

Automatic Generation of Pulmonary Radiology Reports with Semantic Tags

Mengdan Gu, Xin Huang, Yu Fang

College of Electronics and Information Engineering.

Tongji University

Shanghai, China

e-mail: 1732953@tongji.edu.cn

Abstract—The chest X-ray is widely used in clinical practice for diagnosis and treatment. Among all chest diseases, pulmonary disease accounts for the majority, and the description of the lungs in the radiology report is the most complicated. According to this situation, this paper proposes a pulmonary radiology report generation model (SRN+BC-LSTM) based on semantic tags of radiograph. Firstly, for the problem that the image features extracted by CNN do not contain obvious semantic information, this paper selects the high frequency words related to abnormalities in the pulmonary radiology report as semantic tags, and trains the multi-label classifier. Secondly, a binary classifier is combined to improve the BLEU of generated normal reports since the Chinese radiology reports have roughly the same description for the normal samples. The experiment results indicated that our model is 13% higher than the baseline model in BLEU4.

Keywords—component; deep learning; computer-aided detection; radiology report; semantic tag

I. INTRODUCTION

Chest X-ray is an important means of diagnosing chest disease. The writing of the radiology reports usually requires a radiologist with professional medical knowledge and clinical experience, and it is time consuming. Moreover, for remote areas with low medical level, professional radiologists are a scarce resource. A computer-aided radiology report generation system can lighten the workload for radiologists considerably and assist them in decision making.

Most of the radiology report generation methods [1][2][3] are based on the OpenI dataset [4] created by Indiana University, which extracts image information through CNN, and uses LSTM to generate radiology reports composed of finding and impression. Due to the excessive information contained in the long radiology report, the value of BLEU4 generated by the existing method is about 0.2, and the quality of the report is poor. In addition, since the image features extracted by CNN do not contain obvious semantic information, it is difficult for the model to learn the mapping relationship between radiographs and radiology reports. At the same time, in the Chinese radiology report, the description of normal symptoms is very similar. When the normal report is correctly predicted, the BLEU of the generated report is higher, but there is no research on this feature.

To address these issues, this paper proposes a pulmonary radiology report generation model based on semantic tags (SRN+BC-LSTM). Firstly, in view of the problem that the radiology report is so long that the generated report has low value of BLEU, this paper only selects the description related to the lung in the radiology report to research. The reason is that by making statistics of the two largest public datasets, we found that in the ChestX-ray8 dataset [5], the proportion of pulmonary abnormalities is 71.31%, and in the CheXpert dataset [6] is 59.14% (Except Support Devices), which indicates that abnormalities in the lungs account for the majority of all abnormalities. Huang et al. [26] and Yan et al. [27] also verified this in the work of Chinese data sets. Secondly, the image features extracted by CNN do not contain obvious semantic information. To solve this problem, this paper selects the words with high frequency and related to abnormal locations and abnormal symptom in the pulmonary radiology report as the semantic tag of the radiograph. The multi-label classifier is trained to predict semantic tags, and semantic tags is used to generate radiology report. Finally, since Chinese radiology reports usually have similar description of normal symptom, a binary classifier that distinguishes between normal and abnormal is added, and the semantic tags generated by the multi-label classifier are corrected by the binary classifier result, thereby improving the accuracy of predicting normal samples. The experiment was conducted on the Chinese radiology report dataset. The results show that the SRN+BC-LSTM model achieves significant improvements over baseline models according to multiple evaluation metrics.

II. RELATED WORK

A. Radiology Report Generation

In the field of radiology report generation, the public data set is only OpenI [4] of Indiana University, and most of the related research is carried out on this data set. TieNet, designed by Wang et al. [7], uses radiology reports and images as input to predict the type of disease and generate a radiology report. This method improves the accuracy of the disease classification, but the generated radiology report has a low value of BLEU. Jing et al. [1] learned from the hierarchical LSTM of Krause [8] to generate long paragraphs, and used the features and semantic features of the image as input to the sentence-level LSTM to generate multiple topic vectors. The word level LSTM generates a sentence for each topic vector. Li et al. [2] used both the front and lateral chest

radiographs as the input of CNN, combining the search method with the generation method. For the sentences with high frequency, the search method is adopted, and the sentences with low frequency are generated by the hierarchical LSTM method. Finally, all the sentences are composed into a radiology report. The model proposed by Xue et al. [3] combines CNN and LSTM in a cyclical manner. The model combines image features and generated sentences to form inputs to guide the generation of the next sentence.

The objective of existing methods is to generate the finding and impression parts of the radiology report. Since this part contains a large amount of semantic information, it is difficult for model to capture such rich information when generating the report. The radiology report generated by the above method has a lower value of BLEU. Since abnormalities occur mostly in the lungs, the pulmonary report is extracted and generated separately in this paper. The pulmonary report contains less information than the entire report, making it easier for the model to capture valid information when report is generated, resulting in a more accurate report.

B. ImageCaption

The image caption task is designed to automatically generate a short description for a given image. Vinyals et al. [9] and others learned from the ideas in machine translation [10][11][12], using CNN as Encoder to extract image features, and RNN as Decoder to generate a natural language description based on image features. Xu et al. [13] proposed an image caption model based on the attention mechanism. For each time step input, the model pays more attention to the key areas of the image. However, for words such as adverbs and prepositions, visual information cannot be found in the image, and the attention to non-visual words at each time step will reduce the effectiveness of visual information. Therefore, Lu et al. [14] proposed an adaptive attention model with visual markers to determine whether to use the attention mechanism at each time step.

The methods mentioned above all generate captions based on image features, but the image features do not explicitly represent the semantic information of the images. In order to improve the quality of the generated caption, people began to seek ways to obtain semantic information from images to generate captions. Wu et al. [15] proposed using advanced semantic information instead of the previous image features as input to the RNN. They used the high-frequency words in the caption as semantic tags and generated captions using semantic tags. Yao et al. [16] used multi-instance learning to extract image semantic tags, proposed an LSTM [17] model with semantic tags, integrated semantic tags into CNN and RNN using different model structures, and compared the performance of these model structures.

The above two methods require segmentation of the image into small regions to extract image semantic tags. However, some of the image semantic tags selected in this paper are location-dependent. After the image is segmented, the location information may be lost. Therefore, this paper

uses the Spatial Regularization Network (SRN) proposed by Zhu et al. [18] as multi-label classifier to predict semantic tags. The SRN learns the location-related semantic tags by attention mechanism to avoid the loss of position information caused by segmentation.

III. APPROACH

Our approach is summarized in Fig. 1. The SRN+BC-LSTM model includes the semantic tag predictor and the report generator. The model performs semantic tag prediction by combining SRN and binary classifier. A fixed length vector $\mathbf{V}_{att}(\mathbf{I})$ is created by SRN for each image \mathbf{I} , whose length is the size of the semantic tags set. Each bit of the vector represents the predicted probability of a semantic tag. The binary classifier produces an integer $\mathbf{B} \in \{0,1\}$ for each image \mathbf{I} , and multiplies \mathbf{B} by $\mathbf{V}_{att}(\mathbf{I})$ to obtain $\mathbf{V}_{att}(\mathbf{I})$. In the part of report generation, $\mathbf{V}_{att}(\mathbf{I})$ is fed into LSTM [17] to generate reports.

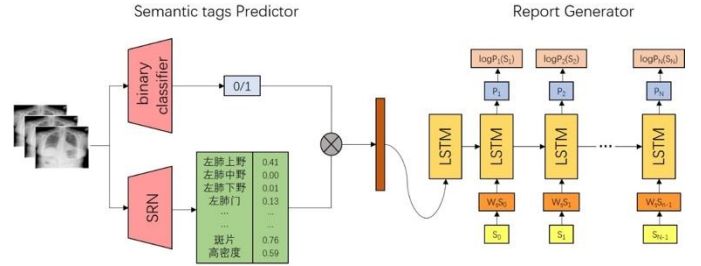


Figure 1. Overall framework of SRN+BC-LSTM model. The left part predicts semantic tags for each image. The right part learns a mapping from the semantic tags to a sequence of words using an LSTM.

A. Semantic tags Predictor

Firstly, a set of semantic tags needs to be built. We choose the words with the highest frequency and related to abnormal locations and abnormal symptom as a set of semantic tags. There will be many synonyms, which will be treated as the same semantic tag, and finally a set of 40 semantic tags can be obtained.

After creating the semantic tag set, the multi-label classifier is trained to generate the semantic tag vector of the image. Our dataset does not mark the ground truth bounding box for each semantic tag, so we cannot use the object detection method to get the predicted tags. The region-based multi-label classification model proposed by Wei et al. [19] can realize multi-label classification without a ground truth bounding box. However, the model needs to segment the image into multiple regions for multi-label classification, which will lose the information of location. Since some semantic tags are location-dependent, we use the SRN [18] to predict semantic tags.

Each image corresponds to a semantic tag vector $\mathbf{y} = [y_1, y_2, \dots, y_c]^T$. Where c is the number of semantic tags. When the image is labeled with the semantic tag j , $y_j = 1$, otherwise $y_j = 0$. When the semantic tag vector is all 0, that is, $\mathbf{y} = [0, 0, \dots, 0]^T$, indicating a normal sample.

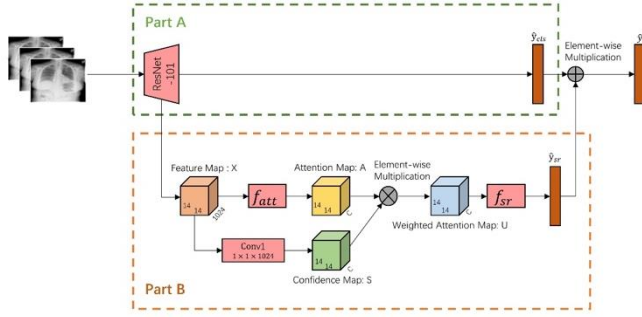


Figure 2." Overall framework of SRN. The part A follows the structure of ResNet-101 and learns one independent classifier for each tag. The part B captures spatial and semantic relations of tags with attention mechanism.

Fig. 2. shows the structure of the SRN. Part A is resnet-101 [20], which uses the block structure proposed in [21]. The output of part A is a predicted semantic tag probability $\hat{y}_{cls} = [\hat{y}_{cls}^1, \hat{y}_{cls}^2, \dots, \hat{y}_{cls}^c]^T$. In part B, the feature map X extracted by the "res4b22 relu" layer of Resnet-101 is used as input to capture the spatial and semantic relationships of the tags. The attention predictor f_{att} consists of a 3-layer convolutional layer, which is used to obtain an attention map A . For each tag T , the area associated with the tag has a higher weight. However, for tags that do not appear in the image, attention map A may highlight the wrong region. Therefore, regularization of the space is required to obtain a weighted attention map $U \in \mathbb{R}^{14 \times 14 \times c}$.

$$U = \sigma(S) \circ A \quad (1)$$

where σ is the sigmoid function, \circ is the element-wise multiplication, and S is the output of the feature map X through a convolutional layer having c convolution kernels and a size of 1×1 . The weighted attention map U is passed through a three-layer convolutional network f_{sr} . The first two layers of f_{sr} can capture the semantic relationship between the tags, and the third layer is used to capture the spatial relationship of the semantic related tags. Finally, the predictive probability vector \hat{y}_{sr} is obtained.

$$\hat{y}_{sr} = f_{sr}(U; \theta_{sr}) \quad (2)$$

where $\hat{y}_{sr} = [\hat{y}_{sr}^1, \hat{y}_{sr}^2, \dots, \hat{y}_{sr}^c]^T \in \mathbb{R}^c$, θ_{sr} is a parameter of a three-layer convolution network.

The tag predictive probability vector is $\hat{y} = 0.5\hat{y}_{cls} + 0.5\hat{y}_{sr}$. The loss function of SRN is,

$$F_1(y, \hat{y}) = \sum_{i=1}^c y^i \log \sigma(\hat{y}^i) + (1 - y^i) \log(1 - \sigma(\hat{y}^i)) \quad (3)$$

In order to improve the accuracy of predicting normal samples, a binary classifier f_{bcls} is added to distinguish

between normal and abnormal. Each image needs to be labeled with normal or abnormal, training a resnet-101 [13] for binary classifier, and the loss function is a binary cross-entropy loss function:

$$F_2(I, y) = y \log p(Y=1|I) - (1-y) \log p(Y=0|I) \quad (4)$$

The input of f_{bcls} is image I , and the output is the predicted probability \hat{y}_{bc} .

$$\hat{y}_{bc} = f_{bc}(I; \theta_{bc}), \hat{y}_{bc} \in \mathbb{R} \quad (5)$$

Set the threshold to λ , when $\hat{y}_{bc} \geq \lambda$, let $B=1$, otherwise $B=0$.

Finally, the semantic tags predictor generates a semantic tags prediction probability vector $V'_{att}(I)$,

$$V'_{att}(I) = B \cdot \hat{y} \quad (6)$$

When the SRN predicts a normal sample as an abnormal sample, if the binary classifier can get the correct result, that is, $B=0$, then the correct tag probability vector can still be obtained.

B. Report Generator

Different from the traditional method of generating reports directly using image features, we use the tag probability vector $V'_{att}(I)$ as the input to the LSTM. $\{S_1, S_2, \dots, S_L\}$ represents a sequence of words in a sentence. The log-likelihood of the words given their context words and the corresponding image can be written as:

$$\log p(S|V'_{att}(I)) = \sum_{t=1}^L \log p(S_t | S_{1:t-1}, V'_{att}(I)) \quad (7)$$

where $p(S_t | S_{1:t-1}, V'_{att}(I))$ is the probability of generating the word S_t given attribute vector $V'_{att}(I)$ and previous words $S_{1:t-1}$. The LSTM model is responsible for generating each word, which takes the tag probability vector $V'_{att}(I)$ and the word sequence $S = \{S_0, S_1, \dots, S_{L+1}\}$ as inputs, where S_0 is a special start word and S_{L+1} is a special END token. Each word has been represented as a one-hot vector S_t of dimension equal to the size of words dictionary. At time step $t=0$, the input of LSTM is $W_{img} V'_{att}(I)$, where W_{img} is the learnable tags embedding weights. Both memory cell c_0 and hidden layer h_0 are initialized to $\vec{0}$. From time step $t=1$ to $t=L$, the input of LSTM is $W_s S_t$, where W_s represents the learnable word embedding weights. At time step $t=L+1$, the generated target word is END.

Our training objective is to learn all parameters in LSTM by minimizing the following cost function:

$$L = -\frac{1}{N} \sum_{i=1}^N \sum_{t=1}^{L^{(i)}+1} \log p_t(S_t^{(i)}) + \lambda_\theta \cdot \|\theta\|_2^2 \quad (8)$$

where N is the number of training samples, and $\mathbf{L}^{(i)}$ is the sentence length of the i -th training sample. $\mathbf{p}_i(\mathbf{S}_i^{(i)})$ represents the activation function of the softmax layer in the LSTM model, θ is the model parameter, and $\lambda_2 \|\theta\|_2^2$ is the regular term.

C. Training Scheme

We train the network in multiple steps. First, we train only the binary classifier on the target dataset, stochastic gradient descend algorithm is used for training, with a batch size of 24, a weight decay of 0.00001, and learning rate is set as 0.0001. Secondly, we train the SRN, the training scheme used is the same as [19], which employs stochastic gradient descend algorithm for training, with a batch size of 24, a momentum of 0.9, and weight decay of 0.0005. The initial learning rate is set as 0.001, and decreased to 1/10 of the previous value whenever validation loss gets saturated, until 10-5. Thirdly, we fix binary classifier and SRN, and focus on training LSTM. The batch size is set to 16. The tags embedding size, word embedding size and hidden state size are all set to 256 in all the experiments. The learning rate is set to 0.001. The dropout rate is set to 0.5.

IV. EXPERIMENTS

A. Dataset

The data set was obtained from the chest X-ray data of a hospital in Shanghai in 2015. After removing the lateral radiographs and some irregular radiographs, a total of 19985 radiographs and corresponding Chinese radiology reports were obtained. We first select the sentences related to the lungs in the radiology report, and preprocess the sentences through tokenizing by Jieba [22] and filtering tokens of frequency no less than 2 as vocabulary, which results in 356 unique tokens. Among the vocabulary, the 40 most frequently occurring tokens associated with abnormal symptoms were selected as the semantic tag for each radiograph. We randomly split the dataset into training, validation and testing by a ratio of 8:1:1.

B. Evaluation Metrics

In our experiments, we adopt three kinds of metrics including BLEU [23], METEOR [24], and ROUGE-L [25], in order to provide a comprehensive evaluation of our model. In general, the higher scores a model achieves under these metrics, its generated reports are more similar to the ground truth reports. BLEU is a metric that has been extensively used for evaluating machine translation algorithms. Concretely, BLEU evaluates the report against the reference report by analyzing their co-occurrences of n -grams. The value of n is usually an integer between 1 and 4. METEOR considers the precision and recall based on the entire corpus, so that the calculated results are more strongly correlated with the results of manual evaluation. ROUGE commonly works for evaluation of text summarization. In this study, we use ROUGE-L version, which basically measures the longest common subsequences between generated report and ground truth report. To summarize, each existing metric has its limitation for evaluating generated reports. Following the

previous works, we calculate the accuracies under all the above metrics and compare them with other models.

C. Baselines

This paper compares SRN+BC-LSTM model with several baseline models, including CNN-RNN [9], Soft Att [13], and AdaAtt [14]. The CNN of all baseline models uses Resnet-101, and the 2048-dimensional vector of the penultimate layer output of Resnet-101 is used as the input of the report generation model. In order to reflect the effectiveness of the binary classifier, this paper also compares the model without the binary classifier (SRN-LSTM) with the SRN+BC-LSTM model.

D. Results and Analyses

Before the start of the experiment, we aim to study the effect of λ , which is the threshold for binary classifier. Since λ determines whether the result of the binary classifier prediction is abnormality or normality, it affects the performance of our model. We tune λ from 0.1 to 0.8 on the TJU X-Ray dataset and the results are shown in Fig. 3. When λ is relatively small, the poor performance is achieved. With the increase of λ , the performance is improved until reaching the peak at $\lambda = 0.3$. Further raising λ will result in a drop. In the following experiments, we set $\lambda = 0.3$.

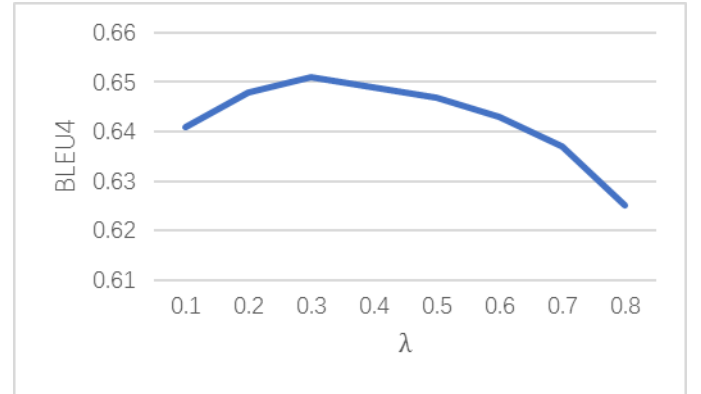


Figure 3. BLEU4 of the generated report when λ is used as the threshold for the binary classifier. The value of λ is in the range of 0.1 to 0.8 and the interval is 0.1.

Table I shows the comparison of the SRN+BC-LSTM with the baseline model on multiple evaluation indicators. Most importantly, SRN+BC-LSTM outperforms all baseline models. The Soft Att [13] and AdaAtt [14] all use the attention mechanism, but the attention is guided by the corresponding report of the radiograph, and they all directly use the image features for the generation of reports. In the SRN-LSTM model, attention is generated based on the semantic tags of the radiographs, which are the keywords in the report. Attention will focus on the areas associated with the semantic tags, so the model can better learn the relationship between the radiographs and the semantic tags. In addition, the semantic tag probability vector generated by SRN is used to generate reports, which can better express the semantic information in the radiograph and finally generate

higher quality reports. The SRN+BC-LSTM model combined with the binary classifiers is slightly better than the SRN-LSTM model in each indicator. It is proved that the addition of the binary classifier improves the accuracy of predicting normal samples, making the generated report more accurate.



The first line of Table II illustrates that the report generated using the semantic tag contains more keywords appearing in the ground truth report than the report generated using the image feature. This comparison demonstrates that

the use of semantic tags to generate reports enables the model to better learn the mapping between radiographs and reports, and to capture more effective semantic information. The second line of the table shows that SRN+BC-LSTM can still get accurate reports when other models generate incorrect reports. It demonstrates that the addition of the binary classifier improves the accuracy of the model prediction for normal samples.

TABLE I. " BLEU-1,2,3,4, ROUGE-L, AND METEOR METRICS COMPARED WITH OTHER BASELINE MODELS AND OUR MODELS ON TJU X-RAY DATASET.

Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE-L	METEOR
CNN-RNN[9]	0.669	0.629	0.602	0.576	0.749	0.436
Soft Att[13]	0.717	0.671	0.638	0.610	0.774	0.461
AdaAtt[14]	0.741	0.687	0.652	0.628	0.789	0.473
SRN-LSTM	0.754	0.706	0.671	0.639	0.801	0.484
SRN+BC-LSTM	0.762	0.716	0.681	0.651	0.809	0.490

TABLE II. " EXAMPLES OF GROUND TRUTH REPORT AND GENERATED REPORTS BY BASELINE MODELS AND OUR MODELS. HIGHLIGHTED WORDS ARE WORDS THAT APPEAR IN THE GROUND TRUTH REPORT. THE FIRST LINE IS A NORMAL SAMPLE, AND THE SECOND LINE IS AN ABNORMAL SAMPLE.

	Ground Truth	CNN-RNN	Soft Att	AdaAtt	SRN-LSTM	SRN+BC-LSTM
 Abnormal	两肺纹理增多，两肺下野见斑片状高密度影，右肺门增浓，结构欠清，右上纵隔及中纵隔见团片状钙化影。	两肺纹理增多，左肺下野见模糊影。	两肺纹理增多，右肺下野见少许模糊影。	两肺纹理增多，右肺门稍增大。	两肺纹理增多，其间见散在斑片状模糊影，右肺门增大、增浓，结构欠清。	两肺纹理增多，右肺门增浓，结构欠清，右肺下野见斑片状模糊影，边缘模糊。
 Normal	两肺纹理增多，未见明显异常实变影。	两肺纹理增多、模糊。	两肺纹理增多，左肺下野见斑片模糊影。	两肺纹理增多，右上纵隔增宽。	两纹理增多，右肺下野见条索影。	两肺纹理增多，未见明显异常实变影。

V. CONCLUSION

In this paper, we study how to automatically generate pulmonary radiology reports for radiographs, with the goal to help radiologists produce reports more accurately and efficiently. We introduce a model that can explore semantic information from radiographs through CNN based on attention mechanism and binary classifier to generate reports. Experiments have shown that reports generated by semantic tags are of higher quality than reports generated by image features. By combining the binary classifier to distinguish between normal and abnormal samples, the model can generate more accurate reports. Experiments on a Chinese chest x-rays dataset demonstrate the effectiveness of our proposed model.

ACKNOWLEDGMENT

We would like to thank reviewers for their comment. This work is supported by the Fundamental Research Funds for the Central Universities (No. 22120180117).

REFERENCES

- [1] B. Jing, P. Xie, and E. Xing, "On the automatic generation of medical imaging reports," arXiv preprint arXiv:1711.08195, 2017.
- [2] Y. Li, X. Liang, Z. Hu, and E. P. Xing, "Hybrid retrieval generation reinforced agent for medical image report generation," in Advances in Neural Information Processing Systems, 2018, pp. 1530–1540.
- [3] Y. Xue, T. Xu, L. R. Long, Z. Xue, S. Antani, G. R. Thoma, and X. Huang, "Multimodal recurrent model with attention for automated radiology report generation," in International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, 2018, pp. 457–466.
- [4] D. Demner-Fushman, M. D. Kohli, M. B. Rosenman, S. E. Shooshan, L. Rodriguez, S. Antani, G. R. Thoma, and C. J. McDonald, "Preparing a collection of radiology examinations for distribution and retrieval," Journal of the American Medical Informatics Association, vol. 23, no. 2, p. ocv080, 2015.
- [5] X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, and R. M. Summers, "Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 2097–2106.
- [6] J. Irvin, P. Rajpurkar, M. Ko, Y. Yu, S. Ciurea-Ilcus, C. Chute, H. Marklund, B. Haghighi, R. Ball, K. Shpankaya et al., "Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison," in Thirty-Third AAAI Conference on Artificial Intelligence, 2019.

- [7] X. Wang, Y. Peng, L. Lu, Z. Lu, and R. M. Summers, "Tienet: Text-image embedding network for common thorax disease classification and reporting in chest x-rays," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 9049–9058.
- [8] J. Krause, J. Johnson, R. Krishna, and L. Fei-Fei, "A hierarchical approach for generating descriptive image paragraphs," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 317–325.
- [9] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 3156–3164.
- [10] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," arXiv preprint arXiv:1409.0473, 2014.
- [11] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using rnn encoder-decoder for statistical machine translation," arXiv preprint arXiv:1406.1078, 2014.
- [12] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in Advances in neural information processing systems, 2014, pp. 3104–3112.
- [13] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in International conference on machine learning, 2015, pp. 2048–2057.
- [14] D. Kornack and P. Rakic, "Cell Proliferation without Neurogenesis in Adult Primate Neocortex," *Science*, vol. 294, Dec. 2001, pp. 2127–2130, doi:10.1126/science.1065467.
- [15] Q. Wu, C. Shen, L. Liu, A. Dick, and A. Van Den Hengel, "What value do explicit high level concepts have in vision to language problems?" in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 203–212.
- [16] T. Yao, Y. Pan, Y. Li, Z. Qiu, and T. Mei, "Boosting image captioning with attributes," in Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 4894–4902.
- [17] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.
- [18] F. Zhu, H. Li, W. Ouyang, N. Yu, and X. Wang, "Learning spatial regularization with image-level supervisions for multi-label image classification," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 5513–5522.
- [19] Y. Wei, W. Xia, J. Huang, B. Ni, J. Dong, Y. Zhao, and Yan, "Cnn: Single-label to multi-label," arXiv preprint arXiv:1406.5726, 2014.
- [20] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.
- [21] K. He, X. Zhang, S. Ren, and J. Sun, "Identity mappings in deep residual networks," in European conference on computer vision. Springer, 2016, pp. 630–645.
- [22] J. Sun, "jieba Chinese word segmentation tool," 2012.
- [23] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in Proceedings of the 40th annual meeting on association for computational linguistics. Association for Computational Linguistics, 2002, pp. 311–318.
- [24] A. Lavie and A. Agarwal, "Meteor: An automatic metric for mt evaluation with improved correlation with human judgments." 2005.
- [25] C. Flick, "Rouge: A package for automatic evaluation of summaries," in Workshop on Text Summarization Branches Out, 2004.
- [26] Huang X, Fang Y, Gu M. Classification of Chest X-ray Disease Based on Convolutional Neural Network[J/OL]. *Journal of System Simulation*. <http://kns.cnki.net/kcms/detail/11.3092.v.20190416.1307.019.html>
- [27] Yan, F., Huang, X., Yao, Y., Lu, M., & Li, M. (2019). Combining LSTM and DenseNet for Automatic Annotation and Classification of Chest X-Ray Images. *IEEE Access*, 7, 74181–74189.