8 side-by-side with equations (1, 2, 3, 4, 5, 6, 7) together with the corresponding color-coded, it is much easier to understand how to compute the final output of the LSTM unit.

$$\Gamma_u = sigmoid\left(\begin{bmatrix} W_{ua} & W_{ux} \end{bmatrix} \begin{bmatrix} a^{<t-1>} \\ x^{<t>} \end{bmatrix} + b_u\right) \quad (1)$$

$$\Gamma_f = sigmoid\left(\begin{bmatrix} W_{fa} & W_{fx} \end{bmatrix} \begin{bmatrix} a^{<t-1>} \\ x^{<t>} \end{bmatrix} + b_f\right) \quad (2)$$

$$\Gamma_o = sigmoid\left(\begin{bmatrix} W_{oa} & W_{ox} \end{bmatrix} \begin{bmatrix} a^{<t-1>} \\ x^{<t>} \end{bmatrix} + b_o\right) \quad (3)$$

$$\tilde{c}^{<t>} = tanh\left(\begin{bmatrix} W_{ca} & W_{cx} \end{bmatrix} \begin{bmatrix} a^{<t-1>} \\ x^{<t>} \end{bmatrix} + b_c\right) \quad (4)$$

$$c^{<t>} = \Gamma_u * \tilde{c}^{<t>} + \Gamma_f * c^{<t-1>} \quad (5)$$

$$a^{<t>} = \Gamma_o * tanh(c^{<t>}) \quad (6)$$

$$\hat{y}^{<t>} = softmax(a^{<t>}) \quad (7)$$

, where $a^{<t-1>}, a^{<t>}, b_u, b_f, b_o, b_c, c^{<t-1>}, c^{<t>}, \tilde{c}^{<t>}, \Gamma_u, \Gamma_f, \Gamma_o$, $\hat{y}^{<t>} \epsilon \mathbb{R}^{m\times 1}; x^{<t>} \epsilon \mathbb{R}^{n\times 1}; W_{ua}, W_{fa}, W_{oa}, W_{ca} \epsilon \mathbb{R}^{m\times m}; W_{ux}, W_{fx}$, $W_{ox}, W_{cx} \epsilon \mathbb{R}^{m\times n}$; t denotes the time step. $sigmoid()$, $tanh()$, and $softmax()$ are functions with the element-wise operations.

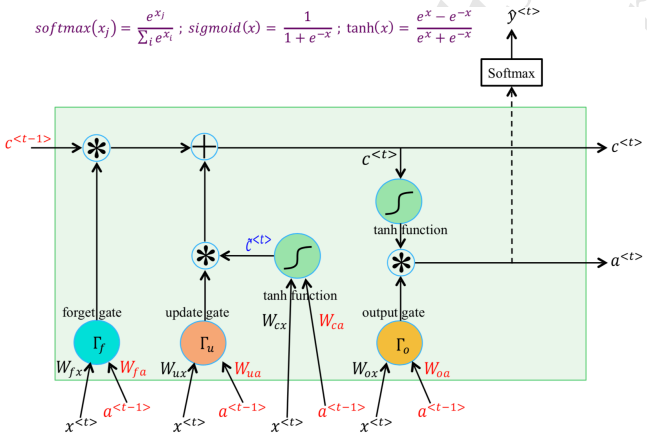


Figure 8: This figure represents the LSTM unit. For the detailed computation, please refer to the Equation (1) to (7). In the figure, Γ_u, Γ_f, and Γ_o denote the update gate, forget gate, and output gate, respectively. $x^{<t>}, a^{<t-1>}$, and $c^{<t-1>}$ represent the inputs and $a^{<t>}, c^{<t>}$ and $\hat{y}^{<t>}$ represent the outputs of the LSTM unit. Note that $\circledast$ and $\oplus$ indicate the element-wise multiplication and element-wise summation, respectively. Please look at the figure and equations from (1) to (7) together for easy understanding.

Table 2: This table shows the quantitative results of different models on the DEN dataset. Then, the VGG16 has the best performance. "Pretrained" indicates the model is initialized from the pretrained 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 both Prec@1 and Prec@5 due to the need to shortlist candidates of diseases in real-world scenarios.

Model	Precision			
	Pretrained		Random init	
	Prec@1	Prec@5	Prec@1	Prec@5
ResNet50 [29]	37.09	63.36	**36.60**	62.87
VGG16 [30]	**54.23**	**80.75**	35.93	**73.73**
VGG19 [30]	32.72	63.75	29.11	60.68

4.2 DNN Visual Explanation Modules

There are some existing DNN visual explanation methods, such as [42, 43]. The authors of [42] 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 diseases treatment procedure, we incorporate the DNN visual explanation module in our proposed method. Also, we exploit this module to help us analyze the effectiveness of the method, referring to EFFECTIVENESS EVALUATION OF THE PROPOSED METHOD section for more details.

4.3 Medical Report Generation

According to [47, 48], 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 mainly 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 [47], to visualize our medical report, referring to Figure 9, to help ophthalmologists get the insight from the above image and text data efficiently and also increase the diagnostic accuracy.

5 EFFECTIVENESS EVALUATION OF THE PROPOSED METHOD

The proposed multimedia-based method mainly contains DNN-based and DNN visual explanation modules. The DNN-based module is composed of the image classification and image captioning sub-modules. In this section, we will evaluate the above modules to show that our proposed method is effective.

5.1 DNN-based Module for Retinal Image Classification Evaluation

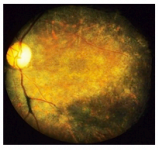
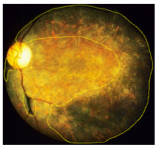
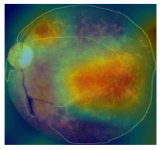
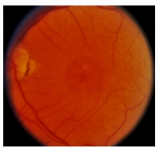
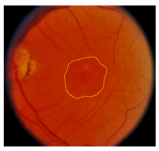
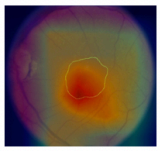
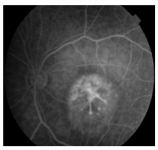
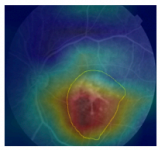
Name of disease	Clinical description	Keywords	Original image	Ground truth labeled by ophthalmologists	CAM result with fine-tuning on our DEN
Bull's Eye Maculopathy Chloroquine	59yr old patient. Had several courses of chloroquine for malaria; native of Africa.	bull's eye maculopathy, chloroquine			
Cone Dystrophy Pattern	69-year-old white male, cone dystrophy pattern.	cone dystrophy			
Bilateral Macular Dystrophy	Fluorescein angiogram of the right eye of a 12-year-old boy with bilateral macular dystrophy.	heredomacular degeneration			

Figure 9: This figure shows the medical reports based on the table-based concept. Since retinal diseases may have some common property or relation, we can put the diseases with the common property or relation together on the table. It can help ophthalmologists get more insights among different retinal diseases. This medical report also can be some diagnosis reference for ophthalmologists and help them to decide on further treatment.

In our experiment, we have three different types of DNNs-based image classification models, such as VGG16, VGG19, and ResNet50. According to the evaluation results of our classification models in Table 2, we find that ResNet50 has the worst performance. Instead, the VGG16 model has better performance than VGG19 and ResNet50 models. It implies that the VGG19 and ResNet50 models may be too complicated for the proposed DEN dataset. To the best of our knowledge, the proposed DEN is the largest retinal dataset. Although it is the large one, the number of training images is still not enough for very deep networks. Note that our proposed DEN dataset has 265 classes, including common and non-common retinal diseases or symptoms, and only 8,512 training images, so it is not easy to achieve high Prec@1 accuracy. That is one of the reasons why we investigate both Prec@1 and Prec@5.

5.2 DNN-based Module for Retinal Image Captioning Evaluation

In [49–51], the authors mention that the evaluation of image captioning results 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-3, BLEU-4 [37], ROUGE [38], and CIDEr [39], to evaluate our retinal image captioning results. In Table 3, it contains the evaluation results of our image captioning models based on the above six different text-to-text similarity metrics. Generally speaking, our image captioning model with the InceptionV3 image feature encoder has the best performance. The performances of image captioning models with VGG16 and VGG19 encoders are very similar. Also, we show some generated image captioning results in Figure 10. Based on Figure 10, we discover that although our image captioning models sometimes cannot generate correct "age" or "gender", the models are capable of generating correct descriptions to important characteristics for retinal images. If we use the commonly used text-to-text similarity metrics to measure the generated captions in Figure 10, the scores will be not that good. However, in practice, the generated captions in Figure 10 are already good and meaningful enough to ophthalmologists. The reason is that "age" and "gender" are not key factors when ophthalmologists analyze a potential retinal disease.

5.3 DNN Visual Explanation Module Evaluation

The main idea of DNN visual explanation module evaluation is that if our DNN visual explanation results generated by CAM [42] are accepted by doctors, it implies that the proposed method is 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 to the ground truth retinal image. We show the qualitative results in Figure 11. In Figure 11, 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 pretrained weights of ImageNet but row (b) has fine-tuning on the DEN dataset and row (c) has no fine-tuning on it. The comparison of row (b) and row (c) shows

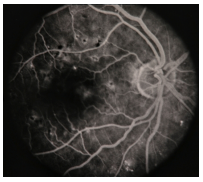
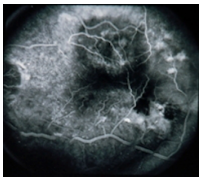
Retinal image	FoV	Ground truth caption	Predicted caption
	30°	67-year-old female with diabetic maculopathy multiple myeloma with retinal detachment.	67 year old patient diabetic maculopathy multiple myeloma with the the in a the to the a retinal detachment.
	25°	75-year-old white male. srnv-md.	60 year old white male. srnv md.

Figure 10: This figure shows some generated retinal image captioning results. Based on this figure, we know that our models can generate meaningful clinical descriptions for ophthalmologists. Note that, in practice, "age" and "gender" are not important factors for ophthalmologists on analyzing a potential retinal disease.

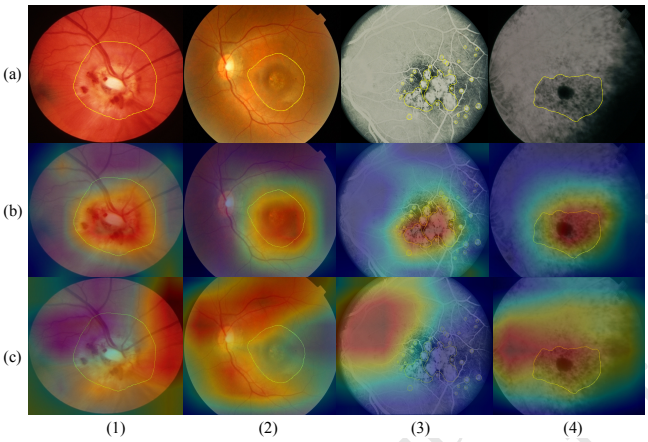


Figure 11: This figure shows the qualitative results of CAM. For the detailed explanation, please refer to subsection 5.3. Note that the results imply that our models think like ophthalmologists in some sense.

that the DNN-based model successfully learns the robust features of retinal disease images by training on the DEN dataset. Also, row (b) indicates that the features learned by DNN agree with the domain knowledge of ophthalmologists. That is to say, our models think like ophthalmologists in some sense. Based on the above experimental results, it implies that our proposed method is effective.

6 DISCUSSION

After the discussion with ophthalmologists, we find that there is another advantage of the proposed method. We know that every medical student needs a lot of training to help themselves increase the diagnostic experience. Since our proposed method can generate a medical report and it includes important medical information, such as retinal diseases prediction and retinal images, this report can be a medical reference for medical students. Also, these medical students

Table 3: This table shows evaluation results of our image captioning models. Note that the results in this table are based on beam-search number 3. "Random" denotes a baseline model with random initialized weights. "Model" denotes our image feature encoder with ImageNet-pretrained weights. Generally speaking, the image captioning model with InceptionV3 image feature encoder has the best performance.

Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4	CIDEr	ROUGE
Random	0.281	0.112	0.069	0.046	0.319	0.218
VGG16	0.626	0.516	0.383	0.349	3.575	0.653
VGG19	0.623	0.521	0.381	0.342	3.562	0.655
InceptionV3	**0.671**	**0.591**	**0.459**	**0.434**	**4.528**	**0.694**

can refine the generated medical report to help the AI-based system becomes better and better. That is to say, in some sense, we also consider the proposed method as a tool for training medical students.

7 CONCLUSION AND FUTURE WORK

To sum up, we propose a novel AI-based or multimedia-based method to automatically generate a medical report for retinal images to improve the efficiency of the traditional retinal diseases treatment procedure. The proposed method is composed of DNN-based and DNN visual explanation modules. The DNN-based module contains two sub-modules including image classification and image captioning sub-modules. To train our deep models and foster the retinal disease research community, we propose a large-scale retinal disease image dataset called DeepEyeNet (DEN) dataset. Also, we provide another retinal image dataset manually labeled by ophthalmologists to evaluate the performance of our DNN visual explanation module. Our experimental results show that the proposed method is effective and successfully improves the traditional treatment procedure of retinal diseases. Finally, we discover that the commonly used text-to-text similarity metrics, such as BLEU, are not that accurate in the clinical description evaluation, so developing a new evaluation metric for medical image captioning is an interesting future work.

REFERENCES

[1] Louis Pizzarello, Adenike Abiose, Timothy Ffytche, Rainaldo Duerksen, R Thulasiraj, Hugh Taylor, Hannah Faal, Gullapali Rao, Ivo Kocur, and Serge Resnikoff. Vision 2020: The right to sight: a global initiative to eliminate avoidable blindness. *Archives of ophthalmology*, 122(4):615–620, 2004.

[2] Melissa H Tukey and Renda Soylemez Wiener. The impact of a medical procedure service on patient safety, procedure quality and resident training opportunities. *Journal of general internal medicine*, 29(3):485–490, 2014.

[3] Nitika Pant Pai, Caroline Vadnais, Claudia Denkinger, Nora Engel, and Madhukar Pai. Point-of-care testing for infectious diseases: diversity, complexity, and barriers in low-and middle-income countries. *PLoS medicine*, 9(9):e1001306, 2012.

[4] SG Vázquez, Brais Cancela, Noelia Barreira, Manuel G Penedo, M Rodríguez-Blanco, M Pena Seijo, G Coll de Tuero, Maria Antònia Barceló, and Marc Saez. Improving retinal artery and vein classification by means of a minimal path approach. *Machine vision and applications*, 24(5):919–930, 2013.

[5] Marcos Ortega, Manuel G Penedo, José Rouco, Noelia Barreira, and María J Carreira. Retinal verification using a feature points-based biometric pattern. *EURASIP Journal on Advances in Signal Processing*, 2009:2, 2009.

[6] Adam Hoover and Michael Goldbaum. Locating the optic nerve in a retinal image using the fuzzy convergence of the blood vessels. *IEEE transactions on medical imaging*, 22(8):951–958, 2003.

[7] Muhammad Moazam Fraz, Paolo Remagnino, Andreas Hoppe, Bunyarit Uyyanonvara, Alicja R Rudnicka, Christopher G Owen, and Sarah A Barman. An ensemble classification-based approach applied to retinal blood vessel segmentation. *IEEE Transactions on Biomedical Engineering*, 59(9):2538–2548, 2012.

[8] Kedir M Adal, Peter G van Etten, Jose P Martinez, Lucas J van Vliet, and Koenraad A Vermeer. Accuracy assessment of intra-and intervisit fundus image registration for diabetic retinopathy screening. *Investigative ophthalmology & visual science*, 56(3):1805–1812, 2015.

[9] J Odstrčilík, Jiri Jan, J Gazárek, and R Kolář. Improvement of vessel segmentation by matched filtering in colour retinal images. In *World Congress on Medical Physics and Biomedical Engineering, Munich, Germany*. Springer, 2009.

[10] Etienne Decencière, Guy Cazuguel, Xiwei Zhang, Guillaume Thibault, J-C Klein, Fernand Meyer, Beatriz Marcotegui, Gwénolé Quellec, Mathieu Lamard, Ronan Danno, et al. Teleophta: Machine learning and image processing methods for teleophthalmology. *Irbm*, 34(2):196–203, 2013.

[11] Meindert Niemeijer, Bram Van Ginneken, Michael J Cree, Atsushi Mizutani, Gwénolé Quellec, Clara I Sánchez, Bob Zhang, Roberto Hornero, Mathieu Lamard, Chisako Muramatsu, et al. Retinopathy online challenge: automatic detection of microaneurysms in digital color fundus photographs. *IEEE transactions on medical imaging*, 29(1):185–195, 2010.

[12] Bashir Al-Diri, Andrew Hunter, David Steel, Maged Habib, Taghread Hudaib, and Simon Berry. A reference data set for retinal vessel profiles. In *2008 30th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 2262–2265. IEEE, 2008.

[13] Retinal Image Computing. Understanding,âĂIJonhsd-optic nerve head segmentation dataset,âĂİ university of lincoln, united kingdom, 2004, 2012.

[14] M Niemeijer, X Xu, A Dumitrescu, P Gupta, B van Ginneken, J Folk, and M Abramoff. Inspire-avr: Iowa normative set for processing images of the retina-artery vein ratio, 2011.

[15] Tomi Kauppi, Valentina Kalesnykiene, Joni-Kristian Kamarainen, Lasse Lensu, Iiris Sorri, A Raninen, R Voutilainen, J Pietilä, H Kälviäinen, and H Uusitalo. Diaretdb1âĂŤstandard diabetic retinopathy database calibration level 1, 2007.

[16] T Kauppi, V Kalesnykiene, et al. Diaretdb0-standard diabetic retinopathy database, calibration level 0. imageret project 2007.

[17] Etienne Decencière, Xiwei Zhang, Guy Cazuguel, Bruno Lay, Béatrice Cochener, Caroline Trone, Philippe Gain, Richard Ordonez, Pascale Massin, Ali Erginay, et al. Feedback on a publicly distributed image database: the messidor database. *Image Analysis & Stereology*, 33(3):231–234, 2014.

[18] Jayanthi Sivaswamy, SR Krishnadas, Gopal Datt Joshi, Madhulika Jain, and A Ujjwaft Syed Tabish. Drishti-gs: Retinal image dataset for optic nerve head (onh) segmentation. In *2014 IEEE 11th International Symposium on Biomedical Imaging (ISBI)*, pages 53–56. IEEE, 2014.

[19] Carlos Hernandez-Matas, Xenophon Zabulis, Areti Triantafyllou, Panagiota Anyfanti, Stella Douma, and Antonis A Argyros. Fire: fundus image registration dataset. *Journal for Modeling in Ophthalmology*, 1(4):16–28, 2017.

[20] Enrique J Carmona, Mariano Rincón, Julián García-Feijoó, and José M Martínez-de-la Casa. Identification of the optic nerve head with genetic algorithms. *Artificial Intelligence in Medicine*, 43(3):243–259, 2008.

[21] Prasanna Porwal, Samiksha Pachade, Ravi Kamble, Manesh Kokare, Girish Deshmukh, Vivek Sahasrabuddhe, and Fabrice Meriaudeau. Indian diabetic retinopathy image dataset: A database for diabetic retinopathy screening research. 2018.

[22] Joes Staal, Michael D Abràmoff, Meindert Niemeijer, Max A Viergever, and Bram Van Ginneken. Ridge-based vessel segmentation in color images of the retina. *TMI*, 23(4):501–509, 2004.

[23] Jeffrey De Fauw, Joseph R Ledsam, Bernardino Romera-Paredes, Stanislav Nikolov, Nenad Tomasev, Sam Blackwell, Harry Askham, Xavier Glorot, Brendan OâĂŹDonoghue, Daniel Visentin, et al. Clinically applicable deep learning for diagnosis and referral in retinal disease. *Nature Medicine*, page 1, 2018.

[24] Andrew Lang, Aaron Carass, Matthew Hauser, Elias S Sotirchos, Peter A Calabresi, Howard S Ying, and Jerry L Prince. Retinal layer segmentation of macular oct images using boundary classification. *Biomedical optics express*, 2013.

[25] Ulrich Gerckens, Lutz Buellesfeld, Edward McNamara, and Eberhard Grube. Optical coherence tomography (oct). *Herz*, 28(6):496–500, 2003.

[26] Ahmet Murat Bagci, Mahnaz Shahidi, Rashid Ansari, Michael Blair, Norman Paul Blair, and Ruth Zelkha. Thickness profiles of retinal layers by optical coherence tomography image segmentation. *American journal of ophthalmology*, 2008.

[27] Domenico Lepore, Graham E Quinn, Fernando Molle, Lorenzo Orazi, Antonio Baldascino, Marco H Ji, Maria Sammartino, Fabio Sbaraglia, Daniela Ricci, and Eugenio Mercuri. Follow-up to age 4 years of treatment of type 1 retinopathy of prematurity intravitreal bevacizumab injection versus laser: fluorescein angiographic findings. *Ophthalmology*, 125(2):218–226, 2018.

[28] Aliaa AA Youssif, Atef Z Ghalwash, Amr S Ghoneim, et al. Comparative study of contrast enhancement and illumination equalization methods for retinal vasculature segmentation. *International Biomedical Engineering Conferemce*, 2006.

[29] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016.

[30] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv:1409.1556*, 2014.

[31] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *CVPR*, pages 3128–3137, 2015.

[32] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *CVPR*, pages 3156–3164, 2015.

[33] Hao Fang, Saurabh Gupta, Forrest Iandola, Rupesh K Srivastava, Li Deng, Piotr Dollár, Jianfeng Gao, Xiaodong He, Margaret Mitchell, John C Platt, et al. From captions to visual concepts and back. In *CVPR*, pages 1473–1482, 2015.

[34] Lianli Gao, Kaixuan Fan, Jingkuan Song, Xianglong Liu, Xing Xu, and Heng Tao Shen. Deliberate attention networks for image captioning. *AAAI*, 2019.

[35] Siqi Liu, Zhenhai Zhu, Ning Ye, Sergio Guadarrama, and Kevin Murphy. Improved image captioning via policy gradient optimization of spider. In *Proceedings of the IEEE international conference on computer vision*, pages 873–881, 2017.

[36] Satanjeev Banerjee and Alon Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72, 2005.

[37] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of ACL*, pages 311–318. Association for Computational Linguistics, 2002.

[38] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. *Text Summarization Branches Out*, 2004.

[39] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575, 2015.

[40] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. Spice: Semantic propositional image caption evaluation. In *European Conference on Computer Vision*, pages 382–398. Springer, 2016.

[41] Lisa Anne Hendricks, Zeynep Akata, Marcus Rohrbach, Jeff Donahue, Bernt Schiele, and Trevor Darrell. Generating visual explanations. In *European Conference on Computer Vision*, pages 3–19. Springer, 2016.

[42] B. Zhou, A. Khosla, Lapedriza. A., A. Oliva, and A. Torralba. Learning Deep Features for Discriminative Localization. *CVPR*, 2016.

[43] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *ICCV*, pages 618–626, 2017.

[44] Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *ECCV*, pages 818–833. Springer, 2014.

[45] Aditya Chattopadhay, Anirban Sarkar, Prantik Howlader, and Vineeth N Balasubramanian. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 839–847. IEEE, 2018.

[46] Kunpeng Li, Ziyan Wu, Kuan-Chuan Peng, Jan Ernst, and Yun Fu. Tell me where to look: Guided attention inference network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9215–9223, 2018.

[47] Jan Zahálka and Marcel Worring. Towards interactive, intelligent, and integrated multimedia analytics. In *2014 IEEE Conference on Visual Analytics Science and Technology (VAST)*, pages 3–12. IEEE, 2014.

[48] Ork De Rooij and Marcel Worring. Efficient targeted search using a focus and context video browser. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 8(4):51, 2012.

[49] Jia-Hong Huang, Modar Alfadly, and Bernard Ghanem. Vqabq: visual question answering by basic questions. *arXiv:1703.06492*, 2017.

[50] Jia-Hong Huang, Cuong Duc Dao, Modar Alfadly, and Bernard Ghanem. A novel framework for robustness analysis of visual qa models. *AAAI Proceeding*, 2019.

[51] Jia-Hong Huang, Modar Alfadly, and Bernard Ghanem. Robustness analysis of visual qa models by basic questions. *arXiv:1709.04625*, 2017.