

Non-local Attention Improves Description Generation for Retinal Images

Jia-Hong Huang^{1†}, Ting-Wei Wu³, C.-H. Huck Yang³, Zenglin Shi¹,
I-Hung Lin⁴, Jesper Tegner², Marcel Worring¹

¹University of Amsterdam, ²King Abdullah University of Science and Technology (KAUST), ³Georgia Institute of Technology,

⁴Department of Ophthalmology, Tri-Service General Hospital, National Defense Medical Center, Taiwan,

† corresponding author: j.huang@uva.nl

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 mainly rely on the input image to generate descriptions. However, many 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 observe 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. Since these keywords commonly exist and are useful for generating medical reports, we incorporate them into automatic report generation. Since we have two types of inputs - expert-defined unordered keywords and images - effectively fusing features from these different modalities is challenging. To that end, we propose a new keyword-driven medical report generation method based on a non-local attention-based multi-modal feature fusion approach, TransFuser, which is capable of fusing features from different types of inputs based on such attention. Our experiments show the proposed method successfully captures the mutual information of keywords and image content. We further show our proposed keyword-driven generation model reinforced by the TransFuser is superior to baselines under the popular text evaluation metrics BLEU, CIDEr, and ROUGE. TransFuser Github: <https://github.com/Jhhuangkay/Non-local-Attention-Improves-Description-Generation-for-Retinal-Images>.

1. Introduction

Automatic medical report generation for retinal images is a challenging computer vision task, falling within the broader task domain of image captioning [43]. In this task, long and semantically coherent medical descriptions for a given image must be generated algorithmically [18, 16, 17,

15]. Several technical features of retinal image report generation [18] complicate this task when compared to the more well-studied domain of natural image captioning, e.g., in [3, 25]. One example is that retinal and natural images have very different characteristics, both in objects' sizes as well as details [36, 18]. As such, existing methods, such as [43, 22], which work well on natural image datasets often do not generalize well to retinal images.

Recently, some methods [19, 28] have been proposed to generate medical reports. These approaches are based on the traditional natural image captioning model and consequently work on image content only. However, many abstract medical concepts or descriptions, [26, 19], cannot be generated based on image information only. To tackle this issue, the authors in [18] propose an average-based method to exploit the expert-defined contextual information, in the form of a keyword sequence, and image content to generate better descriptions. Although the keyword-driven average-based method from [18] improves the medical report generation model, how to effectively fuse the multi-modal information, i.e., expert-defined keywords and image, will become another key issue. Using the average-based method to fuse the multi-modal information in this case probably cannot effectively capture mutual/interactive information between the keywords and image [18]. Losing such interactive information can reduce the quality of the generated descriptions [19, 2, 14, 10, 11]. To generate more accurate and meaningful descriptions for retinal images, we will need a specialized method which is capable of effectively fusing the information of expert-defined keywords and image.

In this paper, we propose a new keyword-driven medical report generation method equipped with a non-local attention-based keyword-image encoder, called TransFuser, illustrated in Figure 1 and 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. The attention mechanism al-

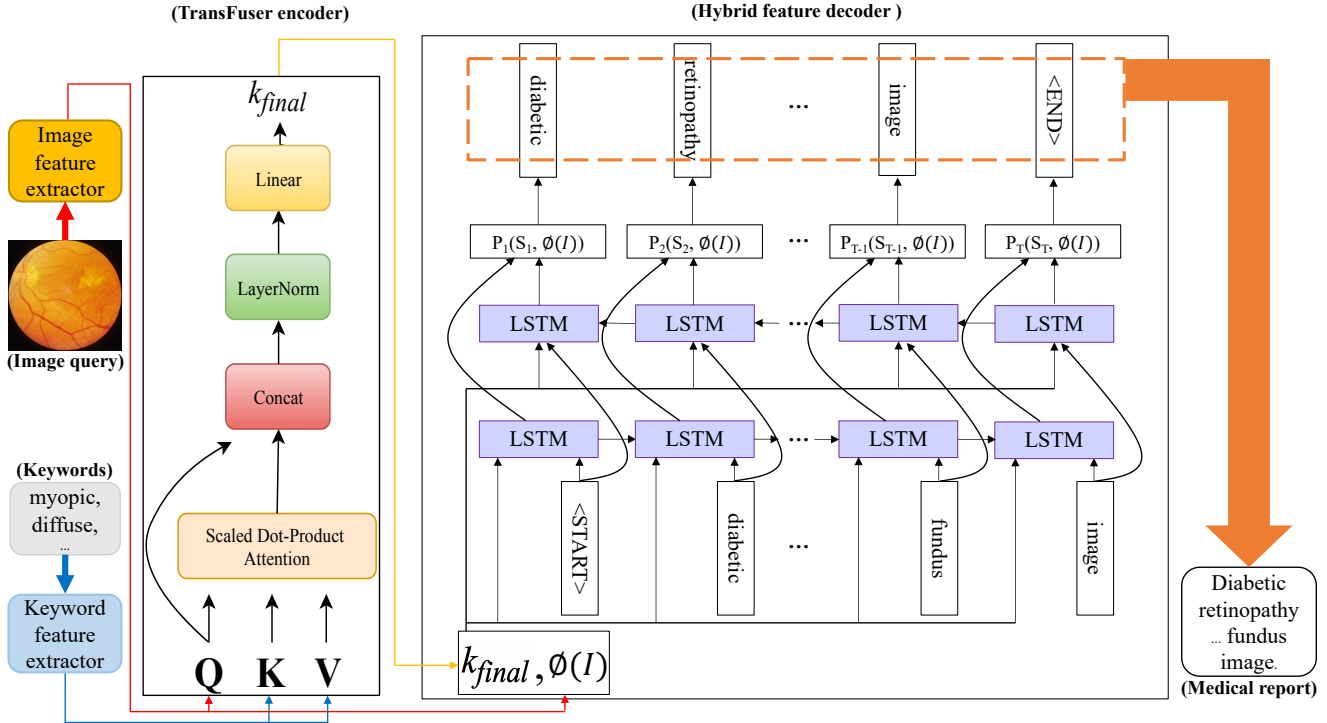


Figure 1: This figure shows the flowchart of our proposed keyword-driven medical report generation model reinforced by our proposed TransFuser. It takes two inputs, a retinal image, and keywords. The purpose of keywords is to reinforce the model to generate more accurate and meaningful descriptions for retinal images. “TransFuser” denotes our proposed multi-modal feature fuser. In the TransFuser, “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 are 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 $i = 1, 2, \dots, T$.

allows the inputs to guide the different modalities to generate more accurate results. Because of the non-local attention mechanism [37], the proposed method is capable of effectively capturing the mutual information between the image and keywords.

The authors [18] have introduced a state of the art model and a large-scale and unique dataset with expert-defined keywords for medical report generation for retinal images. So, we demonstrate the experimental results of our proposed model based on their proposed dataset. We show that our proposed keyword-driven generation model reinforced by the TransFuser is capable of creating more accurate and meaningful descriptions/reports for retinal images than baseline models. This performance is shown in several text evaluation metrics: BLEU-avg (+32%), CIDER (+2.5%), and ROUGE (+25.4%).

2. Related Work

In this section, we review the related image captioning methods for natural and medical images and the existing retinal image datasets.

2.1 Caption Generation for Natural Images

The encoder-decoder based network architecture, [39, 42, 21, 9, 8, 7], is the most popular method to perform image captioning. In these networks, the convolution neural network (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 [31], the authors introduce a text generation method to generate a description for some specific object or region that is called referring expression [24]. The authors of [40] propose a bidirectional LSTM-based method to generate captions. The method exploits past and future information to learn long-term visual language interactions. Attention-based models have shown good performance in image captioning. The authors of [33] introduce an area-based attention model for image captioning. The model predicts the next word and corresponding regions of the image in each RNN time step for generating image descriptions. To the best of our knowledge, most of the existing natural image captioning methods mainly rely on a single image to generate descriptions. However, some abstract concepts or descriptions, [26, 19], cannot be generated based on image information only. So, in this work, we exploit the expert-defined keywords sequence [18] to help

models generate better descriptions.

2.2 Caption Generation for Medical Images

The authors of [28] 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 [19], the authors propose a multi-task learning framework to predict tags and generate captions at the same time. Also, they use an attention-based 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. The authors of [19] note that the generated medical reports based on most of the existing methods are fully-structured or semi-structured, e.g., have tags or use templates. From the medical point of view, radiology images of the chest and retinal images have different properties, such as objects' sizes and details [26, 36, 18]. From the lower level feature perspective, such as color, radiology images of the chest are mainly grey-scale [26] where retinal images are mainly colorful [18]. Most of the methods mentioned above mainly rely on image input to generate captions. In this work, our proposed method starts from the CNN-RNN based framework. To effectively fuse features with different modalities, we introduce TransFuser to reinforce our medical report generation model. Note that a keywords sequence has an unordered nature which normal sentences do not have. How to fuse the input image and keywords with a minimum loss of information, in general, remains an open question [18].

2.3 Retinal Dataset for Medical Report Generation

Since retinal disease research has a long history, many retinal datasets, e.g., [5, 23, 4, 18, 44, 45, 30], have been proposed for computer vision tasks. In [5], the authors have proposed the STARE dataset which consists of 397 retinal images. The dataset is mainly exploited to develop an automatic system for diagnosing diseases of the human eye. In [23], a dataset DIARETDB1 consisting of 89 color fundus images has been introduced. In the DIARETDB1 dataset, 84 retinal images contain at least mild non-proliferative signs of Diabetic Retinopathy (DR), and 5 images are considered normal, i.e., not containing any signs of the DR. The authors of [4] have proposed a dataset MESSIDOR containing 1200 fundus images. In the MESSIDOR dataset, each image has a corresponding text-based clinical description. There is no manual annotation, e.g., of lesions contours or position, on each image. According to [5, 23, 4], the datasets of STARE, DIARETDB1, and MESSIDOR are

composed of retinal images and clinical descriptions, but they do not contain expert-defined keywords. Hence, they are suitable for the task of medical report generation but not for the keyword-driven medical report generation task. Note that the other existing retinal datasets only contain retinal images and they are only used for a pure computer vision task [18]. To validate the keyword-driven idea and train a deep medical report generation model, the authors of [18] have proposed a retinal dataset DeepEyeNet to perform the task of keyword-driven medical report generation. The DeepEyeNet dataset is much larger than the aforementioned retinal datasets. It is composed of 15,709 retinal images. In the DeepEyeNet dataset, each image has the corresponding expert-defined keyword sequence and text-based clinical description. In this work, DeepEyeNet is a proper dataset used to validate the effectiveness of the proposed model. We summarize the aforementioned retinal datasets in Table 1.

3. Methodology

Overview

In this section, we present our keyword-driven medical report generation model reinforced by the TransFuser, referring to Figure 1, and illustrate methods to train the model with supervised keyword knowledge. First, an image and a number of keywords will be fed into modality-specific extractors to acquire an embedded image vector and an embedded keyword vector for each. After this information extraction, the vectors are fed into an encoder in order to obtain a final attention-based embedding vector k_{final} , fusing information from images and keywords, referring to “*Keyword Encoder*” subsection and “*TransFuser Encoder for Multi-modal Feature Fusion*” subsection. Then, we use a bidirectional LSTM-based model to serve as a decoder, referring to “*Hybrid Feature Decoder*” subsection, 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, referring to Figure 1.

3.1 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 to its semantic relationship. Therefore, by treating an indefinite numbers of keywords as a keyword sequence, we add their contribution to the model by introducing a so-called keyword encoder $f(k_n, I)$, which takes as inputs: N 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 next subsection.

Table 1: Summary of retinal datasets. The DeepEyeNet dataset is unique and much larger than the other retinal datasets. According to [18], most of the existing retinal dataset only contains image data, and the dataset size is not large. “Text*” denotes clinical descriptions and keywords. “Text” denotes clinical descriptions only.

Name of Dataset	Field of View	Resolution	Data Type	Number of Images
STARE [5]	$\approx 30^\circ - 45^\circ$	700 * 605	Image + Text	397
DIARETDB1 [23]	50°	1500 * 1152	Image + Text	89
MESSIDOR [4]	45°	1440 * 960 – 2304 * 1536	Image + Text	1,200
DeepEyeNet [18]	$\approx 30^\circ - 60^\circ$	various	Image + Text*	15,709

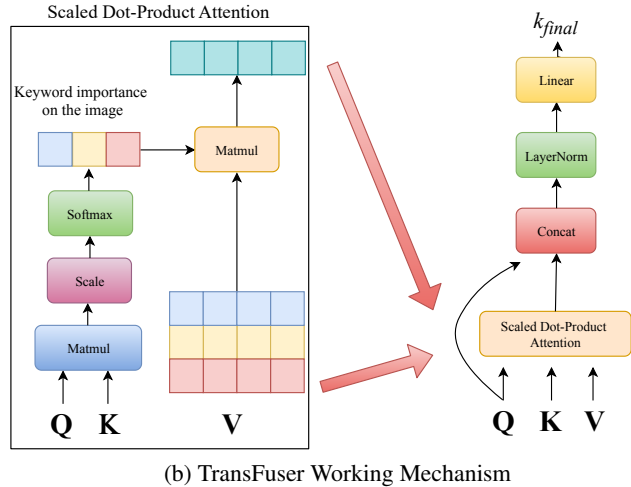
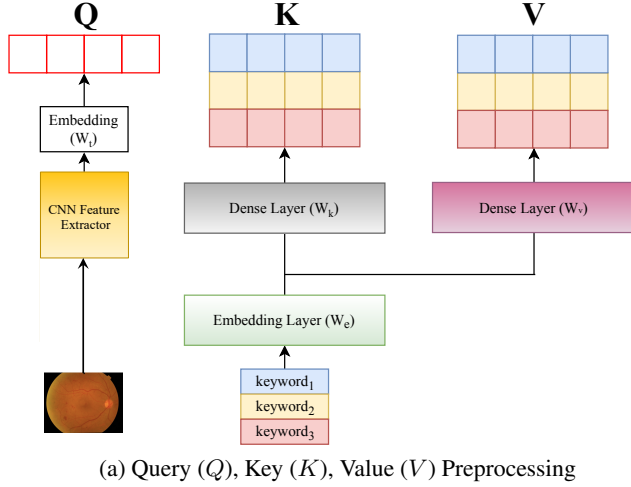


Figure 2: The structure shows in detail the *TransFuser* mechanism. (a), image contents are treated as the query, where keyword embedded vectors are respectively transformed as key and value vector. (b), scaled dot-product attention generates a final weighted embedded vector to represent keyword importance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

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

3.2 TransFuser Encoder for Multi-modal Feature Fusion

Transformer structure [37, 46, 13, 35, 6] 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 put different attention weights on every individual keyword. The so-called scaled dot-product attention mechanism is 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 formulation for $f(k_n, I)$ depicted in Equation-(1) can be interpreted as follows: mapping an image query Q , derived from image I , and a set of keyword key-value pairs (K, V) , derived from keywords k_n , to an output k_{final} . We describe the full procedure as follows.

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

$$\begin{aligned} x_n &= W_e k_n, n \in \{0, \dots, N\} \\ K &= W_k * x_n + b_k \\ V &= W_v * x_n + b_v \end{aligned} \quad (3)$$

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

First, we adopt a CNN image embedder ϕ , [39, 19, 26, 28, 41, 20, 43, 3, 22] to extract image features. Then, we map the image feature vector $\phi(I)$ with the embedding matrix $W_t \in \mathbb{R}^{T_H \times F}$, as shown in Equation-(2). Here, F is the image feature size and T_H is the TransFuser hidden size. The output Q will serve as an image query to interact with the keyword vectors later. Then regardless of the number of keywords, we map the keyword unordered sequence (a number of keywords) with the embedding matrix by $W_e \in \mathbb{R}^{E \times V_k}$. Here E denotes the word embedding size and V_k indicates the number of all vocabulary used in captions, including keywords. Then, we use two linear layers ($W_k, W_v \in \mathbb{R}^{T_H \times E}$) to generate keyword key and value vectors, i.e., K and V , as shown in Equation-(3). The output Z is computed as a weighted sum of the value vectors V , where the assigned weight is every keyword’s 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 exploration of the keyword and image relationship. We skip the positional encoding trick, [37], since we do not wish to include redundant sequential information due to the keyword unordered nature [18].

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

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

Finally, we introduce a residual shortcut with Q to add on to 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, referring to Equation-(5). We can now consistently use the final mixed vector, referring to Equation-(6), to feed it back into our RNN model.

To better understand the *TransFuser* mechanism behind this embedding trick, we 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 K . Therefore, we obtain every keyword weights on the image feature vector. After the scaled and softmax operation, we get probability-like weights for each keyword interpreted as their attention or relationship with the current image. Finally, we multiply the weights with the corresponding value V to denote their hybrid importance in providing attention-weighted image-keyword information.

3.3 Hybrid Feature Decoder

After obtaining the image-keyword hybrid vector k_{final} , we can render our complete image description/report generation model. Here we 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 $S = (S_0, \dots, S_T)$. Note that in each time step $t \in \{0, \dots, T\}$, we have the same image embedding vector e_t and image-keyword hybrid vector k_{final} inputs. We unroll the description generator as follows:

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

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

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

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

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 ϕ , 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 $W_e \in \mathbb{R}^{E \times V}$. Here 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 to alleviate the effect of noises and overfitting. 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 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 minimize the total loss calculated as the sum of the negative log-likelihood at each time step.

For inference, we use “*Beam Search*” [18] to generate a sentence given an image. Instead of greedily choosing the most likely next step as the sequence is constructed, it expands all possible next steps 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.

4. 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, i.e., *TransFuser*, based on the same assumption, mentioned by [18, 19], that an effective deep model is helpful in practice.

4.1 Dataset and Performance Evaluation Metrics

To foster the retinal disease research, the authors [18] introduce a new large-scale retinal image dataset, DEN dataset, with unique keyword labels annotated by experienced retina specialists and also propose a medical report generation model based on the DEN dataset. 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. The DEN dataset 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 follow the same setup of the DEN dataset, i.e., 60%/20%/20% for training/validation/testing, respectively. In the DEN dataset, each retinal image has two corresponding labels, i.e., key-

Table 2: This table shows the evaluation results of our keyword-driven and non-keyword-driven medical report generation models. 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 [43] has the best performance among all the non-keyword-driven models, and the keyword-driven model of [3] has the best performance among all the models. All the keyword-driven models, based on the TransFuser, 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 [39]	w/o	0.054	0.018	0.002	0.001	0.019	0.056	0.083
	w/	0.208	0.124	0.070	0.032	0.109	0.319	0.254
Jing et al. 2018 [19]	w/o	0.130	0.083	0.044	0.012	0.067	0.167	0.149
	w/	0.178	0.107	0.058	0.023	0.092	0.330	0.215
Laserson et al. 2018 [26]	w/o	0.105	0.049	0.009	0.002	0.041	0.064	0.127
	w/	0.148	0.088	0.050	0.023	0.077	0.282	0.198
Li et al. 2018 [28]	w/o	0.066	0.026	0.007	0.001	0.025	0.076	0.091
	w/	0.176	0.106	0.060	0.029	0.093	0.285	0.229
Wang et al. 2018 [41]	w/o	0.081	0.031	0.009	0.004	0.031	0.117	0.134
	w/	0.233	0.152	0.095	0.052	0.133	0.369	0.282
Joshi et al. 2020 [20]	w/o	0.111	0.060	0.026	0.006	0.051	0.066	0.129
	w/	0.166	0.097	0.049	0.023	0.084	0.304	0.199
Xu et al. 2015 [43]	w/o	0.153	0.098	0.058	0.027	0.084	0.211	0.184
	w/	0.194	0.122	0.071	0.033	0.105	0.340	0.238
Cornia et al. 2019 [3]	w/o	0.138	0.080	0.035	0.010	0.066	0.149	0.157
	w/	0.230	0.150	0.094	0.053	0.132	0.370	0.291
Karpathy et al. 2015 [22]	w/o	0.067	0.029	0.005	0.002	0.026	0.031	0.085
	w/	0.200	0.126	0.079	0.041	0.112	0.296	0.244

words and clinical description. The word length in DEN is mainly between 5 and 10 words. Note that we take image and keywords labels as our inputs and clinical description as our ground truth prediction. In our experiment, we exploit the commonly used text evaluation metrics, [32, 29, 38], from the medical report generation field, [18, 27], to evaluate our generated descriptions for retinal images.

4.2 Experimental Settings

We adopt image feature extractors ϕ , pre-trained on ImageNet, to extract our proposed retinal image dataset’s image features. Note that in most of the cases [34, 1] including ours, retinal images are mainly colorful which is similar to ImageNet. So, from the lower level feature perspective, such as color, pre-training on ImageNet can help models’ performances. For each, we first resize the image as the appropriate size to feed into the model. And later on, the layer before the last fully-connected layer is used for embedding features ready to feed into 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 only 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 of 50. For the word embedding layer, we use an embedding size of $E = 300$ to encode words, and

we use a hidden layer size $H_{LSTM} = 256$. Subsequently first in training, for each image feature set extracted from ϕ , 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. 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.

4.3 Effectiveness Analysis

Keywords. Since the characteristics of medical images are different from general images, 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 2, the model of [43] has the best performance among all the non-keyword-driven models, and the keyword-driven model of [3] has the best performance among all the models. Based on Table 2, 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

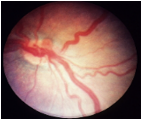
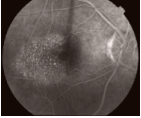
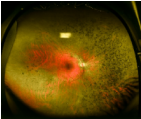
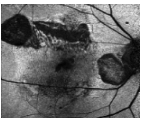
	Ground Truth Keywords	Ground Truth Caption	Non-keyword-driven Model	Keyword-driven Model
	retinopathy, prematurity, rop	Baby born at [age] weeks gestation. The right eye reveals a large arteriole venous shunt in the peripheral retina. The venules leading away from the shunt appear to be larger than normal. Superimposed on the venules and arterioles are a number of small, round, reddish pink 'bulbs' that are preretinal.	[age] year old [gender], 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 [age] year old [gender] 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.	[gender] patient, best family.	Left fundus of a [age] year old [gender] with bilateral retinitis pigmentosa. [gender] has progressive visual complaints starting at age [age], 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 [age] year old [gender] with an age related macular degeneration on [gender] both eyes.	Macular hole.	Autofluorescence to image of the right eye of a [age] year old [gender] with acute decrease in vision mainly right eye is with pigment clumping and optic nerve drusen in the right eye.

Figure 3: In this figure, we randomly show some generated medical reports based on the keyword-driven and non-keyword-driven models. Based on this figure, we see 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 the “Discussion” section and “Qualitative Results and Analysis” subsection for more details and the explanation of the [age] and [gender].

Table 3: This table is to show that our proposed TransFuser performs better than the baseline models 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 [3] in Table 2.

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
DeepOpht [18]	0.184	0.114	0.068	0.032	0.100	0.361	0.232
Ours (TransFuser)	0.230	0.150	0.094	0.053	0.132	0.370	0.291

Table 4: 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 [3] in Table 2.

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	0.230	0.150	0.094	0.053	0.132	0.370	0.291

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 the image while subtly alluding to 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 2.

TransFuser. Since our keywords have an unordered nature, the intuitive ways to fuse the keywords and image features are summation and element-wise multiplication. The authors [18] have proposed another average method, i.e., DeepOpht, to fuse the keywords and image features. According to Table 3, we can see that our proposed TransFuser beats the summation and element-wise multiplica-

tion baselines and DeepOpht. We discover that TransFuser and DeepOpht perform much better than the summation and element-wise multiplication baselines. The reason is that since a keyword sequence is unordered, summation and element-wise multiplication feature fusion methods probably capture the input order of words which could be a wrong/fake order, i.e., bias information. TransFuser and DeepOpht are less affected by such bias information. Moreover, our TransFuser performs better than DeepOpht. Based on Table 3, the performance increases about 32% in BLEU-avg, 2.5% in CIDEr, and 25.4% in ROUGE, respectively. The reason is that our TransFuser captures better mutual/interactive information between keywords and image. See the next subsection.

Interaction between keywords and image. According to Table 4, the performance of “Image”-only and “Keywords”-only baselines are worse than the “Image + Keywords” method. It implies 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 again the result is based on the best model [3] in Table 2.

4.4 Qualitative Results and Analysis

We present some qualitative results generated by our medical report generation model in Figure 3. 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 of important characteristics for retinal images. Note that, ideally, “age” and “gender” would be part of the dataset and that a system should make it part of the description by post-processing or slot filling [18].

4.5 Evaluation with Retinal Specialists

We use 5-level report/description 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. Note that the ground-truth reports are generated by ophthalmologists. We ask the other five different retinal specialists to score the quality of the model-generated report and the corresponding 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 3.6/5.0 for our model-generated reports and an average score of 4.5/5.0 for the ground-truth reports. Since the ground-truth reports are defined by ophthalmologists, the above results show that the proposed model obtains competitive performance against the human expert baseline.

5. Discussion

Reasoning ability. According to Figure 3, we see the non-keyword-driven model sometimes cannot generate a long

and correct conceptual description for retinal images. Also, Figure 3 shows the proposed keyword-driven model could have 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 3, our proposed keywords-driven model is capable of partially understanding input keywords. For example, in the fourth example of Figure 3, our proposed model generates “acute decrease” in the description based on the understanding of the keyword “degeneration”. Similarly, the first example of Figure 3 demonstrates some keyword understanding ability of our model. It implies that our proposed method brings us closer to the goal of automatic medical report generation for retinal images.

Are features from deeper models good in our task?

Based on our experimental result in Table 2 and [12, 18], we see 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.

6. Conclusion and Future Work

To sum up, we propose a new keyword-driven medical report generation method for automatic report generation for retinal images. The proposed method is equipped with a non-local attention-based mechanism, called TransFuser, which is capable of effectively fusing features with different modalities. Our experiments show that the proposed model can generate more accurate and meaningful descriptions/reports for retinal images, and the performance increases about 32% in BLEU-avg, 2.5% in CIDEr, and 25.4% in ROUGE. Our experiments also show that the proposed keyword-driven method, reinforced by the TransFuser, is superior to the non-keyword-driven one. Since TransFuser is a multi-modal feature fuser, applying it to other different combinations of modalities, such as image and speech, will be interesting future work.

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] Kalyan Acharjya, Girija Shankar Sahoo, and Sudhir Kr Sharma. An extensive review on various fundus databases use for development of computer-aided diabetic retinopathy screening tool. In *Soft Computing and Signal Processing*, pages 407–418. Springer, 2019.
- [2] Aishwarya Agrawal, Jiasen Lu, Stanislaw Antol, Margaret Mitchell, C Lawrence Zitnick, Devi Parikh, and Dhruv Batra. Vqa: Visual question answering. *International Journal of Computer Vision*, 123(1):4–31, 2017.
- [3] Marcella Cornia, Lorenzo Baraldi, Rita Cucchiara, Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. Show, control and tell: A framework for generating controllable and grounded captions. In *Proceedings of the IEEE Conference on CVPR*, 2019.
- [4] Etienne Decencière, Xiwei Zhang, Guy Cazuguel, Bruno Lay, Béatrice Cochener, Caroline Trone, Philippe Gain, Richard Ordonez, Pascale Massin, Ali Erginay, et al. Feedback on a publicly distributed image database: the messidor database. *Image Analysis & Stereology*, 33(3):231–234, 2014.
- [5] Adam Hoover and Michael Goldbaum. Locating the optic nerve in a retinal image using the fuzzy convergence of the blood vessels. *IEEE transactions on medical imaging*, 22(8):951–958, 2003.
- [6] 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.
- [7] Jia-Hong Huang. Robustness analysis of visual question answering models by basic questions. *King Abdullah University of Science and Technology, Master Thesis*, 2017.
- [8] Jia-Hong Huang, Modar Alfadly, and Bernard Ghanem. Vqabq: Visual question answering by basic questions. *VQA Challenge Workshop, CVPR*, 2017.
- [9] 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.
- [10] Jia-Hong Huang, Modar Alfadly, Bernard Ghanem, and Marcel Worring. Assessing the robustness of visual question answering. *arXiv preprint arXiv:1912.01452*, 2019.
- [11] 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.
- [12] Jia-Hong Huang, Cuong Duc Dao, Modar Alfadly, and Bernard Ghanem. A novel framework for robustness analysis of visual qa models. *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence*, 2019.
- [13] 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.
- [14] Jia-Hong Huang and Marcel Worring. Query-controllable video summarization. In *Proceedings of the 2020 International Conference on Multimedia Retrieval*, pages 242–250, 2020.
- [15] 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.
- [16] 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.
- [17] 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.
- [18] 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.
- [19] Baoyu Jing, Pengtao Xie, Eric Xing, Baoyu Jing, Pengtao Xie, and Eric Xing. On the automatic generation of medical imaging reports. *ACL*, 2018.
- [20] Ketan Joshi, Vikas Tripathi, Chitransh Bose, and Chaitanya Bhardwaj. Robust sports image classification using inceptionv3 and neural networks. *Procedia Computer Science*, 167:2374–2381, 2020.
- [21] Tae Joon Jun, Jihoon Kweon, Young-Hak Kim, and Daeyoung Kim. T-net: Nested encoder–decoder architecture for the main vessel segmentation in coronary angiography. *Neural Networks*, 2020.
- [22] 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*, 2015.
- [23] Tomi Kauppi, Valentina Kalesnykiene, Joni-Kristian Kamarainen, Lasse Lensu, Iris Sorri, A Raninen, R Voutilainen, J Pietilä, H Kälviäinen, and H Uusitalo. Diaretdb1—standard diabetic retinopathy database calibration level 1, 2007.
- [24] Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. Referitgame: Referring to objects in photographs of natural scenes. In *Proceedings of the 2014 conference on EMNLP*, pages 787–798, 2014.
- [25] Dong-Jin Kim, Jinsoo Choi, Tae-Hyun Oh, and In So Kweon. Dense relational captioning: Triple-stream networks for relationship-based captioning. In *CVPR*, 2019.
- [26] Jonathan Laserson, Christine Dan Lantsman, Michal Cohen-Sfady, Itamar Tamir, Eli Goz, Chen Brestel, Shir Bar, Maya Atar, and Eldad Elnekave. Textray: Mining clinical reports to gain a broad understanding of chest x-rays. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 553–561. Springer, 2018.
- [27] Christy Y Li, Xiaodan Liang, Zhiting Hu, and Eric P Xing. Knowledge-driven encode, retrieve, paraphrase for medical image report generation. *arXiv preprint arXiv:1903.10122*, 2019.

- [28] Yuan Li, Xiaodan Liang, Zhiting Hu, and Eric P Xing. Hybrid retrieval-generation reinforced agent for medical image report generation. In *Advances in Neural Information Processing Systems*, pages 1530–1540, 2018.
- [29] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. *Text Summarization Branches Out*, 2004.
- [30] 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.
- [31] Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L Yuille, and Kevin Murphy. Generation and comprehension of unambiguous object descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 11–20, 2016.
- [32] 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.
- [33] Marco Pedersoli, Thomas Lucas, Cordelia Schmid, and Jakob Verbeek. Areas of attention for image captioning. In *ICCV*, pages 1242–1250, 2017.
- [34] Prasanna Porwal, Samiksha Pachade, Ravi Kamble, Manesh Kokare, Girish Deshmukh, Vivek Sahasrabudhe, and Fabrice Meriaudeau. Indian diabetic retinopathy image dataset (idrid): A database for diabetic retinopathy screening research. *Data*, 3(3):25, 2018.
- [35] Riccardo Di Sipio, Jia-Hong Huang, Samuel Yen-Chi Chen, Stefano Mangini, and Marcel Worring. The dawn of quantum natural language processing, 2021.
- [36] David M Tierney, Joshua S Huelster, Josh D Overgaard, Michael B Plunkett, Lori L Boland, Catherine A St Hill, Vincent K Agboto, Claire S Smith, Bryce F Mikel, Brynn E Weise, et al. Comparative performance of pulmonary ultrasound, chest radiograph, and ct among patients with acute respiratory failure. *Critical Care Medicine*, 48(2):151–157, 2020.
- [37] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NIPS*, pages 5998–6008, 2017.
- [38] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on CVPR*, 2015.
- [39] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on CVPR*, 2015.
- [40] Cheng Wang, Haojin Yang, Christian Bartz, and Christoph Meinel. Image captioning with deep bidirectional lstms. In *MM*, pages 988–997. ACM, 2016.
- [41] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, and Ronald M Summers. Tienet: Text-image embedding network for common thorax disease classification and reporting in chest x-rays. In *CVPR*, pages 9049–9058, 2018.
- [42] Xinyu Xiao, Lingfeng Wang, Kun Ding, Shiming Xiang, and Chunhong Pan. Deep hierarchical encoder–decoder network for image captioning. *IEEE Transactions on Multimedia*, 21(11):2942–2956, 2019.
- [43] 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. *arXiv preprint arXiv:1502.03044*, 2015.
- [44] Chao-Han Huck Yang, Jia-Hong Huang, Fangyu Liu, Fang-Yi Chiu, Mengya Gao, Weifeng Lyu, I-Hung Lin, and Jesper Tegner. A novel hybrid machine learning model for auto-classification of retinal diseases. *Workshop on Computational Biology, ICML*, 2018.
- [45] C-H Huck Yang, Fangyu Liu, Jia-Hong Huang, Meng Tian, MD 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.
- [46] Chao-Han Huck Yang, Yun-Yun Tsai, and Pin-Yu Chen. Voice2series: Reprogramming acoustic models for time series classification. *International Conference on Machine Learning (ICML)*, 2021.