Multimodal Recurrent Model with Attention for Automated Radiology Report Generation

Yuan Xue[1], Tao Xu[2], L. Rodney Long[3], Zhiyun Xue[3], Sameer Antani[3], George R. Thoma[3], and Xiaolei Huang[1(✉)]

[1] College of Information Sciences and Technology, Penn State University, University Park, PA, USA
sharon.x.huang@gmail.com
[2] Department of Computer Science and Engineering, Lehigh University, Bethlehem, PA, USA
[3] National Library of Medicine, National Institutes of Health, Bethesda, MD, USA

Abstract. Radiologists routinely examine medical images such as X-Ray, CT, or MRI and write reports summarizing their descriptive findings and conclusive impressions. A computer-aided radiology report generation system can lighten the workload for radiologists considerably and assist them in decision making. Although the rapid development of deep learning technology makes the generation of a single conclusive sentence possible, results produced by existing methods are not sufficiently reliable due to the complexity of medical images. Furthermore, generating detailed paragraph descriptions for medical images remains a challenging problem. To tackle this problem, we propose a novel generative model which generates a complete radiology report automatically. The proposed model incorporates the Convolutional Neural Networks (CNNs) with the Long Short-Term Memory (LSTM) in a recurrent way. It is capable of not only generating high-level conclusive impressions, but also generating detailed descriptive findings sentence by sentence to support the conclusion. Furthermore, our multimodal model combines the encoding of the image and one generated sentence to construct an attention input to guide the generation of the next sentence, and henceforth maintains coherence among generated sentences. Experimental results on the publicly available Indiana U. Chest X-rays from the Open-i image collection show that our proposed recurrent attention model achieves significant improvements over baseline models according to multiple evaluation metrics.

1 Introduction

A radiologist completes a radiology report, by analyzing images from an examination, recognizing both normal and abnormal findings, and coming to a diagnosis. This process of medical image interpretation and reporting can be error-prone, however, even for experienced specialists. Where the discrepancies can

A. F. Frangi et al. (Eds.): MICCAI 2018, LNCS 11070, pp. 457–466, 2018.
https://doi.org/10.1007/978-3-030-00928-1_52

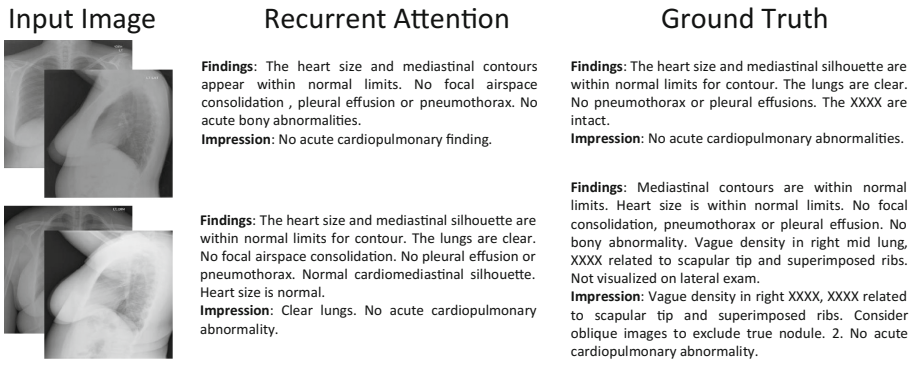

Fig. 1. Examples of original reports vs. reports generated by our recurrent attention model. Note that, Findings is a paragraph containing some descriptive sentences; Impression is a conclusive sentence. XXXXs are wrongly removed keywords due to de-identification.

come from the lack of knowledge or faulty reasoning by radiologists, staff shortage and excess workload also contribute to the errors in radiology reports [1]. To reduce workload and error occurrences, an automated or computer-aided reporting system can be helpful. An illustration of the automated report generation problem is shown in Fig. 1. The inputs are medical images of the same human subject from different views. In the resulting report, *Impression* is a single-sentence conclusion or diagnosis, and *Findings* is a paragraph containing multiple sentences that describe the radiologist's observations and findings regarding different regions in the images.

Most of the existing literature related to the report generation problem are based on deep learning technologies, following the encoder-decoder architecture originally used for machine translation [2]. While generation of the conclusive impression can be done by existing image captioning models that describe an image with a single sentence [15,19,20], Recurrent Neural Networks (RNNs) used by these existing models are known to be incapable of handling long sequences or paragraphs due to vanishing or exploding gradients [17]. Long Short-term Memory (LSTM) [8] alleviates this issue to some degree with a gating mechanism to learn long-term dependencies, but it still cannot completely prevent gradient from vanishing and thus is hard to model a very long sequence.

To generate a paragraph description, which is a very long sequence, some pioneering works have been done in the domain of natural image captioning, with hierarchical recurrent networks [12,13,21]. Mostly they use two levels of RNNs for paragraph generation: first, a paragraph-level RNN generates some topics, then a sentence-level RNN takes the topics as input and generates corresponding sentences. In [12,13], the authors utilize a pre-trained dense-captioning model [10] to detect semantic regions of the images. However, such pre-trained models are often not available for medical images. Toward the goal of medical image annotation, Shin *et al.* [18] proposed a deep learning

framework to automatically annotate chest x-rays with Medical Subject Headings (MeSH) annotations for the first time. They use a CNN to classify the x-ray images with different disease labels. RNNs are then trained to describe the contexts of a detected disease with more details. Furthermore, a cascade model is applied to combine image and text contexts to improve annotation performance. Zhang *et al.* [22] establish a direct multimodal mapping from medical images to diagnostic reports. They use an auxiliary attention sharpening (AAS) module to learn the image-language alignments more efficiently. However, their generated diagnostic reports are limited to describing five types of cell appearance features, which makes their problem less complex than general radiology report generation. Jing *et al.* [9] adopt the hierarchical generation framework from [12] to generate detailed descriptions of medical images along with a co-attention model which can simultaneously attend to both visual and semantic features. Their work achieved impressive results on the IU chest x-ray dataset [4], although some repetitions can be found in their generated reports because their hierarchical model does not take contextual coherence into consideration.

In this paper, we focus on the generation of a findings paragraph. We break down the paragraph generation task into easier subtasks, where a subtask is concerned with generating one sentence at a time. To guarantee the intra-paragraph dependency and coherence among sentences, we develop a recurrent model, in which a first sentence is generated and then each succeeding sentence is generated by taking the encodings of both its preceding sentence and the image, as joint inputs. The main contributions of our work toward automated radiology report generation are: (1) we propose a novel recurrent generation model to generate the findings paragraph, sentence by sentence, whereby a succeeding sentence is conditioned upon multimodal inputs that include its preceding sentence and the original images, (2) we adopt an attention mechanism for our proposed multimodal model to improve performance. Extensive experiments on the Indiana U. Chest x-rays dataset demonstrate the effectiveness of our proposed methods.

2 Methodology

Assume we are generating a findings paragraph that contains L sentences. The probability of generating the i-th sentence with length T satisfies:

$$\begin{aligned}&\mathbb{P}(S_i = w_1, w_2, ..., w_T | V; \theta) \\ &= \mathbb{P}(S_1|V) \prod_{j=2}^{i-1} \mathbb{P}(S_j|V, S_1, ...S_{j-1}) \mathbb{P}(w_1|V, S_{i-1}) \prod_{t=2}^{T} \mathbb{P}(w_t|V, S_{i-1}, w_1, ...w_{t-1}),\end{aligned} \tag{1}$$

where V is the given medical image, θ is the model parameter (we omit the θ in the right hand side), S_i represents the i-th sentence and w_t is the t-th token in the i-th sentence. Similar to the n-gram assumption in language models, we adopt the Markov assumption for sentence level generation with a "2-gram" model, which means the current sentence being generated depends only on its

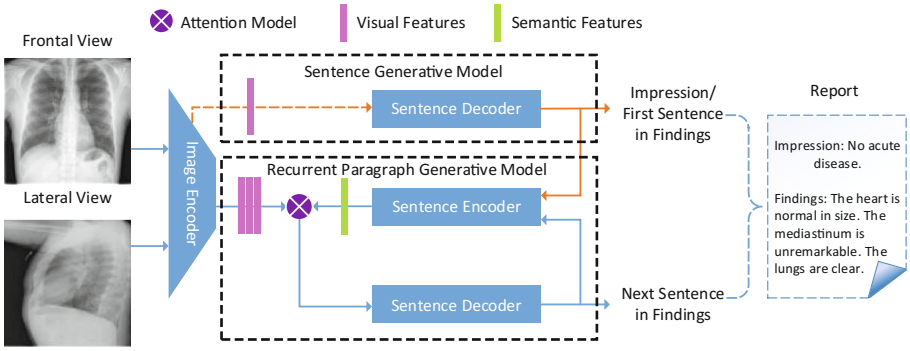

Fig. 2. The architecture of the proposed multimodal recurrent generation model with attention for radiology reports. Best viewed in color.

immediately preceding sentence and the image. This simplifies the estimated probability to be:

$$\hat{\mathbb{P}}(S_i = w_1, w_2, ...w_T|V;\theta) = \underbrace{\mathbb{P}(S_1|V)}_{1} \underbrace{\prod_{j=2}^{i-1} \mathbb{P}(S_j|V, S_{j-1})}_{2} \underbrace{\mathbb{P}(w_1|V, S_{i-1}) \prod_{t=2}^{T} \mathbb{P}(w_t|V, S_{i-1}, w_1, ...w_{t-1})}_{3}. \tag{2}$$

Our goal is to find the optimal parameter for the Maximum Log-likelihood Estimate (MLE) as

$$\theta^* = \underset{\theta}{\text{argmax}} \sum_{i=1}^{L} \log \hat{\mathbb{P}}(S_i = G_i|V;\theta), \tag{3}$$

where G_i is the ground truth for the i-th sentence in the findings paragraph.

As shown in Eq. 2, we separate that equation into 3 parts denoted by underbraces and introduce our model part by part. The overall architecture of our framework that takes medical images from multiple views as input and generates a radiology report with impression and findings is shown in Fig. 2. In order to generate the findings paragraph, we first use an encoder-decoder model which takes an image pair as input and generates the first sentence. Then the first sentence is fed into a sentence encoding network to output the semantic representation of that sentence. After that, both visual features of the image and semantic features of the preceding sentence are combined as the input to the multimodal recurrent generation network that generates the next sentence. This process is repeated until the model generates the last sentence in the paragraph. More details will be explained in the next subsections.

2.1 Image Encoder

In our model (Fig. 2), an image encoder is first applied to extract both global and regional visual features from the input images. The image encoder is a Convolutional Neural Network (CNN) that automatically extracts hierarchical visual features from images. More specifically, our image encoder is built upon the pre-trained resnet-152 [7]. We resize the input images to 224×224 to keep consistent with the pre-trained resnet-152 image encoder. Then, the local feature matrix $f \in \mathbb{R}^{1024 \times 196}$ (reshaped from $1024 \times 14 \times 14$) are extracted from the "res4b35" layer of resnet-152. Each column of f is one regional feature vector. Thus each image has 196 sub-regions. Meanwhile, we extract the global feature vector $f \in \mathbb{R}^{2048}$ from the last average pooling layer of resnet-152. For multiple input images from several views (e.g., frontal and lateral views as demonstrated in this paper), their regional and global features are concatenated accordingly before feeding into the following layers. For efficiency, all parameters in layers built from the resnet-152 are fixed during training.

2.2 Sentence Generative Model

In general, both the one-sentence impression and the first sentence in the findings paragraph contain some high level descriptions of the image. Thus, we develop a sentence generative model that takes the global visual features learned by the image encoder as input. Such a model can be trained to generate the impression. It can also be jointly trained with the recurrent generative model to generate the first sentence in the findings as an initialization of the recurrent model (formulated in part 1 of Eq. 2). In the sentence generative model, a single layer LSTM [8] is used for sentence decoding. The initial hidden states and cell states of the LSTM are set to be zero. The visual feature vector is used as the initial input of the LSTM to predict the first word of the sentence and then the whole sentence is produced word by word. Before being fed into the LSTM, a fully connected layer is utilized to transform the visual feature vector so that it has the same dimension as the word embedding. In all LSTM modules used in this paper, the dimensions of word embedding and the dimensions of hidden states are 512 and 1024, respectively.

2.3 Recurrent Paragraph Generative Model

As shown in Fig. 2, our recurrent paragraph generative model takes the sentence and regional image features as input and generates findings paragraph sentence by sentence. It has two main components: sentence encoder and attentional sentence decoder.

Sentence Encoder is used to extract semantic vectors from text descriptions. Two types of well-known text encoders are explored in this paper. The first one is a Bi-directional Long Short-Term Memory (Bi-LSTM) [6] which can encode better context information than the conventional one-directional LSTM. In the Bi-LSTM, each word corresponds to two hidden states, one for each

direction. Inspired by [3], the 1D convolution neural network is also applied for sentence encoding. Our CNN model takes the 512 dimensional word embedding as input and has three convolution layers to learn hierarchical features. Each convolution layer has the kernel size 3, stride 1 and 1024 feature channels. The max-pooling operation is applied over feature maps extracted from each convolution layer, yielding a 1024 dimensional feature vector. The final sentence feature is the concatenation of feature vectors from different layers. We compare these two proposed encoder networks in Sect. 3.

Attentional Sentence Decoder takes regional visual features and the previously generated sentence as a multimodal input, and generates the next sentence. This solves both part 2 and part 3 of Eq. 2. The sentence decoder is a stacked 2-layer LSTM. The image pair V are converted as input to the 2-layer LSTM, then the learned encoding of the preceding sentence guides our model to generate the next sentence. We repeat this process until an empty sentence is generated, which indicates the end of the paragraph. In this way, the consistence of context in the paragraph is guaranteed.

To make different sentences focus on different image regions and capture the dependency between sentences, we propose a sentence based visual attention [15,20] mechanism for our recurrent generative model. Semantic features of the preceding sentence and regional visual representations are fed through a fully connected layer followed by a softmax layer to get the attention distribution over $k = 196$ image regions. First, we compute the attention weights over k regions as

$$\boldsymbol{a} = \boldsymbol{W}_{\text{att}} \tanh(\boldsymbol{W}_v \boldsymbol{v} + \boldsymbol{W}_s \boldsymbol{s} \mathbf{1}^k), \tag{4}$$

where $\boldsymbol{v} \in \mathbb{R}^{d_v \times k}$ are the regional visual features learned by the image encoder, $s \in \mathbb{R}^{d_s}$ represents the encoding of the preceding sentence. $\boldsymbol{W}_{\text{att}} \in \mathbb{R}^{1 \times k}$, $\boldsymbol{W}_v \in \mathbb{R}^{k \times d_v}$ and $\boldsymbol{W}_s \in \mathbb{R}^{k \times d_s}$ are parameters of the attention network. $\mathbf{1}^k \in \mathbb{R}^{1 \times k}$ is a vector with all ones. $d_v = 1024$ is the dimension of the regional visual feature; d_s is the dimension of the sentence feature ($d_s = 2048$ for Bi-LSTM and $d_s = 3072$ for CNN sentence encoder). Next, we normalize it over all regions to get the attention distribution:

$$\alpha_i = \frac{\exp(a_i)}{\sum_i \exp(a_i)}, \tag{5}$$

where a_i is the i-th dimension in $\boldsymbol{a}$. Finally, we compute the weighted visual representation as

$$\boldsymbol{v}_{\text{att}} = \sum_{i=1}^{k} \alpha_i \boldsymbol{v}_i. \tag{6}$$

The input of the sentence decoder are now the weighted visual representation. When generating different sentences, the attention model focuses on different regions of the image based on the context of the preceding sentence. Features or regions which are not relevant to current sentence are filtered out and the model cannot see the sentence encoding directly so it is less likely to overfit to the semantic input. Performance comparison for the model with and without attention module can be found in Sect. 3.

All our proposed models are trained by the Adam optimizer [11]. The initial learning rate is set to be 1e−4 and learning rate decay is 0.9 for every 5 epochs.

The batch size is 460 for training. During the training, we adopt a teacher forcing policy, i.e., we always feed our decoder with ground truth word or sentence for the generation in the next timestep. During testing, greedy search is used for generating words and sentences in every timestep for efficiency. Previously generated words/sentences will be fed into the decoder as part of the input for the next word/sentence. The recurrent generative model will keep generating sentences until it generates an empty sentence. All modules are trained jointly in an end-to-end fashion by minimizing the cross entropy loss.

3 Experiments

We evaluate our model on the Indiana University Chest X-Ray collection [4]. The dataset contains 3,955 radiology reports from 2 large hospital systems within the Indiana Network for Patient Care database, and 7,470 associated chest x-rays from the hospitals' picture archiving systems. Each report is associated with a pair of images which are the frontal and lateral views, and contains comparison, indication, findings, and impression sections. All reports are fully anonymized for de-identification; however, 2.5% of findings/impression words are also removed during the de-identification, resulting in some keywords missing in the report. Since the original data are from multiple hospitals and are inconsistent, there are some images or findings missing in the original dataset. For our experiments, we filtered out reports without two complete image views or without complete sections of findings and impression, resulting in a smaller dataset with 2,775 reports associated with 5,550 images.

For data preparation, we tokenized all the words in the findings and impression in the dataset and obtained 2,218 unique words. Considering that the c size is already very small, we decided not to drop infrequent words with only once or twice appearances. We also added two special tokens, ⟨S⟩ and ⟨/S⟩, into the vocabulary to indicate the start and the end of a sentence. To evaluate our models, we randomly picked 250 reports to form the testing set. All evaluations are done on the testing set.

We use some common evaluation metrics for image captioning to provide a quantitative comparison. We report BLEU [16], METEOR [5] and ROUGE [14] scores of all proposed models and compare them with baseline models in Table 1. However, these evaluation metrics are not specially designed for medical report generation tasks. Hence we suggest another complementary metric. We construct a keyword dictionary from MTI annotations of the original dataset and some manual annotations. The dictionary contains 438 unique keywords, and we compute the keywords accuracy (KA) metric as the ratio of the number of keywords correctly generated by a model to the number of all keywords in the groundtruth findings. An example result can be found in Fig. 1.

For comparison, we reimplemented two baseline models [12,19] for radiology report generation. We use the same pre-trained resnet-152 image encoder for all models. Note that we do not have a pre-trained dense captioning model for medical images, thus we only use features learned by the image encoder

directly for hierarchical generation [12]. Since Bi-LSTM encoding achieves better performance than convolution encoding in experiments, we adopt Bi-LSTM encoding for our final model. We also implemented a baseline model without attention module. In the recurrent generative model without attention, the sentence encoding learned by the sentence encoder are used as the initial hidden state and cell state of the sentence decoder. From Table 1, we can see that our final model with attention shows significant improvements over baseline models in all evaluation metrics. Moreover, although the hierarchical model [12] achieves reasonably high evaluation scores, the generated reports contain some repetitions and the paragraphs are not very coherent. In comparison, reports generated by our proposed model contain fewer repetitions and have more coherent context.

Table 1. Evaluation of generated reports on our testing set using BLEU, METEOR, ROUGE and KA metrics. We compare our models with two baseline models including a baseline implementation of the hierarchical generation model [12].

Methods	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE	KA
Vanilla CNN-RNN [19]	0.273	0.144	0.116	0.082	0.125	0.226	0.435
Hierarchical generation [12]	0.437	0.323	0.221	0.172	0.244	0.325	0.568
Ours-recurrent-conv	0.416	0.298	0.217	0.163	0.227	0.309	0.532
Ours-recurrent-BiLSTM	0.423	0.307	0.223	0.165	0.236	0.322	0.543
Ours-recurrent-attention	**0.464**	**0.358**	**0.270**	**0.195**	**0.274**	**0.366**	**0.596**

4 Discussions

In this paper, our main focus is on generating detailed findings for a report. For impression generation, a classification based method may be better at distinguishing abnormal cases and then giving the final conclusion. From the example results in Fig. 1 we can see that in the first row, both findings and impression are in accord with the groundtruth. However, in the second row of Fig. 1, both the generated findings and impression missed some abnormal descriptions. The main reason may be that we are training on a small training set, the training samples for abnormal cases are even fewer, and there are some inconsistency as well as noise in the original groundtruth reports. Moreover, our model does not create very well new sentences that have never appeared in the training set. This could be due to the difficulty in learning correct grammar from a small corpus since the objective function for training does not consider syntactic correctness. We expect that addressing the above limitations would require a larger and better annotated dataset, a new training strategy and a new evaluation metric which takes both keyword accuracy and grammar correctness into account.

5 Conclusions

In summary, we have proposed a multimodal recurrent model with attention for radiology report generation. A long and detailed paragraph can be generated recurrently sentence by sentence. Such a model can provide interpretable justifications as part of a computer-aided reporting system to assist clinicians in making decisions. We have shown that generating long and detailed paragraphs of findings is not only theoretical feasible but also practically useful. Experiments on a chest x-rays dataset demonstrate the effectiveness of our proposed method.

References

1. Brady, A., Laoide, R.Ó., McCarthy, P., McDermott, R.: Discrepancy and error in radiology: concepts, causes and consequences. Ulster Med. J. **81**(1), 3 (2012)
2. Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., Bengio, Y.: Learning phrase representations using rnn encoder-decoder for statistical machine translation. arXiv preprint arXiv:1406.1078 (2014)
3. Conneau, A., Kiela, D., Schwenk, H., Barrault, L., Bordes, A.: Supervised learning of universal sentence representations from natural language inference data. arXiv preprint arXiv:1705.02364 (2017)
4. Demner-Fushman, D., Kohli, M.D., Rosenman, M.B., Shooshan, S.E., Rodriguez, L., Antani, S., Thoma, G.R., McDonald, C.J.: Preparing a collection of radiology examinations for distribution and retrieval. J. Am. Med. Inform. Assoc. **23**(2), 304–310 (2015)
5. Denkowski, M., Lavie, A.: METEOR universal: Language specific translation evaluation for any target language. In: Proceedings of the Ninth Workshop on Statistical Machine Translation, pp. 376–380 (2014)
6. Graves, A., Schmidhuber, J.: Framewise phoneme classification with bidirectional lstm and other neural network architectures. Neural Netw. **18**(5–6), 602–610 (2005)
7. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR, pp. 770–778 (2016)
8. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Comput. **9**(8), 1735–1780 (1997)
9. Jing, B., Xie, P., Xing, E.: On the automatic generation of medical imaging reports. arXiv preprint arXiv:1711.08195 (2017)
10. Johnson, J., Karpathy, A., Fei-Fei, L.: Densecap: Fully convolutional localization networks for dense captioning. In: CVPR, pp. 4565–4574 (2016)
11. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
12. Krause, J., Johnson, J., Krishna, R., Fei-Fei, L.: A hierarchical approach for generating descriptive image paragraphs. In: CVPR, pp. 3337–3345 (2017)
13. Liang, X., Hu, Z., Zhang, H., Gan, C., Xing, E.P.: Recurrent topic-transition GAN for visual paragraph generation. CoRR, abs/1703.07022 2 (2017)
14. Lin, C.Y.: ROUGE: A package for automatic evaluation of summaries. Text Summarization Branches Out (2004)
15. Lu, J., Xiong, C., Parikh, D., Socher, R.: Knowing when to look: Adaptive attention via a visual sentinel for image captioning. In: CVPR, pp. 375–383 (2017)

16. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: BLEU: a method for automatic evaluation of machine translation. In: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, pp. 311–318. Association for Computational Linguistics (2002)
17. Pascanu, R., Mikolov, T., Bengio, Y.: On the difficulty of training recurrent neural networks. In: International Conference on Machine Learning, pp. 1310–1318 (2013)
18. Shin, H.C., Roberts, K., Lu, L., Demner-Fushman, D., Yao, J., Summers, R.M.: Learning to read chest x-rays: Recurrent neural cascade model for automated image annotation. In: CVPR, pp. 2497–2506 (2016)
19. Vinyals, O., Toshev, A., Bengio, S., Erhan, D.: Show and tell: A neural image caption generator. In: CVPR, pp. 3156–3164 (2015)
20. Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., Zemel, R., Bengio, Y.: Show, attend and tell: Neural image caption generation with visual attention. In: International Conference on Machine Learning, pp. 2048–2057 (2015)
21. Yu, H., Wang, J., Huang, Z., Yang, Y., Xu, W.: Video paragraph captioning using hierarchical recurrent neural networks. In: CVPR, pp. 4584–4593 (2016)
22. Zhang, Z., Xie, Y., Xing, F., McGough, M., Yang, L.: Mdnet: A semantically and visually interpretable medical image diagnosis network. In: CVPR, pp. 6428–6436 (2017)