

Automatic Generation of Medical Imaging Diagnostic Report with Hierarchical Recurrent Neural Network

Changchang Yin*, Buyue Qian†, Jishang Wei‡, Xiaoyu Li*, Xianli Zhang*, Yang Li*, Qinghua Zheng†

*School of Computer Science and Technology, Xi'an Jiaotong University, Xi'an, Shaanxi, China

Email: {lenty, xiaoyuli, xlbryant, vigilee}@stu.xjtu.edu.cn

†National Engineering Lab for Big Data Analytics, Xian Jiaotong University, Xi'an, Shaanxi, China.

Email: {qianbuyue, qhzheng}@xjtu.edu.cn

‡HP Labs, 1501 Page Mill Rd, Palo Alto, CA 94304, USA

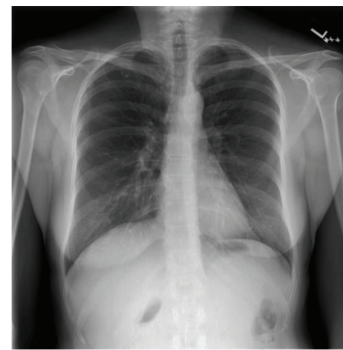
Email: weijishang@gmail.com

Abstract—Medical images are widely used in the medical domain for the diagnosis and treatment of diseases. Reading a medical image and summarizing its insights is a routine, yet nonetheless time-consuming task, which often represents a bottleneck in the clinical diagnosis process. Automatic report generation can relieve the issues. However, generating medical reports presents two major challenges: (i) it is hard to accurately detect all the abnormalities simultaneously, especially the rare diseases; (ii) a medical image report consists of many paragraphs and sentences, which are longer than natural image captions. We present a new framework to accurately detect the abnormalities and automatically generate medical reports. The report generation model is based on hierarchical recurrent neural network (HRNN). We introduce a topic matching mechanism to HRNN, so as to make generated reports more accurate and diverse. The soft attention mechanism is also introduced to HRNN model. Experimental results on two image-paragraph pair datasets show that our framework outperforms all the state-of-art methods.

I. INTRODUCTION

Medical images are widely used in the medical domain for the diagnosis and treatment of diseases. The medical images are usually read by highly trained experts. They write textual reports (e.g., Figure 1) to narrate the findings of the abnormalities and diseases of the patients. It costs a lot of time of the experts. For example, it takes 5 to 20 minutes for a professional radiologist to read a CT (Computed Tomography) image and type the findings. Many radiologists read hundreds of radiology images and write the findings for each image per day. The reading of the medical images and the writing of medical reports almost occupy most of the radiologists' work time. Automatic generating medical image reports will improve the experts' efficiency and save their work time.

Automatic report generation can be considered as image captioning. There has been great progress in natural image captioning with deep learning [1]–[3]. However, in the medical domain, most existing methods [4]–[6] are based on some traditional image captioning models and do not perform very well. Only several studies [7]–[9] try to leverage state-of-art deep learning based models to generate high-quality medical reports. There are two main challenges in the task of medical



Indication: No acute cardiopulmonary abnormality.

Findings: Lungs are clear without focal infiltrates. Calcified right upper lobe granuloma unchanged from prior. No pneumothorax or pleural effusion. Normal heart size. Normal pulmonary vascularity. Bony thorax intact.

Impression: No acute cardiopulmonary abnormality.

Tags: Calcified Granuloma

Fig. 1. Example of an image and the corresponding information provided in the training set of IU X-Ray dataset.

image report generation. The first is that it is hard to accurately detect patients' abnormalities. The abnormal cases are usually much rarer than healthy controls. The second challenge is that the medical reports are very long. Many report paragraphs consist of more than 60 words. Thus it is difficult to produce proper and detailed medical reports.

In order to overcome the two challenges mentioned above, we propose a new framework, consisting of two steps, to generate medical imaging reports. The first step is to detect abnormalities from a medical image, which can be addressed as a multi-label classification problem. A deep convolutional neural network (CNN) is used to detect the biomedical abnormalities. It is universal that a patient is associated with several abnormalities at the same time. Different abnormalities often are indicated in different parts of the same image. We propose a **global label pooling (GLP)** mechanism, which directly predicts abnormalities heat maps from the feature maps, as shown in Figure 2. A max pooling operation is followed to generate the final abnormality detection result. This method improves abnormality detection accuracy.

The second step is to generate a long annotation for each image. We adopt a hierarchical recurrent neural network

(HRNN) to generate long medical annotations. The HRNN is composed of two recurrent neural networks (RNN): a sentence RNN and a word RNN. The sentence RNN takes detected abnormalities and the features maps extracted in CNN as inputs, and then generates several topic vectors. Given a topic vector, the word RNN produces an appropriate sentence. Most sentences are only related to a single disease or a part of the location of the medical image, so intuitively we introduce topic attention mechanism to sentence RNN. Each time the RNN generates a topic vector, it attends to feature maps and detected abnormalities simultaneously. Besides, we observe that there are many repeated sentences in the generated reports. The reason may be that the sentence RNN always generates similar topic vectors. We propose a topic matching mechanism to project the ground truth sentences and topic vectors into the same semantic space, and then make the corresponding topic-sentence pair closer in the semantic space. When paired topic and sentence share the same semantics, the sentence RNN will produce more diverse topic vectors for the different sentences.

To validate the effectiveness of our model, several experiments are conducted on two public datasets. The experiment results show that our model outperforms all the state-of-art methods.

The contributions of the paper can be summarized as:

- 1) We present a novel framework which can simultaneously detect medical abnormalities and produce long captions.
- 2) We propose to use global label pooling to replace global feature pooling in multi-label classification CNN, which improves the accuracy and robustness of the CNN.
- 3) We introduce the topic-attention mechanism to the sentence RNN, which attends to the several regions and tags related to the object sentence.
- 4) We propose the topic matching mechanism, which enhances the diversity and accuracy of the topic vectors.

The rest of the paper is organized as follows. We review the related studies in Section II. In Section III, we describe our model in detail. In Section IV, we conduct experiments on two real-world EHR datasets. Section V concludes our work.

II. RELATED WORK

In this section, we review the works related to image captioning and medical report generation.

A. Image Captioning

Image captioning, targeting at bridging the visual and linguistic domain, aims to automatically generate descriptive sentences for given images. The problem is posed as a ranking task in early works [10], [11]. For a given image, the methods rank the existing captions and retrieve the relevant captions. Therefore, the approaches cannot describe unseen compositions for new images. Another solution is to treat the problem as a template retrieval task [11], [12], but the generated captions are heavily hand-designed and rigid.

In recent years, deep learning based models achieve excellent performances in many different tasks, including machine translation [13]–[15], image recognition [16]–[18] and so on.

Inspired by the successes of sequence generation in machine translation, Vinyals et al. [1] present a generative model *Show and Tell* based on a deep recurrent architecture that combines recent advances in computer vision and machine translation and that can be used to generate natural sentences describing an image. It is the first work that leverages deep learning method to generate image captions, and significantly improve the captioning performance. The model is based on the encoder-decoder framework, which uses a CNN as encoder network and a RNN as a decoder network. Many variants [2], [19], [20] based on the encoder-decoder framework are proposed and perform very well. After the attention mechanism is introduced to the image captioning models, it has shown to be useful to generate descriptions for images. After the attention mechanism [2], [21] is introduced to the image captioning models, it has shown to be useful to generate descriptions for images. Xu and et al. [2] present a soft deterministic attention and a hard stochastic attention mechanism to automatically attend a small region rather than a whole image, which significantly improves the performance of image captioning. You and et al. [21] propose to selectively attend to semantic proposals. However, these models are typically trained to predict the next word given the previous words. If the previous words are wrong, the errors will accumulate along the way. To make the models more robust, reinforcement learning (RL) are introduced to the image captioning task [3], [22], [23] propose to introduce . The models based on RL can generate more human-like and fluent captions.

B. Paragraph generation

The traditional image captioning models can properly describe an image with only a single sentence. However, one sentence's description is high-level and coarse. By only describing images with a single high-level sentence, there is a fundamental upper-bound on the quantity and quality of information approaches can produce [24]. Thus, paragraph generation task is introduced, which aims to describe images in detail with a paragraph rather than a sentence. Karpathy and Fei-fei [25] propose a new model *Image-Flat* generate dense descriptions of images. *Image-Flat* runs an image captioning model on regions and aligns the generated captions to various regions. Based on *Image-Flat*, Johnson and et al. [26] propose *DenseCap* to joint train the object detection model *Faster R-CNN* [27] and image captioning model in an end-to-end way. However, the two models suffer the same limitation: descriptions generated for dense captioning are not coherent. They do not produce a cohesive whole paragraph to describe the entire image. A hierarchical recurrent network (HRNN) [24] is proposed to generate a paragraph. HRNN is also based on the encoder-decoder framework, where CNN as the encoder to extract the image information, and then given the extracted feature, a two-level LSTM model is adopted to generate paragraph captions. The two-level LSTM consists of a sentence RNN and a word RNN. The sentence RNN is used to generate topic vectors. Each topic vector is used to produce a sentence to describe the image. Based on the hierarchical

recurrent network, Liang and et al. [28] propose to introduce a generative adversarial network, which can help make the generated paragraph more fluent and human-like.

C. Medical report generation.

Automatically medical report generation is a significant and difficult task, which need to interpret and summarize the insights gained from medical images such as radiography or biopsy samples [4]. It needs accurate abnormality detection and state-of-art detailed image caption generation. It can be treated as the image captioning models' application to the medical domain. There are some existing studies to automatically generate medical reports or annotations for medical images [4], [5], [7]–[9]. Most existing studies are based on two public medical image-caption pair datasets. The first is ImageCLEFcaption [4], [5]¹. ImgeCLEFcaption is an evaluation campaign about medical concept detection and report generation. The organization provides a large scale dataset of 184,000 image-caption pairs. However, due to the low quality of the dataset (there are more than 110,000 kinds of medical concept and less than 100 images for per concept), the submitted models of the campaign perform not very well. The model [6] with the highest BLEU [29] score just matches the test image with training set images, and return the caption of the most similar image, but not generates a new caption.

The second dataset is the public IU X-Ray dataset [7], which contains 7,470 pairs of radiology images and reports. There are several other studies which adopt paragraph generation models to generate medical reports for X-Ray images. Shin and et al. [7] firstly present a deep learning model to efficiently detect abnormalities from an image and annotate its contexts. Jing and et al. [8] propose to leverage CNN to detect tags, and then combine the hierarchical recurrent network and attention mechanism to generate detailed medical reports. However, the models usually generate some similar but useless sentences. The main reason is that the sentence RNN generates similar topic vectors. Therefore, we propose a topic matching mechanism to diversify the topics and the generated results, which improve the model's performance.

III. METHOD

In this part, we discuss the new framework to automatically generate medical reports from a medical image. Firstly, a deep convolutional neural network is used to detect the biomedical abnormalities. Then, we propose a new image captioning model with attention and topic matching mechanism to produce a detailed description of a given image.

A. Abnormality Detection with Global Label Pooling

The abnormality detection is treated as a multi-label classification task. Given an image x , the model predicts the probabilities of all the abnormalities. For every single tag, we optimize a weighted binary cross entropy loss. We compute

the postive sample's weight w_p and negative sample's weight w_n as follow:

$$w_p = \frac{P + N}{P} \quad (1)$$

$$w_n = \frac{P + N}{N} \quad (2)$$

, where P denotes the number of postive samples for the given tag and N is the number of negative samples. We compute the weights (i.e., 572 times' computation for 572 tags in IU X-Ray dataset) for each tag.

Our abnormality detection model is based on DenseNet [18], which uses a global pooling operation before the last fully connected layer. The operation is not suitable for an abnormality detection task. When a patient is associated with several abnormalities simultaneously, the global pooling operation may do harm to feature extraction for two reasons: (i) abnormality detection usually relies on a local part image's feature rather than the global feature; (ii) because different diseases' features may be distributed on various regions, the global pooling operation could mix several diseases' features and lose some spacial information.

Therefore, we tune the DenseNet model with a global label pooling operation (**GLP**). Firstly, we use the same convolutional layers as DenseNet, except the global pooling layer and the last fully connected layer. The last output channel of DenseNet is 1024. Therefore, we obtain a feature map with the shape $[W, H, 1024]$ for each input image, where W and H denote the width and height of the feature map, as shown in Figure 2. Then, we leverage a convolutional layer to directly predict the abnormalities based on regions rather than the whole image feature. The convolutional layer's parameter shape is $[1, 1, 1024, 572]$. A label map, with the shape as $[W, H, 572]$, is obtained, which indicates the heat maps for all the abnormalities. Finally, a global max pooling layer is followed to generate the multi-label classification's output vector. Because we only know the labels of abnormalities, but not the specific location of each abnormality, we have to use the output vector to compute the cross entropy loss and then train the whole detection model. We call this adjustment as **global label pooling (GLP)**, while the global pooling on feature maps is called as **global feature pooling (GFP)**. GFP, same as traditional CNN, adopt a global pooling on feature maps and then leverage fully connected layer to predict the abnormality tags. Besides, it is also easier for GLP to find the abnormal regions from the heat maps.

B. Caption Prediction

After CNN detects the possible abnormalities for each patient, we present a new image captioning model to generate the paragraph to describe a medical image. Our captioning model is based on HRNN [24]. The HRNN structure is composed of two levels: a sentence RNN and a word RNN. The sentence RNN receives the image features and decides how many sentences to generate in the resulting paragraph, and produces an input topic vector for each sentence. Given this topic vector, the word RNN infer the words of a single

¹<https://www.imageclef.org/2019/caption>

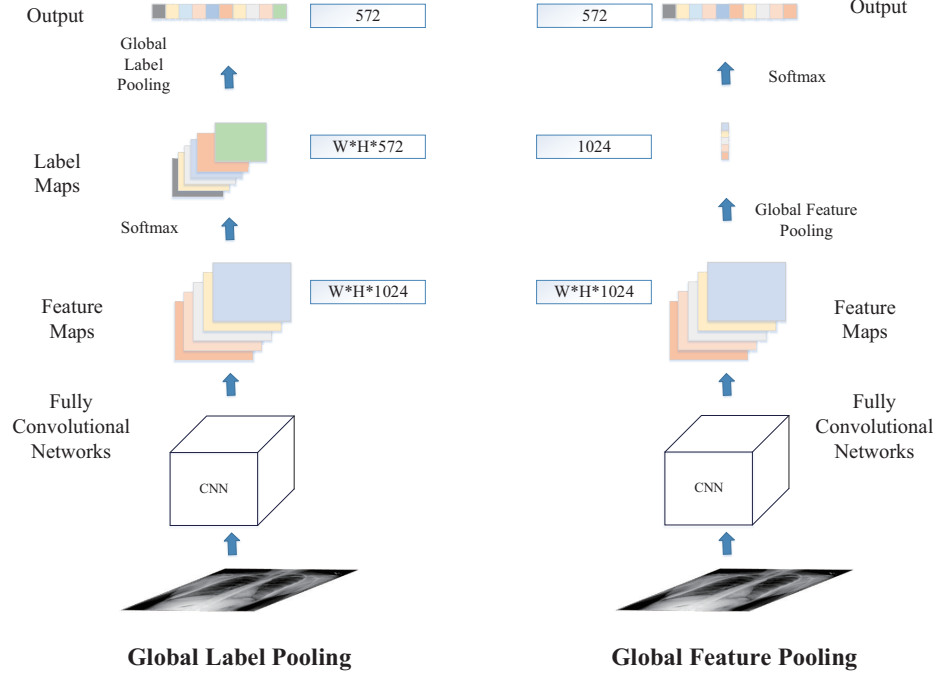


Fig. 2. Illustrations of global feature pooling and global label pooling. The main difference is that global label pooling predicts a heat map for the input image, and then conducts a max pooling operation over the heat maps.

sentence. The outputs of the word RNN are concatenated to the final paragraph. Figure 3 provides an overview of the framework.

Sentence RNN. We use a single-layer Long Short-Term Memory networks (LSTM) to generate topic vectors. The initial hidden and cell states are set to zero. The sentence RNN takes as input feature maps and predicted tags, both of which are extracted by CNN classifier from initial medical images. Most sentences in medical reports usually describe a part of the location of the image or a kind of abnormality of the patient. Intuitively, an attention mechanism can be introduced to the sentence RNN. For each location i and tag j , the mechanism generates the corresponding positive weights β_i and α_j . The weights can be interpreted as the probabilities that the corresponding location or tag are focused to produce the next topic vector. The weights are computed as follow:

$$\beta_i = \frac{\exp(\gamma_i)}{\sum_i^{W \times H} \exp(\gamma_i)} \quad (3)$$

$$\gamma_i = W_{va} \tanh(W_v f_i + W_{v,h} h^s)$$

$$\alpha_j = \frac{\exp(\theta_j)}{\sum_j^{|t|} \exp(\theta_j)} \quad (4)$$

$$\theta_j = W_{ta} \tanh(W_t t_j + W_{t,h} h^s)$$

,where $f \in R^{W \times H \times k}$ denotes extracted feature maps, and $f_i \in R^k$ is the i^{th} region's feature vector, $t_j \in R^d$ denotes the embedding of the j^{th} predicted tag, $h^s \in R^d$ is the hidden state of the sentence RNN, $W_{v,h}$, W_t , $W_{t,h} \in R^{d \times d}$, $W_v \in R^{d \times k}$, W_{ta} , $W_{va} \in R^d$ are learned parameters initialized randomly.

We adopt a soft attention on the feature map f and the predicted tags t . At time step s , the next visual attention vector z_{vatt} , tag attention vector z_{tatt} and the context attention vector z are computed by

$$z_{vatt}^s = \sum_{i=1}^{W \times H} \beta_i * f_i \quad (5)$$

$$z_{tatt}^s = \sum_{j=1}^{|t|} \alpha_j * t_j \quad (6)$$

$$z^s = W_{vatt} z_{vatt}^s + W_{tatt} z_{tatt}^s \quad (7)$$

We use an output layer to compute the next topic vector v^s .

$$v^s = W_h h^s + W_{ctx} z^s \quad (8)$$

$W_{vatt} \in R^{d \times k}$, $W_{tatt} \in R^{d \times d}$ in Eq. 7 and $W_h \in R^{d \times d}$, $W_{ctx} \in R^{d \times d}$ in Eq. 8 are learned parameters initialized randomly.

At the same time, the sentence RNN adopts a linear projection from h^s to produce a stop control probability p^s over the two states $\{ \text{CONTINUE}=0, \text{STOP}=1 \}$. The stop control probability determines whether to produce the next topic vector. The topic vector represents the semantics of a sentence to be generated by the word RNN.

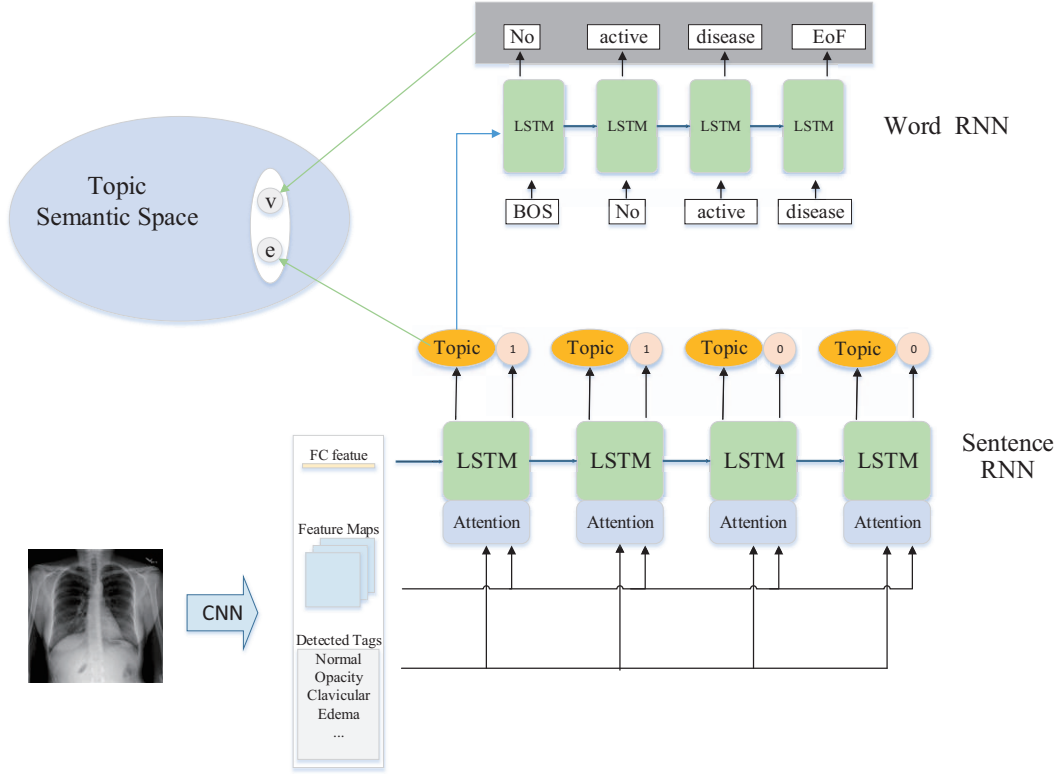


Fig. 3. Framework of the hierarchical recurrent neural network. We use a multi-label classification CNN to detect the abnormal tags of the patients. The extracted feature maps and the embeddings of the detected tags are attended by the Sentence RNN. Every topic generated by Sentence RNN and its corresponding ground truth sentence are mapped into the topic semantic space, so as to make them share the same semantics.

In order to avoid that the model always focuses several regions and tags, we add a regularization loss over the weight parameters:

$$l_{attention} = \lambda_{tag} \sum_{j=1}^{|t|} (1 - \alpha_j)^2 \quad (9)$$

$$+ \lambda_{cxt} \sum_{i=1}^{W \times H} (1 - \beta_i)^2$$

where λ_{cxt} and λ_{tag} are adjustable hyper-parameters.

Word RNN. The word RNN takes topic vectors as input and aims to generate meaningful and detail sentences based on the topic vectors. At the first step, the word RNN is fed with the topic vector and a special START token, and then predicts a probability distribution over the words in the vocabulary and a special END token. The generated words are sampled from the probability distribution. Next, the last sampled word (or the last ground truth word in the training phase) is fed into the word RNN to produce the next words. The process is repeated until the END token is sampled.

The model is trained with a weighted sum of two cross-entropy loss: a sentence loss l_{sent} on the stopping distribution over {STOP, CONTINUE} and a word loss l_{word} on the word distribution and the END token:

$$l_{hrnn}(x, y) = \lambda_{sent} \sum_{i=1}^S l_{sent}(p_i, I[i = S]) \quad (10)$$

$$+ \lambda_{word} \sum_{i=1}^S \sum_{j=1}^{O_i} l_{word}(p_{ij}, y_{ij})$$

, where p_i denotes the predicted result whether i^{th} sentence is the last sentence, p_{ij} and y_{ij} respectively denote the predicted word distribution and the ground truth of the j^{th} word of the i^{th} sentence, λ_{sent} and λ_{word} are adjustable hyper-parameters, O_i denotes the length of i^{th} sentence, S denotes the max number of generated sentences of a report.

C. Topic Matching

When the images are fed into the model, the parameters of sentence RNN are trained with the gradients that are propagated from the loglikelihood loss of the word RNN. Relative to word RNN's parameters, which are directly trained with the loglikelihood loss, it is inefficient to train the sentence RNN. It is possible that the sentence RNN is not fully trained, and produces similar topic vectors for different sentences. Based on these topic vectors, the word RNN generates repeated sentences in one report and similar medical reports for most patients in the validation set. Thus we propose a topic

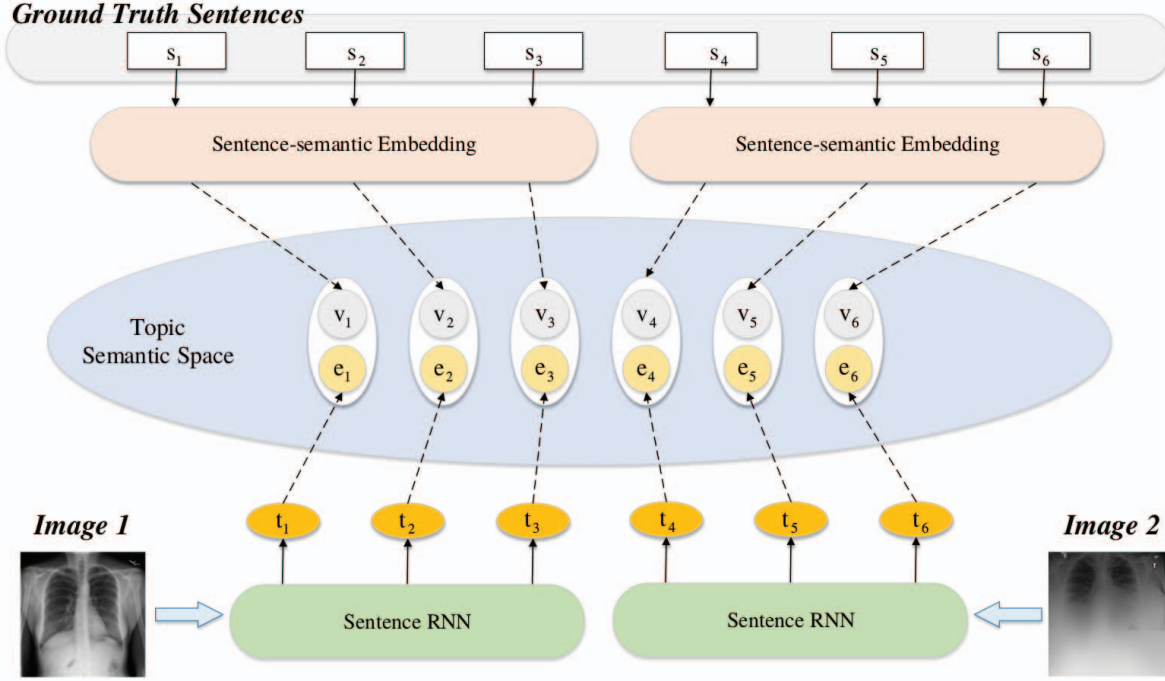


Fig. 4. Topic matching mechanism. There are two images and the corresponding ground truth captions. The left image's ground truth captions are s_1 , s_2 and s_3 . The sentence RNN generates three topics for the left image, t_1 , t_2 and t_3 . Then, the sentences and the topics are mapped to semantic vectors (e.g., v_2 and e_2) in the same topic semantic space. Finally, we train the model to make paired vectors more similar with Eq. (11).

matching mechanism to train the sentence RNN better and to increase the diversity of the topic vectors.

The topic matching mechanism maps the topic vectors and the sentences into a same semantic space. Since one topic vector is paired with one sentence, which narrates the corresponding information of a medical image, we assume they share the same semantics. Both the topic vectors and the sentences are fed into two different mapping networks to learn the semantic space. The topic mapping network is composed of two fully connected layers followed by ReLU layer. The sentence mapping network embeds each word into a vector. The vector is projected into the semantic space with a fully connected layer. A sentence is represented by a 2-D matrix. We use a 1-D global max-pooling layer along the words dimension. After the topic vector and its corresponding sentence are jointly embedded into a common space, the topic vector is mapped to an embedding vector e and its corresponding sentence is mapped to another embedding vector v . We can compute the cosine similarity between e and v in the semantic space. The model is trained by minimizing a contrastive loss:

$$l_{topic}(e, v) = \lambda_{topic} \left(\sum_{v' \in V} \max(0, \gamma - ev + ev') \right) \quad (11)$$

$$+ \sum_{e' \in E} \max(0, \gamma - ev + e'v)$$

where e (or v) is the embedding vector of the topic vector generated by sentence RNN (or the corresponding sentence of

the ground truth), e' and v' are the embedding of the negative paired topic (or sentence) samples, λ_{topic} is an adjustable hyper-parameter. In our research, 127 negative samples are randomly chosen from training set for each positive sample. γ denotes the contrastive margin.

Figure 4 gives a toy example to illustrate the mechanism, there are two images with the corresponding ground truth sentences. The sentences s_1 , s_2 and s_3 belong to the left image, while the sentences s_4 , s_5 and s_6 belong to the right image. The sentence RNN generates the six topic vectors (i.e., t_1 to t_6) for the two images. Then we map the sentences and topic vectors into the same semantic space, and obtain the corresponding semantic vectors (i.e., v_1 to v_6 and e_1 to e_6). Then we compute the contrastive loss l_{topic} for every pair (e , v) with the Eq. (11).

We assume that the paired topic and sentence share the same semantics in the learned space and similar sentences are projected into similar vectors. Therefore, the topic vectors, which share the same semantics, will become closer and more accurate to describe the semantic information of the sentences. It will be easy for the word RNN to learn to generate better sentences from diverse and accurate topic vectors.

Therefore, the objective function is defined as the sum of three loss functions: l_{hrnn} , $l_{attention}$ and l_{topic} .

IV. EXPERIMENT

In this section, we evaluate our framework on two public datasets: a medical image-report pair dataset IU X-Ray [30]

and a natural image-paragraph pair dataset GENOME [24].

A. Medical Report Generation

The Indiana University Chest X-Ray Collection (IU X-Ray) [30] is a publicly available radiology dataset of chest x-rays and reports. The dataset contains 7,470 pairs of radiology images and reports. Each report is structured as *impression*, *findings*, *tags*, *comparison* and *indication*. We treat the tags as the multi-label classification labels. The concatenation of impression and findings are regarded as the ground truth report. After preprocessing the data by tokenizing, converting to lower-cases, and filtering tokens of frequency no less than 3, there are 572 unique tags remained. On average, each image is associated with 2.2 tags, 5.7 sentences and each with an average of 6.5 words.

Implementation details. We implement our model on tensorflow and train on a GeForce GTX TITAN GPU. The abnormality detection CNN is trained first. Following the previous studies [8], [9], we divide the data into 6,470 training, 500 validation and 500 testing images. Although there is a great gap between the medical data and natural image data, a superior initialization obtained from a pre-trained model can help train the abnormality detection model. Our abnormality detection CNN takes the weights from the pre-trained DenseNet [18] and continue to train with IU X-Ray dataset [30]. For each image, 20 top tags with the highest probabilities and a feature map shaped as [7, 7, 1024] are used to generate medical captions. In the training process of caption prediction model, we set $k = 1024$, $d = 512$, $\lambda_{sent} = 5.0$, $\lambda_{word} = 1.0$, $\lambda_{topic} = 10.0$, $\lambda_{tag} = 5.0$, $\lambda_{cxt} = 5.0$, $\gamma = 0.1$. The settings of λ_{sent} and λ_{word} come from [24]. We adopt a grid search on other hyper-parameters to get the best γ setting and observe λ_{topic} , λ_{tag} , λ_{cxt} are not sensitive in relatively large ranges. Our model produces 6 sentences at most, and the length of each sentence is less than 30.

Abnormality Detection. We use a DenseNet variant with global label pooling (GLP) to predict the tags for each image. To evaluate the performance of GLP, we compare it with the initial DenseNet, which use global feature pooling (GFP). For each patient's image, we select the n tags with the highest probability as predicted tags, as shown @ n (e.g., @5 and @10) in I. The table displays the recalls and precisions of GLP and GFP. The results show that our GLP model performs better than conventional GFP.

We speculate that the outperformance of GLP relies on three reasons: (i) without any global pooling before the last classification layer, GLP can retain more spacial information; (ii) most of the abnormalities are indicated in some local area of an image, GLP detects abnormalities based on local features rather than a global feature; (iii) the features of different diseases may be influenced by each other after the global pooling in GFP.

Caption Prediction. In the training process of the captioning model, we divide the data into training, validation and testing set, the same as the abnormality detection CNN. We conduct our experiments 20 times and present the average results in

TABLE I
TAG PREDICATION RECALL AND PRECISION ON IU X-RAY DATASET.

	@5		@10		@20	
	Recall	Precision	Recall	Precision	Recall	Precision
GFP	0.483	0.203	0.591	0.131	0.697	0.077
GLP	0.552	0.245	0.646	0.145	0.741	0.083

Table II. The annotation generation performance is assessed on the basis of BLEU- $\{1,2,3,4\}$ [29], CIDEr [31], ROUGE-L [32], METEOR [33]. We compare our method with several state-of-the-art image captioning models: CNN-RNN [1], LRCN [34], Soft ATT [2], ATT-RK [21] and HRGR-Agent [9]. The baseline results are obtained from [9] and [8].

In order to better show the effectiveness of our method, we implement our model with 4 versions: Ours-no-Attention (basic Hierarchical RNN), Ours-Attention-only (HRNN with attention mechanism), Ours-Topic-only (HRNN with topic matching mechanism), Ours (HRNN with attention and topic matching mechanism). The results show that both the attention mechanism and the topic matching mechanism greatly improve the medical annotation generation performance. Compared with the baselines, our model performs the best in most evaluation criteria except CIDEr, which demonstrates our model's effectiveness. HRGR-Agent [9] outperforms ours on CIDEr. The reason may be that HRGR-Agent introduces RL directly guided by CIDEr reward. Our model performs better than other methods without RL on CIDEr metric. In the computation of CIDEr, the unusual words have bigger weights than other words, for example the description of a specific disease. However, our model, which is trained with the cross-entropy loss, ignores the situation. In this respect, RL can greatly improve the generation performance. In the future, we will also introduce reinforcement learning to our model.

Examples of the generated reports on the input images are shown in Figure 6. We use underlining to emphasize the descriptions of abnormalities. We can find that the descriptions generated by our model are more qualitative and diverse, and have more detailed information about abnormal findings. Our model's generated paragraphs are more similar to the reports written by human experts.

In order to qualitatively demonstrate the effectiveness of the proposed topic matching mechanism, we visualize the generated topic vectors in test set, as shown in Figure 5. The topics are selected as follow. Firstly, we obtain each patient image's abnormality tags (e.g., pleural effusion). Then, we compare the generated sentences with the tags. If there is one sentence which describes a given tag, it means the topic vector used to generate the sentence is related to the tag, and we select the topic vector. All the topic vectors describing the same tag have the same color. We can observe that in some regions of Figure 5 (a), the topics that describe the same tag (e.g., pulmonary disease, nodule) are clustered fairly close together, which indicates that sentence RNN learn to represent sentence topics more accurately, compared with Figure 5 (b)

TABLE II
AUTOMATIC EVALUATION RESULTS ON IU X-RAY DATASET.

Model	CIDEr	METEOR	ROUGE-L	BLEU-1	BLEU-2	BLEU-3	BLEU-4
CNN-RNN	0.110	0.159	0.267	0.316	0.211	0.140	0.095
LRCN	0.190	0.155	0.278	0.369	0.229	0.149	0.099
Soft ATT	0.302	0.167	0.323	0.399	0.251	0.168	0.118
ATT-RK	0.155	0.171	0.323	0.369	0.226	0.151	0.108
HRGR-Agent	0.381	-	0.341	0.436	0.278	0.197	0.150
Ours-no-Attention	0.279	0.162	0.320	0.403	0.252	0.173	0.126
Ours-Attention-only	0.337	0.172	0.328	0.428	0.263	0.183	0.137
Ours-Topic-only	0.330	0.169	0.329	0.417	0.278	0.188	0.147
Ours	0.342	0.175	0.344	0.445	0.292	0.201	0.154

TABLE III
AUTOMATIC EVALUATION RESULTS ON VISUAL GENOME DATASET.

Model	METEOR	CIDEr	BLEU-1	BLEU-2	BLEU-3	BLEU-4
Sentence-Concat	0.121	0.068	0.0311	0.151	0.076	0.040
Image-Flat	0.128	0.111	0.340	0.200	0.122	0.077
Regions-Hierarchical	0.159	0.135	0.419	0.241	0.142	0.087
RTT-GAN	0.184	0.204	0.421	0.254	0.149	0.092
Ours-no-Attention	0.156	0.134	0.410	0.238	0.140	0.084
Ours-Attention-only	0.159	0.144	0.420	0.243	0.144	0.088
Ours-Topic-only	0.160	0.140	0.418	0.241	0.142	0.087
Ours	0.164	0.146	0.430	0.256	0.153	0.094
Human	0.192	0.285	0.429	0.257	0.156	0.097

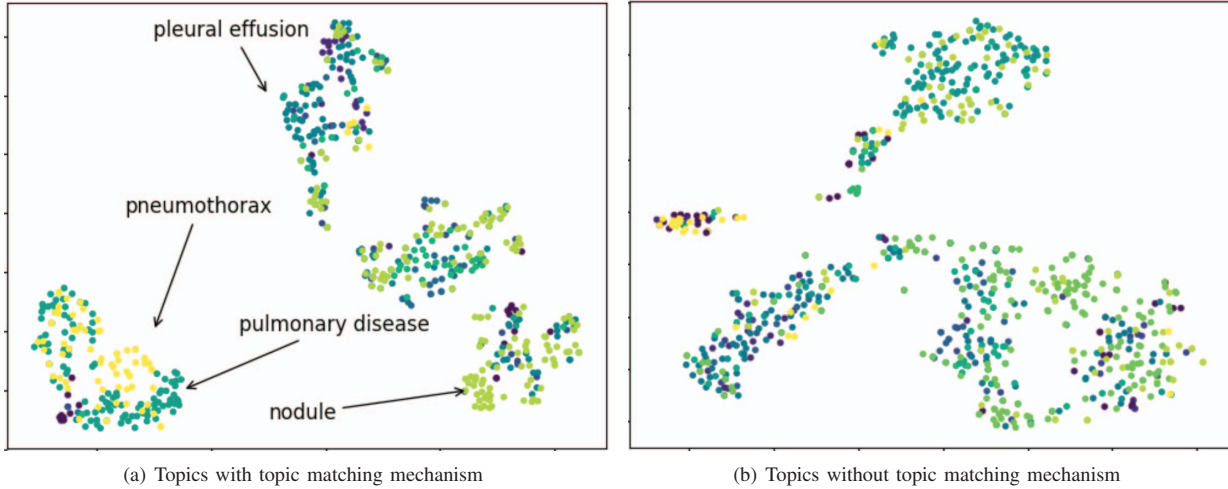


Fig. 5. t-SNE Scatterplots of topic vectors generated by sentence RNN.

where there is no obvious cluster of same colored points.

B. Natural Image Captioning

In order to demonstrate our model's general applicability, we conduct other experiments on a natural image-paragraph pair dataset VISUAL GENOME [24]. The dataset is comprised of 19,551 images and the corresponding paragraph descriptions. For fair comparison with the state-of-art methods [24] and [28], we split all the images-paragraphs pairs into three parts: 14,575 for training, 2,487 for validation and 2,489 for testing. We train the model on the training set, and the best

model tested on the validation set is used to measure the performance on the test set. The experiments are repeated 20 times and the average performances are shown in Table III. Same with [28] and [24], we use the dense captioning model [26] to extract semantic features. The features of the top 50 ROI regions are used in the paragraph generation networks.

We compare our method with Sentence-Concat, Image-Flat [35], Regions-Hierarchical [24], RTT-GAN [28]. We obtain the results of the baselines from [24] and [28]. As shown in Table III, our model outperforms all the baselines on the BLEU metrics, which demonstrates that our model can be generalized

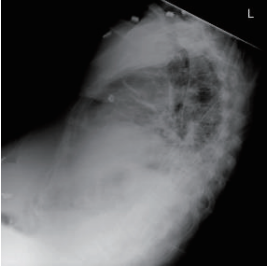
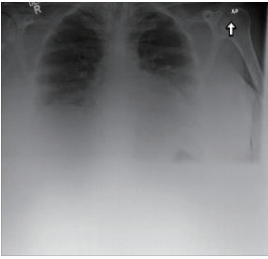
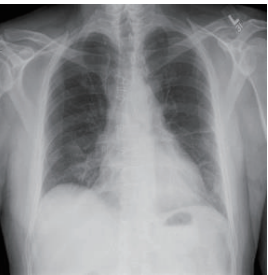
	Ground Truth	Hierarchical RNN	Ours
	Bibasilar airspace disease greater on the right atelectasis or right midlung atelectasis. There is <u>mild cardiomegaly</u> . <u>The thoracic aorta is tortuous</u> . Lung volumes are low with asymmetric elevation of the right hemidiaphragm. There is <u>platelike atelectasis</u> in the right midlung along with bibasilar airspace disease atelectasis or infiltrate. No pneumothorax.	No acute cardiopulmonary abnormality. There is a moderate rightsided pneumothorax with large pleural effusion. There is a moderate rightsided pneumothorax with large pleural effusion. There is no pneumothorax. There is no pneumothorax. There is no pneumothorax.	<u>Cardiomegaly</u> and pulmonary vascular congestion and interstitial edema of the cardiac silhouette of the thoracic spine of the. <u>The thoracic spine is tortuous</u> . There is a moderate moderate hiatal hernia to the right upper lobe. There is a moderate sized right pleural effusion of the right costophrenic.
	Chest tube tip projects outside the thoracic cavity. <u>No residual residual pneumoperitoneum</u> consistent with known colonic. Chest tube tip now projects outside the thoracic cavity. No definite residual pneumothorax. <u>Stable cardiomeastinal silhouette</u> . There are low lung volumes. <u>No large pleural effusion</u> . No focal airspace consolidation. Small amount of subdiaphragmatic free air.	Cardiomegaly with pulmonary vascular congestion and interstitial edema. There is a moderate rightsided pneumothorax with tip in the right atrium. There is a moderate rightsided pneumothorax with large pleural effusion. No pneumothorax masses. <u>No pneumothorax masses</u> .	Cardiomegaly and pulmonary vascular congestion and interstitial edema of the <u>cardiomeastinal silhouette</u> of the left upper of the pleural. There is a moderate sized right pleural effusion of the right costophrenic. <u>No pneumothorax or pleural effusion of the costophrenic</u> .
	Scattered bilateral subsegmental atelectasis. Decreased from prior mild cardiomegaly. There are <u>postoperative changes of sternotomy and cabg</u> . There is stable mild cardiomegaly. There are scattered of subsegmental atelectasis decreased from the prior chest radiograph. <u>No focal airspace consolidation</u> . <u>No pleural effusion or pneumothorax</u> . There are minimal <u>degenerative changes of the spine</u> .	No acute cardiopulmonary abnormality. There is a moderate rightsided pneumothorax with large pleural effusion. There is a moderate rightsided pneumothorax with large pleural effusion. There is <u>no pneumothorax</u> . There is <u>no pneumothorax</u> . There is <u>no pneumothorax</u> .	No acute cardiopulmonary abnormality. The heart is normal in size. The lungs are clear. <u>Degenerative changes of the spine</u> . <u>No focal airspace consolidation pleural effusion or pneumothorax</u> . <u>No pleural effusion or pneumothorax</u> .

Fig. 6. Examples of report generations compared to the ground truths for input images in the testing set. There are many repeated sentences in Hierarchical RNN's results.

to natural image captioning domain. On CIDEr and METEOR, RTT-GAN performs the best. RTT-GAN pre-trains the model with cross-entropy loss and then uses a generative adversarial network (GAN) to fine-tune it. The discriminator is based on CIDEr reward. Therefore, as the only method based on GAN, RTT-GAN performs better than our model on CIDEr and METEOR. Compared with the other models without RL or GAN, our model outperforms them on all the metrics.

V. CONCLUSION

In this paper, we propose a new framework to learn to detect disease, and generate medical reports from the initial images. We propose a GLP mechanism and the abnormality detection experiment shows that GLP performs better than GFP. We introduce the topic matching mechanism, context and semantic attention to hierarchical RNN, which can help to enhance the diversities of the generated sentences. We conduct medical annotation generation experiments on IU X-Ray dataset, and the results demonstrate that our model can produce better

annotations than the baselines. The experiments on GENOME dataset show that our model can be extended to the narrative paragraph generation of natural images.

VI. ACKNOWLEDGEMENT

This work is sponsored by "China Northwest Cohort Study" under the National Key Research and Development Program of China with grant number 2018YFC130078; "Multi-model Based Patient Similarity Learning for Medical Data Modelling and Learning" under National Natural Science Foundation of China General Program with grant number 61672420; Project of China Knowledge Center for Engineering Science and Technology; National Natural Science Foundation of China Innovation Research Team No. 61721002; Ministry of Education Innovation Research Team No. IRT_17R86; Key Project of Natural Science Foundation of China under grant No. 61532015.

REFERENCES

- [1] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015*, 2015. [Online]. Available: <https://doi.org/10.1109/CVPR.2015.7298935>
- [2] K. Xu, J. Ba, R. Kiros, K. Cho, A. C. Courville, R. Salakhutdinov, R. S. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015*, 2015. [Online]. Available: <http://jmlr.org/proceedings/papers/v37/xuc15.html>
- [3] S. J. Rennie, E. Marcheret, Y. Mroueh, J. Ross, and V. Goel, "Self-critical sequence training for image captioning," in *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, 2017.
- [4] C. Eickhoff, I. Schwall, A. G. S. de Herrera, and H. Müller, "Overview of imageclef2017 - image caption prediction and concept detection for biomedical images," in *Working Notes of CLEF 2017 - Conference and Labs of the Evaluation Forum, Dublin, Ireland, September 11-14, 2017.*, 2017. [Online]. Available: http://ceur-ws.org/Vol-1866/invited_paper_7.pdf
- [5] S. A. Hasan, Y. Ling, O. Farri, J. Liu, H. Müller, and M. Lungren, "Overview of imageclef 2018 medical domain visual question answering task," in *Working Notes of CLEF 2018 - Conference and Labs of the Evaluation Forum, Avignon, France, September 10-14, 2018.*, 2018. [Online]. Available: http://ceur-ws.org/Vol-2125/paper_212.pdf
- [6] S. Liang, X. Li, Y. Zhu *et al.*, "ISIA at the imageclef 2017 image caption task," in *Working Notes of CLEF 2017 - Conference and Labs of the Evaluation Forum, Dublin, Ireland, September 11-14, 2017.*, 2017.
- [7] H. Shin, K. Roberts, L. Lu, D. Demner-Fushman, J. Yao, and R. M. Summers, "Learning to read chest x-rays: Recurrent neural cascade model for automated image annotation," in *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016*, 2016.
- [8] E. P. Xing, P. Xie, and B. Jing, "On the automatic generation of medical imaging reports," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018*, 2018.
- [9] C. Y. Li, X. Liang, Z. Hu, and E. P. Xing, "Hybrid retrieval-generation reinforced agent for medical image report generation," *CoRR*, vol. abs/1805.08298, 2018. [Online]. Available: <http://arxiv.org/abs/1805.08298>
- [10] M. Hodosh, P. Young, and J. Hockenmaier, "Framing image description as a ranking task: Data, models and evaluation metrics," *J. Artif. Intell. Res.*, vol. 47, 2013. [Online]. Available: <https://doi.org/10.1613/jair.3994>
- [11] A. Farhadi, S. M. M. Hejrati, M. A. Sadeghi, P. Young, C. Rashtchian, J. Hockenmaier, and D. A. Forsyth, "Every picture tells a story: Generating sentences from images," in *Computer Vision - ECCV 2010*, 2010. [Online]. Available: https://doi.org/10.1007/978-3-642-15561-1_2
- [12] G. Kulkarni, V. Premraj, S. Dhar, S. Li, Y. Choi, A. C. Berg, and T. L. Berg, "Baby talk: Understanding and generating simple image descriptions," in *The 24th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2011*, 2011. [Online]. Available: <https://doi.org/10.1109/CVPR.2011.5995466>
- [13] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.
- [14] K. Cho, B. van Merriënboer, Ç. Gülçehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using RNN encoder-decoder for statistical machine translation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, 2014, pp. 1724–1734.
- [15] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, 2014, pp. 3104–3112.
- [16] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *CoRR*, vol. abs/1409.1556, 2014.
- [17] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, 2016, pp. 770–778. [Online]. Available: <https://doi.org/10.1109/CVPR.2016.90>
- [18] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, 2017. [Online]. Available: <https://doi.org/10.1109/CVPR.2017.243>
- [19] T. Chen, Y. Liao, C. Chuang, W. T. Hsu, J. Fu, and M. Sun, "Show, adapt and tell: Adversarial training of cross-domain image captioner," in *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, 2017, pp. 521–530.
- [20] X. Liu, H. Li, J. Shao, D. Chen, and X. Wang, "Show, tell and discriminate: Image captioning by self-retrieval with partially labeled data," in *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part XV*, 2018, pp. 353–369.
- [21] Q. You, H. Jin, Z. Wang, C. Fang, and J. Luo, "Image captioning with semantic attention," in *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016*, 2016. [Online]. Available: <https://doi.org/10.1109/CVPR.2016.503>
- [22] M. Ranzato, S. Chopra, M. Auli, and W. Zaremba, "Sequence level training with recurrent neural networks," *CoRR*, 2015. [Online]. Available: <http://arxiv.org/abs/1511.06732>
- [23] B. Dai, S. Fidler, R. Urtasun, and D. Lin, "Towards diverse and natural image descriptions via a conditional GAN," in *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, 2017, pp. 2989–2998. [Online]. Available: <https://doi.org/10.1109/ICCV.2017.323>
- [24] J. Krause, J. Johnson, R. Krishna, and L. Fei-Fei, "A hierarchical approach for generating descriptive image paragraphs," in *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, 2017. [Online]. Available: <https://doi.org/10.1109/CVPR.2017.356>
- [25] A. Karpathy and F. Li, "Deep visual-semantic alignments for generating image descriptions," in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, 2015, pp. 3128–3137. [Online]. Available: <https://doi.org/10.1109/CVPR.2015.7298932>
- [26] J. Johnson, A. Karpathy, and L. Fei-Fei, "Densecap: Fully convolutional localization networks for dense captioning," in *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016*, 2016. [Online]. Available: <https://doi.org/10.1109/CVPR.2016.494>
- [27] S. Ren, K. He, R. B. Girshick, and J. Sun, "Faster R-CNN: towards real-time object detection with region proposal networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, 2017.
- [28] X. Liang, Z. Hu, H. Zhang, C. Gan, and E. P. Xing, "Recurrent topic-transition GAN for visual paragraph generation," in *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, 2017, pp. 3382–3391. [Online]. Available: <https://doi.org/10.1109/ICCV.2017.364>
- [29] K. Papineni, S. Roukos, T. Ward, and W. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, 2002, pp. 311–318. [Online]. Available: <http://www.aclweb.org/anthology/P02-1040.pdf>
- [30] D. Demner-Fushman, M. D. Kohli, M. B. Rosenman, S. E. Shooshan, L. Rodriguez, S. K. Antani, G. R. Thoma, and C. J. McDonald, "Preparing a collection of radiology examinations for distribution and retrieval," *JAMIA*, vol. 23, no. 2, pp. 304–310, 2016. [Online]. Available: <https://doi.org/10.1093/jamia/ocv080>
- [31] R. Vedantam, L. Zitnick, and D. Parikh, "Cider: Consensus-based image description evaluation," *CoRR*, vol. abs/1411.5726, 2014. [Online]. Available: <http://arxiv.org/abs/1411.5726>
- [32] C.-Y. Lin, "Rouge: A package for automatic evaluation of summaries," *Text Summarization Branches Out*, 2004.
- [33] M. J. Denkowski and A. Lavie, "Meteor universal: Language specific translation evaluation for any target language," in *Proceedings of the Ninth Workshop on Statistical Machine Translation, WMT@ACL 2014, June 26-27, 2014, Baltimore, Maryland, USA*, 2014, pp. 376–380. [Online]. Available: <http://aclweb.org/anthology/W14/W14-3348.pdf>
- [34] J. Donahue, L. A. Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, T. Darrell, and K. Saenko, "Long-term recurrent convolutional networks for visual recognition and description," in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, 2015, pp. 2625–2634. [Online]. Available: <https://doi.org/10.1109/CVPR.2015.7298878>
- [35] A. Karpathy and L. Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 4, pp. 664–676, 2017. [Online]. Available: <https://doi.org/10.1109/TPAMI.2016.2598339>