Multi-view Radiology Report Generation

Anonymous EMNLP submission

Abstract

The generation of radiology report is very time-consuming. Sometimes, for inexperienced physicians, they easily makes erroneous diagnosis. To address these issues, we aims to study the automatic medical imaging reports. We adopted attention mechanism to narrow down the regions to focus for that timestamp. Besides, we utilized characteristic of radiology image to learn comprehensively from both frontal features and lateral ones for a specific patient. We demonstrate the effectiveness by evaluating on one publicly available dataset - IU X-Ray.

1 Introduction

In radiology practice, the most critical part is to generate a text description, or report according to the clinical radiographs. However, many diagnosis and treatment of many diseases such as pneumonia and pneumothorax heavily rely on these medical images. Although the interpretation of medical images can presently be delivered by experienced radiologists and pathologists, the process to generate report is still too time-consuming to reach max efficiency. To be more specific, if a doctor in a large Metropolis city have to read hundreds of radiographs per day, it will take most of their working time by spending about 5-10 minutes on each of them. It then creates a motivation to provide automatic support for this task. However, this automatic report generation task (Jing et al., 2017) poses many technical challenges. Traditional image captioning approaches (Xu et al., 2015) (Vinyals et al., 2015). These models integrate spatial-visual attention newworks to produce shorter and less complex text whereas the radiology report should put the accuracy of clinical description as the top priority. This motivate us to investigate on how to integrate the features and style of radiographs into traditional image captioning model towards much high-quality reports.

2 Related Works

2.1 Image Captioning

Image captioning is a task aiming at automatically generating description for the given images. Withing the last few years, many image captioning model rely mostly on the combination of convolutional neural network (CNN) and recurrent neural network (RNN). CNN extracts features from images and RNN encodes the features into text. Recently, attention mechanism has received with the success of Show and Tell and its follow up (Xu et al., 2015). These models integrate spatial-visual attention mechanism over image features learned from CNN. To effect utilize the spatial features of radiographs, we propose a parallel CNN-RNN structures and leverage both on each timestamp.

3 Methods

3.1 Overview

A common radiographs are usually composed of the frontal-view images and lateral-view images. It seldom miss any of them. These different view medical images aids on the interpretation. The organs and tissues in human body is definitely not two-dimensional. To fully check the health status of human body, it is not helpful for radiologists to use only one view point. They may wants to reference images shot from different angle to make more accurate diagnosis. In our task we remove those case without both frontal-view images and lateral-view images. Given this two distinct views for a specific patient, we use two parallel CNN encoders to perform features extraction separately. Each groups of features are then sent to two corresponding attention networks and also the LSTM decoders. To predict the word, we then concatenate the hidden outputs for that timestamp and performs the corresponding words.

3.2 Encoders: Image Features Extraction

In many successful image captioning approaches, CNN is believed to effectively and accurately extract meaningful features. Therefore, for our encoder part, our model general adopts pretrained ResNet50 (He et al., 2016) with fine-tune, but with two parallel identical encoder. The two encoders process the frontal and the lateral radiology image of the same patient. They both extract L vectors, each of which is a D-dimensional representation corresponding to that image. We refers it as annotation vectors for the rest of the paper.

$$a^f = \{a^f_1, a^f_2, ..., a^f_L\} \tag{1}$$

$$a^l = \{a^l_1, a^l_2, ..., a^l_L\} \tag{2}$$

3.3 Decoders: Report Generation

RNN is a state-of-the-art deep learning model for modeling sequential information. However, the normal RNN may suffer the vanishing gradient problem. Here, we use a special kind of RNN - Long Short Term Memory network (LSTM) (Hochreiter and Schmidhuber, 1997), which is capable to avoid vanish gradient issue and able to learn long-term dependencies. The LSTM is made up of a memory cell, an input gate, an output gate and a forget gate.

Below is the general LSTM implementation.

$$f_t = \sigma_g(W_f x_t + U_f h^v_{t-1} + b_f) \tag{3}$$

$$f_t = \sigma_g(W_i x_t + U_i h^v_{t-1} + b_i) \tag{4}$$

$$o_t = \sigma_g(W_o x_t + U_o h^v_{t-1} + b_o) \tag{5}$$

$$g_t = \tanh(W_c x_t + U_c h^v_{t-1} + b_c) \tag{6}$$

$$c_t = f_t \circ c_{t-1} + i_t \circ g_t \tag{7}$$

$$h^v_t = o_t \circ \tanh(c_t) \tag{8}$$

where x_t: input vector to the LSTM unit; f_t: forget gate's activation vector; i_t: input/update gate's activation vector; o_t: output gate's activation vector; h_t: hidden state vector also known as output vector of LSTM unit; c_t: cell state vector; W, U, b: learn-able weight matrices and bias vector; $\circ$: element-wise product; v: $\{frontal, lateral\}$..

To process two encoded annotations at each time stamp, our model becomes h^v_t where $v \in \{f, l\}$. f stands for the frontal annotation, while l for the lateral annotation.

Besides, we utilize the visual mechanism in [1] as ϕ operation to compute $\hat{z}_t$ from the annotation vectors $a_i, i = 1, 2, .., L$. The attention weight α_i of each annotation vector i is computed by the attention model f_{att} for which is composed of multilayer perceptron conditioned on the previous hidden state h_{t-1}. In our model, we have two attention network f^v_{att} where $v \in \{f, l\}$ processed the frontal annotations and the lateral annotations.

$$e^v_{ti} = f^v_{att}(a^v_i, h^v_{t-1}) \tag{9}$$

$$\alpha^v_{ti} = \frac{exp(e^v_{ti})}{\sum_{k-1}^{L} exp(e^v_{tk})} \tag{10}$$

$$\hat{z}^v_t = \phi(\{a^v_i\}, \{\alpha^v_i\}) \tag{11}$$

For hidden state in each timestamp, we will do the concatenate operation. The concatenated hidden vectors will be used to compute the output word probability $p(y_t|a, y_{t-1})$ given the context vector and the previous word.

3.4 Loss Function

We rewrite original loss function defined in [1] to considered both attention weights as below.

$$L = -log(P(y|x)) + \lambda \sum_{i}^{L} (1 - \sum_{t}^{C} \alpha_{ti}^{frontal})^2 + \lambda \sum_{i}^{L} (1 - \sum_{t}^{C} \alpha_{ti}^{lateral})^2 \tag{12}$$

4 Experiment

4.1 Dataset

We use one publicly medical image dataset to evaluate our proposed model.

IU X-Ray The Indiana University Chest XRay Collection (IU X-Ray) (Demner-Fushman et al., 2016) is a set of chest x-ray images paired with their corresponding diagnostic reports. The dataset contains 7,470 pairs of images and reports. Each report consists of the following sections: impression, findings, tags1, comparison, and indication. In this paper, for text preprocessing, we treat the contents in findings as the target captions to be generated. We preprocessed the data by converting all tokens to lower cases, removing all of non-alpha tokens. For image preprocessing part, we firstly resize both height and width to 256. After using a trained

ResNet50 model to classify between frontal-view images and the lateral views, we get 3425 pairs of both frontal-view and lateral-viewed images. We split those pairs into 2055,685and 685 pairs for the training, validation and testing sets, respectively. If there are multiple frontal-view image or multiple lateral-view image of the patient. We choose the one of them for our model.

4.2 Evaluation

We use BLEU (Papineni et al., 2002), or the Bilingual Evaluation Understudy, as an evaluation for our tasks. It is a score for comparing a candidate translation of text to one or more reference translations. We will report BLEU-4 scores in the next section.

4.3 Results

We evaluate our model on the BLEU-4 metrics. Besides, we use teacher-forcing technique in both of training and validation part, regardless of the word last generated. In other words, we still supplying the ground truth as the input at each decode-step. This process can speed-up the convergence. However, in the validation stage, we should mimic real inference conditions as much as possible. Thus, we also provide results without teacher-forcing, named real BLEU-4 scores. Below are figures (Figure1, Figure2) showing the evaluations of our model and the baseline model. The pink lines are our model, while the orange lines are the baseline model.

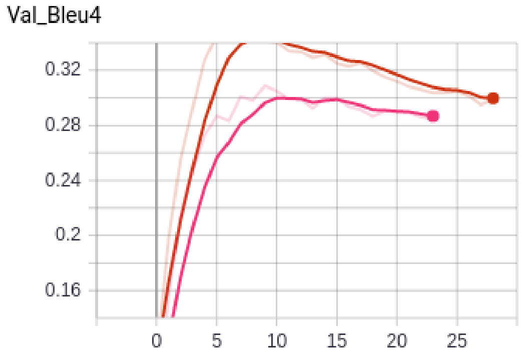

Figure 1: BLEU-4 scores on the Validation set.

5 Conclusion

We can seen that if we use more spatial information to generation radiology report, we can generally improve the performance of model. In other words, the process combining the frontal-view and lateral view is helpful while making generating radiology report.

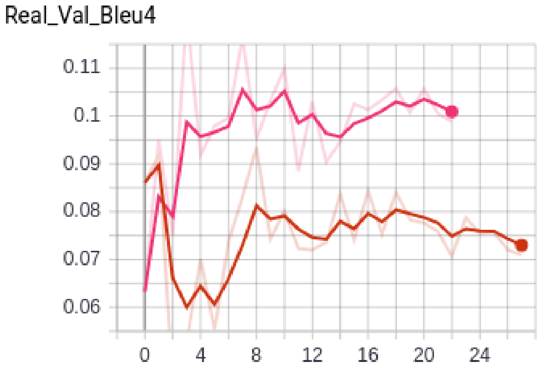

Figure 2: Real BLEU-4 scores on the Validation set.

References

Dina Demner-Fushman, Marc D Kohli, Marc B Rosenman, Sonya E Shooshan, Laritza Rodriguez, Sameer Antani, George R Thoma, and Clement J McDonald. 2016. Preparing a collection of radiology examinations for distribution and retrieval. *Journal of the American Medical Informatics Association*, 23(2):304–310.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. *CVPR*.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation*.

Baoyu Jing, Pengtao Xie, and Eric Xing. 2017. On the automatic generation of medical imaging reports. *arXiv preprint arXiv:1711.08195*.

K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. *ACL*.

O. Vinyals, A. Toshev, S. Bengio, and D. Erhan. 2015. Show and tell: A neural image caption generator. *CVPR*.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron C. Courville, Ruslan Salakhutdinov, Richard S. Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. *CoRR*, abs/1502.03044.