Expert-defined Keywords Improve Interpretability of Retinal Image Captioning

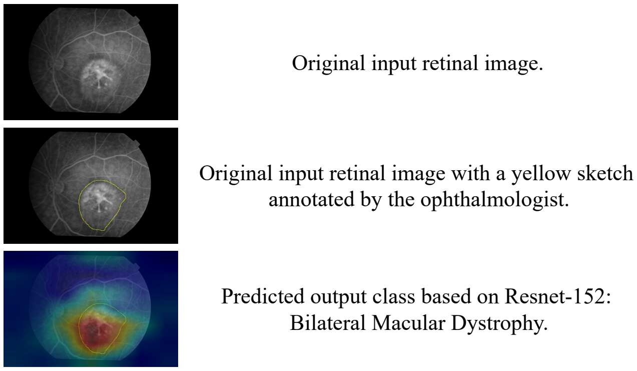
Ting-Wei Wu[2†], Jia-Hong Huang[1†‡], Joseph Lin[3‡], Marcel Worring[1]
[1]University of Amsterdam, [2]Georgia Institute of Technology, [3]University of California, Los Angeles / MediaTek Inc.,
waynewu@gatech.edu, j.huang@uva.nl, joee624@g.ucla.edu / Jachie.Lin@mediatek.com, m.worring@uva.nl
† equal contribution ‡ work conducted during visiting MediaTek Inc.

Abstract

Automatic machine learning-based (ML-based) medical report generation systems for retinal images suffer from a relative lack of interpretability. Hence, such ML-based systems are still not widely accepted. The main reason is that trust is one of the important motivating aspects of interpretability and humans do not trust blindly. Precise technical definitions of interpretability still lack consensus. Hence, it is difficult to make a human-comprehensible ML-based medical report generation system. Heat maps/saliency maps, i.e., post-hoc explanation approaches, are widely used to improve the interpretability of ML-based medical systems. However, they are well known to be problematic. From an ML-based medical model's perspective, the highlighted areas of an image are considered important for making a prediction. However, from a doctor's perspective, even the hottest regions of a heat map contain both useful and non-useful information. Simply localizing the region, therefore, does not reveal exactly what it was in that area that the model considered useful. Hence, the post-hoc explanation-based method relies on humans who probably have a biased nature to decide what a given heat map might mean. Interpretability boosters, in particular expert-defined keywords, are effective carriers of expert domain knowledge and they are human-comprehensible. In this work, we propose to exploit such keywords and a specialized attention-based strategy to build a more human-comprehensible medical report generation system for retinal images. Both keywords and the proposed strategy effectively improve the interpretability. The proposed method achieves state-of-the-art performance under commonly used text evaluation metrics BLEU, ROUGE, CIDEr, and METEOR. Project website: https://github.com/Jhhuangkay/Expert-defined-Keywords-Improve-Interpretability-of-Retinal-Image-Captioning.

1. Introduction

Automatic machine learning-based (ML-based) medical systems, e.g., medical report generation for retinal images, are still not widely accepted [7, 4]. The main reason is that such ML-based systems suffer from a relative lack of explainability/interpretability. Hence, it is hard for humans to understand or at least get an explanation for the machine-made decision. As mentioned in [7], some high-level definitions of interpretability have been proposed by various researchers [6, 31]. For example, the authors of [4] define that interpretability, e.g., in the form of attribute importance, conveys a sense of causality to a system's target group. This concept of causality can only be grasped when the system points out the underlying input-output causal effect relationship. However, precise technical definitions of interpretability still lack consensus. Hence, making a human-comprehensible ML-based medical report generation system is challenging.


Original input retinal image.

Original input retinal image with a yellow sketch annotated by the ophthalmologist.

Predicted output class based on Resnet-152: Bilateral Macular Dystrophy.

Figure 1: A heat map based on CAM [53], i.e., a post-hoc explanation method. Brighter colors (red) indicate regions with higher levels of importance according to the deep neural network, and darker colors (blue) indicate regions with lower levels of importance. "Bilateral Macular Dystrophy" is the predicted retinal disease by an ML-based model.

To improve the interpretability, methods [53, 42] based on heat maps/saliency maps [35] are widely used to highlight or explain how much each region of a medical image contributed to a decision given by an ML-based medical system. However, such methods are well known to be problematic in the broader interpretability literature [1, 7]. Take the result based on CAM [53] in Figure 1 as an example. From an ML-based medical model's perspective, the highlighted areas of the retinal image are deemed most important for the diagnosis/classification of retinal disease. However, from the perspective of ophthalmologists, even the hottest regions (in red) of the heat map contain both non-useful and useful information. Hence, simply localizing the region does not reveal exactly what it was in that area that the ML-based model considered useful. That is to say, the ophthalmologist cannot know if the ML-based model properly established that the presence of the macular pattern was important in the decision, if the vessels were the deciding factor, or if the model had relied on an inhuman feature, such as a particular texture or pixel value that might have more to do with the image acquisition process than the underlying retinal disease. This explainability gap of this widely used interpretability approach, such as [53, 42], relies on humans to decide what a given heat map, i.e., a given explanation, might mean. Unfortunately, human is biased and tends to ascribe a positive interpretation [3, 7]. The same aforementioned issue also happens to an ML-based medical report generation model.

Textual data, e.g., a sequence of expert-defined keywords, is human-comprehensible. Hence, it is an effective carrier of expert domain knowledge. As described in [23], ophthalmologists have usually written down a small set of keywords denoting important information in the early diagnosis process. Hence, they can be collected without much effort [23, 20, 18, 19, 25, 33, 50]. In this work, we propose to exploit interpretability boosters, in particular expert-defined keywords, and a specialized attention-based strategy to build a more human-comprehensible medical report generation system for retinal images. Since the human-comprehensible keywords carry the domain knowledge of ophthalmologists, we exploit them to teach an ML-based model to generate more explainable results. The proposed attention-based strategy is to describe the salient combination of local features that match with keywords in a certain modality, referring to Section 3 for details. Both the expert-defined keywords and the proposed strategy help improve the interpretability.

Contributions

i We propose a more **explainable** retinal image captioning model based on **interpretability boosters**, in particular **expert-defined keywords**.

ii A **novel attention-based strategy** in the transformer decoder is proposed to match human-comprehensible keywords with local image patches. The strategy effectively **reinforces the interpretability** of the proposed method.

iii According to the extensive experiments on the existing large retinal image captioning dataset, when equipped with the context-aware transformer decoder, **performance improvements on the baselines are witnessed in all commonly used metrics**. This demonstrates that the semantic-grounded image representations are effective and can generalize to a wide range of models.

2. Related Work

2.1 Current Methods for Improving Interpretability

Typically, attempts to produce human-comprehensible explanations for an ML-based model's decision have been mainly divided into two categories: inherent interpretability and post-hoc interpretability [7]. A simple ML-based method modeling input data usually has inherent explainability. Take a linear regression model as an example where a simple coefficient measures the direction and strength of the relationship. However, in modern AI use cases, models describing complex data distributions cannot be explained by a simple relationship between inputs and outputs. In such scenarios, many works focus on dissecting the ML-based model's decision-making process, i.e., post-hoc interpretability [53, 42, 40, 43, 16, 17, 22, 10]. In [53], the authors propose a class activation mapping (CAM) technique based on the global average pooling layer proposed in [30]. The proposed CAM builds a generic localizable deep representation that exposes the implicit attention of a convolutional neural network (CNN) on an image. [42] propose a gradient-weighted class activation mapping (Grad-CAM) technique to exploit the gradients of any target concept flowing into the final convolutional layer to generate a coarse localization map, highlighting important image regions. Heat maps are popular and widely used in medical imaging-related fields. They provide a simple means of understanding some of the limitations of post-hoc interpretability techniques [53, 42, 7]. Hence, they are illustrative. However, heat maps are well known to be problematic in the broader interpretability literature [1]. The concerns also extend to other well known post-hoc explanation approaches, e.g., locally interpretable model-agnostic explanations (LIME) [40] and Shapley values (SHAP) [43].

Keywords are meant to represent the important image content while subtly alluding to its semantic relationship. Also, they are effective expert domain knowledge carriers. Hence, in this work, they are used to improve the interpretability gap of the heat map-based explainability methods.

2.2 Natural Image Captioning

The encoder-decoder paradigm is a popular network architecture for image captioning [47, 27], which leads to promising results. A convolution neural network (CNN) is first utilized to encode the image and a recurrent neural network (RNN) is adopted to generate the output word sequence. In [48], a bidirectional LSTM-based approach is proposed to create image descriptions. Both past and future information are utilized at the same time to learn long-term interactions between vision and language. In [39], an area-based attention model is introduced for image captioning. The area-based model predicts the next word and corresponding regions of the image in each RNN time step for creating image captions. The authors of [51] propose to exploit graph convolutional networks (GCN) [41] and Long Short-Term Memory (LSTM) [9] to build an encoder-decoder architecture for image captioning. The graphs are built over the detected objects in an image based on their spatial and semantic connections. In [24], the authors propose an attention on attention (AoA) module to determine the relevance between attention results and queries. The AoA module is based on conventional attention mechanisms, both applied to the encoder and the decoder of an image captioning model. In [37], the authors introduce a unified attention block that employs bilinear pooling to selectively capitalize on visual information. The attention blocks are integrated into the image encoder and sentence decoder to leverage higher-order interactions of multi-modal features.

The aforementioned methods are mainly based on natural images to generate simple/rough image descriptions. Retinal and natural images have very different characteristics, both in objects' sizes and details [23]. Hence, when those natural image-based approaches are directly used to generate captions for retinal images, the quality of the generated medical descriptions still needs improvement.

2.3 Retinal Image Captioning

Medical description generation for a given retinal image, i.e., retinal image captioning, is a challenging computer vision task. In retinal image captioning, long and semantically coherent medical descriptions for a given retinal image must be generated algorithmically [46, 36, 23, 20, 18]. In [46], the authors introduce an ML-based clinical decision support system to assist ophthalmologists more effectively. The proposed system is mainly based on an LSTM-based image captioning model. In [36], an automatic medical description generation model based on CNN and self-trained bidirectional LSTM is proposed. In [23], the authors propose an AI-based method to improve the traditional retinal disease treatment procedure. The proposed model consists of a retinal disease identifier, a clinical description generator, and a CAM-based deep network visual explanation module. Also, the authors propose a large-scale retinal image captioning dataset DeepEyeNet to train and validate their method. The authors of [20] propose a context-driven encoding network to generate more accurate and meaningful medical reports for retinal images. The proposed method is composed of a multi-modal input encoder and a fused-feature decoder. In [18], the authors propose an end-to-end transformer-based model for retinal image description generation. The model is mainly based on the non-local attention mechanism, feature reinforcement module, and masked self-attention.

ML-based models have been proposed for retinal image captioning. However, none of them is clearly interpretable. To build a more human-comprehensible retinal image captioning system, we start from the encoder-decoder based framework. Then, the expert-defined keywords and specialized attention-based strategy, referring to Section 3, are used to reinforce the interpretability of the proposed method.

3. Methodology

In this section, we present the proposed explainable retinal image captioning model as shown in Figure 2. The proposed model is driven by interpretability boosters, i.e., expert-defined keywords. Overall, the model generates a long and semantically coherent medical description from a given retinal image and a list of corresponding expert-defined keywords. In Section 3.1, a more general scenario is also considered, i.e., without expert-defined keywords as input. Given a retinal image, we use a CNN to learn visual features from the image patches, which will be first fed into a multi-label classifier to predict relevant keywords. Note that the predicted keywords are considered as "pseudo" expert-defined keywords. These keywords' embedding vectors will serve as semantic features for the retinal images. After the information extraction, the visual and semantic features are fed to a contextual transformer decoder to sample output words as medical descriptions sequentially. The contextual transformer decoder resembles a pervasive transformer decoder [44] except the input for encoder-decoder attention module is different. We introduce an image-keyword attention-based encoder to fuse information both from images and keywords.

3.1 Interpretability Booster Prediction

According to [23, 20, 18], early in the diagnosis process, ophthalmologists have usually written down a small set of keywords denoting important information. Hence, expert-defined keywords commonly exist in that case. However, expert-defined keywords may not always commonly exist in other fields, e.g., biology, chemistry, or physics. Hence, besides directly using ground truth expert-defined keywords for report generation, we also introduce a multi-label classifier to predict these keywords beforehand of the given image. Note that the correctness of keyword prediction affects the model performance, referring to Section 5.1. Given an image I, we extract its features $\mathbf{v} \in \mathbb{R}^{N \times H_I}$ with a CNN extractor $\phi(\cdot)$ [8] and then feed them in a multi-layer perceptron (MLP) classifier to predict one or more keywords from L vocabulary with a distribution:

$$p(\mathbf{l}_i = 1|\mathbf{v}) \propto e^{(W_i^{MLP}(\mathbf{v}))}, \tag{1}$$

where $\mathbf{l} \in \mathbb{R}^L$ is a keyword vector, $\mathbf{l}_i$ denoting the presence and absence of the i-th keyword. W_i^{MLP} refers to the weight of MLP classifier associated with i-th output. We select the keywords with $p(\mathbf{l}_i = 1|\mathbf{v}) > \tau$ (confidence threshold) as the used keywords to reinforce the decoding process.

3.2 Image & Interpretability Booster Fusion

After generating corresponding keywords, to exploit interactions between the keywords and image, we embed keyword sequences with image content and draw different attention weights on every individual keyword with a self-attention mechanism. To be more specific, for a given set of keywords $\{k_i\}_{i=1}^K$, K is the number of keywords, we first preprocess them by adding a special token "[SEP]" between each keyword to form a complete sequence. We adopt a glove embedding layer W_e to obtain the keyword embedded vector $\mathbf{k} \in \mathbb{R}^{K \times H_e}$, where H_e is the embedding size. Then, we introduce an attention feature mapping $f(\mathbf{v}, \mathbf{k})$, referring to Equation (5). It could be interpreted as mapping an image query Q from image I and a set of keyword key-value pairs K, V from keywords $\mathbf{k}$ to an output Z. Here we leverage the dot-product mechanism for much faster and more space-efficient in exploring the keyword and image relationship. The positional encoding trick is skipped since we do not wish to include redundant sequential information with keywords' unordered nature.

$$Q = W_q\phi(I) + b_q \tag{2}$$

$$K = W_k\mathbf{k} + b_k \tag{3}$$

$$V = W_v\mathbf{k} + b_v \tag{4}$$

$$\begin{aligned} Z &= Attention(Q, K, V) \\ &= softmax(\frac{QK^T}{\sqrt{d_k}})V \end{aligned} \tag{5}$$

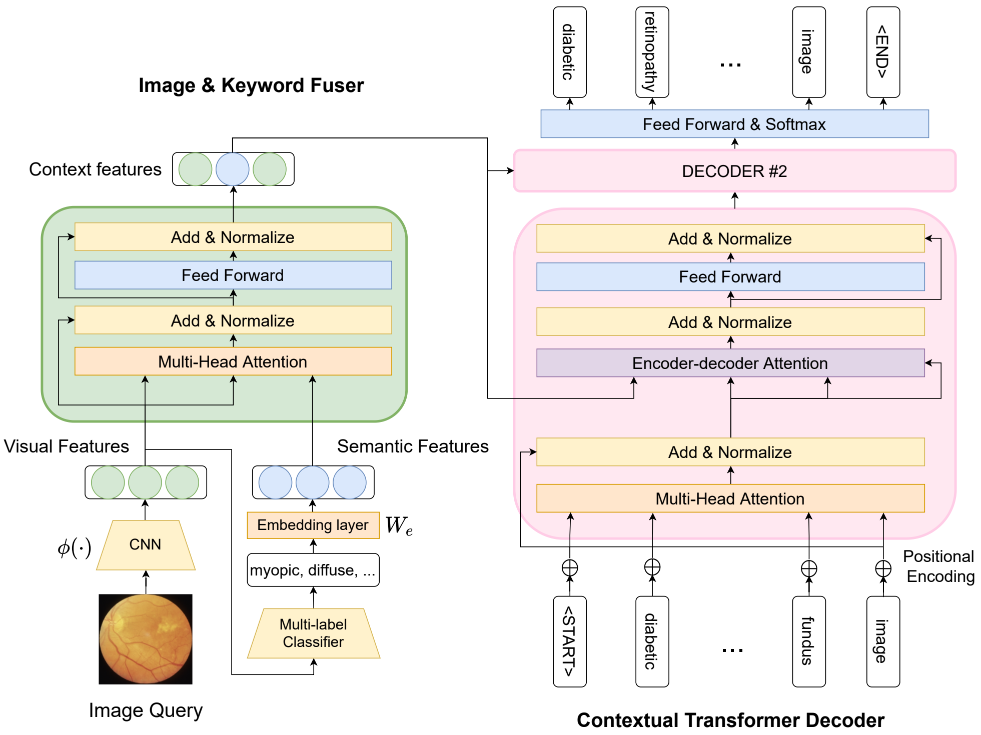


Figure 2: This figure shows the flowchart of the proposed method. It contains an image/keyword fuser and a contextual transformer decoder for medical description generation. Visual and semantic features are respectively retrieved from a CNN extractor and a multi-label classifier. These two features will be fused within a transformer block to weight patch importance. Then the context features will serve as an encoder output for a transformer decoder to generate adequate medical descriptions. “Image & Keyword” deontes “Image & Interpretability Booster”

Similarly, we also employ a residual connection, followed by layer normalization and position-wise feed-forward layers to enhance the model performance.

$$Z_{Norm} = LayerNorm(Q + Z) \tag{6}$$

$$k_{final} = max(0, W_1 Z_{Norm} + b_1)W_2 + b_2 \tag{7}$$

During the matrix multiplication QK^T, the image query Q is respectively interacted (multiplied) with every keyword embedded vector denoted as every key K. Therefore, we could obtain every keyword weights on the image vector. After scaled and softmax operation, we could get probability-like weights for each keyword interpreted as their attention or relationship with the current image. Finally, we multiply the weights back with the corresponding value V to denote their hybrid importance for providing attention-weighted image-keyword information.

3.3 Contextual Transformer Decoder

Transformer is one of the state-of-the-art approaches in sequence modeling and transduction problems [44]. Its attention mechanism allows language modeling of global dependencies between input and output, preventing the memory constraint limits of conventional recurrent models. Inspired by the transformer’s architecture and in view of its parallelization for attention-weighted positions, its nature is deployed for our main output decoder. A contextual transformer decoder cell could be expressed in Figure 2. It comprises a masked self-attention unit, an encoder-decoder attention unit, and a final feed-forward layer, similar to a conventional counterpart. We similarly exploit the encoder-decoder structure [44] where the encoder follows the attention function $f(\mathbf{v}, \mathbf{k})$ directly. We can then illustrate the decoding process as the following:

$$\mathbf{x} = W_e S \tag{8}$$

$$C_1 = \mathbf{x} + PE(\mathbf{x}) \tag{9}$$

$$C'_{l-1} = MultiHeadAtt([C_{l-1}, C_{l-1}, C_{l-1}]) \tag{10}$$

$$C_l = FCN(MultiHeadAtt([k_{final}, k_{final}, C'_{l-1}])) \tag{11}$$

In Equation (8), we denote a true sentence describing the image as $S = (S_0, ..., S_T)$ and map the bag-of-word ids into word vectors $\mathbf{x} \in \mathbb{R}^{T \times H_e}$ with the same glove embedding layer W_e. Then, we add positional embedding in equation (9) on top of $\mathbf{x}$ to introduce sequential information. The semantic vector will then repeatedly visit the multiple attention layer block. For each layer of the decoder, we feed the input into a self-attention layer and an encoder-decoder layer to further attend on image-keyword fusion contexts. We also use the dropout technique to alleviate the effect of noises and overfitting. Finally, we send the output of the final layer C_L into a fully connected layer to obtain the joint distribution of decoding words.

$$P_L = W_v C_L \tag{12}$$

$$L(P|S, I, K) = \mathbb{E}_{S \sim P_I}[logP_L(S, I, K)] \tag{13}$$

If we denote P_I as the true medical descriptions for I provided in the training set and $P_L(S, I, K)$ as the final probability distribution after one fully-connected layer and softmax function, we could have the overall likelihood function $L(P|S, I, K)$ depending on our medical descriptions and the given image shown in Equation (13). Finally, we could minimize the total loss calculated as the sum of the negative log-likelihood at each time step. For inference, for each step we perform *"Greedy Search"* where we sample the words based on the maximum likelihood of each word output P_t on a predicted distribution P_{t+1} until P_{t+1} = special end-of-sentence token.

4. Experiments

In this section, we describe the commonly used retinal image captioning dataset and evaluation metrics. Summaries of baseline models and experimental setup are provided.

4.1 Dataset

DeepEyeNet [23] is a commonly used benchmark for retinal image captioning. The total amount of retinal images is $15,709$. Each retinal image has two corresponding labels, i.e., expert-defined keywords and clinical description. The word length is mainly between 5 and 10 words. The labels are annotated by experienced retinal specialists based on retinal image analysis and conversation with patients. In this work, we extend the DeepEyeNet dataset with $3,145$ expert-annotated retinal images based on the same data collection method as DeepEyeNet [23]. Hence, the size of the used dataset for experiments is $18,854$. We separate the whole dataset into $80\%/10\%/10\%$, i.e., $15,083/1,885/1,886$, for training/validation/testing, respectively.

4.2 Performance Evaluation Metrics

In the experiments, we exploit the commonly used text evaluation metrics, [38, 29, 45, 2, 14, 15, 12, 13, 11, 5], used in retinal image captioning field, [28], to evaluate the generated medical descriptions for retinal images. Although these automatic evaluation metrics are popular in natural and retinal image captioning tasks, these metrics' innate properties [38, 29, 45, 2, 23, 21] make them more suitable for natural image captioning not retinal image captioning. Hence, in this work, we also conduct a human expert evaluation for the proposed method, referring to Section 5.3.

4.3 Baseline Models

We compare the proposed method with several competitive image captioning models.

- **LSTM** [48] builds on a deep CNN and BiLSTM structure for image captioning.
- **Show and tell** [49] adopts the attention mechanism on several patches of the original image to focus on particular area when generating descriptions.
- **Semantic Att** [52] predicts a list of visual attributes which are attended with hidden states both at inputs and outputs in a RNN caption generator.
- **CoAtt**[26] adopts co-attention mechanism to produce joint context vectors for generating medical descriptions based on topics.
- **H-CoAtt**[34] proposes a co-attention model for visual question answering tasks which hierarchically reasons the questions based on visual features.
- **ContexGPT** [18] adopts a non-local attention mechanism, masked self-attention, and feature reinforcement module to build a retinal image captioning network.
- **DeepContex** [20] proposes a context-driven encoding network for retinal image captioning.
- **MIA**[32] presents a mutual iterative attention to jointly consider interactions between images and keywords for image captioning and visual question answering.

4.4 Experimental Setup

ResNet50 [8], pre-trained on ImageNet, is used as our retinal image feature extractor ϕ. We first resize the image to the appropriate size to feed in the model. The layer before the last fully-connected layer is adopted for embedding visual features. To process the annotations and keywords in the dataset, non-alphabet characters are removed, all remaining characters are converted to lower-case, and all the words that appear only once are replaced by a special token $\langle UNK \rangle$. As a result, our vocabulary size is $3,524$. All sentences are truncated or padded with a max length 50. For keyword prediction, we set the threshold $\tau = 0.5$. For the word embedding layer, we use an embedding size $H_e = 300$ to encode words. We use two transformer blocks with 8 attention heads, $2,048$ hidden size of the fully connected layer, and 64 hidden size. Finally, we set the mini-batch size to 64 and the learning rate to 1e-4 to train all the models with 10 epochs.

5. Results and Discussion

5.1 Quantitative Analysis

With expert-defined keywords. We first report the results of medical description generation by providing retinal images and corresponding expert-defined keywords, i.e., ground truth keywords, jointly, referring Table 1. It is clear that vanilla LSTM decoder performs much worse than other models with attention mechanisms, which is non-surprising for its deficiency in capturing image dependencies. By introducing expert-defined keywords in the generation process, keyword-driven models starting from Semantic Att [52] render a large increase in every metric which validates the benefits of keywords that guide the model for accurate predictions. The expert-defined keywords are human-comprehensible and hence provide improved interpretability, referring to Section 5.3 for human expert evaluation. Improvements in the co-attention mechanism between images and keywords

Table 1: This table shows the evaluation results of the proposed model compared with several competitive baselines by using expert-defined keywords, i.e., ground truth keywords. “BLEU-avg” denotes the average score of BLEU-1, BLEU-2, BLEU-3, and BLEU-4. All the keyword-driven models are superior to the non-keyword-driven models.

Model	**BLEU-1**	**BLEU-2**	**BLEU-3**	**BLEU-4**	**B-avg**	**ROUGE**	**CIDEr**	**METEOR**
LSTM [48]	0.2273	0.1650	0.1224	0.1017	0.1541	0.2533	0.1102	0.2437
Show and tell [49]	0.4234	0.3583	0.3002	0.2757	0.3394	0.4463	0.3029	0.4335
Semantic Att [52]	0.5904	0.5100	0.4360	0.3969	0.4833	0.6228	0.4460	0.6056
ContexGPT [18]	0.6254	0.5500	0.4758	0.4344	0.5214	0.6602	0.4951	0.6390
CoAtt [26]	0.6712	0.5950	0.5211	0.4817	0.5673	0.6988	0.5419	0.6798
H-CoAtt [34]	0.6718	0.5956	0.5201	0.4829	0.5676	0.7045	0.5417	0.6864
DeepContex [20]	0.6749	0.6036	0.5307	0.4890	0.5745	0.7020	0.5496	0.6835
MIA [32]	0.6877	0.6138	0.5421	0.5000	0.5859	0.7195	0.5596	0.7006
Ours	**0.6969**	**0.6195**	**0.5496**	**0.5008**	**0.5892**	**0.7252**	**0.5650**	**0.7044**

Table 2: This table shows the evaluation results of the proposed model compared with several competitive baselines by using predicted keywords, i.e., pseudo expert-defined keywords. “BLEU-avg” denotes the average score of BLEU-1, BLEU-2, BLEU-3, and BLEU-4. All the keyword-driven models are superior to the non-keyword-driven models.

Model	**BLEU-1**	**BLEU-2**	**BLEU-3**	**BLEU-4**	**B-avg**	**ROUGE**	**CIDEr**	**METEOR**
LSTM [48]	0.2273	0.1650	0.1224	0.1017	0.1541	0.2533	0.1102	0.2437
Show and tell [49]	0.4234	0.3583	0.3002	0.2757	0.3394	0.4463	0.3029	0.4335
H-CoAtt [34]	0.4465	0.3822	0.3285	0.2969	0.3636	0.4788	0.3409	0.4564
ContexGPT [18]	0.4493	0.3744	0.3109	0.2800	0.3536	0.4771	0.3171	0.4588
Semantic Att [52]	0.4541	0.3771	0.3117	0.2777	0.3552	0.4785	0.3118	0.4610
CoAtt [26]	0.4647	0.4038	0.3479	0.3162	0.3831	0.4906	0.3563	0.4759
DeepContex [20]	0.4683	0.3966	0.3302	0.2969	0.3730	0.4941	0.3341	0.4803
MIA [32]	0.5077	0.4446	0.3861	0.3514	0.4224	0.5326	0.3897	0.5163
Ours	**0.5268**	**0.4600**	**0.3915**	**0.3634**	**0.4354**	**0.5482**	**0.4105**	**0.5316**

further strengthen our belief that models pay large attention to integrated representation collections from both the visual and semantic concepts. By leveraging the mutual attention weights from image and keywords, our model with the transformer decoder replacing the LSTM decoder outperforms all other baselines, where previous tokens and fusion concepts could be fully referenced to generate the next token. Overall, we see an increase of 74% in BLEU average, 63% in ROUGE, 87% in CIDEr, and 63% in METEOR, compared with non-keyword-driven attention models [49].

Keyword prediction. To simulate a more general setting, we also report the experimental results by predicted keywords, i.e., pseudo expert-defined keywords, using our pre-trained multi-label classifier in Table 2 and Table 3. We can see the benefit of keyword fusion is degraded due to some erroneously predicted keywords. This is particularly challenging in the medical description generation task since there are $3,465$ keyword options in the DeepEyeNet dataset where the number of keywords to select is undetermined for a given image. But still, we can observe an overall improvement in all metrics by inducing the predicted keyword contexts. Our approach especially outperforms several co-attention based approaches, which intrinsically overfit the training data resulting from their model complexity. The single cross attention embedded in our transformer decoder between the visual and semantic concepts could be more robust to the keyword noise during the several layer transitions.

Co-attention between image and keywords. According to Table 4, we find that the performance of “Image only” and “Keyword only” baselines are worse than the “Image+Keywords” methods. It implies that the interaction between keywords and image is crucial for medical report generation. To further demonstrate the benefits of attention mechanisms, we provide another baseline where image and keyword features are concatenated without further fusion. We can see a large performance degradation without using attention mechanisms.

5.2 Qualitative Results and Analysis

Comparison with the classic attention model. We present some qualitative results generated by three medical generation models including ours and [49, 18] in Figure 3. Show and tell [49] does not apply any keywords and ContexGPT [18] serves as a keyword-oriented baseline. In the first two images, by semantically attending to the correct predicted keywords, both ContexGPT and our model generate descriptions related to keywords. But our model matches the ground truth identically. Show and tell model [49] without explicit textual attributes seem to diverge from accurate symptom names and detailed illustrations. ContexGPT also loses track of accurate semantic information besides the keyword guidance. It substantiates the need for our pre-trained keyword predictor to first coarsely tag a given retinal image then extend the details, which is more imperative and intuitive in medical fields compared with common domains.

For the third image, we can observe the keywords are not explicitly involved in the ground truth, intead symptom illustration. Both two baselines provide irrelevant image descriptions, while our model provides more details of il-

Table 3: The table is to show the performance drop when expert-defined keywords are not available, i.e., the case "With predicted keywords".

Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4	B-avg	ROUGE	CIDEr	METEOR
With predicted keywords	0.5268	0.4600	0.3915	0.3634	0.4354	0.5482	0.4105	0.5316
With expert-defined keywords	**0.6969**	**0.6195**	**0.5496**	**0.5008**	**0.5892**	**0.7252**	**0.5650**	**0.7044**

Table 4: The table is to demonstrate the ablation study of the proposed model structure. "Image only" and "Keywords only" refer to input either feature only into our model. "Image+Keywords (concat)" indicates we only concatenate image and keyword vectors and send them into the transformer decoder. "Image+Keywords (coatt)" is the complete structure of the proposed method.

Input	BLEU-1	BLEU-2	BLEU-3	BLEU-4	B-avg	ROUGE	CIDEr	METEOR
Image only	0.4357	0.3651	0.3041	0.2773	0.3455	0.4608	0.3067	0.4454
Keywords only	0.5568	0.4970	0.4322	0.3971	0.4708	0.6110	0.4618	0.5881
Image+Keywords (concat)	0.6527	0.5752	0.4984	0.4626	0.5472	0.6783	0.5166	0.6643
Image+Keywords (coatt)	**0.6969**	**0.6195**	**0.5496**	**0.5008**	**0.5892**	**0.7252**	**0.5650**	**0.7044**

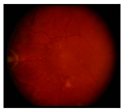
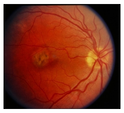
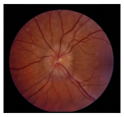
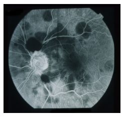
Retinal Image	Keywords	Ground Truth	Show and Tell	ContexGPT	Our Method
	pigment epithelial detachment (ped)	62 year old male armd with ped partly organized.	13 year old patient dusn / optic papillitis.	Pigment epithelial lesions.	62 year old male armd with ped partly organized.
	presumed ocular histoplasmosis syndrome	32 year old woman with presumed ocular histoplasmosis syndrome with choroidal neovascular membrane.	23 year old white female pseudo pohs / mewds.	Presumed ocular histoplasmosis syndrome with large subretinal new vessel membrane in the fovea.	32 year old woman with presumed ocular histoplasmosis syndrome with choroidal neovascular membrane.
	papilledema	Os optic nerve with frank swelling.	29 year old female pohs with cnvm.	No history.	Os with subtle central pigment epithelial changes presumably presumed ocular histoplasmosis syndrome.
	sub-arachnoid hemorrhage	60 year old white female was found unconscious in her home she was rushed to the hospital where a cat scan of her head revealed a large sub arachnoid hemorrhage a carotid angiogram showed a ruptured aneurysm of the posterior communicating artery on the right side the next day.	This eight year old white female who was in perfect health complaining of a visual disturbance in the left eye the right eye was completely normal the left eye had a exudative detachment of the macula.	The patient a 29 year old white female developed idiopathic thrombocytopenis purpura itp in 1964.	A 60 year old white female was found unconscious in her home she was rushed to the hospital where a cat scan of her head revealed a large sub arachnoid hemorrhage a carotid angiogram showed a ruptured aneurysm of the posterior communicating artery on the right side the next day.

Figure 3: Illustration of descriptions generated by the proposed model and two baseline models [49, 18].

lustrating phenomenon related to the symptoms. We further demonstrate the robustness of our model to generate long descriptions based on the context fusion in the last image, compared to less structured expressions from other baselines.

Does our model fully understand the fused concepts? To better understand how our model utilizes the fused visual and semantic concepts for token sampling, we visualize the attention weights on the input image at each time step, referring to Figure 4. Each image consists 64 patches and each patch has a weight of the current word and corresponding keyword-fused image patch. We can see our model is less sensitive (showing minor saliency on overall regions) to words specifying number (i.e. 29, 55) or color (in white) by solely relying on the input image and keywords. But our model heavily depends on some specific image regions to predict a medical keyword. We can see a trend of similar saliency between consecutive words which our model diagnoses to be the abnormality of the particular image. These highlighted regions allude some promising interpretability of how our model understands keywords and image patches to generate an adequate sequence.

5.3 Evaluations with Retinal Specialists

We use 5-level report quality evaluation, i.e., from 1 to 5, the higher the better. Since our research resource is limited, we are only able to randomly select 100 samples from our model-generated reports and the corresponding ground-truth report. We ask five different retinal specialists to score the quality of the model-generated report and the correspond-

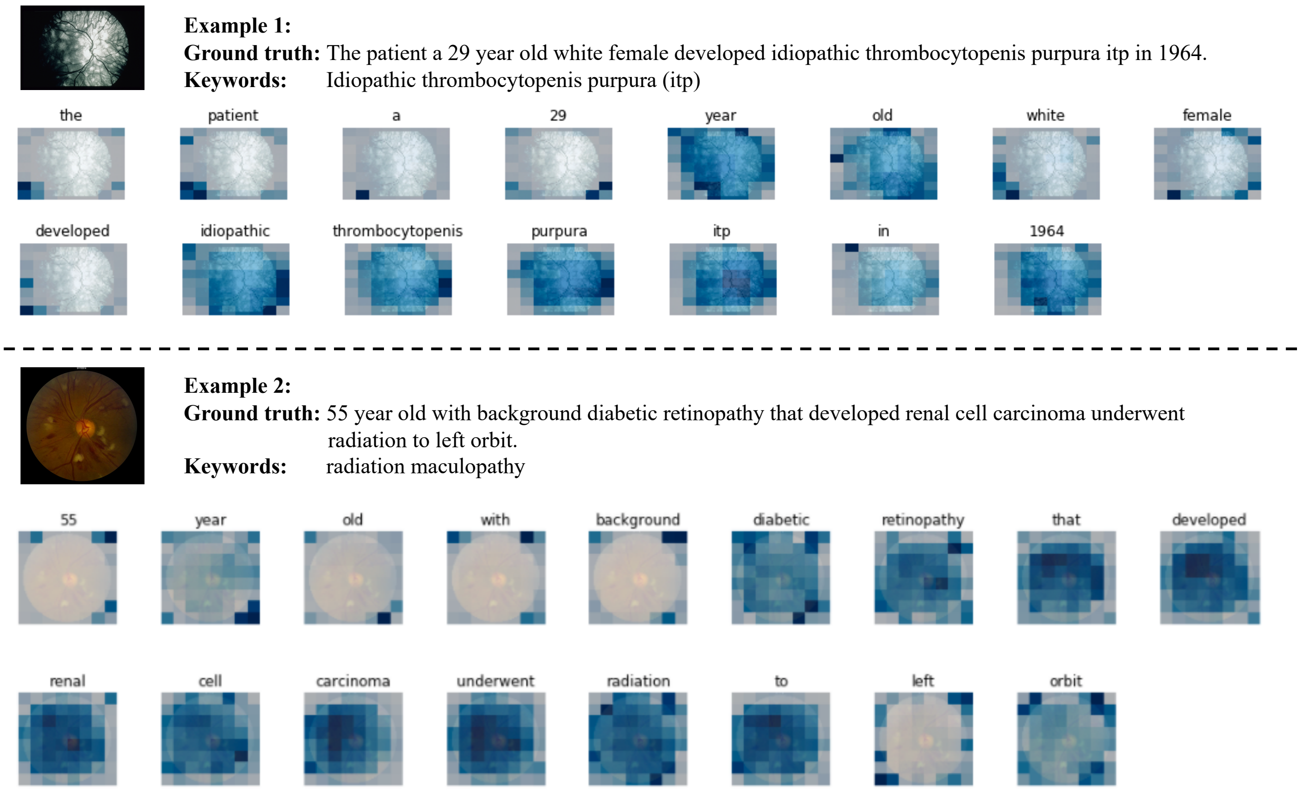

Example 1:
Ground truth: The patient a 29 year old white female developed idiopathic thrombocytopenis purpura itp in 1964.
Keywords: Idiopathic thrombocytopenis purpura (itp)
the
patient
a
29
year
old
white
female
developed
idiopathic
thrombocytopenis
purpura
itp
in
1964
Example 2:
Ground truth: 55 year old with background diabetic retinopathy that developed renal cell carcinoma underwent radiation to left orbit.
Keywords: radiation maculopathy
55
year
old
with
background
diabetic
retinopathy
that
developed
renal
cell
carcinoma
underwent
radiation
to
left
orbit

Figure 4: Visualization of image attention propagating through the text generation. Each example consists of an input image, ground truth descriptions, and predicted keywords. Each predicted word is shown on top of image attention at each time step.

ing ground-truth report, respectively. Note that these five retinal specialists do not know whether a report is model-generated or expert-generated. Finally, we get an average score of $4.0/5.0$ for our model-generated reports and an average score of $4.3/5.0$ for the ground-truth reports. Since the ground-truth reports are defined by ophthalmologists, the above results show that the proposed method obtains competitive performance against the human expert baseline. We use the same above setup to conduct interpretability evaluation. The first case is presenting 100 generated reports without corresponding keywords to the five retinal specialists. The second case is presenting the same 100 generated reports with corresponding keywords. In the first case, the interpretability evaluation score is $4.0/5.0$. In the second case, the interpretability evaluation score is $4.6/5.0$. Hence, the interpretability is improved by keywords.

5.4 Main Limitation of the Proposed Approach

If the expert-defined keywords are not available in some domains, then the performance of the proposed model will decrease. Also, the keywords probably cannot be always generated accurately by the proposed method. The reason is that one of the main purposes of expert-defined keywords is to teach a model to predict correct keywords.

6. Conclusion and Future Work

To sum up, an explainable medical report generation method for retinal images is proposed based on expert-defined keywords and a novel attention-based strategy. The proposed method is capable of predicting required technical keywords and fusing them for advanced word sampling. The experiments show that the proposed model can generate more accurate and meaningful descriptions for retinal images, and the performance increases about 74% in BLEU average, 63% in ROUGE, 87% in CIDEr, and 63% in METEOR compared with non-keyword attention-based baselines. Attention visualization denotes some intriguing patterns of potential symptoms in specific image regions. To help our research community develop a more explainable ML-based model for retinal image captioning, proposing an automatic metric to measure explainability is an interesting future direction.

7. Acknowledgments

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

References

[1] Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. Sanity checks for saliency maps. pages 9508–9518, 2018.

[2] Satanjeev Banerjee and Alon Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with hu-

man judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72, 2005.

[3] Aaron M Bornstein. Is artificial intelligence permanently inscrutable? 2016.

[4] Nadia Burkart and Marco F Huber. A survey on the explainability of supervised machine learning. *Journal of Artificial Intelligence Research*, 70:245–317, 2021.

[5] Riccardo Di Sipio, Jia-Hong Huang, Samuel Yen-Chi Chen, Stefano Mangini, and Marcel Worring. The dawn of quantum natural language processing. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8612–8616. IEEE, 2022.

[6] Finale Doshi-Velez and Been Kim. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*, 2017.

[7] Marzyeh Ghassemi, Luke Oakden-Rayner, and Andrew L Beam. The false hope of current approaches to explainable artificial intelligence in health care. *The Lancet Digital Health*, 3(11):e745–e750, 2021.

[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[9] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[10] Tao Hu, Pascal Mettes, Jia-Hong Huang, and Cees GM Snoek. Silco: Show a few images, localize the common object. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5067–5076, 2019.

[11] Jia-Hong Huang. Robustness analysis of visual question answering models by basic questions. *King Abdullah University of Science and Technology, Master Thesis*, 2017.

[12] Jia-Hong Huang, Modar Alfadly, and Bernard Ghanem. Vqabq: Visual question answering by basic questions. *VQA Challenge Workshop, CVPR*, 2017.

[13] Jia-Hong Huang, Modar Alfadly, and Bernard Ghanem. Robustness analysis of visual qa models by basic questions. *VQA Challenge and Visual Dialog Workshop, CVPR*, 2018.

[14] Jia-Hong Huang, Modar Alfadly, Bernard Ghanem, and Marcel Worring. Assessing the robustness of visual question answering. *arXiv:1912.01452*, 2019.

[15] Jia-Hong Huang, Cuong Duc Dao, Modar Alfadly, and Bernard Ghanem. A novel framework for robustness analysis of visual qa models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 8449–8456, 2019.

[16] Jia-Hong Huang, Luka Murn, Marta Mrak, and Marcel Worring. Gpt2mvs: Generative pre-trained transformer-2 for multi-modal video summarization. In *Proceedings of the 2020 International Conference on Multimedia Retrieval*, pages 580–589, 2021.

[17] Jia-Hong Huang and Marcel Worring. Query-controllable video summarization. In *Proceedings of the 2020 International Conference on Multimedia Retrieval*, pages 242–250, 2020.

[18] Jia-Hong Huang, Ting-Wei Wu, and Marcel Worring. Contextualized keyword representations for multi-modal retinal image captioning. In *Proceedings of the 2020 International Conference on Multimedia Retrieval*, pages 645–652, 2021.

[19] Jia-Hong Huang, Ting-Wei Wu, C-H Huck Yang, Zenglin Shi, I Lin, Jesper Tegner, Marcel Worring, et al. Non-local attention improves description generation for retinal images. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 1606–1615, 2022.

[20] Jia-Hong Huang, Ting-Wei Wu, Chao-Han Huck Yang, and Marcel Worring. Deep context-encoding network for retinal image captioning. In *2021 IEEE International Conference on Image Processing (ICIP)*, pages 3762–3766. IEEE, 2021.

[21] Jia-Hong Huang, Ting-Wei Wu, Chao-Han Huck Yang, and Marcel Worring. Longer version for" deep context-encoding network for retinal image captioning". *arXiv preprint arXiv:2105.14538*, 2021.

[22] Jia-Hong Huang, Chao-Han Huck Yang, Pin-Yu Chen, Andrew Brown, and Marcel Worring. Causal video summarizer for video exploration. In *2022 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE, 2022.

[23] Jia-Hong Huang, C-H Huck Yang, Fangyu Liu, Meng Tian, Yi-Chieh Liu, Ting-Wei Wu, I Lin, Kang Wang, Hiromasa Morikawa, Hernghua Chang, et al. Deepopht: medical report generation for retinal images via deep models and visual explanation. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 2442–2452, 2021.

[24] Lun Huang, Wenmin Wang, Jie Chen, and Xiao-Yong Wei. Attention on attention for image captioning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4634–4643, 2019.

[25] C-H Huck Yang, Fangyu Liu, Jia-Hong Huang, Meng Tian, I-Hung Lin, Yi Chieh Liu, Hiromasa Morikawa, Hao-Hsiang Yang, and Jesper Tegner. Auto-classification of retinal diseases in the limit of sparse data using a two-streams machine learning model. In *Asian Conference on Computer Vision*, pages 323–338. Springer, 2018.

[26] Baoyu Jing, Pengtao Xie, and Eric Xing. On the automatic generation of medical imaging reports. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2577–2586, Melbourne, Australia, July 2018. Association for Computational Linguistics.

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

[28] Christy Y. Li, Xiaodan Liang, Zhiting Hu, and Eric P. Xing. Knowledge-driven encode, retrieve, paraphrase for medical image report generation, 2019.

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

[30] Min Lin, Qiang Chen, and Shuicheng Yan. Network in network. *arXiv preprint arXiv:1312.4400*, 2013.

[31] Zachary C Lipton. The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue*, 16(3):31–57, 2018.

[32] Fenglin Liu, Yuanxin Liu, Xuancheng Ren, Xiaodong He, and Xu Sun. Aligning visual regions and textual concepts for semantic-grounded image representations, 2019.

[33] Yi-Chieh Liu, Hao-Hsiang Yang, C-H Huck Yang, Jia-Hong Huang, Meng Tian, Hiromasa Morikawa, Yi-Chang James Tsai, and Jesper Tegner. Synthesizing new retinal symptom images by multiple generative models. In *Asian Conference on Computer Vision*, pages 235–250. Springer, 2018.

[34] Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. Hierarchical question-image co-attention for visual question answering, 2017.

[35] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Proceedings of the 31st*

international conference on neural information processing systems, pages 4768–4777, 2017.

[36] Sanjukta Mishra and Minakshi Banerjee. Automatic caption generation of retinal diseases with self-trained rnn merge model. In *Advanced Computing and Systems for Security*, pages 1–10. Springer, 2020.

[37] Yingwei Pan, Ting Yao, Yehao Li, and Tao Mei. X-linear attention networks for image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10971–10980, 2020.

[38] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics, 2002.

[39] Marco Pedersoli, Thomas Lucas, Cordelia Schmid, and Jakob Verbeek. Areas of attention for image captioning. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1242–1250, 2017.

[40] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144, 2016.

[41] Franco Scarselli, Marco Gori, Ah Chung Tsoi, Markus Hagenbuchner, and Gabriele Monfardini. The graph neural network model. *IEEE transactions on neural networks*, 20(1):61–80, 2008.

[42] 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 *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.

[43] Dylan Slack, Sophie Hilgard, Emily Jia, Sameer Singh, and Himabindu Lakkaraju. Fooling lime and shap: Adversarial attacks on post hoc explanation methods. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pages 180–186, 2020.

[44] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

[45] 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.

[46] Sivamurugan Vellakani and Indumathi Pushbam. An enhanced oct image captioning system to assist ophthalmologists in detecting and classifying eye diseases. *Journal of X-Ray Science and Technology*, 28(5):975–988, 2020.

[47] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3156–3164, 2015.

[48] Cheng Wang, Haojin Yang, Christian Bartz, and Christoph Meinel. Image captioning with deep bidirectional lstms. In *Proceedings of the 24th ACM international conference on Multimedia*, pages 988–997. ACM, 2016.

[49] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhutdinov, Richard Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention, 2016.

[50] C-H Huck Yang, Jia-Hong Huang, Fangyu Liu, Fang-Yi Chiu, Mengya Gao, Weifeng Lyu, Jesper Tegner, et al. A novel hybrid machine learning model for auto-classification of retinal diseases. *Workshop on Computational Biology, ICML*, 2018.

[51] Ting Yao, Yingwei Pan, Yehao Li, and Tao Mei. Exploring visual relationship for image captioning. In *Proceedings of the European conference on computer vision (ECCV)*, pages 684–699, 2018.

[52] Quanzeng You, Hailin Jin, Zhaowen Wang, Chen Fang, and Jiebo Luo. Image captioning with semantic attention, 2016.

[53] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2921–2929, 2016.