

ImageSem at ImageCLEF 2018 Caption Task: Image Retrieval and Transfer Learning

Yu Zhang¹, Xuwen Wang¹, Zhen Guo, Jiao Li*

Institute of Medical Information, 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 ImageCLEF 2018 caption task. We participated in both of the concept detection and the caption prediction tasks, with submitting 15 runs in total. In this study, we applied LIRE, an open source Lucene Image Retrieval, to index 222,314 images in training and 9,938 images in test sets. In concept detection subtask, we retrieved the similar images in the training set and applied Latent Dirichlet Allocation (LDA) for clustering concepts of the similar images. The transfer learning method was integrated to solve multi-label annotation in the concept detection task. In caption prediction, we used image retrieval strategies by tuning the parameters: the top similar images and number of candidate concepts. In the evaluation, ImageSem achieved the best F1 Score of 0.0928 in the concept detection subtask and the Mean BLEU score of 0.2501 in the caption prediction subtask.

Keywords: Concept Detection; Caption Prediction; LDA; Transfer Learning; Multi-label Classification; Image Retrieval.

1 Introduction

The corpus of annotated medical images, interpreting and summarizing the insights of images, are important for medical image processing and machine learning technology application [1,2]. ImageCLEF task aims to promote the computational method development for machine understandable medical images, starting from visual content and textual descriptor alignment [3]. ImageCLEF 2018 caption task [4], part of ImageCLEF 2018 [5], includes two subtasks, namely concept detection and caption prediction [6]. Our team, ImageSem, participated in both tasks. Fig. 1 shows our workflow in ImageCLEF 2018 Caption Task.

The concept detection subtask aims to identify the UMLS [7] Concept Unique Identifiers (CUIs) for a given medical image from the biomedical literature. We proposed approaches including multi-label classification, information retrieval and topic modeling. Convolutional Neural Networks (CNNs) is applied to train multi-label annotation

¹ Yu Zhang and Xuwen Wang contributed equally

of medical images [8,9]. The LIRE search engine is employed for the information retrieval approach [10,11]. The Latent Dirichlet Allocation (LDA) is used for CUIs topic modeling [12].

The caption prediction subtask aims to predict and generate natural language caption for a given medical image. We proposed a retrieval-based method using LIRE on the training set and combined with preferred concepts recognized from the preceding subtask.

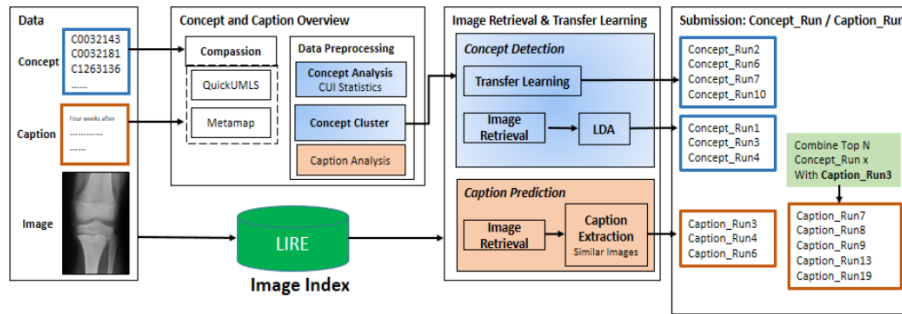


Fig. 1. Workflow of ImageSem team in the Caption Task

This paper is organized as follows: Section 2 introduces the task data and our data preprocessing method. Section 3 describes our methods for concept detection. Section 4 presents our methods for caption prediction. Section 5 summarizes all the runs submitted by our team. Section 6 makes a brief conclusion.

2 Data Preprocessing

2.1 Data overview

The training and test datasets contained 222,314 and 9,938 biomedical images respectively. The images were extracted from scholarly articles in PubMed Central (PMC) [13]. In the concept detection subtask, a set of UMLS CUIs was provided for each image. The image captions were provided in caption generation task. Fig. 2 shows two figures with captions in PMC and assigned concepts (note that the UMLS terms and semantic types were extracted by our team but were not provided by task).

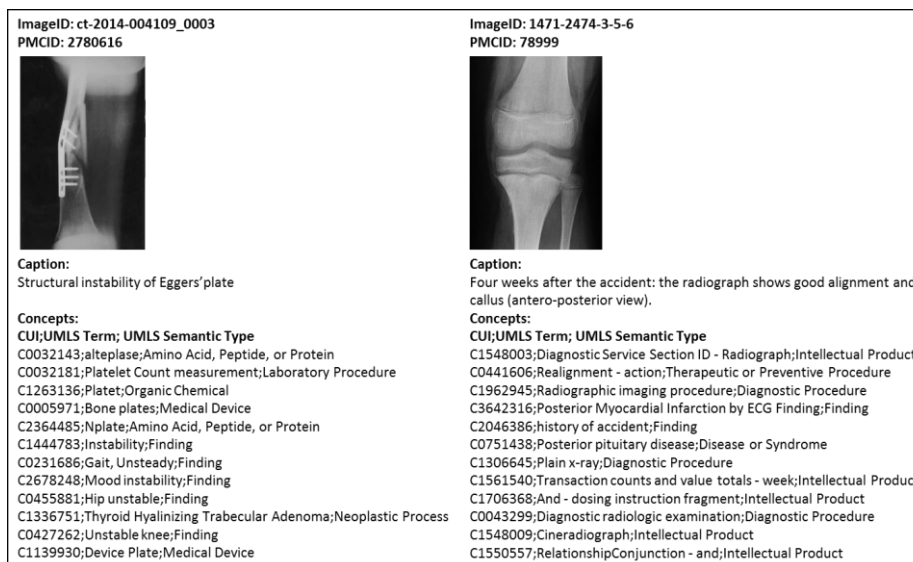


Fig. 2. The task images, its and pre-annotated concepts.

We firstly analyzed the annotated concept frequency distribution in order to better understand the task images. The distribution is important for multi-label training object selection and similar image measurement. The training data includes 222,314 images associated with 111,156 CUIs. Table 1 shows the concept distribution. It can be seen most annotated CUIs (92.19%) were used less than 100 times among 222,314 images. Thus, it is challenging to train a model to learn the annotation patterns of these 102,480 concepts. For the frequent concepts, there are 1312 concepts with frequency greater than 1000. Table 2 shows the top ranked concepts, their annotated image number, and their corresponding UMLS terms. Some general concepts like *medical image* (C1704254) and *image* (C1704922) were highly used but meaningless.

Table 1. Statistics of the concepts in the medical images of training set.

Frequency	Number	Proportion
0-100	102480	92.19%
100-500	6189	5.57%
500-1000	1175	1.06%
1000+	1312	1.18%
Total	111156	100.00%

Table 2. Top frequent concepts in the training set.

CUI	Associated Image	UMLS Term
C1550557	77,003	Relationship Conjunction - and
C1706368	77,003	And - dosing instruction fragment
C1704254	20,165	Medical Image
C1696103	20,164	image - dosage form
C1704922	20,164	Image
C3542466	20,164	Image (foundation metadata concept)
C1837463	19,491	Narrow face
C0376152	19,253	Marrow
C1546708	19,253	Marrow - Specimen Source Codes
C0771936	19,079	Yarrow flower extract

2.2 Image indexing

We used LIRE [10,11] to index the medical images released by ImageCLEF 2018. Table 3 shows the six features used in LIRE, including color and texture features.

Table 3. Descriptors of the features.

Letter	Name
A	Color and edge directivity descriptor
B	Fuzzy color and texture histogram
C	Color histograms
D	Auto color correlation
E	Tamura texture features
F	Gabor texture features

2.3 Concept selection for transfer learning

As for transfer learning, the problem of detecting concepts from medical images was treated as a multi-label classification task. However, too many CUIs without sufficient medical images for training were not feasible for multi-label classification (222314 medical images with 111156 CUIs in the training data). Therefore, we chose to only use the most frequent CUIs for training the model. Eventually, 1312 CUIs, each of which appears in more than 1000 images of the training data, were selected for the multi-label classification.

Further, after analyzing the training data, we found that a number of CUIs co-occur in almost the same set of medical images. To make use of this characteristic, we clustered the CUIs according to their similar scores based on their co-occurrence. The formulation of calculating the similar scores between CUIs is shown as follows:

$$\text{SIMILAR_SCORE}(A, B) = \frac{\text{images}_A \cap \text{images}_B}{\text{images}_A \cup \text{images}_B}$$

Where, $\text{SIMILAR_SCORE}(A, B)$ denotes the similar score between the CUI A and the CUI B. images_A and images_B separately represents the set of medical images in which the CUI A and the CUI B appears. CUIs with similar score more than 0.8 are clustered into the same group. Accordingly, 1312 CUIs are clustered into 459 groups and the first CUI of each group is selected as the representation CUI. Appearance of the representation CUI is the same as the appearance of all the CUIs in its group. Eventually, just the 459 representation CUIs are fed into the model for multi-label classification.

The medical images which contains at least one of the 1312 CUIs were selected for training. And for each image, we re-built its corresponding set of CUIs, only retaining the CUIs inside the 1312 CUIs and mapping them to the representation CUIs of their corresponding groups. Finally, 208595 medical images with 459 representation CUIs were used to train the transfer learning model for multi-label classification.

3 Concept Detection Methods

For the concept detection sub task, we employed three methods to find multiple CUIs for a specific image, including the multi-label classification method, retrieval-based method and the topic modeling method.

In the multi-label classification method, Convolutional Neural Networks (CNNs) was applied to assign one or multiple CUIs from the predefined CUIs label set.

In the retrieval-based method, we used LIRE (Lucene Image Retrieval) to retrieve the most similar images and corresponding CUIs from the training set.

In the topic modeling method, Latent Dirichlet Allocation (LDA) was used to analyze the topic distribution of CUIs from retrieved similar images and their CUIs.

3.1 Multi-label classification with CNN

3.1.1 Inception-v3

In recent years, deep neural network such as convolutional neural networks(CNN) and recurrent neural networks(RNN) have made great success in large-scale image processing, image content recognition, and image caption generation. Inception-v3, a convolutional neural network(CNN) model of Google, is an architecture that often achieves superior performance with low computational cost. The key advantage of Inception-v3 is the factorization of convolution kernel, for example, it can decompose a 7x7 convolution kernel into two one-dimensional kernels(a 1x7 kernel and a 7x1 kernel). Through the factorization of convolution kernels, it can accelerate the training and increase the

depth of the network. In this study, the Inception-v3 model is pre-trained on the ImageNet datasets with more than 1 million images and 1000 classes[14].

3.1.2 Transfer learning for concept detection

However, for our concept detection task which is treated as a multi-label classification problem, directly retraining the whole Inception-v3 model based on the training set needs to take at least a few days. Therefore, we used the pre-trained Inception-v3 based transfer learning method to identify the concepts from medical images. Specifically, we froze the parameters of all the previous layers, removed the last softmax layer and added a fully-connected layer and a sigmoid layer. While training, only the last two layers need to be trained to map the medical images to the CUIs. Totally, 208595 medical images with 459 representation CUIs were fed into the model. Eventually, after getting the predicting results of the test set, we extended the results through replacing the representation CUIs with all the CUIs in their corresponding groups according to the clustering result.

3.2 Image retrieval

LIRE is an open source Java library that provides a simple way to retrieve images and photos based on color and texture characteristics. We used LIRE to create a Lucene index of image features on the whole training set for content based image retrieval (CBIR).

We submitted each query image from the test set to LIRE and selected top 50 visually similar images from the training set. For a given test image, we combined related CUIs of similar images as candidate concepts, then computed a concept score $s(c)$ to determine which concepts to be assigned as semantic labels. In the following concept score equation, α_j denotes the normalization weight of similar figure j , $P(j)$ denotes the probability of figure j and $P(c|j)$ represents the probability of concept c that is assigned to figure j .

$$\text{Concept score: } s(c) = \sum_{j=1}^k \alpha_j \cdot P(j) \cdot P(c|j)$$

$$\text{In which, } P(c|j) = \frac{\text{count}(c,j)}{|c_j|}, \alpha_j = 1 - \frac{f_j - f_{\min}}{f_{\max} - f_{\min}}$$

Then candidate concepts were ranked according to their concept score $s(c)$. We set a threshold τ and select top K CUIs as final related concepts.

Besides, we also considered the method of applying QuickUMLS or Metamap tools to label CUIs on retrieved similar images captions, but by testing, we found some difference between automatic tagging and original provided CUIs in the training set, which may due to different parameter settings or unknown concept expanding strategies. To avoid this uncontrollable noise, we focus on the analysis of provided CUIs.

3.3 Image retrieval with topic model

On the basis of retrieved similar images and candidate CUIs, we employed topic modeling method to select more relevant concept for a given test image. Latent Dirichlet Allocation (LDA) is a widely used generative statistical topic model in natural language processing. In this subtask, we assume concepts related to each image are collected into documents, so each document is a mixture of a number of topics and each concept is attributable to one of the document's topics.

We applied Gensim, a topic modeling Python package to modeling topic distribution on retrieved similar images and candidate CUIs. For a given test image with its retrieved 50 similar images, we collected 50 documents of CUIs as the input of LDA model. According to the topic distribution θ of the current document set, we picked the topic with the highest probability $p(z|D)$ as the candidate topic, and finally selected CUIs from the candidate topic that with probabilities $p(c|z)$ above the threshold ϕ_0 as the final result.

Before submitting the runs we carried out experiments on the training data using highly related concepts detected in CNNs method as a hint for choosing better candidate topics. However, it didn't provide better results. So we submit the normal runs of topic modeling.

4 Caption Prediction Methods

We used retrieval-based method in the caption prediction. The basic assumption is that similar images have similar lingual descriptions/captions.

4.1 Caption selection and combination

For each test image, we used LIRE to retrieve similar images from the training set, and combined the captions of similar images as a new caption of the test image. We tuned the parameter, the number of top similar images, to determine the candidate captions for further combination.

4.2 Concept selection and combination

We combined preferred concepts detected in the preceding concept detection subtask. The preferred concepts including CUIs from CNNs' high score output, as well as CUIs from the output of LDA model. We extracted all the UMLS terms of each CUIs, and combined them with captions generated in the previous section.

5 Submitted Runs

This section provided a detailed description of our runs submitted to ImageCLEF 2018 caption task.

5.1 Runs of concept detection

We submitted the following 7 runs to the Concept Detection subtask (see Table 4):

Table 4. ImageSem performance in the concept detection

Run	Submission ID	Mean F1 Score
Concept_Run10	5583	0.092849
Concept_Run2	5556	0.090867
Concept_Run4	5561	0.090705
Concept_Run1	5554	0.089415
Concept_Run6	5574	0.082799
Concept_Run7	5575	0.066151
Concept_Run3	5558	0.000000

Concept_Run1_submission_ID_5554: We exploited the LIRE to retrieve the 50 most similar images to each medical image of the test set. Then, for a given test image, all the CUIs of its corresponding 50 most similar images were taken as the input of LDA model. Further, we pick the topic with the highest probability $p(z|D)$ as the candidate topic. Finally, we select CUIs from the candidate topic that with probabilities $p(c|z)$ above 0.005 as the CUIs of that test image.

Concept_Run2_submission_ID_5556: Multi-label classification using transfer learning model, based on the pre-trained Inception-v3. The batch size was set to 20 and the learning rate was set to 0.003. While, the training steps were set to 15000. Finally, for a test image, top 10 representation CUIs of the predicting results were selected as the preliminary result, and we extended the preliminary result through replacing the representation CUIs with all the CUIs in their corresponding groups according to the clustering result as the final result of that test image.

Concept_Run3_submission_ID_5558: The same as the Concept_Run1_submission_ID_5554 except that, in the final submission file, all the CUIs were separated by comma.

Concept_Run4_submission_ID_5561: The same as the Concept_Run1_submission_ID_5554 except that of all the CUIs of the given test image's corresponding 50 most similar images, only the CUIs which appears in more than 1000 images of the training data were taken as the input of LDA model.

Concept_Run6_submission_ID_5574: The same as the Concept_Run2_submission_ID_5556 except that the training steps were set to 5000

Concept_Run7_submission_ID_5575: The same as the Concept_Run2_submission_ID_5556 except that for a test image, top 20 representation CUIs of the predicting results were selected as the preliminary result.

Concept_Run10_submission_ID_5583: The same as the Concept_Run2_submission_ID_5556 except that the training steps were set to 25000.

5.2 Runs of caption prediction

We submitted the following 8 runs to the Concept Detection subtask (see Table 5):

Table 5. ImageSem performance in the caption generation

Run	Submission ID	Mean BLEU Score
Caption_Run4	5527	0.250086
Caption_Run9	5546	0.234312
Caption_Run13	5548	0.227806
Caption_Run19	5552	0.227065
Caption_Run13	5526	0.22443
Caption_Run7	5531	0.222768
Caption_Run8	5545	0.222081
Caption_Run6	5528	0.196338

Caption_Run3_submission_ID_5526: We exploited the LIRE to retrieve the 50 most similar images to each medical image of the test set. Then, for a given test image, the captions of its top 2 most similar images were concatenated together as the result of that test image.

Caption_Run4_submission_ID_5527: The same as the Caption_Run3_submission_ID_5526 except that for a given test image, the captions of its top 3 most similar images were concatenated together as the result of that test image.

Caption_Run6_submission_ID_5528: The same as the Caption_Run3_submission_ID_5526 except that of the top 2 most similar images to the test image, only the captions of images with similar distance less than 5 to the test image were concatenated together as the result of that test image.

Caption_Run7_submission_ID_5531: The same as the Caption_Run3_submission_ID_5526 except that, for a given test image, we added the top 1 CUI of the predicting result of that test image from the concept detection task based on transfer learning into the final result.

Caption_Run8_submission_ID_5545: The same as the Caption_Run3_submission_ID_5526 except that, for a given test image, we added the top 2 CUI of the predicting result of that test image from the concept detection task based on transfer learning into the final result.

Caption_Run9_submission_ID_5546: The same as the Caption_Run3_submission_ID_5526 except that, for a given test image, we added the top 3 CUI of the predicting result of that test image from the concept detection task based on transfer learning into the final result.

Caption_Run13_submission_ID_5548: The same as the Caption_Run3_submission_ID_5526 except that, for a given test image, we added the top 1 CUI of the predicting result of that test image from the concept detection task based on the retrieval method and LDA into the final result.

Caption_Run19_submission_ID_5552: The same as the Caption_Run3_submission_ID_5526 except that, for a given test image, both the top 1 CUI of the predicting resu

It of the concept detection task based on the retrieval method and LDA and the top 1 CUI of the predicting result based on transfer learning are added into the final result.

6 Conclusions

This paper presents the participation of the Image Semantics group (ImageSem) at the ImageCLEF 2018 caption task. We submitted 7 runs in the concept detection and 8 runs in the caption prediction tasks. The evaluation results showed that we achieved the best F1 Score of 0.0928 in the concept detection subtask and the Mean BLEU score of 0.2501 in the caption prediction subtask. Our methods mainly relied on image retrieval and transfer learning.

In our experiments, we found the ground truth concept annotations were not exactly represent the semantics of the images. It is difficult for error analysis from either computational view or clinical/biomedical view. In the future work, we would like to contribute the corpus construction together with the ImageCLEF committee.

7 Acknowledgement

This study was supported by the National Key Research and Development Program of China (Grant No. 2016YFC0901901 and No. 2017YFC0907500), the Key Laboratory of Medical Information Intelligent Technology Chinese Academy of Medical Sciences, the National Population and Health Scientific Data Sharing Program of China, and the Knowledge Centre for Engineering Sciences and Technology (Medical Centre).

References

1. Interagency Working Group on Medical Imaging Committee on Science, National Science and Technology Council, Roadmap for medical imaging research and development, 2017.12.
2. Litjens, G., Kooi, T., Bejnordi, B. E., Aaa, S., Ciompi, F.: A survey on deep learning in medical image analysis. *Medical Image Analysis* 42(9), 60 (2017).
3. Eickhoff, C., Schwall, I., García Seco de Herrera, A., Müller, H.: Overview of ImageCLEFcaption 2017 - image caption prediction and concept detection for biomedical images. In: *CLEF 2017 Labs Working Notes. CEUR Workshop Proceedings, CEUR-WS.org* <<http://ceur-ws.org>>, Dublin, Ireland (September 11-14 2017).
4. 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 (September 10-14 2018).
5. Ionescu, B., Müller, H., Villegas, M., García Seco de Herrera, A., Eickhoff, C., Andrearczyk, V., Dicente Cid, Y., Liauchuk, V., Kovalev, V., Hasan, SA., Ling, Y., Farri, O., Liu, J., Lungren, M., Dang-Nguyen, DT., Piras, L., Riegler, M., Zhou, LT., Lux, M., Gurrin, C.: Overview of ImageCLEF 2018: Challenges, Datasets and Evaluation. In: *Experimental IR*

- Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Ninth International Conference of the CLEF Association (CLEF 2018). Lecture Notes in Computer Science, Springer, Avignon, France (September 10-14 2018).
6. ImageCLEFcaption Homepage, <http://www.imageclef.org/2018/caption>, last accessed 2018/5/30.
 7. UMLS (Unified Medical Language System) Homepage, <https://www.nlm.nih.gov/research/umls/>, last accessed 2018/5/30.
 8. Jacques C.: Special Issue: Digital Libraries. *Commun. ACM* 39(11), (1996).
 9. Razavian, A. S., Azizpour, H., Sullivan, J., Carlsson, S.: CNN Features Off-the-Shelf: An Astounding Baseline for Recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 512-519. IEEE, Columbus, OH, USA (2014).
 10. LIRE (Lucene Image Retrieval) Homepage, <http://www.lire-project.net/>, last accessed 2018/5/30.
 11. Gan, R., Yin, J.: Using LIRe to Implement Image Retrieval System Based on Multi-feature Descriptor. In: Third International Conference on Digital Manufacturing & Automation, pp. 1014-1017. IEEE, Guilin, China (2012).
 12. Blei DM., Ng AY., Jordan MI.: Latent dirichlet allocation. *J Machine Learning Research Archive* 3, 993-1022 (2003).
 13. PubMed Homepage, <https://www.ncbi.nlm.nih.gov/pmc/>, last accessed 2018/5/30.
 14. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M.: Imagenet large scale visual recognition challenge. *International Journal of Computer Vision* 115(3), 211-252 (2015).