

ImageSem at ImageCLEFmed Caption 2019 Task: a Two-stage Medical Concept Detection Strategy

Zhen Guo¹, Xuwen Wang¹, Yu Zhang¹, Jiao Li^{1*}

¹ Institute of Medical Information / Medical Library, Chinese Academy of Medical Sciences & Peking Union Medical College, Beijing 100020, China
li.jiao@imicams.ac.cn

Abstract. This paper presents the participation of the Image Semantics group (ImageSem) of the Institute of Medical Information at the ImageCLEFmed Caption task, which was launched by ImageCLEF 2019. The Concept Detection sub-task aims at identifying 5,528 semantic concepts from 70,786 training images and 10,000 test images. In this study, we proposed the two-stage concept detection strategy, including the medical image pre-classification based on body parts and the transfer learning-based multi-label classification model. We totally submitted 10 runs in the final evaluation. The evaluation results showed that we achieved an F1 Score of 0.2235, which ranked 8th overall. There is still a great room for improving the performance of concept detection from large-scale medical images.

Keywords: Concept Detection; Transfer Learning; Multi-label Classification; Pre-classification.

1 Introduction¹

The ImageCLEF task [1] contributes to enhancing the computational methods for machine understandable medical images [2, 3]. ImageCLEFmed Caption 2019 [4] focus on the concept detection task, which aims to identify the UMLS [5] Concept Unique Identifiers (CUIs) for a given medical image from the biomedical literature. On behalf of the Institute of Medical Information, Chinese Academy of Medical Sciences, our Image Semantics group (ImageSem) participated in the concept detection task of ImageCLEFmed Caption 2019, and submitted 10 runs to the final evaluation.

Fig. 1 shows the workflow and submissions of ImageSem in ImageCLEFmed Caption 2019. On the basis of data analysis and preprocessing, we applied two kinds of concept detection methods. For one thing, we referenced our previous work in ImageCLEFcaption 2018 task [6], and applied the transfer learning-based multi-label classification model to the overall training set to predict high-frequency concepts. For another thing, we proposed a two-stage medical concept detection strategy. Specifically, for a given medical image, a pre-classification model was used to determine which body

¹ Copyright (c) 2019 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0). CLEF 2019, 9-12 September 2019, Lugano, Switzerland.

part the image belongs to, and multiple labels were predicted based on the corresponding multi-label classification model. Finally we collected useful concepts using different concept selection strategies.

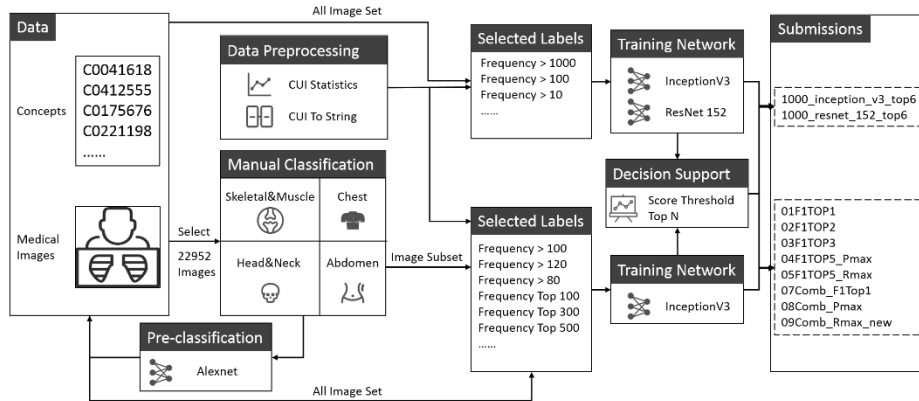


Fig. 1. Workflow of ImageSem at the ImageCLEFmed Caption 2019 Task

This paper is organized as follows. Section 2 analyses the concept detection data set of the ImageCLEFmed Caption 2019 task, and describes our work of data preprocessing. Section 3 presents the methods for concept detection. Section 4 lists all of our submitted runs. Section 5 makes a brief summarization.

2 Data

2.1 Data analysis

The ImageCLEFmed Caption 2019 task provides a subset of the Radiology Objects in Context (ROCO) dataset [7]. To focus on radiology images and non-compound figures, automatic filtering with deep learning systems as well as manual revisions were applied, reducing the dataset to 70,786 radiology images of several medical imaging modalities. It is further divided into a training set (56,629 images) and a validation set (14,157 images). In the concept detection task, a set of CUIs was provided for each image, totally 5,528 annotated concepts (CUIs). Table 1 shows the concept distribution in the overall dataset, and Table 2 shows the top ranked concepts in the training set. It is observed that the high-frequency concepts account for most proportion (about 97.7%) of the overall occurrence in the dataset.

2.2 Data preprocessing

Selecting concepts and images for multi-label classification models. Considering the uneven concept distribution in table 1, we define the problem of detecting high-frequency concepts from medical images as a multi-label classification task. For training the multi-label classification model, we selected 87 CUIs appeared in more than 1,000

medical images, 548 CUIs appeared in 100 to 1,000 images, and 1,263 CUIs appeared in 10 to 100 images, respectively. Then we extracted all the medical images containing high-frequency CUIs from the training set and constructed corresponding subsets, namely F1000, F100, and F10. For each medical image, we filtered out low-frequency CUIs.

Backtracking semantic types of CUIs and manual image annotation for pre-classification. To realize the pre-classification of medical images based on different body parts, we backtracked the semantic types of all CUIs from the UMLS and selected useful TUIs for automatically assigning images to different body parts, e.g. T023, which stands for “Body Part, Organ, or Organ Component” includes multiple body-related CUIs. Then concepts with T023 were automatically extracted and manually classified to corresponding body parts. We extracted images annotated with pre-defined concepts of corresponding body parts, and manually check each image subset.

Table 1. Statistics of the concepts from the training set and the validation set.

Frequency	Number	Proportion of Num	Occurrence	Proportion of occur
0-10	3630	65.67%	9987	2.31%
10-100	1263	22.85%	45630	10.54%
100-1000	548	9.91%	173472	40.09%
1000+	87	1.57%	203664	47.06%
Total	5528	100.00%	432753	100%

Table 2. Top10 high-frequency concepts in the training set.

CUI	Associated Image	UMLS Term
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	5678	visible lesion
C0772294	5677	alesion

3 Methods

In the ImageCLEFcaption 2018 task, we applied two methods to identify multiple concepts for a specific image, including the transfer learning-based multi-label classifica-

tion model and the image retrieval-based topic model [6]. The experimental results indicated that the transfer learning-based multi-label classification method was robust on high-frequency concept detection across different data sets, while the image retrieval-based topic models identified the high-frequency concepts and low-frequency concepts at the same time, but depended heavily on the quality of the retrieved images.

In the ImageCLEFmed Caption 2019 task, for one thing, we continued to use the transfer learning-based multi-label classification model to identify high frequency concepts, for another thing, we paid more attention to the distinction of labels between images of different body parts, and classified medical images based on body parts before the concept detection process.

3.1 Transfer learning-based multi-label classification

The problem of detecting high-frequency concepts from medical images was viewed as a multi-label classification task, and Convolutional Neural Networks (CNNs) was employed to assign one or multiple CUIs to a specific medical image. We used the Inception-V3 [8] and ResNet152 [9], which were pre-trained on the ImageNet datasets including 1.2 million images with more than 1,000 common object classes [10]. The fully-connected layer before the last softmax layer was replaced and the parameters of the pre-trained CNN model were transferred as the initial parameters of our multi-label classification model.

During the training process, we selected 87 CUIs appeared in more than 1,000 medical images in the training set as high-frequency labels, and collected corresponding medical images from the training set, namely F1000 subset. Then we fine-tuned network weights layer by layer and adjust parameters based on the validation set. For a given test image, top N concepts which prediction probability higher than the threshold were selected as the predicted labels.

3.2 Medical image pre-classification based on body parts

By observing the radiology images of the ROCO dataset from the ImageCLEFmed Caption 2019 task, and analyzing the semantic type of some concept CUIs, we were inspired to cluster the images into different categories based on different kinds of body parts.

First, we summarized four body part-related categories based on the medical imaging reading diagnostic atlas [11], including “abdomen”, “chest”, “head and neck” and “skeletal muscle”. Second, we cluster concepts in the training set according to their semantic type, e.g., concepts with the TUI number T023 (Body Part, Organ, or Organ Component) or T029 (Body Location or Region) were automatically extracted and classified to corresponding categories. Third, some part of medical images with annotated concepts in the training set were classified into different categories. We manually double check the images being assigned to different categories and created four body part-based image-concepts subset. Finally, we employed the AlexNet [12] model to automatically classify the rest of medical images in the training set to different categories, as well as the validation set and the test set, which achieved the best accuracy of 84.73%

on the validation set. We had also applied other networks to perform pre-classification, such as the ResNet152 and the Inception V3, however, the complex network structure showed no significant advantage in the classification performance. Table 3 shows the distribution of medical images in different body part categories. Then we could train multi-label classification models on different medical image categories, respectively

Table 3. Statistics of medical images pre-classified into different body part categories.

Dataset	Abdomen	Chest	Head and neck	Skeletal Muscle	Total
Manual annotated	7546	5406	6000	4000	22952
Training	19430	12458	15445	9296	56629
Validation	4802	3040	4003	2312	14157
Test	3578	2277	2607	1538	10000

3.3 Two-stage medical concept detection

On the basis of the above works, we proposed a two-stage medical concept detection model. For a given medical image, the computer will firstly determine which body part the given image belongs to, after the pre-classification step, multiple labels will be predicted based on the corresponding multi-label classification model, the Inception V3 model we used in this study. Different concept selection strategies were also applied to different categories, such as using concept of frequency higher than 100, output top N concepts, or concepts with score above a specific threshold, etc. Then we combined the best output of different categories, which evolved plenty of combinations.

4 Submitted Runs

We submitted the following 10 runs of concept detection to the ImageCLEFmed Caption 2019 task (see Table 4):

Table 4. Submission runs by the ImageSem group in ImageCLEFmed Caption 2019 task

Submission Run	Rank overall	Mean F1 Score
F1TOP1.txt	8	0.2235690
F1TOP2.txt	9	0.2227917
F1TOP5_Pmax.txt	10	0.2216225
F1TOP3.txt	11	0.2190201
07Comb_F1Top1.txt	12	0.2187337
F1TOP5_Rmax.txt	13	0.2147437
08Comb_Pmax.txt	18	0.1912173
09Comb_Rmax_new.txt	40	0.1121941
yu_1000_inception_v3_top6.csv	52	0.0009450
yu_1000_resnet_152_top6.csv	53	0.0008925

Run1 (F1TOP1): This submission employed the two-stage concept detection strategy, in which medical images were firstly pre-classified into different body parts using Alexnet, then multiple concepts were predicted for the given image using multi-label classification models trained on the corresponding image subset. The max epoch was set to 30 and the learning rate was set to 0.001. For the images in the test set, we selected concepts with frequency above 100 in the training set as the training labels. If the given image was classified to the “abdomen” or the “chest” subset, output the top7 concepts of corresponding multi-label classification model. If the given image belongs to the “head & neck” or the “skeletal muscle” subset, output the top 5 concepts. Finally, we combined all of the selected concepts as overall results.

Run2 (F1TOP2): The same training process as the F1TOP1 except that we selected the top 5 concepts for the images belongs to the “abdomen” subset, concepts with score above 0.2 for the “chest” subset, top 7 concepts for the “head & neck” subset and concepts with score above 0.1 for the “skeletal muscle” subset.

Run3 (F1TOP5_Pmax): The same training process as the F1TOP1 except that we selected the top 5 concepts for the images belongs to the “abdomen”, the “chest” and the “head & neck” subset, and the top 3 concepts for the “skeletal muscle” subset.

Run4 (F1TOP3): The same training process as the F1TOP1 except that we selected the top 10 concepts for the images belongs to the “abdomen”, the top 5 concepts for the “chest” subset, concepts with score above 0.1 for the “head & neck” subset, and the top 7 concepts for the “skeletal muscle” subset.

Run5 (07Comb_F1Top1): The same training process as the F1TOP1 except that we selected the top 7 concepts for the images belongs to the “abdomen”, concepts with score above 0.3 for the “chest” subset, the top 5 concepts for the “head & neck” subset, and concepts with score above 0.25 for the “skeletal muscle” subset.

Run6 (F1TOP5_Rmax): The same training process as the F1TOP1 except that we selected the top 10 concepts for the images belongs to the “abdomen”, concepts with score above 0.1 for the “chest” subset, the top 7 concepts for the “head & neck” subset, and the top 10 concepts for the “skeletal muscle” subset.

Run7 (08Comb_Pmax): The same training process as the F1TOP1 except that we selected the top 3 concepts for the images belongs to the “abdomen”, the “chest”, the “head & neck” and the “skeletal muscle” subset. The above combination of parameters achieved the best precision rate in our validating experiments.

Run8 (09Comb_Rmax_new): The same training process as the F1TOP1 except that we selected the concepts with score above 0.05 for the images belongs to the “abdomen”, the “chest”, the “head & neck” and the “skeletal muscle” subset. The above combination of parameters achieved the best recall rate in our validating experiments.

Run9 (yu_1000_inception_v3_top6): This submission utilized the transfer learning-based multi-label classification method, which is using the Inception V3 model pre-trained on the ImageNet dataset to perform multi-label classification. The batch size was set to 20, the max epoch was set to 30 and the learning rate was set to 0.003. For the images in the test set, we selected 87 concepts with frequency above 1000 in the training set as the training labels, and output the top 6 concepts for each test image.

Run10 (yu_1000_resnet_152_top6): This submission employed the transfer learning-based concept detection using the ResNet152 model pre-trained on the ImageNet data

set. The batch size was set to 20, the max epoch was set to 30 and the learning rate was set to 0.003. For the images in the test set, we also selected 87 concepts with frequency above 1000 in the training set as the training labels, and output the top 6 concepts for each test image.


Image ID: ROCO_CLEF_16018		
	GT Concepts C0043299; Diagnostic radiologic examination C1962945; Radiographic imaging procedure C1548003; Diagnostic Service Section ID - Radiograph C0025066; Mediastinum C0087086; Thrombus C0817096; Chest C0003842; Arteries	Predict Concept C0817096; Chest C0024109; Lung C0043299; Diagnostic radiologic examination C1962945; Radiographic imaging procedure C1548003; Diagnostic Service Section ID - Radiograph

Fig. 2. An example of concept detection from the validation set of the ImageCLEFmed Caption 2019 task. The GT concepts were ground truth provided by the ImageCLEF organizers, while the Predict Concepts were results of our two-stage medical concept detection model.

Fig. 2 shows an example of concept detection from the validation set of the ImageCLEFmed Caption 2019 collection. The predicted concepts matched four labels (in red) with the ground truth concepts, while the unmatched concept (C0102410983129; Lung) was also meaningful to the given image. The good data quality, as well as the pre-classification based on body parts contribute to the preferable performance on detecting semantic concepts from large-scale medical images. In summarization, we achieved an F1 score of 0.2235, ranked 8th in the overall submission results, but there is still a great room for improvement in the further research.

5 Conclusions

This paper presents the participation of the Image Semantics group (ImageSem) at the ImageCLEFmed Caption 2019 task. We submitted 10 runs in the concept detection task. Multiple concepts were identified for interpreting medical images by the two-stage concept detection strategy, including the medical image pre-classification based on body parts and the transfer learning-based multi-label classification. The evaluation results showed that we achieved an F1 Score of 0.2235, which was superior to our former achievement in ImageCLEFcaption 2018. The reason for the improvement may due to the good data quality, as well as the pre-classification of medical images based on pre-defined body part categories.

However, the work of semantic concept detection on large-scale open medical images still needs further research, and we will try to seek more useful semantic clues from external labelled data.

6 Acknowledgement

This study was supported by the Non-profit Central Research Institute Fund of Chinese Academy of Medical Sciences (Grant No. 2018-I2M-AI-016, Grant No. 2017PT63010 and Grant No. 2018PT33024); the National Natural Science Foundation of China

(Grant No. 81601573) and the Fundamental Research Funds for the Central Universities (Grant No. 3332018153).

References

1. Ionescu, B., Muller, H., Peteri, 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., Rodriguez, C.C., Vasilopoulos, 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).
2. Eickhoff, C., Schwall, I., García Seco de Herrera, A., Müller, H.: Overview of ImageCLEFcaption 2017 - the image caption prediction and concept extraction tasks to understand biomedical images. In: CLEF 2017 Working Notes. CEUR Workshop Proceedings, CEUR-WS.org <<http://ceur-ws.org>>, Dublin, Ireland (2017).
3. García Seco de Herrera, A., Eickhoff, C., Andrearczyk, V., Müller, H.: Overview of the ImageCLEF 2018 caption prediction tasks. In: CLEF 2018 Working Notes. CEUR Workshop Proceedings, CEUR-WS.org <<http://ceur-ws.org>>, Avignon, France (2018).
4. Pelka, O., Friedrich, C.M., García Seco de Herrera, A., Müller, H.: Overview of the ImageCLEFmed 2019 concept detection task. In: CLEF 2019 Working Notes. CEUR Workshop Proceedings, CEUR-WS.org <<http://ceur-ws.org/Vol-2380/>>, ISSN 1613-0073, Lugano, Switzerland (2019).
5. Aronson, A.R.: Effective mapping of biomedical text to the umls metathesaurus: the metamap program. In: Proceedings.AMIA Symposium, pp. 17-21 (2001).
6. Zhang, Y., Wang, X., Guo, Z., Li, J.: ImageSem at ImageCLEF 2018 caption task: image retrieval and transfer learning. In: CLEF 2018 Working Notes. CEUR Workshop Proceedings, CEUR-WS.org <<http://ceur-ws.org>>, Avignon, France (2018).
7. Pelka, O., Koitka, S., Rückert, J., Nensa, F., Friedrich, C.M.: Radiology Objects in COntext (ROCO): a multimodal image dataset. In: Stoyanov D. et al. (eds) Intravascular Imaging and Computer Assisted Stenting and Large-Scale Annotation of Biomedical Data and Expert Label Synthesis. LABELS 2018, CVII 2018, STENT 2018. Lecture Notes in Computer Science, vol.11043. Springer, Cham (2018).
8. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition, pp. 2818-2826 (2016).
9. He K, Zhang X, Ren S, et al.: Deep Residual Learning for Image Recognition. Computer Vision and Pattern Recognition, 70-778(2016).
10. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, ZH., Karpathy, A., Khosla, A., Bernstein, M., Berg, A.C., Li, FF.: Imagenet large scale visual recognition challenge. International Journal of Computer Vision 115(3), pp.211-252 (2015).
11. Ding, J., Wang, X.: Medical imaging reading diagnostic atlas. 2nd edn. People's Medical Publishing House, Beijing (2013).
12. Ding, L., Li, H., Hu, C., Zhang, W., Wang, S.: ALEXNET feature extraction and multi-kernel learning for object oriented classification. J ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, pp.277-281(2018).