Concept detection based on multi-label classification and image captioning approach - DAMO at ImageCLEF 2019

Jing Xu[1], Wei Liu[2], Chao Liu[2], Yu Wang[2], Ying Chi[2], Xuansong Xie[2], and Xiansheng Hua[2]

[1] Beihang University, Beijing, China
xujing212@buaa.edu.cn
[2] Alibaba Group DAMO Academy AI Center, China
{vivi.lw,maogong.lc,tonggou.wangyu,xinyi.cy, xingtong.xxs,xiansheng.hxs}@alibaba-inc.com

Abstract. Medical image captioning is an important and challenging task, which covers computer vision and natural language processing. This ImageCLEF 2019 [6] Caption competition is dedicated to research this field. The purpose of this year challenge is using radiological images to detect the concepts representing the key information. In this paper, we illustrate the proposed method to address the issue, based on multi-label classification model and CNN-LSTM architecture with attention mechanism. We also perform a detailed analysis and processing for the overall dataset and demonstrate performance with the baseline in the caption prediction task. In final evaluation, we completed 9 submissions and ranked second among 12 participants with our best mean F1-score.

Keywords: Radiology · Image caption · Concept detection · Multi-label classification · Encoder-decoder.

1 Introduction

Medical images, such as radiological images, are widely used in hospital diagnosis and disease treatment. The reading and summarization of medical images is usually performed by experienced medical professionals, and obtaining information from radiological medical images is a time-consuming and laborious task. Therefore, it is essential to automatically and efficiently extract vital information from medical images. ImageCLEF 2019 Caption [11] is the third year of the challenge, starting in 2017, to analyze and solve the problem of medical image caption. The organizing committee provided a large corpus of medical radiology images and UMLS (Unified Medical Language System) [1] concepts pairs, and the purpose of this task is to detect the relevant concepts based on the visual

radiology images. Evaluation criteria is conducted in terms of F1-score between concepts predicted and ground truth concepts.

Inspired by the recent successes of convolutional architectures on other end-to-end frameworks [3, 5, 14, 16], we study convolutional architectures for the task of image concept detection. Specifically, we handle each concept sequence corresponding to each radiological image as a set of labels, and attempt to build a multi-label classification network to solve the task. Furthermore, increasing research has been devoted to image captioning, and almost all of the current proposed methods are under the framework of CNN+RNN [9, 10, 15, 17]. To imitate the human visual attention mechanism, the attention module has been applied. Hence, we adopt the encoder-decoder network, in which a basic CNN is used for the vision feature extractor, and an LSTM is employed to generate sentences due to the ability of learning long term dependencies through a memory cell.

The paper is organized as follow: Section 2 describes the analysis of the overall data, Section 3 introduces the method for the concept detection task, Section 4 demonstrates the details of the experiments and results, and Section 5 discusses and concludes our work.

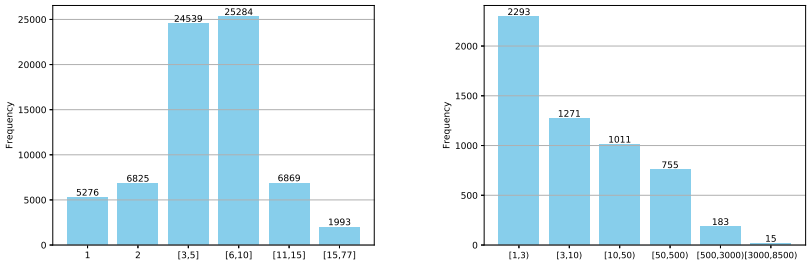
2 Data analysis

In the ImageCLEF 2019 Concept Detection Task, the overall dataset contains 70,786 radiology images of several medical imaging modalities. The images are collected from open access biomedical journal articles (PubMed Central) [13], and the corresponding UMLS concepts that totals 5,528 are extracted from the original image caption. Training dataset includes 56,629 images, and the number of associated concepts is 5216. Validation dataset includes 14,157 images, and the concepts related is 3233. It is worth mentioning that the sequences of the training dataset does not include the total concepts, and 312 concepts appear only in the validation dataset.

To further understand the datasets, we performed statistical analysis to reveal the overall data distribution. The Top-10 concepts descriptions, and the statistics of the length of concept sequence corresponding to each image and the distribution of concept frequency are shown in Table 1 and Fig. 1, respectively. From Table 1, we can see that among the 10 concepts of high frequency, the C0043299 and C1962945, or the C0040395 and C0040405 have similar meanings. Thus, we conducted correlation analysis for all pairs of concepts. See section 4.1 for details. We counted the length of the concept sequences corresponding to each images in dataset. Only one or two concepts of the sequence account for 17.09% of the overall sample. From Fig. 1, the total number of concepts with a frequency less than 3 is 2,293, accounting for 41.48% of the entire concept dictionary, while there are only 15 concepts with a frequency of more than 3,000 times.

Table 1: Top-10 Concept description

Concept ID	Number	Concept
C0441633	8425	diagnostic scanning
C0043299	7906	x-ray procedure
C1962945	7902	radiogr
C0040395	7697	tomogr
C0034579	7564	pantomogr
C0817096	7470	thoracics
C0040405	7164	x-ray computer assisted tomography
C1548003	6428	radiograph
C0221198	5679	visible lesion
C0772294	5677	alesion



(a) Concept length distribution (b) Concept frequency distribution

Fig. 1: (a) We count the length of the concept sequences corresponding to each images in overall dataset, only one or two concepts of the sequence account for 17.09% and the maximum length is 77 and only appears once. (b) The figure shows the frequency distribution of all concepts, the horizontal axis represents the word frequency interval, and the word frequency less than 10 times accounts for 64.47%.

3 Methodology

We design two distinct methods to address the concept detection issue, one is to transform the issue into a multi-label classification problem, the other is to treat it as an image captioning task, using the encoder-decoder network to generate the concepts.

3.1 Multi-label classification approach

Since there is no strong contextual correlation between the concepts of an image, we transform this task into a multi-label classification problem. That is, an image

has several labels. Let l_i be the label of i-th image, as follows:

$$l_i = [c_{i,1}, c_{i,2}, \ldots, c_{i,j}, \ldots c_{i,n}] \tag{1}$$

where n is the total number of labels. If the i-th image has the j-th label, the $c_{i,j}$ is set to 1, else 0.

We utilize the latest deep learning method to solve this problem, which has achieved great success on the field of image processing, such as classification, captioning. Empirically the deeper the network is, the richer the features extracted on different levels. While the drawbacks of gradient vanishing and explosion make it difficult to converge. To overcome this problem, He et al. proposed ResNet [3], which reformulates the layers as learning residual functions with reference to the layer inputs, instead of learning unreferenced functions, and had won the first prize on ImageNet competitions. We choose the pre-trained ResNet-101 model on the ImageNet dataset [8] as backbone in our multi-label classification experiment. The overall process is shown in Fig. 2. An image is firstly preprocessed to adapt to the input of the net, and feed forward to the net to get the output feature vectors. Then passing by a fully connection layer with sigmoid activation function to calculate the probability of each class. If the probability is greater than 0.5, we assert the input image belongs to that class. Finally the predicted labels obtained, which can be reflected back to the original concepts.

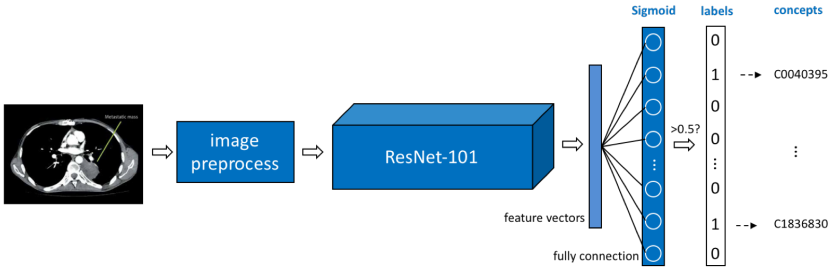


Fig. 2: The overall process of multi-label classification

3.2 Medical image captioning approach

Considering that the concept detection task is to generate text information from the corresponding radiology images, we attempt to address it with the CNN-RNN model framework with attention mechanism. Typically, the model that generates a sequence of concepts will use an encoder to encode the input into a fixed form and use a decoder to decode it into a sequence verbatim.

In our approach, the encoder built upon the pre-trained ResNet-101 is first applied to extract visual features from the input images. We resize the input images normalized by the mean and standard deviation to 224×224 for uniformity,

and then fine-tuned the convolutional blocks on the given medical dataset with a smaller learning rate. We utilize the visual features captured by the *conv_5* convolution block in the ResNet to better describe the local information. Meanwhile, the model combine soft attention mechanism [17] to dynamically select spatial characteristic of the input image.

In decoder, We apply a long short-term memory (LSTM) network [4] that produces a caption by generating one word at every time step conditioned on a context vector capturing the visual information, the previous hidden state and the previously generated concepts. After extracting visual features in CNN, we transform the encoded image to create the initial hidden state h and cell state c for the LSTM decoder. At each decode step, the encoded image and the previous hidden state is used to generate weights for each pixel in the attention network. Finally, the previous generated concept and the weighted average of the encoded image are fed to the LSTM decoder to generate the next concept with the highest score. In addition, we also perform beam search with different beam sizes instead of sampling the maximum probability words.

4 Experiments and Results

4.1 Data preprocessing

Concept association mining It has been found that some concepts have a certain relevance as they often appear simultaneously in different radiological images. Thus, we filter out the high-correlation concept combinations from the high-frequency concepts in training dataset. First, we utilize association rule mining to search for relationships between all concepts defined as a set of items $C = \{c_1, c_2, c_3, \dots c_M\}$, and $I = \{i_1, i_2, i_3, \dots i_N\}$ represents a collection of all training samples, where i_m is the concept sets corresponding to each image. Obviously, $i_N \subset C$. The form of association rule for the concept sets X and Y can be written as: $X \to Y$, where $X \subset C$, $Y \subset C$, and $X \cap Y = \emptyset$. The *support* is the fraction of training set that contain both X and Y, and the $confidence$ represents the measure that how often concepts in Y appear in sample sets that contain X. We suppose σ represents the frequency of occurrence of an item-set. Specifically,

$$support(X \to Y) = \frac{\sigma(X \cup Y)}{N} \tag{2}$$

$$confidence(X \to Y) = \frac{\sigma(X \cup Y)}{\sigma(X)} \tag{3}$$

Second, the concept subsets divided with $support > 0.02$ is a total of 99, and from which we select the combinations contain with the most elements with $confidence > 0.9$. Finally, we define 9 different concept combinations as 9 new concepts, called concept grouping, as shown in Table 2. During the training process, we replace the concepts involved with the 9 new concepts and define the dataset changed as C_g, and then map them in predicted results.

Table 2: Associated concept combination

Concept ID	Concept sets
C1	C0034579;C0040405;C0040395
C2	C0043299;C1962945;C1548003
C3	C0221198;C0772294
C4	C0009924;C0449900
C5	C0817096;C0024109
C6	C0412555;C0041618
C7	C0007876;C1552858;C0728940;C0184905;C0015252
C8	C0013516;C0183129
C9	C0003842;C0002978

Data filtering It is obvious that the dataset is extremely unbalanced through the statistics above. The low-frequency concepts would not only not be learned, but bring great bias to the model. Therefore, we filter out the concepts which are indeed rare. Firstly, we pick out all the concepts which only occurs once and get the corresponding images. Then checking all the related concepts on each image one by one, if the frequencies of all related concepts are once either, the image would be moved out of the dataset. The filtered dataset denotes as $D_{f,1}$ with 163 concepts and 98 images omitted.

At the same time, we also roughly filter out the concepts with frequency less than 3 or 5 times defined $D_{f,3}$ and $D_{f,5}$, to avoid these noises affecting the overall dataset distribution.

Data redivision Since the pre-divided training dataset provided by organizer dose not contain all the concepts need to be learned, we re-divided all the data as follows:

(a) Picking out the images form validation dataset, as mentioned above, which has the concepts that are not occurred in training dataset.

(b) Changing these images slightly by random transformation, such as mirroring, rotation, etc.

(c) Appending these transformed images to the training dataset, while the original validation dataset keeps unchanged.

Image preprocess and normalization In order to make full use of the provided data, several data augmentation operations are introduced in our experiment. Specifically, the image is firstly random flipped horizontally or vertically with a probability of 0.5, and then resized to different scale ranging from 0.6 to 1.2 with bilinear interpolation. Finally, the transformed image is random cropped into the size of 224×224 before input into the backbone net.

4.2 Training parameters

Multi-label classification method Parameters of the last fully connection layer are initialized by MSRA method [2] and the F1-score is utilized as the criteria. The batch size and the max iteration epochs are set to 64 and 100 respectively. We apply the Adam [7] optimizer to fine-tune the model with an initial learning rate of 0.001. The training procedure is shown in the Fig. 3.

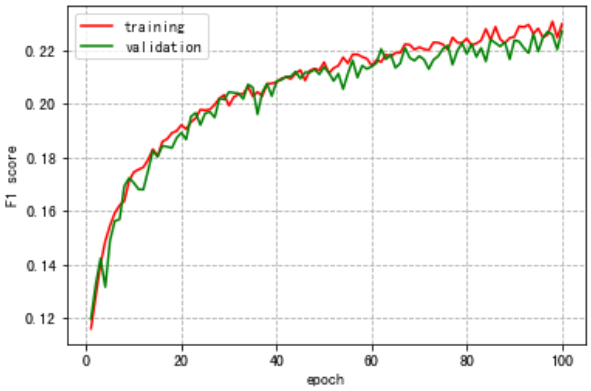


Fig. 3: The training procedure of multi-label classification

CNN-RNN with attention mechanism method With fine-tuning the encoder, the model was trained with cross entropy loss for 30 epochs, batch size of 20 and dropout rate of 0.5. In concept generation, we set the dimensions of all hidden states and word embeddings as 512. We used the Adam optimizer and the learning rates for the CNN and the LSTM were $1e^{-4}$ and $4e^{-4}$ respectively. Early stopping was used to prevent over-fitting when performance on a validation dataset started to degrade. The best model saved was used to predict the sequence of concepts in the test images.

4.3 Specific description in each run

We completed a total of 9 graded submissions before the deadline, the evaluation results for our submitted runs is shown in Table 3 and the specific method for each run is as follows:

Run ID_27103: In this run, we applied multi-label classification model introduced in Section 3.1 and the dataset trained was filtered dataset $D_{f,1}$. We chose the pre-trained ResNet-101 model as backbone in our experiment and performed concept grouping C_g in data preprocessing. Meanwhile, we applied the Adam

Table 3: The results of our submitted runs

Run ID	Method	Mean F1-score	Rank
27103	MLC+$D_{f,1}$+C_g	0.2655	4
27184	MLC+$D_{f,1}$	0.2614	6
26786	CNN+RNN+att+$D_{f,3}$	0.2316	7
27107	27103$\bigcup$26786	0.2135	14
27106	27184$\bigcup$26786	0.2116	15
27188	CNN+RNN+att+RL	0.0590	41
26877	CNN+RNN+att+RL	0.0585	42
27111	CNN+RNN+att+RL	0.0567	43
27158	CNN+RNN+att+RL	0.0537	44

optimizer to fine-tune the model with an initial learning rate of $1e^{-3}$, and the max epoch was 100 in training procedure.

Run ID_27184: This process is similar to ID_27103. We chose the pre-trained ResNet-101 model and the learning rate is $1e^{-3}$. The multi-label classification model was trained with filtered dataset $D_{f,1}$ except for the concept grouping strategy and the max epoch was 60 in training procedure.

Run ID_26786: This run we utilized the CNN-RNN architecture with attention mechanism, based on pre-trained ResNet-101 and LSTM. We used the Adam optimizer and the learning rates for the CNN and the LSTM were $1e^{-4}$ and $4e^{-4}$ respectively. In the training dataset $D_{f,3}$, concepts occurring less frequently than 3 was ignored. Early stopping was used, and the best model saved was used to predict the sequence of concepts in the test images.

Run ID_27107: We combined the predicted results of ID_27103 and ID_26786, that is, the final results in test dataset was the union of two methods for each sample.

Run ID_27106: Similarly, the final results in this run was the union of the predicted results of ID_27184 and ID_26786.

Run ID_27188&26877&27111&27158: These process were based on ID_26786, the pre-trained ResNet-101 was used for the vision model, and an LSTM was employed to generate sentences. The dataset trained was filtered dataset $D_{f,5}$. Otherwise, we made an attempt to apply reinforcement learning [12] in decoder, and the experimental results performed well on validation dataset but were poorly effective on the test dataset.

Table 4: Top-5 groups in Concept Detection Task

Gruop name	Mean F1-score	Rank
AUEB NLP Group	0.2823	1
damo(ours)	0.2655	2
ImageSem	0.2236	3
UA.PT_Bioinformatics	0.2059	4
richard_ycli	0.1952	5

5 Discussion and Conclusion

The evaluation for the caption detection task is conducted using the mean F1-score. As shown in Table 3, among the 61 results submitted by all participants, Run ID_27103 based multi-label classification model has achieved the better performance with the mean F1-score of 0.2655. We mitigate the impact of extreme data imbalance on the model by setting threshold culling noise data, and utilize association rule mining to search for the high-correlation concept combinations. For CNN-LSTM network method, the model did not perform well on the test dataset. Since in the sequences corresponding to the radiology images, the concept exists independently, although some frequent concepts have slight correlation.

Overall, we have completed this challenge in the medical image concept detection task and our group rank second among 12 participants (see Table 4). The method adopted has achieved preliminary results and we will further investigate the medical image captioning task based on higher quality datasets and advanced deep learning algorithms.

References

1. Bodenreider, O.: The unified medical language system (umls): integrating biomedical terminology. Nucleic acids research **32**(suppl_1), D267–D270 (2004)
2. He, K., Zhang, X., Ren, S., Sun, J.: Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In: Proceedings of the IEEE international conference on computer vision. pp. 1026–1034 (2015)
3. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
4. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural computation **9**(8), 1735–1780 (1997)
5. Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4700–4708 (2017)
6. Ionescu, B., Müller, H., Péteri, R., Cid, Y.D., Liauchuk, V., Kovalev, V., Klimuk, D., Tarasau, A., Abacha, A.B., Hasan, S.A., Datla, V., Liu, J., Demner-Fushman, D., Dang-Nguyen, D.T., Piras, L., Riegler, M., Tran, M.T., Lux, M., Gurrin, C.,

Pelka, O., Friedrich, C.M., de Herrera, A.G.S., Garcia, N., Kavallieratou, E., del Blanco, C.R., Rodríguez, C.C., Vasillopoulos, N., Karampidis, K., Chamberlain, J., Clark, A., Campello, A.: ImageCLEF 2019: Multimedia retrieval in medicine, lifelogging, security and nature. In: Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the 10th International Conference of the CLEF Association (CLEF 2019), LNCS Lecture Notes in Computer Science, Springer, Lugano, Switzerland (September 9-12 2019)
7. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
8. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Advances in neural information processing systems. pp. 1097–1105 (2012)
9. Kulkarni, G., Premraj, V., Ordonez, V., Dhar, S., Li, S., Choi, Y., Berg, A.C., Berg, T.L.: Babytalk: Understanding and generating simple image descriptions. IEEE Transactions on Pattern Analysis and Machine Intelligence **35**(12), 2891–2903 (2013)
10. Lu, J., Xiong, C., Parikh, D., Socher, R.: Knowing when to look: Adaptive attention via a visual sentinel for image captioning. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 375–383 (2017)
11. Pelka, O., Friedrich, C.M., García Seco de Herrera, A., Müller, H.: Overview of the ImageCLEFmed 2019 concept prediction task. In: CLEF2019 Working Notes. CEUR Workshop Proceedings, CEUR-WS.org, Lugano, Switzerland (September 09-12 2019)
12. Rennie, S.J., Marcheret, E., Mroueh, Y., Ross, J., Goel, V.: Self-critical sequence training for image captioning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 7008–7024 (2017)
13. Roberts, R.J.: Pubmed central: The genbank of the published literature (2001)
14. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
15. Vinyals, O., Toshev, A., Bengio, S., Erhan, D.: Show and tell: A neural image caption generator. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3156–3164 (2015)
16. Xie, S., Girshick, R., Dollár, P., Tu, Z., He, K.: Aggregated residual transformations for deep neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1492–1500 (2017)
17. Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhutdinov, R., Zemel, R., Bengio, Y.: Show, attend and tell: Neural image caption generation with visual attention. arXiv preprint arXiv:1502.03044 (2015)