

ImageSem Group at ImageCLEFmed Caption 2021 Task: Exploring the Clinical Significance of the Textual Descriptions Derived from Medical Images

Xuwen Wang¹, Zhen Guo¹, Chunyuan Xu², Lianglong Sun¹ and Jiao Li^{1*}

¹ Institute of Medical Information and Library, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing, 100020, China

² School of Life Science, Beijing Institute of Technology, Beijing, 100081, China

Abstract

This paper presents the work of ImageSem group in the ImageCLEFmed Caption 2021 task. In the concept detection subtask, we employed the transfer learning-based multi-label classification model as our baseline. We also trained multiple fine-grained MLC models based on manually annotated semantic categories, such as Imaging Type, Anatomic Structure, and Findings, which may reveal clinical insights of radiology images. We submitted 9 runs to the concept detection subtask, and achieved the F1 Score of 0.419, which ranked 3rd in the leader board. In the caption prediction subtask, our first method simply combines detected concepts according to the sentence patterns. The second method used a dual path CNN model for matching images and captions. We submitted 4 runs to the caption prediction subtask, and achieved the BLEU score of 0.257, which ranked 6th among the participating teams.

Keywords

Concept detection, caption prediction, multi-label classification, fine-grained semantic labelling

1. Introduction

The medical track of ImageCLEF[1] aims at promoting the research of computer-aided radiology image analysis and interpretation. ImageCLEFmed Caption 2021[2] is one of the ImageCLEFmedical tasks, which focus on mapping visual information of radiology images to textual descriptions. It consists of two subtasks, namely Concept Detection and Caption Prediction. On behalf of the Institute of Medical Information and Library, Chinese Academy of Medical Sciences, our Image Semantics group (ImageSem) participated in both of the two subtasks.

The concept detection subtask aims to identify the UMLS [3] Concept Unique Identifiers (CUIs) for a given radiology image. Following our previous work on ImageCLEF 2019 [4], we employed transfer learning-based multi-label classification (MLC) [5],[6] as our first method for modeling all the concepts in the training set. In order to annotate each image with more meaningful concepts, we manually classified the concepts into three categories according to their UMLS semantic types, namely Imaging Type, Anatomical Structure, and Findings. Then we trained MLC sub models separately for different concept categories as our second method.

The caption prediction subtask asks participants to generate coherent captions for the entirety of an image, which requires higher accuracy and semantic interpretability of expression. We also employed two methods for caption prediction. The first method was the pattern-based combination of concepts identified in the previous task. The second method was based on the dual path CNN model [7], which

CLEF 2021 – Conference and Labs of the Evaluation Forum, September 21–24, 2021, Bucharest, Romania

✉ wang.xuwen@imicams.ac.cn (X. Wang); li.jiao@imicams.ac.cn (J. L.)

🌐 <https://www.imicams.ac.cn> (X. Wang); <https://www.imicams.ac.cn> (J. L.)

🆔 [0000-0003-3022-6513](https://orcid.org/0000-0003-3022-6513) (X. Wang); [0000-0002-7454-0750](https://orcid.org/0000-0002-7454-0750) (Z. Guo); [0000-0001-6391-8343](https://orcid.org/0000-0001-6391-8343) (J. L.)



© 2021 Copyright for this paper by its authors.
Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).



CEUR Workshop Proceedings (CEUR-WS.org)

is commonly used in the image-text retrieval field to match images and captions for instance-level retrieval. This paper is organized as follows. Section 2 describes the data set of the ImageCLEFmed Caption 2021 task. Section 3 presents the methods for concept detection and caption prediction. Section 4 lists all of our submitted runs. Section 5 makes a brief summarization.

2. Dataset

The ImageCLEFmed Caption 2021 task is in its 5th edition this year. Compared with previous years, the released images were strictly limited to radiology, and the number of images and associated UMLS concepts were reduced. There were 222,314 images with 111,156 concepts in 2018 [8], 70,786 radiology images with 5,528 concepts in 2019 [9], 80,747 radiology images with 3,047 concepts in 2020 [10], and 3,256 radiology images with 1,586 concepts and 3,256 captions in 2021. Another improvement of the dataset is that the validation set and test set include real radiology images annotated by medical doctors, which increased the medical context relevance of the UMLS concepts. For one thing, the reduction of concept scope and size lowered the difficulty of concept identification. For another thing, the reduction of image size is not conducive to training large-scale neural networks.

The organizers provided UMLS concepts along with their imaging modality information, for training purposes. We observed that most images were assigned with concepts indicating the diagnostic procedure or medical device, and some images were accompanied by concepts indicating the body part, organ or clinical findings. As shown in Table 1, the high-frequency concepts are concentrated in several specific semantic types. For our experiments, we utilized this feature and manually classified three concept categories for building fine-grained multi-label classification models.

Table 1 High-frequency concepts in the training and validation set of ImageCLEFmed caption 2021 task.

CUI	#Num	Term String	TUI	Semantic Type
C0040398	1400	Tomography, Emission-Computed	T060	Diagnostic Procedure
C0024485	796	Magnetic Resonance Imaging	T060	Diagnostic Procedure
C1306645	627	Plain x-ray	T060	Diagnostic Procedure
C0041618	373	Ultrasonography	T060	Diagnostic Procedure
C0009924	283	Contrast Media	T130	Indicator, Reagent, or Diagnostic Aid
C0577559	274	Mass of body structure	T033	Finding
C0002978	119	angiogram	T060	Diagnostic Procedure
C0221198	108	Lesion	T033	Finding
C1322687	107	Endoscopes, Gastrointestinal Tract, Upper Tract	T074	Medical Device
C0205400	92	Thickened	T033	Finding
C1881358	78	Large Mass	T033	Finding
C0202823	60	Chest CT	T060	Diagnostic Procedure
C0005910	59	Body Weight	T032	Organism Attribute
C0150312	55	Present	T033	Finding
C0180459	53	Disks (device)	T073	Manufactured Object
C0003617	52	Appendix	T023	Body Part, Organ, or Organ Component
C0228134	50	Spinal epidural space	T030	Body Space or Junction
C0016658	47	Fracture	T037	Injury or Poisoning
C0005889	47	Body Fluids	T031	Body Substance
C0227613	47	Right kidney	T023	Body Part, Organ, or Organ Component

3. Methods

This section describes methods we used in two subtasks. Fig. 1 shows the workflow and submissions of ImageSem in ImageCLEFmed Caption 2021.

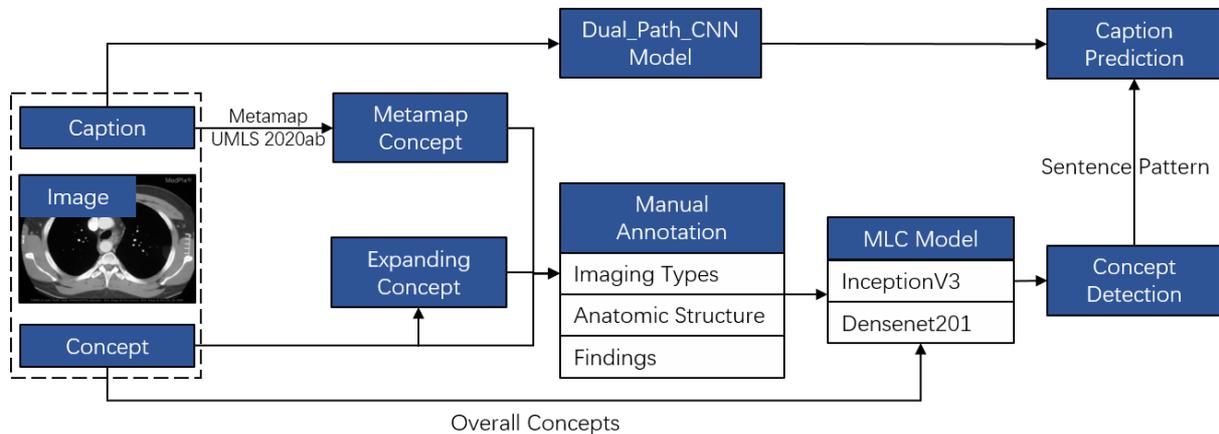


Figure 1: Workflow of ImageSem in the ImageCLEFmed Caption 2021 task

3.1. Concept detection

In the concept detection subtask, for one thing, we employed the transfer learning-based multi-label classification model to identify overall concepts; for another thing, we paid more attention to the distinction of labels with different semantic types, and focus on three major categories of concepts, which may reveal clinical insights of radiology images.

3.1.1. Transfer learning-based multi-label classification

In our previous work, we employed a transfer learning-based multi-label classification model to assign multiple CUIs to a specific medical image. This is a classic approach under the condition of limited tag size and high frequency concepts. In our first method, for modeling overall concepts, we applied the Inception-V3[5] and DenseNet 201[6] which were pre-trained on the ImageNet datasets [11]. 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 MLC model.

During the training process, we collected 1,586 CUIs from both of training set and validation set as our labels. Then we fine-tuned the models on the validation set. For a given test image, concepts of high probabilities above the threshold were selected as the prediction labels. Empirically, we adjusted the threshold gradually from 0.1 to 0.7 on the basis of the validation set.

3.1.2. Fine-grained multi-label classification

In this method, according to the UMLS semantic types, we go further to divide ImageCLEF concepts into four semantic categories, namely Imaging Type (IT), Anatomic Structure (AS), Findings (FDs) and others. Based on the official training set and validation set, we reprocessed the images and associated concepts via our medical image annotation platform.

As shown in Figure 2, for a given radiology image, there are three sources of related concepts. The first one is ImageCLEF concepts annotated by concept extraction tools and medical doctors. These concepts are semantically related, but often incomplete, since many images having only one concept. The second source of concepts are automatically annotated from the given image captions, using the Metamap tool [12] together with UMLS 2020ab. These concepts are more comprehensive, but also introduce noise words. The third source is the expanding concepts that we summarize manually based on the high-frequency ImageCLEF concepts, for labelling convenience purpose.

We invited graduate students majoring in medical imaging to label images with reference to visual information, caption descriptions and the above three sources of concepts. The labeling protocol is that each radiology image was assigned with at least one IT label, zero or more AS labels, and zero or more FDs labels. Specifically, ImageCLEF concepts that are difficult to be classified to the above categories, can be assigned to the ‘Others’.

Then we build three image-concept sub collections for training fine-grained MLC models. These collections have same training and validation images, but differentiate in related concepts. Table 2 shows the distribution of different concept categories.

We verified our MLC models based on the re-annotated validation set. The experimental results showed that our model performs well on the prediction of Imaging Type labels, with F1 score of 0.9273. However, the predictions for the other two kinds of labels are far from satisfactory. One possible reason is that there are few images but too many labels for training. It is intuitively understandable that images of the same or similar cases would have a similar anatomic structure or medical findings label. Whereas the data characteristics of this subtask are obviously not suitable for specific diseases, which raised the difficulty to predict accurate body part, organ, or findings.

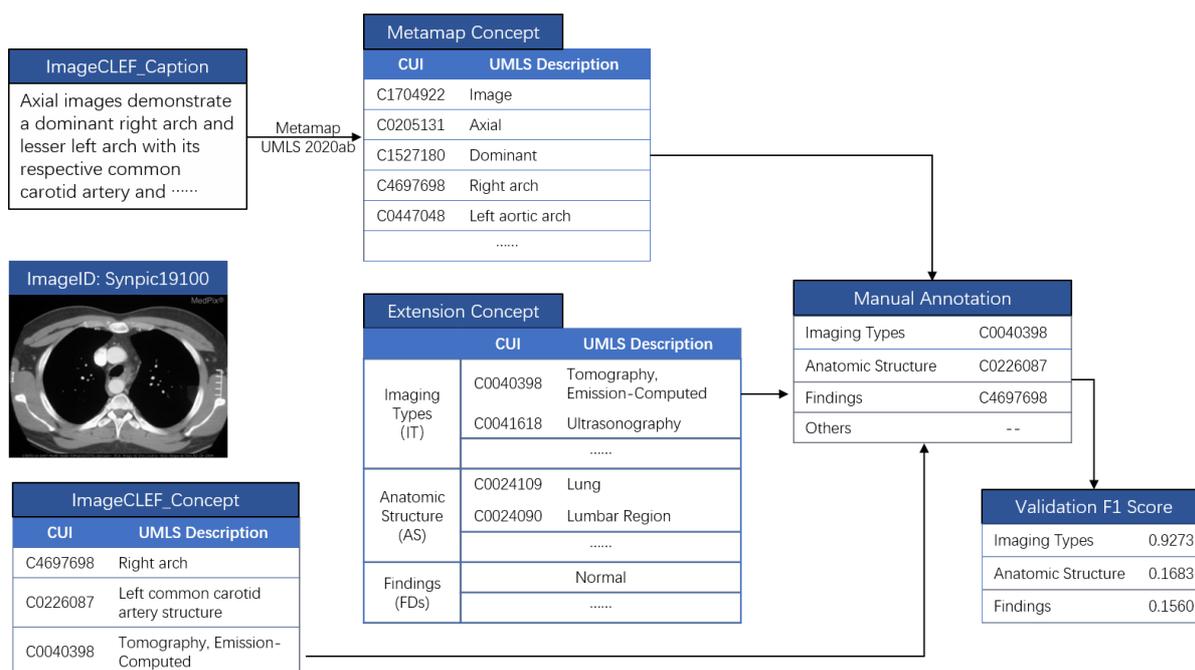


Figure 2: Process of manual re-annotation and fine-grained MLC model training and validation

Table 2 Distribution of concepts from different semantic categories

Category	#Concepts	Concept Sample
Imaging Types	99	C0040398 Tomography Emission-Computed
Anatomic Structure	786	C0228134 Spinal epidural space
Findings	854	C0577559 Mass of body structure

3.2. Caption Prediction

3.2.1. Pattern-based caption generation

For generating reasonable image captions, the first method was the pattern-based combination of concepts identified in the previous task. We designed a simple sentence pattern based on the

characteristic of captions in the training and validation set, see Table 3. Obviously, the accuracy of concept detection results would directly determine the quality of sentence generation.

Table 3 Sentence pattern for caption generation

Pattern	Sample
<image> of <body> demonstrate / show / suggest <findings>	synpic24243 : Sagittal T1-weighted image of the cervical spine demonstrates cord expansion.
<image> demonstrate / show / suggest <findings> in/of/within <body>	synpic19193 : Lateral radiograph of the skull shows lytic lesions in the temporoparietal region.

3.2.2. Image matching for caption prediction

In this method, we employed the algorithm commonly used in the image-text retrieval field to match images and captions for instance-level retrieval. It is based on an unsupervised assumption that every image/test group can be viewed as one class, so each category is equivalent to 1+m (1 image vs m descriptions) samples.

We use the model proposed by Zheng[7], which contains two convolutional neural networks to learn visual and textual representations simultaneously. When testing, we first extract the image feature by image CNN and the text feature by text CNN, and then use the cosine distance to evaluate the similarity between the image and candidate sentences.

- **Data Preparation**

In this field, most existing works use two generic retrieval datasets (Flickr30k and MSCOCO), which have more than 30,000 images. Each image in these datasets is annotated with around five sentences. So we expanded the caption from 1 to 5 sentences per image. Specifically, we first translate the caption into Chinese, Japanese, German, French and then translate back to English. We use GoogleNews-vectors word2vec model trained by Google, which contains 2,000,000 words to get our dictionary. Our dictionary ultimately have 6039 words, each has a 1*300 vector corresponding to it.

- **Train**

Given a sentence, we convert it into code T of size $n * d$, where n is the length of the sentence, and d denotes the size of the dictionary. T is used as the input for the text CNN. Given an image, we resize it to 224×224 pixels, which are randomly cropped.

The training process includes two stages: in the first stage, we use the instance loss to learn fine-grained differences between intra-modal samples with similar semantics. in the second stage, we use the ranking loss to focus on the distance between the two modalities to build the relationship between the image and text.

- **Test**

In this experiment, we use 16,280 sentences from training set and validation set as candidate captions, each sentence is corresponding to its text feature extracted by text CNN. For each test image, we first extract the image feature by image CNN, and then use the cosine distance to evaluate the similarity between the image and candidate sentences.

When we use the model trained on ImageCLEF datasets, we get the almost same top 10 sentences from 16,280 candidate captions, because the features learned by text CNN between each captions is not discriminative. However, when we test it on the model trained by MSCOCO datasets, each query image can get different sentences, but they do not match either.

4. Submitted runs

Table 4 presents the 9 runs we submitted to the concept detection subtask, along with the official rankings. We take the Inception-V3 model trained on overall concepts as a baseline. We tried to submit concepts of the three semantic categories predicted by sub MLC models. The submissions were either by categories or by combining with the baseline results. To our surprise, the

concepts of Imaging Types achieved the best F1 score of 0.419, indicating the high precision and coverage of this kind of concepts in radiology images. As to the concepts from other types and baseline results, they introduce more unmentioned words and reduce the overall score. However, in view of our experience on manual labeling, we believe that some unmentioned words may also be helpful in interpreting the given image. Figure 3 shows two examples of our method on the validation set.

Table 5 shows the 4 runs we submitted to the caption prediction subtask. We take the Dual path CNN model as our baseline, which achieved a BLEU score of 0.137. The pattern-based method achieved a BLEU score of 0.257, still far from satisfactory descriptions that are readable and interpretable for doctors.

Table 4 Submissions of ImageSem in the concept detection subtask

Approach	F1 Score	Ranking
03ImagingTypes	0.419	14
02Comb_ImagingTypes_Baseline	0.400	16
07Intersect_06_baseline	0.396	17
04Comb_ImagingTypes_AnatomicStructure	0.370	19
05Comb_ImagingTypes_MedicalFindings	0.355	22
06Comb_ImagingTypes_AnatomicStructure_Findings	0.327	24
08AnatomicStructure	0.037	28
09Findings	0.019	29
01baseline	0.380	18

Table 5 Submissions of ImageSem in the caption prediction subtask

Approach	BLEU
04pattern1+ImagingTypes_AnatomicStructure_Findings	0.203
05pattern2+ImagingTypes_AnatomicStructure_Findings	0.181
06pattern3+ImagingTypes_AnatomicStructure_Findings	0.257
03baseline_Dual_Path_CNN Model	0.137

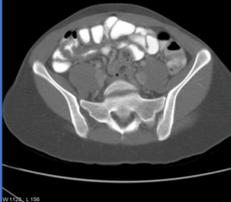
Image	Task	Ground Truth		Prediction		
 <p>Figure ID: synpic33490</p>	Concept	CUI	UMLS Terms	Imaging Types	CUI	UMLS Terms
		C0040398	Tomography, Emission-Computed		C0040398	Tomography, Emission-Computed
		C0030797	Pelvis		C0030797	Pelvis
		C3686616	Structure of mesenteric adipose tissue		C0030797	Pelvis
		C0458420	Region of bladder		C0577559	Mass of body structure
	C0013687	effusion				
Caption	mesenteric fat stranding in the region of the bladder. free fluid in left pelvis.		The Tomography, Emission-Computed image of the pelvis demonstrate mass of body structure.			
 <p>Figure ID: synpic40743</p>	Concept	CUI	UMLS Terms	Imaging Types	CUI	UMLS Terms
		C0040398	Tomography, Emission-Computed		C0040398	Tomography, Emission-Computed
		C0205400	Thickened		C0003617	Appendix
	C0734003	Wall of ileum	C0577559	Mass of body structure		
Caption	Diffuse wall thickening of the terminal ileum.		The Tomography, Emission-Computed image of the appendix demonstrate mass of body structure.			

Figure 3: Examples of concepts and captions predicted by ImageSem on the validation set

5. Conclusions

This paper presents the participation of the ImageSem Group at the ImageCLEFmed Caption 2021 task. We tried different strategies for both subtasks. In the concept detection subtask, we used the transfer learning-based MLC model to detect overall 1,586 concepts. We also trained multiple fine-grained MLC models based on manually annotated semantic categories. One of the lessons is that we have become much clearer about which concepts are clinically relevant to radiology images, and in order to obtain better predictions, the semantic labels of images should be more focused and specific. Furthermore, how to generate a readable description based on clear and clinically meaningful concepts, is still worth exploring.

6. Acknowledgements

This work has been supported by the National Natural Science Foundation of China (Grant No. 61906214), the Beijing Natural Science Foundation (Grant No. Z200016).

References

- [1] B. Ionescu, H. Müller, R. Peteri, A. Ben Abacha, D. Demner-Fushman, S. Hasan, M. Sarrouti, O. Pelka, C. Friedrich, A. Herrera, J. Jacutprakart, V. Kovalev, S. Kozlovski, V. Liauchuk, Y. Dicente Cid, J. Chamberlain, A. Clark, A. Campello, H. Moustahfid, A. Popescu, The 2021 ImageCLEF Benchmark: Multimedia Retrieval in Medical, Nature, Internet and Social Media Applications, 2021, pp. 616–623.
- [2] O. Pelka, C. M. Friedrich, A. Herrera, H. Müller, Overview of the ImageCLEFmed 2021 concept & caption prediction task, in: CLEF2021 Working Notes, 'CEUR' Workshop Proceedings, CEUR-WS.org, Bucharest, Romania, 2021.
- [3] O. Bodenreider, The unified medical language system (umls): integrating biomedical terminology, *Nucleic Acids Research* 32 (2004) 267–270.
- [4] Z. Guo, X. Wang, Y. Zhang, J. Li, Imagesem at imageclefmed caption 2019 task: a two-stage medical concept detection strategy, Lugano, 2019.
- [5] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, Z. Wojna, Rethinking the inception architecture for computer vision, in: *IEEE*, 2016, pp. 2818–2826.
- [6] G. Huang, Z. Liu, V. Laurens, K. Q. Weinberger, Densely connected convolutional networks, in: *IEEE Computer Society*, 2016.
- [7] Z. Zheng, L. Zheng, M. Garrett, Y. Yang, M. Xu, Y.-D. Shen, Dual-path convolutional image-text embedding with instance loss, *ACM Transactions on Multimedia Computing, Communications, and Applications* 2 (2020) 1–23. doi: <https://doi.org/10.1145/3383184> .
- [8] Y. Zhang, X. Wang, Z. Guo, J. Li, Imagesem at imageclef 2018 caption task: Image retrieval and transfer learning. in: *Clef2018 working notes*, Avignon, France, 2018.
- [9] V. Kougia, J. Pavlopoulos, Androutopoulos, Aueb nlp group at imageclefmed caption 2019. in: *Clef2019 working notes*, CEUR-WS.org, Lugano, Switzerland (2019), 2019.
- [10] I. B. H. Müller, R. Péteri, A. Abacha, C. B., M. G., Overview of the imageclef 2020: Multimedia retrieval in medical, lifelogging, nature, and internet applications, *Lecture Notes in Computer Science* (2020).
- [11] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, Bernstein, Imagenet large scale visual recognition challenge, *International Journal of Computer Vision* (2014) 1–42.
- [12] A. R. Aronson, Effective mapping of biomedical text to the umls metathesaurus: the metamap program, 2001, pp. 17–21.