

# TransFuser: Keyword-driven Medical Report Generation for Retinal Images

AAAI Press

Association for the Advancement of Artificial Intelligence  
2275 East Bayshore Road, Suite 160  
Palo Alto, California 94303

## Abstract

Automatically generating medical reports from retinal images is a difficult task in which an algorithm must generate semantically coherent descriptions for a given retinal image. Existing methods are based on the conventional natural image captioning model which mainly relies on the input image to generate descriptions. However, some abstract medical concepts or descriptions cannot be generated based on image information only. In this work, we integrate additional information to help solve this task; we highlight that early in the diagnosis process, ophthalmologists have usually written down a small set of keywords denoting important information. These keywords are then subsequently used to aid the later creation of medical reports for a patient. In practice, since these keyword labels commonly exist and are useful for generating medical reports, we incorporate them in automatic report generation. We propose a keyword-driven medical report generation model that generates more accurate and meaningful descriptions for retinal images. Since we have two types of inputs - keywords and images - how to effectively fuse features from these different modalities is challenging. To that end, we introduce a new multi-modality fusion approach, TransFuser which is capable of fusing features from different types of inputs. To foster retinal disease research and validate our keyword-driven method, we introduce a new large-scale image dataset. This dataset contains 15,709 retinal images and corresponding keywords and descriptions. Our experiments show that the proposed method successfully captures the mutual information of keywords and image. We find that our proposed keyword-driven medical report generation model is superior to non-keyword-driven baselines under the popular text evaluation metrics BLEU, CIDEr, and ROUGE.

## Introduction

Automatic medical report generation for retinal images is a challenging computer vision task, falling within the broader task domain of image captioning (Xu et al. 2015). In this task, long and semantically coherent medical descriptions for a given image must be generated algorithmically. Several technical features of retinal image report generation complicate this task when compared to the more well studied domain of natural image captioning, e.g., in (Cornia et al. 2019; Kim et al. 2019). One example is that retinal and natural images have very different characteristics, both in objects' sizes as well as details. As such, existing methods, such as (Xu et al. 2015; Karpathy et al. 2015), which work well on

natural image datasets often do not generalize well to retinal images. Recently, some related methods (Jing et al. 2018; Li et al. 2018) have been proposed to generate medical reports. These approaches are based on the traditional natural image captioning model which mainly relies on the input image to generate descriptions. However, some abstract medical concepts or descriptions cannot be generated based on image information only. To generate more accurate and meaningful descriptions for retinal images, we will need a new specialized method which is capable of using input contextual information and image information simultaneously.

In this paper, we propose a keyword-driven medical report generation model, illustrated in Figure 1. The model is inspired by the framework of the Visual Question Answering (VQA) model with an attention mechanism presented in (Lu et al. 2016). In the VQA model, there are two types of inputs with different modalities, an image, and a question sentence. The attention mechanism allows the inputs to guide each other to generate more accurate results. In our proposed keyword-driven medical report generation model, we also use two types of inputs with different modalities: namely images and keywords. The technical problem our proposed method tackles is fusing the input image and keywords with minimum loss of information. It has been shown that, in similar multi-modal contexts, the performance of models can decrease if models are poorly designed; how to solve this issue in general remains an open question (Ben-Younes et al. 2017; Fukui et al. 2016). To address this issue here, we introduce a novel keyword-image encoder, called TransFuser, illustrated in Figure 2. In the TransFuser encoder, feature vectors of different modalities are fused to perform the automatic medical report generation task. Generally speaking, it encodes unordered keyword sequences with image content and draws different attention weights on every individual keyword. Because of the transformer-based structure (Vaswani et al. 2017), our TransFuser is capable of effectively capturing the mutual information between keywords and image.

To build a reporting system for retinal images, using multiple information sources requires dedicated studies. We undertake such a study in this work. We establish a realistic retinal image dataset on which to build our models. Such datasets are difficult to create, as they require many hours of dedicated annotation from skilled ophthalmologists. With such a dataset in place, it is the right time to research how to exploit deep learning-based models. We highlight that in many cases, ophthalmologists will have written down a

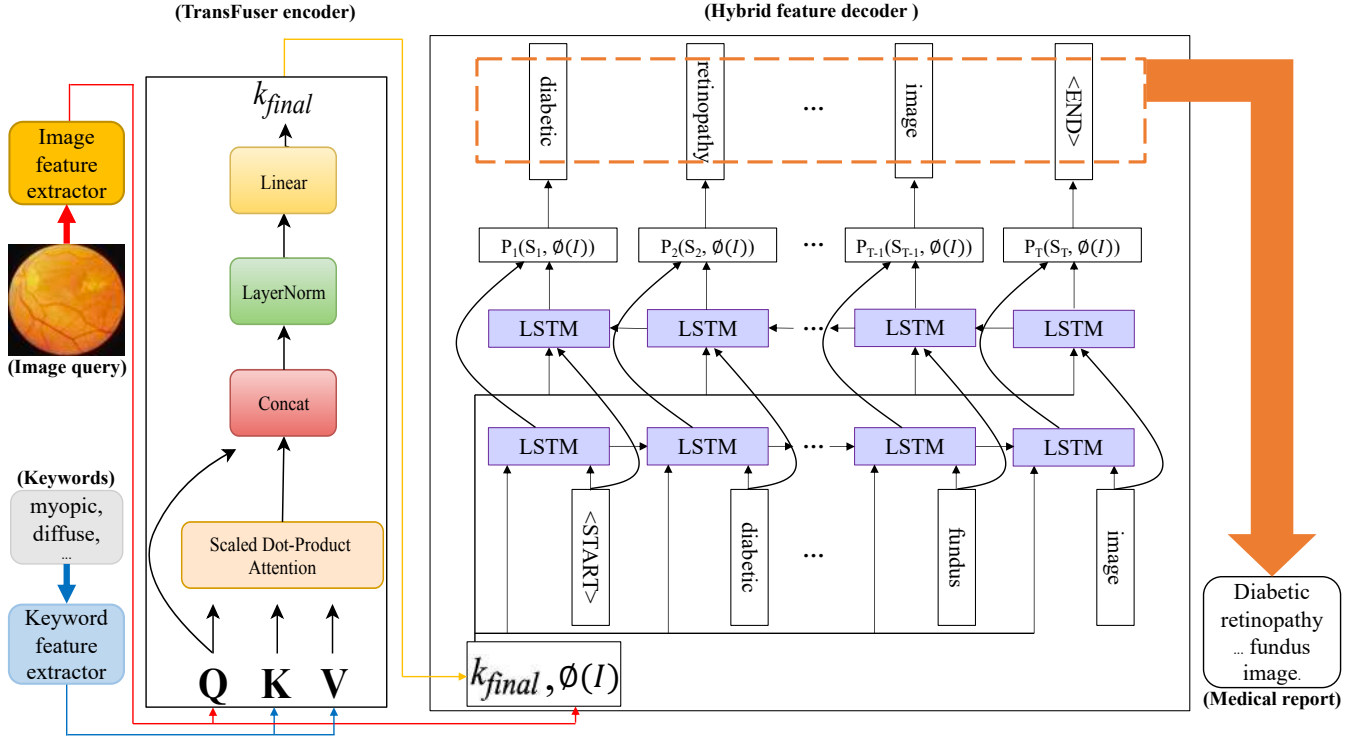


Figure 1: This figure shows the flowchart of our proposed keyword-driven medical report generation method. It contains two modalities of inputs, a retinal image, and keywords. The purpose of keywords is to reinforce the model to generate more accurate and meaningful descriptions for retinal images, denoting by orange color. “TransFuser” denotes our proposed multi-modal feature fuser. “Concat” denotes concatenation, “LayerNorm” denotes layer normalization, “Linear” denotes a fully-connected layer, and  $k_{final}$  denotes an attention-based embedding vector.  $Q$  is a transformed image query,  $K$  is key vectors,  $V$  denotes weight value vectors,  $\phi(I)$  denotes an image feature vector, and  $P_i(S_i, \phi(I))$  is a probability distribution where  $i = 1, 2, \dots, T$ .

few keywords describing a retinal image in a cursory analysis. Typically, keywords were created from an ophthalmologist’s analysis of a retinal image and a conversation with the patient. Medical reports are then created for patients at a later time. We aim to leverage the keywords in a deep learning approach for automatic report generation. To the best of our knowledge, we are the first to investigate methods using these keywords to aid the generation of medical reports on retinal images.

We demonstrate experimental results of TransFuser on our newly introduced retinal image dataset. We show that our proposed keyword-driven medical report generation model is capable of creating more accurate and meaningful descriptions for retinal images than a non-keyword-driven model baseline. This performance is shown in several text evaluation metrics: BLEU-avg (+58%), CIDEr (+75%), and ROUGE (+58%).

### Contributions

- i We propose a novel keyword-driven medical report generation model which is superior to non-keyword-driven baselines when measured with several popular text evaluation metrics.
- ii We propose a new approach, called TransFuser, to effectively fuse features from images and text, and use them to perform the medical report generation task.
- iii We introduce a new real-world retinal image dataset

with keyword labels annotated by experienced ophthalmologists. The dataset contains 15,709 retinal images, and corresponding keywords and captions.

### Related Work

In this section, we review the related image captioning methods with the most popular encoder-decoder architecture and some work inspiring us to propose the keyword-driven medical report generation model.

#### Caption Generation for Natural and Medical Images

The encoder-decoder based network architecture, (Vinyals et al. 2015; Karpathy et al. 2015; Li et al. 2018), is the most popular method to perform the image captioning task. In these networks, the convolution neural networks (CNN) is considered as an encoder and used to extract global image features, and the recurrent neural networks (RNN) is regarded as a decoder and used to generate a sequence of words. In (Mao et al. 2016), the authors introduce a text generation method to generate a description for some specific object or region that is called referring expression (Kazemzadeh et al. 2014). The authors of (Wang et al. 2016) propose a bidirectional LSTM-based method to generate captions. The method exploits past and future information at the same time to learn long-term visual language interactions. Attention-based models have shown good performance in the image captioning task. The authors of (Pedersoli et al. 2017) introduce an area-based attention

model for image captioning. The model can predict the next word and corresponding regions of the image in each RNN time step for generating image descriptions. The authors of (Li et al. 2018) introduce a Hybrid Retrieval-Generation Reinforced Agent to incorporate human prior knowledge with learning-based generation for medical image captioning. The agent exploits a retrieval policy module to decide between using a generation module to generate sentences and retrieving specific sentences from the template database, which is built based on prior human knowledge. Based on hierarchical decision-making, it then sequentially generates multiple sentences. In (Jing et al. 2018), the authors propose a multi-task learning framework to predict tags and generate captions at the same time. Also, they use a co-attention mechanism to localize regions which contain abnormalities and generate long descriptions for those regions via a hierarchical LSTM model. The above works try to generate a medical report for radiology images of the chest. In addition, the authors of (Jing et al. 2018) note that the generated medical reports based on most of the existing methods are fully-structured or semi-structured, e.g., tags, templates. From the medical point of view, radiology images of the chest and retinal images have different properties, such as objects' sizes and details. Most of the methods mentioned above mainly rely on the image input to generate captions. However, some abstract concepts or descriptions cannot be generated only based on image information. In this work, our proposed method starts from the CNN-RNN based framework. To effectively fuse features with different modalities, we exploit the attention (Vaswani et al. 2017) to develop our proposed TransFuser.

### Visual Question Answering (VQA)

VQA (Agrawal et al. 2017; 2018) is a computer vision task with multi-modal inputs, a query question and an image. A VQA model with attention mechanism is capable of attending to local image regions related to the query question (Shih, Singh, and Hoiem 2016; Chen et al. 2016). The authors of (Lu et al. 2016) propose a co-attention mechanism that jointly performs image and language attention in the VQA task. In (Huang et al. 2019), the authors propose a new metric to evaluate the robustness of attention-based VQA models. Our proposed keyword-driven medical report generation model is inspired by the architecture of the VQA model with attention mechanism. In the proposed keyword-driven model, we also use two types of inputs with different modalities: images and keywords. Note that a question sentence has an ordered nature which our keywords do not have. How to fuse the input image and keywords with a minimum loss of information is challenging. It has been shown that how to solve this issue, in general, remains an open question (Ben-Younes et al. 2017; Fukui et al. 2016).

## Methodology

### Overview

In this section, we present our keyword-driven medical report generation model and illustrate methods to train the model with supervised keyword knowledge as shown in Figure 1. First, an image and a number of keywords will be fed in their own extractors to acquire an embedded image vector and an embedded keyword vector each. After this information extraction, the embedded image vector and the embedded keyword vector are then fed to the *TransFuser* encoder in order to obtain a final attention-based embedding vector  $k_{final}$ , fusing information both from images and

keywords. Then, we use a bidirectional LSTM-based model to serve as a *Hybrid Feature* decoder and sample output words to form medical descriptions. This LSTM-based decoder would have the image vector extracted from the image feature extractor,  $k_{final}$  as mentioned, and a decoder output token from the last time step as inputs for the final sentence generation. We repeat the above for the whole training set.

### Keyword Encoder

In this subsection, we further explore the keyword's effect and its mechanism in our proposed model for automatic medical report generation. Keywords are meant to represent the important image content while subtly alludes its semantic relationship. Therefore, by treating an indefinite numbers of keywords as a keyword sequence, we add on their contribution to the model by introducing a so-called keyword encoder  $f(k_n, I)$ , which takes inputs: a number  $N$  of keywords  $k_n$  and an image  $I$ , referring to Equation (1). We name this non-linear feature mapping procedure  $f(k_n, I)$  "*TransFuser*", which serves as an image-keyword hybrid approach and will be further depicted in the following.

$$k_{final} = f(k_n, I), n \in \{0, \dots, N\} \quad (1)$$

### TransFuser for Multi-modality Features Fusion

Transformer structure (Vaswani et al. 2017) has been firmly established as one of the state-of-the-art approaches in sequence modeling and transduction problems. Its attention mechanism allows language modeling of global dependencies between input and output, preventing the memory constraint limits of traditional recurrent models. Inspired by its structure and in view of its parallelization for attention-weighted positions, we deploy its nature to embed keyword sequences with image content and draw different attention weights on every individual keyword. The so-called scaled dot-product attention mechanism is also used for computing keyword importance on the image embedded vector. But instead of treating the last decoder output as the query in an encoder-decoder attention cell, we use the image vector directly. So, the detailed idea for  $f(k_n, I)$  depicted in Equation (1) would be interpreted as mapping an image query  $Q$  from image  $I$  and a set of keyword key-value pairs  $K, V$  from keywords  $k_n$  to an output  $Z$ .

$$Q = W_t \times \phi(I) \quad (2)$$

$$x_n = W_e k_n, n \in \{0, \dots, N\}$$

$$K = W_k * x_n + b_k \quad (3)$$

$$V = W_v * x_n + b_v$$

$$Z = \text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (4)$$

First, we adopt a CNN image embedder  $\phi$ , (Vinyals et al. 2015; Jing et al. 2018; Laserson et al. 2018; Li et al. 2018; Wang et al. 2018; Li et al. 2019; Xu et al. 2015; Cornia et al. 2019; Karpathy et al. 2015) to extract image features. Then, we map the image feature vector  $\phi(I)$  with the embedding matrix  $W_t \in \mathbb{R}^{T_H \times F}$ , where  $F$  is the image feature size and  $T_H$  is the TransFuser hidden size as shown in Equation (2). The output  $Q$  will serve as an image query to interact with the following keyword vectors. Then regardless of the number of keywords, we map the keyword unordered sequence (a number of keywords) with the embedding matrix by  $W_e \in \mathbb{R}^{E \times V_k}$ .  $E$  means the word embedding size and  $V_k$  means the number of all vocabulary in captions, including keywords. Then, we use two linear

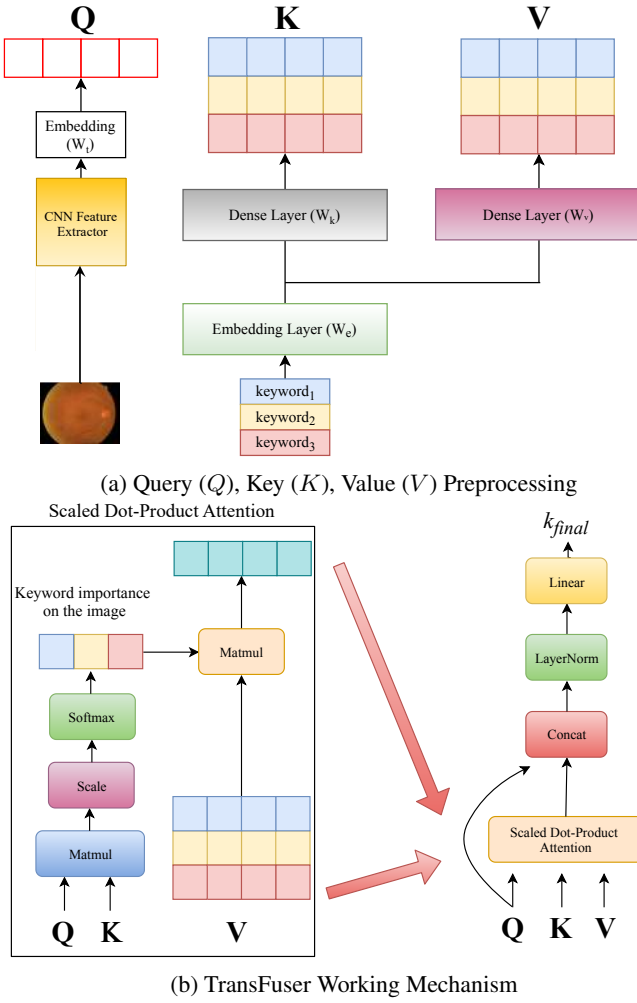


Figure 2: The structure shows the detailed *TransFuser* mechanism. In Figure 2 (a), image contents are treated as the query, where keyword embedded vectors are respectively transformed as key and value vector. In Figure 2 (b), scaled dot-product attention generates a final weighted embedded vector to represent keyword importance on the image vector. Finally, the final keyword vector is generated after a fully-connected layer and layer normalization.

layers ( $W_k, W_v \in \mathbb{R}^{T_H \times E}$ ) to generate keyword key and value vectors  $K, V$  as shown in Equation (3). The output  $Z$  would be computed as a weighted sum of the value vectors  $V$ , where the assigned weight is every keyword importance calculated by dot-product attention on a single image query  $Q$  and the key  $K$  as shown in Equation (4). We leverage the dot-product mechanism for much faster and more space-efficient in exploring the keyword and image relationship. We skip positional encoding trick since we do not wish to include redundant sequential information with keyword unordered nature.

$$Z_{Norm} = \text{LayerNorm}(Q + Z) \quad (5)$$

$$k_{final} = \max(0, W_1 Z_{Norm} + b_1) W_2 + b_2 \quad (6)$$

Finally, we introduce residual shortcut with  $Q$  to add on

attention output  $Z$ . Then the output  $Z_{Norm}$  is obtained after layer normalization and fed into position-wise feed-forward networks similarly connected after the attention sub-layer. Then, we could consistently use the final mixed vector to feed it back into our RNN model, referring to Equation (5) and Equation (6).

To better understand *TransFuser* mechanism behind this embedding trick, we could refer to Figure 2 for detailed descriptions. During the matrix multiplication  $QK^T$ , 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.

### Hybrid Feature Decoder

After obtaining the image-keyword hybrid vector  $k_{final}$ , we could render our complete image description generation model. Here we will feed  $k_{final}$  and image embedding vector  $e_t$  in each time step of a subsequent bidirectional LSTM decoder model, as well as preceding tokens as defined by  $p(S_t|I, S_0, \dots, S_{t-1})$ , where we denote a true sentence describing the image as  $S = (S_0, \dots, S_T)$ . To notice, in each time step  $t \in \{0, \dots, T\}$ , we will have same image embedding vector  $e_t$  and image-keyword hybrid vector  $k_{final}$  inputs. Finally, we could expect to unroll the description generator as follows:

$$e_t = W_d \times \phi(I), t \in \{0, \dots, T\} \quad (7)$$

$$x_t = W_e S_t, t \in \{0, \dots, T\} \quad (8)$$

$$P_t = \text{BiLSTM}([e_t, k_{final}, x_t]), t \in \{0, \dots, T\} \quad (9)$$

$$L(P|I, S) = \mathbb{E}_{S \sim P_I} [\log P(S, I)] \quad (10)$$

In Equation (7) and Equation (8), we represent each word as a bag-of-words id  $S_t$ . Then words  $S$  and image vector  $I$  are mapped to the same space: the image by using an image encoder  $\phi$ , i.e., a deep convolutional neural network connected with a fully-connected layer  $W_d \in \mathbb{R}^{E \times F}$  and the words by word embedding  $W_e \in \mathbb{R}^{E \times V}$ .  $E$  represents the word embedding size,  $F$  is the image feature size, and  $V$  is the number of all vocabulary in captions. In Equation (9), for each time step, we feed the network with image contents  $e_t$ , image-keyword hybrid vector  $k_{final}$  and ground truth word vector  $x_t$  to strengthen its memory of images. We also use dropout technique to alleviate the effect of noises and over-fitting. Finally, if we denote  $P_I$  as the true medical descriptions for  $I$  provided in the training set and  $P(S, I)$  as the final probability distribution after one fully-connected layer and softmax function, we could have the overall likelihood function  $L(P|I, S)$  depending on our medical descriptions and the given image shown in Equation (10). Then finally we could minimize the total loss calculated as the sum of the negative log-likelihood at each time step.

For inference, we use two approaches to generate a sentence given an image. The first one is “*Greedy Search*” where we sample the words based on the maximum likelihood of each word output  $P_t$ , and provide the corresponding embedding vector and sample out  $P_{t+1}$  until  $P_{t+1}$  = special end-of-sentence token. The second one is “*Beam Search*”. Instead of greedily choosing the most likely next step as the sequence is constructed, it expands all possible next steps

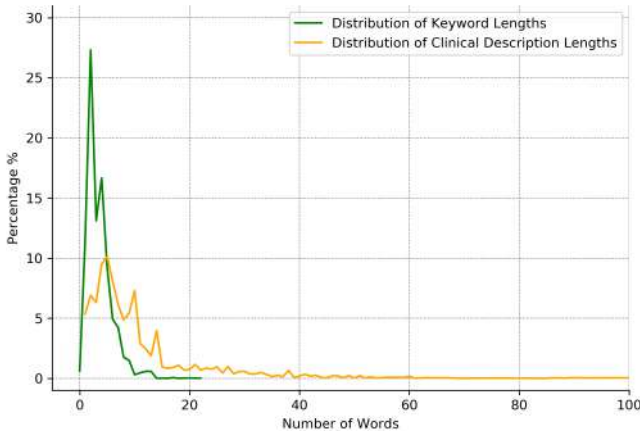


Figure 3: The word length distribution of the keyword and clinical description labels. Based on this figure, we observe the longest word length in our dataset is more than 15 words for keywords and 50 words for clinical descriptions. Note that the longest word length of existing datasets for natural image captioning or VQA (Lin et al. 2014; Chen et al. 2015; Agrawal et al. 2017) is only around 10 words. It implies that our proposed dataset is challenging.

and keeps the  $k$  most likely sentences, where  $k$  is a user-specified parameter and controls the number of beams or parallel searches through the sequence of probabilities. That is, we consider the set of  $k$  sentences up to time  $t$  to be candidates and generate  $P_{t+1}$ . Then we keep maintaining the best  $k$  sentences with the maximum overall probabilities. Multiple candidate sequences will increase the likelihood of better matching a target sequence. However, this increased performance results in a decrease in decoding speed.

## Experiments and Analysis

In this section, we will evaluate our proposed keyword-driven medical report generation method based on the commonly used metrics to see whether our method is capable of generating more accurate and meaningful descriptions for retinal images. We will also analyze the effectiveness of the proposed keyword-image encoder, TransFuser.

### Dataset preparation

To foster the retinal disease research and validate our claim, we introduce a new large-scale retinal image dataset with keyword labels annotated by experienced retina specialists. The keywords labels contain important information about potential diseases and patients based on retinal image analysis and conversation with patients. In practice, keywords are valuable for ophthalmologists to write medical reports for patients. Our proposed retinal image dataset, from thousands of patients, contains two types of images, grey-scale Fluorescein Angiography (FA) and colorful Color Fundus Photography (CFP). The total amount of images is 15,709, including 1,811 FA and 13,898 CFP. We separate the whole dataset into 60%/20%/20%, i.e., 9425/3142/3142, for training/validation/testing, respectively. In our dataset, each retinal image has two corresponding labels, keywords, and clinical description. The word length in our proposed dataset is mainly between 5 and 10 words, referring to Figure 3 for the word length distribution of the keyword and clinical description labels. Note that we take image and key-

words labels as our inputs and clinical description as our ground truth prediction. To better understand our dataset, we show some examples and provide the Venn-style word cloud visualization results for keywords and clinical description labels in **supplementary**.

### Performance evaluation metrics

In our experiment, we exploit the commonly used text evaluation metrics, (Papineni et al. 2002; Lin 2004; Vedantam, Lawrence Zitnick, and Parikh 2015), from the medical report generation field, (Li et al. 2019), to evaluate our generated descriptions for retinal images. Bilingual evaluation understudy (BLEU) is a popular and pioneer metric in automatically evaluating the generated results of machine translation. However, it has some limitation, which is that BLEU scores are useful only if the length of the generated caption is short (Callison-Burch, Osborne, and Koehn 2006). Recall-Oriented Understudy for Gisting Evaluation (ROUGE) is another set of metrics used for evaluating the quality of the generated text. It compares word pairs, word sequences, and n-grams with a collection of human-created reference summaries. Consensus-based Image Description Evaluation (CIDEr) is an automatic consensus metric for measuring the quality of generated image captions, and it achieves human consensus using term frequency-inverse document frequency (TF-IDF) (Robertson 2004).

### Experimental settings

We adopt image feature extractors  $\phi$ , pre-trained on ImageNet, to extract our proposed retinal image dataset’s image features. For each, we first resize the image as the appropriate size to feed in the model. And later on, the layer before the last fully-connected layer is used for embedding features ready to feed in the main LSTM model. To process the annotations and keywords in the dataset, we remove non-alphabet characters, convert all remaining characters to lower-case, and replace all the words that appear once with a special token  $\langle UNK \rangle$ . As a result, our vocabulary size  $V = 4007$  and vocabulary size, including keywords  $V_k = 4292$ . All sentences are truncated or padded with a max length 50. For word embedding layer, we use an embedding size  $E = 300$  to encode words, and we use a hidden layer size  $H_{LSTM} = 256$ . Subsequently first in training, for each image feature set extracted from  $\phi$ , we feed them with word embedded vectors simultaneously in an LSTM. Later on, we start to include keywords fused from our embedded model. In our TransFuser model, we use hidden size  $T_H = 64$  for representation learning. Finally, for every model, we set the mini-batch size to 64 and the learning rate to 0.001 to train the model with two epochs.

### Effectiveness analysis

**Keywords.** Since the characteristic of medical image is different to the general image, and different CNN models have different capabilities to capture the character of the image, we exploit different CNN architectures without and with keywords to demonstrate the effectiveness of our keyword-driven method. In our experiment, we have two types of models, the keyword-driven, and non-keyword-driven. According to Table 1, the model of (Xu et al. 2015) has the best performance among all the non-keyword-driven models, and the keyword-driven model of (Cornia et al. 2019) has the best performance among all the models. Based on Table 1, we notice that all the keyword-driven models are superior to the non-keyword-driven models. Also, we discover that different CNN architectures do have different capabilities to capture the characteristics of the image, especially in the case of our retinal images. Generally speaking, the best keyword-driven model performance, comparing to the best



Table 1: This table shows the evaluation results of our keyword-driven and non-keyword-driven medical report generation models. Note that we highlight the best scores of keyword-driven and non-keyword-driven models in each column, respectively. “w/o” denotes non-keyword-driven baseline models, and “w/” denotes our proposed keyword-driven models. “BLEU-avg” denotes the average score of BLEU-1, BLEU2, BLEU-3, and BLEU-4. Note that the model of (Xu et al. 2015) has the best performance among all the non-keyword-driven models, and the keyword-driven model of (Cornia et al. 2019) has the best performance among all the models. All the keyword-driven models are superior to the non-keyword-driven models.

Model		BLEU-1	BLEU-2	BLEU-3	BLEU-4	BLEU-avg	CIDEr	ROUGE
(Vinyals et al. 2015)	w/o	0.054	0.018	0.002	0.001	0.019	0.056	0.083
	w/	<b>0.208</b>	<b>0.124</b>	<b>0.070</b>	<b>0.032</b>	<b>0.109</b>	<b>0.319</b>	<b>0.254</b>
(Jing et al. 2018)	w/o	0.130	0.083	0.044	0.012	0.067	0.167	0.149
	w/	<b>0.178</b>	<b>0.107</b>	<b>0.058</b>	<b>0.023</b>	<b>0.092</b>	<b>0.330</b>	<b>0.215</b>
(Laserson et al. 2018)	w/o	0.105	0.049	0.009	0.002	0.041	0.064	0.127
	w/	<b>0.148</b>	<b>0.088</b>	<b>0.050</b>	<b>0.023</b>	<b>0.077</b>	<b>0.282</b>	<b>0.198</b>
(Li et al. 2018)	w/o	0.066	0.026	0.007	0.001	0.025	0.076	0.091
	w/	<b>0.176</b>	<b>0.106</b>	<b>0.060</b>	<b>0.029</b>	<b>0.093</b>	<b>0.285</b>	<b>0.229</b>
(Wang et al. 2018)	w/o	0.081	0.031	0.009	0.004	0.031	0.117	0.134
	w/	<b>0.233</b>	<b>0.152</b>	<b>0.095</b>	<b>0.052</b>	<b>0.133</b>	<b>0.369</b>	<b>0.282</b>
(Li et al. 2019)	w/o	0.111	0.060	0.026	0.006	0.051	0.066	0.129
	w/	<b>0.166</b>	<b>0.097</b>	<b>0.049</b>	<b>0.023</b>	<b>0.084</b>	<b>0.304</b>	<b>0.199</b>
(Xu et al. 2015)	w/o	0.153	0.098	0.058	0.027	0.084	0.211	0.184
	w/	<b>0.194</b>	<b>0.122</b>	<b>0.071</b>	<b>0.033</b>	<b>0.105</b>	<b>0.340</b>	<b>0.238</b>
(Cornia et al. 2019)	w/o	0.138	0.080	0.035	0.010	0.066	0.149	0.157
	w/	<b>0.230</b>	<b>0.150</b>	<b>0.094</b>	<b>0.053</b>	<b>0.132</b>	<b>0.370</b>	<b>0.291</b>
(Karpathy et al. 2015)	w/o	0.067	0.029	0.005	0.002	0.026	0.031	0.085
	w/	<b>0.200</b>	<b>0.126</b>	<b>0.079</b>	<b>0.041</b>	<b>0.112</b>	<b>0.296</b>	<b>0.244</b>

Table 2: This table is to show that our proposed TransFuser performs better than the baselines under the “Image + Keywords” situation. Note that “mul” denotes element-wise multiplication, and “sum” denotes summation. The results are based on the best keyword-driven model (Cornia et al. 2019) in Table 1.

Fusing method	BLEU-1	BLEU-2	BLEU-3	BLEU-4	BLEU-avg	CIDEr	ROUGE
Baseline-1 (sum)	0.014	0.002	0.001	0.000	0.004	0.019	0.023
Baseline-2 (mul)	0.077	0.031	0.004	0.001	0.028	0.042	0.102
Ours (TransFuser)	<b>0.230</b>	<b>0.150</b>	<b>0.094</b>	<b>0.053</b>	<b>0.132</b>	<b>0.370</b>	<b>0.291</b>

Table 3: The table is to show that the proposed TransFuser is capable of capturing not only the original information of keywords and image but also the interactive information between them. The results are based on the best keyword-driven model (Cornia et al. 2019) in Table 1.

Input	BLEU-1	BLEU-2	BLEU-3	BLEU-4	BLEU-avg	CIDEr	ROUGE
Keywords	0.057	0.029	0.017	0.005	0.027	0.168	0.091
Image	0.153	0.098	0.058	0.027	0.084	0.211	0.184
Image + Keywords	<b>0.230</b>	<b>0.150</b>	<b>0.094</b>	<b>0.053</b>	<b>0.132</b>	<b>0.370</b>	<b>0.291</b>

non-keyword-driven model, increases about 58% in BLEU-avg, 75% in CIDEr, and 58% in ROUGE, respectively. The reason is that keywords are meant to represent the important content of image while subtly alludes its semantic relationship. So, in the above case, we can consider keywords as extra information for the models. Our experimental results show that the proposed keyword-driven method is superior to the non-keyword-driven method in the sense of the commonly used metrics, referring to Table 1 for more details.

**TransFuser.** Since our keywords have unordered nature, the intuitive ways to fuse the keywords and image features are the summation and element-wise multiplication. According to Table 2, we discover that our proposed TransFuser beats the summation and element-wise multiplication baselines. It implies that our TransFuser is effective. Note that the results are based on the best model (Cornia et al. 2019) in Table 1.

**Interaction between keywords and image.** According to Table 3, we discover that the performance of “Image” and “Keywords” only baselines are worse than the “Image + Keywords” method. It implies that the interaction between

keywords and image is crucial for medical report generation and our proposed TransFuser is capable of capturing this interaction, i.e., the relation between keywords and image. Note that the results are based on the best model (Cornia et al. 2019) in Table 1.

#### Qualitative results and analysis

We present some qualitative results generated by our medical report generation model in Figure 4. Although our models cannot create correct “age” or “gender” as these are not present in the content, the models are capable of generating correct descriptions to important characteristics for retinal images. Note that “age” and “gender” are not key factors when ophthalmologists analyze a retinal disease. It means that the proposed keyword-driven method is capable of creating more accurate and meaningful descriptions for retinal images. Also, we discover that when we take the commonly used text evaluation metrics, such as BLEU, to measure the quality of the generated captions, the evaluation scores are not that good. That is because of the innate property of the BLEU metric (Papineni et al. 2002; Huang et al. 2019; 2018; Huang, Alfadly, and Ghanem 2017). However, note


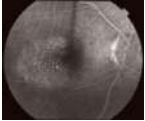
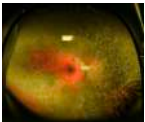
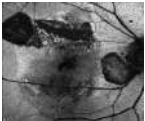
	Ground Truth Keywords	Ground Truth Caption	Non-keyword-driven Model	Keyword-driven Model
	retinopathy, prematurity, rop	Baby born at 26 weeks gestation. The right eye reveals a large arteriole venous shunt in the peripheral retina. The venules leading away from the shunt appear to be larger than normal. Superimposed on the venules and arterioles are a number of small, round, reddish pink 'bulbs' that are preretinal.	13 year old female, solar retinopathy / familial.	Red free photo showing traumatic photo revealed traumatic retinal of vision of optic nerve head and color photo of the macula that remained hole.
	idiopathic, macular, telangiectasia, parafoveal, juxtafoveal, telangiectasis	The fellow eye was unremarkable on this red free image.	Fundus autofluorescence faf image of the left eye of a 25 year old woman with bilateral macular colobomata and pigmentary retinopathy similar to the optic nerve with proliferative diabetic retinopathy.	The telangiectasis occurs unilaterally in the macula be of choroidal folds peripheral nevus shows cnvm in this eyes shows resolved with myopia over the macula and lung remained and remained over the exam was remained cnvm. Resolved and pdt.
	retinitis, pigmentosa	Autosomal dominant retinitis pigmentosa.	Female patient, best family.	Left fundus of a 32 year old lady with bilateral retinitis pigmentosa. She has progressive visual complaints starting at age 5, and is the offspring of a consanguineous marriage. Marked disc pallor, retinal arteriolar attenuation, pigment disturbance and macular degeneration are classic features.
	autofluorescence, imaging, age-related, macular, degeneration, amd	An autofluorescence image of a 78 year old man with an age related macular degeneration on his both eyes.	Macular hole.	Autofluorescence to image of the right eye of a 28 year old woman with acute decrease in vision mainly right eye is with pigment clumping and optic nerve drusen in the right eye.

Figure 4: This figure shows the generated medical reports for retinal images based on the keyword-driven and non-keyword-driven models. Note that age and gender are not important factors for ophthalmologists on analyzing a potential retinal disease in practice. Based on this figure, we know that the keyword-driven models are capable of generating more accurate descriptions of important characteristics for retinal images. The blue color is to denote the keywords understanding of our proposed model. Please refer to “Discussion” section for more details.

that since our proposed real-world dataset is challenging and exhibits complex structures - e.g., intrinsic relations among discrete entities under nature’s law (Strubell et al. 2018; Hu et al. 2016) - the baseline methods have difficulty to work well on it. Thus, our methods should be viewed as a first promising step for this difficult challenge.

## Discussion

**Reasoning ability.** According to Figure 4, we discover that the non-keyword-driven model sometimes cannot generate a long and correct conceptual description for retinal images. Also, Figure 4 shows that the proposed keyword-driven model has better reasoning ability than the non-keyword-driven one since it can create a long and conceptually correct medical report for retinal images.

**Does our model fully understand input keywords?** The answer probably is no. However, based on Figure 4, our proposed keywords-driven model is capable of partially understanding input keywords. For example, in the fourth example of Figure 4, our proposed model generates “acute decrease” in the description based on the understanding of “degeneration” keyword-driven. Similarly, the first example of Figure 4 also demonstrates some keyword understanding ability of our model. It implies that our proposed method makes us closer to the goal of automatic medical report generation for retinal images on our proposed real-world dataset.

**Are features from deeper models good in any task?** Based on our experimental result in Table 1 and some VQA task papers, (Huang et al. 2019; Huang, Alfadly, and Ghanem

2017), we see that image features extracted by deeper networks do not imply better performance in a task with multi-modal inputs, even though they are good in most of the pure computer vision tasks such as object detection and activity recognition. We conjecture that the description generation in an LSTM unit still needs other transformations based on image features, so it probably hurts the final performance of the medical report generation task even when the best image features extracted by deeper networks are used.

## Conclusion and Future Work

To sum up, we propose a novel keyword-driven medical report generation method to perform the automatic report generation task for retinal images. We introduce a new approach, called TransFuser, which is capable of effectively fusing features with different modalities. Also, we introduce a new large-scale retinal image dataset with keyword labels annotated by ophthalmologists. Our experiments show that the proposed model can generate more accurate and meaningful descriptions for retinal images, and the performance increases about 58% in BLEU-avg, 75% in CIDEr, and 58% in ROUGE. The above shows our proposed keyword-driven method is superior to the non-keyword-driven one. In Figure 4, we discover there exist inconsistency (Li et al. 2018), referring to our **supplementary** for more examples and details, between the ophthalmologists’ opinions and the evaluation result based on the commonly used text evaluation metrics. So, developing a proper medical report evaluation metric will be interesting future work.

## References

- Agrawal, A.; Lu, J.; Antol, S.; Mitchell, M.; Zitnick, C. L.; Parikh, D.; and Batra, D. 2017. Vqa: Visual question answering. *International Journal of Computer Vision* 123(1):4–31.
- Agrawal, A.; Batra, D.; Parikh, D.; and Kembhavi, A. 2018. Don't just assume; look and answer: Overcoming priors for visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4971–4980.
- Ben-Younes, H.; Cadene, R.; Cord, M.; and Thome, N. 2017. Mutan: Multimodal tucker fusion for visual question answering. In *Proceedings of the IEEE international conference on computer vision*, 2612–2620.
- Callison-Burch, C.; Osborne, M.; and Koehn, P. 2006. Re-evaluation the role of bleu in machine translation research. In *11th Conference of the European Chapter of the Association for Computational Linguistics*.
- Chen, X.; Fang, H.; Lin, T.-Y.; Vedantam, R.; Gupta, S.; Dollár, P.; and Zitnick, C. L. 2015. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*.
- Chen, K.; Wang, J.; Chen, L.-C.; Gao, H.; Xu, W.; and Nevatia, R. 2016. Abc-cnn: An attention based convolutional neural network for visual question answering. In *CVPRW*.
- Cornia, M.; Baraldi, L.; Cucchiara, R.; Cornia, M.; Baraldi, L.; and Cucchiara, R. 2019. Show, control and tell: A framework for generating controllable and grounded captions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 8307–8316.
- Fukui, A.; Park, D. H.; Yang, D.; Rohrbach, A.; Darrell, T.; and Rohrbach, M. 2016. Multimodal compact bilinear pooling for visual question answering and visual grounding. *arXiv preprint arXiv:1606.01847*.
- Hu, Z.; Ma, X.; Liu, Z.; Hovy, E.; and Xing, E. 2016. Harnessing deep neural networks with logic rules. *arXiv preprint arXiv:1603.06318*.
- Huang, J.-H.; Alfadly, M.; and Ghanem, B. 2017. Vqabq: visual question answering by basic questions. *VQA Challenge Workshop in CVPR*.
- Huang, J.-H.; Dao, C. D.; Alfadly, M.; Yang, C. H.; and Ghanem, B. 2018. Robustness analysis of visual qa models by basic questions. *VQA Challenge and Visual Dialog Workshop in CVPR*.
- Huang, J.-H.; Dao, C. D.; Alfadly, M.; and Ghanem, B. 2019. A novel framework for robustness analysis of visual qa models. *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence*.
- Jing, B.; Xie, P.; Xing, E.; Jing, B.; Xie, P.; and Xing, E. 2018. On the automatic generation of medical imaging reports. *ACL*.
- Karpathy, A.; Fei-Fei, L.; Karpathy, A.; Fei-Fei, L.; Karpathy, A.; and Fei-Fei, L. 2015. Deep visual-semantic alignments for generating image descriptions. In *CVPR*, 3128–3137.
- Kazemzadeh, S.; Ordonez, V.; Matten, M.; and Berg, T. 2014. Referitgame: Referring to objects in photographs of natural scenes. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 787–798.
- Kim, D.-J.; Choi, J.; Oh, T.-H.; and Kweon, I. S. 2019. Dense relational captioning: Triple-stream networks for relationship-based captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 6271–6280.
- Laserson, J.; Lantsman, C. D.; Cohen-Sfady, M.; Tamir, I.; Goz, E.; Brestel, C.; Bar, S.; Atar, M.; and Elnekave, E. 2018. Texttray: Mining clinical reports to gain a broad understanding of chest x-rays. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 553–561. Springer.
- Li, Y.; Liang, X.; Hu, Z.; and Xing, E. P. 2018. Hybrid retrieval-generation reinforced agent for medical image report generation. In *Advances in Neural Information Processing Systems*, 1530–1540.
- Li, C. Y.; Liang, X.; Hu, Z.; and Xing, E. P. 2019. Knowledge-driven encode, retrieve, paraphrase for medical image report generation.
- Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*, 740–755. Springer.
- Lin, C.-Y. 2004. Rouge: A package for automatic evaluation of summaries. *Text Summarization Branches Out*.
- Lu, J.; Yang, J.; Batra, D.; and Parikh, D. 2016. Hierarchical question-image co-attention for visual question answering. In *Advances In Neural Information Processing Systems*, 289–297.
- Mao, J.; Huang, J.; Toshev, A.; Camburu, O.; Yuille, A. L.; and Murphy, K. 2016. Generation and comprehension of unambiguous object descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 11–20.
- Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W.-J. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, 311–318. Association for Computational Linguistics.
- Pedersoli, M.; Lucas, T.; Schmid, C.; and Verbeek, J. 2017. Areas of attention for image captioning. In *Proceedings of the IEEE International Conference on Computer Vision*, 1242–1250.
- Robertson, S. 2004. Understanding inverse document frequency: on theoretical arguments for idf. *Journal of documentation* 60(5):503–520.
- Shih, K. J.; Singh, S.; and Hoiem, D. 2016. Where to look: Focus regions for visual question answering. In *CVPR*, 4613–4621.
- Strubell, E.; Verga, P.; Andor, D.; Weiss, D.; and McCallum, A. 2018. Linguistically-informed self-attention for semantic role labeling. *arXiv preprint arXiv:1804.08199*.
- Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. In *Advances in neural information processing systems*, 5998–6008.
- Vedantam, R.; Lawrence Zitnick, C.; and Parikh, D. 2015. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4566–4575.
- Vinyals, O.; Toshev, A.; Bengio, S.; and Erhan, D. 2015. Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3156–3164.
- Wang, C.; Yang, H.; Bartz, C.; and Meinel, C. 2016. Image captioning with deep bidirectional lstms. In *Proceedings of the 24th ACM international conference on Multimedia*, 988–997. ACM.
- Wang, X.; Peng, Y.; Lu, L.; Lu, Z.; and Summers, R. M. 2018. Tienet: Text-image embedding network for common thorax disease classification and reporting in chest x-rays. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 9049–9058.
- Xu, K.; Ba, J.; Kiros, R.; Cho, K.; Courville, A.; Salakhutdinov, R.; Zemel, R.; and Bengio, Y. 2015. Show, attend and tell: Neural image caption generation with visual attention. *arXiv preprint arXiv:1502.03044*.