Mathematical Biosciences and Engineering

Mathematical Biosciences and Engineering, 16(4): 1978–1991
DOI: 10.3934/mbe.2019097
Received: 18 December 2018
Accepted: 18 February 2019
Published: 08 March 2019

http://www.aimspress.com/journal/MBE

Research article

Identifying concepts from medical images via transfer learning and image retrieval

Xuwen Wang, Yu Zhang, Zhen Guo and Jiao Li*

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

*** Correspondence:** Email: li.jiao@imicams.ac.cn; Tel: +86-10-5232-8740; Fax: +86-10-5232-8610.

Abstract: Automatically identifying semantic concepts from medical images provides multimodal insights for clinical research. To study the effectiveness of concept detection on large scale medical images, we reconstructed over 230,000 medical image-concepts pairs collected from the ImageCLEFcaption 2018 evaluation task. A transfer learning-based multi-label classification model was used to predict multiple high-frequency concepts for medical images. Semantically relevant concepts of visually similar medical images were identified by the image retrieval-based topic model. The results showed that the transfer learning method achieved F1 score of 0.1298, which was comparable with the state of art methods in the ImageCLEFcaption tasks. The image retrieval-based method contributed to the recall performance but reduced the overall F1 score, since the retrieval results of the search engine introduced irrelevant concepts. Although our proposed method achieved second-best performance in the concept detection subtask of ImageCLEFcaption 2018, there will be plenty of further work to improve the concept detection with better understanding the medical images.

Keywords: concept detection; transfer learning; multi-label classification; medical image retrieval; LDA

1. Introduction

Medical images such as Computed Tomography (CT), X-ray and pathological images have become the key evidence for clinical diagnosis. Interpreting the insights gained from medical images requires adequate medical knowledge and clinical experiences. With the rapid growth of digital medical images, automatically identifying semantic concepts from medical images provides useful

multimodal information for clinical research.

Inspired by recent success of deep learning models in image analysis [1], many researchers exploit various models to interpret medical images for clinical applications, such as disease detection and lesion recognition, e.g., Kong et al. put forward three kinds of convolutional neural networks (CNNs) models and integrated transverse plane images, coronal plane images, and annotations information to improve the accuracy of breast tumor classification, and achieved the accuracy of 75.11% and AUC of 0.8294 on a dataset containing 880 images [2]. However, due to the limited available medical images with semantic annotation, especially for rare diseases, most of the previous studies focus on the single-label prediction or a few of multi-label classification on small datasets.

To address the problem of limited training data, Pan et al. introduced the transfer learning method to transform knowledge learned from one domain to another [3]. For similar tasks such as image analysis, previous layers of deep neural networks have the same functions. So deep models such as convolutional neural networks (CNNs) can be trained and transformed efficiently between different datasets by sharing and fine-tune similar parameters. Esteva et al. trained a deep learning model on more than 1.28 million images of common items, and then successfully trained a human-level skin cancer detection model by transfer learning on 120,000 manually labeled skin cancer images [4]. Yu et al. proposed a hybrid transfer learning method for recognizing 30 labels from composite biomedical images and achieved the F1 value of 0.488 [5].

To explore automatic methods mapping from visual information to condensed textual descriptions, the CLEF Cross-Language Image Retrieval Track (ImageCLEF) launched the ImageCLEFcaption evaluation task since 2017 [6]. The recent ImageCLEFcaption 2018 task contains two subtasks, namely concept detection and caption prediction [7]. The concept detection subtask aims to identify the Unified Medical Language System (UMLS) Concept Unique Identifiers (CUIs) [8,9] for a given medical image from biomedical literature. It can be seen from the overview [6,10] that most researchers used some form of CNNs to represent visual information, fewer researchers used a traditional bag of visual words model. On the basis of the visual representation, additional methods such as attention mechanism were also used to identify useful medical concepts. On average, concepts detected by CNNs models were more robust, while the use of very deep residual networks did not introduce significant improvements over shallower networks [11–13]. As another popular method, several works used image retrieval to obtain visually similar images of given medical images and then detected concepts from the captions of retrieved images [14,15]. Zhang et al. presented the participation of our ImageSem group at the ImageCLEFcaption 2018 task, briefly introduced concept detection methods based on CNNs models and image retrieval, and achieved the second-best F1 score of 0.092 in the concept detection task [16]. Pinho et al. achieved a best mean F1 score of 0.1102 in the same concept detection task, using two kinds of classification algorithms over the feature spaces learned from a variant of generative adversarial networks with an auto-encoding process [17]. Although the overall performance is too far from the application, it is generally believed that the task of concept detection on large-scale heterogeneous medical images is challenging but meaningful.

To better understanding and describing the semantic content of medical images, we reconstructed a dataset of medical image-concepts pairs for concept detection on the basis of the ImageCLEFcaption 2018 collection. Based on the new dataset, we identified multiple concepts from large scale medical images by complementary methods, including the transfer learning-based multi-label classification models for high-frequency concepts, the image retrieval-based topic models

for latent relevant concepts from visually similar images, and fusion strategies combining concepts identified by both methods.

This paper is organized as follows: Section 2 introduces the material and methods, including dataset reconstruction, data analysis, data preprocessing, as well as multiple concept detection methods. Section 3 describes the experiments of concept detection on medical images. Section 4 shows the results of different methods and fusion strategies. Section 5 discusses errors and makes a brief conclusion.

2. Materials and method

2.1. Data

2.1.1. Data reconstruction

The corpus of annotated medical images is important for understanding the insights of medical images. The ImageCLEFcaption 2018 task [6] released a collection of medical image-caption pairs collected from scholarly articles in PubMed Central (PMC) [18]. Images were classified automatically to select useful radiology or clinical images, and the QuickUMLS toolkit [19] was used to annotate UMLS concepts in image captions. Each image was assigned with multiple concepts represented by Concept Unique Identifiers (CUIs). The collection of the ImageCLEFcaption 2018 concept detection subtask comprises a training set of 222,314 medical image-concepts pairs, and a test set of 9,938 image-concepts pairs. However, due to automatic labeling and unknown expanding strategies, the collection contains totally 111,156 concepts, in which mixed with lots of noise words or irrelevant concepts. Table1 shows the top 10 high-frequency CUIs in the ImageCLEFcaption 2018 training set.

Table 1. Top 10 high-frequency concepts in the concept detection training set of the ImageCLEFcaption 2018 collection.

CUIs	Associated images	UMLS terms
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

By observing concept CUIs and corresponding UMLS terms (backtracked from UMLS), we found that instead of medical terminology, the concept most commonly used to interpret medical images was a meaningless conjunction “AND”. Synonyms such as ‘Medical Image’, ‘image-dosage form’, ‘image’, etc., were assigned to the same image repeatedly. In addition, some unreasonable

matching strategies may lead to the abnormal quantity of concepts, e.g., a term 'Arrow' was mapped to multiple concepts with similar lexical form but the inconsistent meaning (such as 'Narrow face', 'Marrow', 'Yarrow flower extract'). To sum up, this ground truth provides plenty of inappropriate concepts for interpreting medical images. It is difficult for analyzing the semantic association between concepts and images from either computational view or biomedical view.

In this study, to reduce the influence of uneven noisy data and interpret medical images with more useful concepts, we reconstructed the concept detection dataset based on the image-caption pairs from the ImageCLEFcaption 2018 collection. The reconstructed collection includes a training set (Rec-training) and a test set (Rec-test) containing 222,314 and 9,938 medical images respectively. We used MetaMap [20] to recognize concepts in image captions, chose the strict strategy to guarantee the quality of concepts. The new dataset is referred to as the ImageSem collection.

2.1.2. Data analysis

The Rec-training set includes 222,314 images annotated with 76,938 non-repetitive concepts (CUIs), which are significantly different from the ImageCLEFcaption 2018 collection in concepts quantity and frequency, as shown in Table 2.

Table 2. Top 10 high-frequency CUIs in the Rec-training set of the ImageSem collection.

CUIs	Quantity of associated images	UMLS terms
C1547282	69,808	Show
C0336721	27,060	Arrow
C1704922	19,121	Image
C0449911	15,882	View
C0523207	10,916	Hematoxylin and eosin stain method
C0030705	10,786	Patients
C0205091	10,082	Left
C0205090	8,626	Right
C4489445	8,127	Magnification

Figure 1 shows a medical image with its corresponding caption and concepts. Compared with concepts annotated by the ImageCLEFcaption 2018 task, the new concepts from the ImageSem collection are more loyal to the image caption and concise enough for interpreting the given image.

Table 3 shows the distribution of medical concepts in the Rec-training set. The CUIs frequency is equivalent to the quantity of associated images of a specific concept. It is observed that most concepts (92.87%) appear in less than 50 images. The overall occurrence of the CUIs in the Rec-training set is 2,241,191, in which concepts with the frequency higher than 1,000 account for 40% of the overall occurrence, and concepts with the frequency higher than 500 account for 50%.

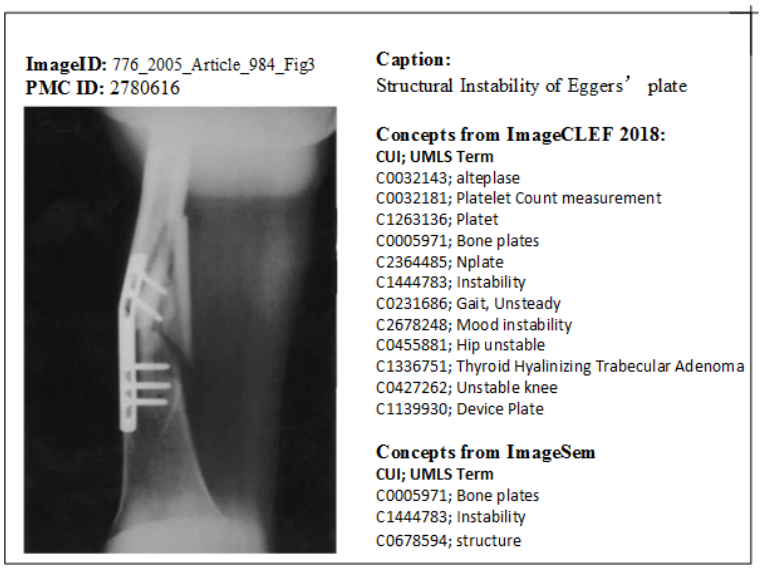

Figure 1. An example of a medical image with its caption and concepts.

Table 3. Statistics of concepts assigned to medical images in the Rec-training set.

CUIs frequency	CUIs quantity	Proportion
0–10	60,152	78.18%
10–50	11,303	14.69%
50–100	2,345	3.05%
100–500	2,413	3.14%
500–1000	393	0.51%
1000–10000	325	0.42%
10000+	7	0.009%
Total	76,938	100.00%

2.1.3. Data preprocessing

2.1.3.1. Selecting concepts and images for transfer learning

Considering the uneven concept distribution in table 3, it is too hard to build a transfer learning model for so many low-frequency concepts, and a mass of concepts may give rise to a significant increase in training time. As a compromise, 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 separately selected 332 CUIs appeared in more than 1,000 medical images and 725 CUIs appeared in more than 500 images in the Rec-training set, namely TL_F1000 subset and TL_F500 subset. Then we extracted all the medical images containing high-frequency CUIs from the Rec-training set. Totally 192,478 medical images for the TL_F1000 subset and 200,662 medical images for the TL_F500 subset. For each medical image, we filtered out low-frequency CUIs.

2.1.3.2. Image indexing

For the image retrieval-based method, we employed LIRE (Lucene Image Retrieval) [21,22] to perform content-based image retrieval (CBIR). LIRE is an open source Java library that provides a

simple way to retrieve images and photos based on color and texture characteristics. We created the Lucene index for medical images as well as corresponding captions and concepts in the Rec-training set. Then we retrieved visually similar images and collected image-concepts pairs for each target image.

2.2. Concept detection methods

In this section, we describe complementary methods to identify multiple concepts for a specific image, including the transfer learning method, the image retrieval-based topic modeling method and also the fusion strategies of the two methods.

2.2.1. Transfer learning for detecting high-frequency concepts

We used the transfer learning method to identify multiple high-frequency concepts for medical images. We applied Inception-V3, a CNNs model released by Google, to perform multi-label classification. Profit from improvements in the factorization of convolution kernel, the Inception-V3 model can decompose a 7×7 convolution kernel into two one-dimensional convolution kernels(a 1×7 kernel and a 7×1 kernel), which speed up the calculations and increase the network depth.

In this work, the Inception-V3 model was pre-trained on the ImageNet datasets including 1.2 million images with more than 1,000 common object classes [23,24]. Specifically, all the parameters of previous layers were frozen and the last softmax layer was replaced with a fully-connected layer and a sigmoid layer. During the re-training step, only the last two new layers were trained to map medical images to concept CUIs, which cost a very short time. We retrained the CNNs model on both of the TL_F1000 and the TL_F500 subset, namely normal transfer learning. As medical images in the ImageSem collection vary a lot with the ImageNet dataset, we also tried to retrain more layers of the CNNs model and fine-tune weights layer by layer, which may cost longer training time, namely a global fine-tune transfer learning.

2.2.2. Image retrieval-based topic model for identifying relevant concepts

Different from the transfer learning method that focuses on high-frequency concepts, the image retrieval-based method identifies relevant concepts from visually similar images, which contain both high and low-frequency concepts. In this section, we used the topic model to analyze the topic distribution of concepts collected from retrieved similar images, and selected topical relevant concepts for a specific medical image.

Firstly, we submitted a query image to the search engine to retrieve similar images from the Rec-training set. Then we collected concepts from retrieved images as relevant documents. Each document is assumed to be a mixture of a number of topics, and each concept belongs to one of the topics. We employed the Latent Dirichlet Allocation (LDA) model [25] to perform the topic modeling process.

Let I be an image, D be the documents collected from similar images of I, and $\mathbf{c} = \{c_1, \ldots c_T\}$ be a sequence with T concepts. The objective of a concept detection model is to maximize the log-likelihood of the concept sequence of a given image, which is

$$\log p(\mathbf{c}|I) \propto logp(\mathbf{c}|D) = \sum_{t=1}^{T} \log p(c_t|c_{t-1}, \dots, c_1, D) \quad (1)$$

Let $z \in \{z_1, \dots, z_K\}$ be the topics of a relevant document, K is the size of the topic set. Based on the above hypothesis, the objective function is converted to compute the log-likelihood of a joint distribution $p(\mathbf{c}, z|D)$, which can be approximated as follows.

$$\log p(\mathbf{c}, z|D) = logp(\mathbf{c}|z, D)p(z|D) = logp(\mathbf{c}|z, D) + logp(z|D) \quad (2)$$

$$logp(\mathbf{c}|z, D) = \sum_{t=1}^{T} \log p(c_t|c_{t-1}, \dots, c_1, z, D) \quad (3)$$

Let $\boldsymbol{C} = \{c_1, \dots, c_V\}$ be the vocabulary with V concepts, and $\boldsymbol{d} = \{d_1, \dots, d_M\}$ be M documents containing concepts of similar images. Then document d is generated as follows.

- Choose $\theta \sim Dirichlet(\alpha)$.
- For each of the T concepts c_t in d:
  - Choose a topic $z_t \sim Multinomial\ (\theta)$.
  - Choose a concept c_t from $P(c_t|z_t, \beta)$.

Where α and β are hyper-parameters for the symmetric Dirichlet distributions, the mixing proportion θ is drawn from a Dirichlet prior with parameter α. The probability of d is defined as follows.

$$p(d|\alpha, \beta) = \int_{\theta} p(\theta|\alpha)(\prod_{t=1}^{T} \sum_{z_k} p(z_k|\theta)p(c_t|z_k, \beta))\, d\theta \quad (4)$$

Then we can learn $p(c|z)$, concept probabilities given a topic, and $p(z_k|d)$, topic probabilities given a document, which provides clues for choosing useful concepts.

2.2.3. Fusion strategies for concept detection

To make better use of the results from both methods, we proposed three fusion strategies to cover as many useful concepts as possible. The first approach combined the results of the transfer learning method and the image retrieval-based topic model directly. The second one used high-frequency concepts detected in transfer learning method as a hint for choosing better candidate topics in the image retrieval-based method. The third one filtered the input CUIs documents of the topic model with high-frequency concepts detected in transfer learning method.

3. Experiments

3.1. Experimental setup

An experimental study was performed to verify the effectiveness of proposed concept detection

methods. As for the collection, we randomly selected 10,000 samples from the Rec-training set as the validation set for regulating parameters. The rest of 212,314 image-concepts pairs remained as the training set. The overall 9,938 medical images in the test set were used for evaluating the performance of different methods.

As a baseline method, we combined concept CUIs of top 10 similar images directly for a given test image. As for training the transfer learning model, medical images were resized to 299 x 299 pixels, the batch size was set to 20, the learning rate was set to 0.003, the training steps was set to 25000. As for the retrieval-based topic model, we applied Gensim [26], a Python package for modeling the topical distribution of concepts. For a given image, we collected CUIs of retrieved similar images as the input of LDA model. According to the topic distribution of the retrieved CUIs documents, we picked the topic with the highest probability as the candidate topic, and selected CUIs with probabilities above the threshold φ_0 from the candidate topic as the final output. The hyper-parameters α and β were learned automatically from corpora, the number of topics K was set to 20, the iteration was set to 10,000, the number of similar images was set to 10, the threshold φ_0 of term probability in each topic was 0.01, and the gamma was set to 0.05. Then we combined the best results of the above methods with three different fusion strategies.

3.2. Evaluation criteria

The performance evaluation follows the ImageCLEFcaption 2018 task. The balanced precision and recall trade-off were measured in terms of F1 scores, which were computed by the Python's scikit-learn library. Specifically, we computed the micro F1 score for each medical image in the test set, and the average of micro F1 scores across all the test images was regarded as the final measure of the model.

4. Results

4.1. Results of transfer learning model

Table 4 shows the effectiveness of multi-label classification models on the modified ground truth of the Rec-test set, namely "GT_F500" and "GT_F1000", in which only concepts with the frequency above 500 and 1,000 remained. "TL_F500" and "TL_F1000" separately denote the results of transfer learning models trained on the TL_F500 subset and TL_F1000 subset. "TL_F500_gft" and "TL_F1000_gft" denote the results of global fine-tuned transfer learning models trained on the TL_F500 subset and TL_F1000 subset. It can be observed that, although more concepts were fed into the classification models (725 concepts in "TL_F500" VS 332 concepts in "TL_F1000"), models trained on the TL_F1000 subset achieved better results than the same one trained on the TL_F500 subset. This indicates to some extent that the CNNs model performs better on recognizing concepts with larger training samples, and too many labels may result in the reduction of classification. What we can also learn is that compared with normal transfer learning models, the global fine-tuned models such as "TL_F1000_gft" improved significantly, either in precision, recall or the F1 score.

Table 4. Results of concept detection by transfer learning models on the modified GT_F500 and GT_F1000 of the Rec-test set.

Model	Ground Truth	P	R	F1
TL_F500	GT_F500	0.0968	0.1939	0.1178
TL_F500_gft	GT_F500	0.1384	0.2725	0.1667
TL_F1000	GT_F1000	0.1015	0.2554	0.1334
TL_F1000_gft	GT_F1000	0.1489	0.3686	**0.1942**

Table 5 shows the results of concept detection by transfer learning models on the ground truth of the Rec-test set. The overall performance of transfer learning models declined due to the additional low-frequency concepts in the ground truth. However, the global fine-tuned transfer learning model "TL_F1000_gft" showed robustness and achieved the best F1 score of 0.1298, which is comparable with the state of art in large scale concept detection tasks.

Table 5. Results of concept detection by transfer learning models on the Rec-test set.

Model	P	R	F1
TL_F500	0.0918	0.0978	0.0874
TL_F500_ft	0.1313	0.1413	0.1245
TL_F1000	0.0931	0.0991	0.0885
TL_F1000_ft	0.1365	0.1486	**0.1298**

4.2. Results of image retrieval-based topic model

As for the image retrieval-based topic models, experiments were performed on the Rec-test set and the corresponding CUIs of retrieved similar images, as shown in Table 6. The baseline was "ReSim_10", which combined concepts of retrieved top 10 similar images of a given image. The "RT" represents the results of the image retrieval-based topic model with default parameters. The "RT_10+" used the same parameters as the "RT" model but remain CUIs with the frequency higher than 10 in the Rec-training set, and achieved F1 score of 0.0515.

Table 6. Results of concepts detection by image retrieval-based methods on the Rec-test set.

Model	P	R	F1
ReSim_10	0.0209	0.1867	0.0363
RT	0.0344	0.0754	0.0428
RT_10+	0.0411	0.0906	**0.0515**

It can be seen that image retrieval-based models achieved a recall of 0.0906, which was approximate with normal transfer learning methods. However, the low precision of the retrieval-based models indicated that noise concepts account for a large proportion in results. Inspired by this, the image retrieval based method should be improved from two aspects: on the one hand, due to the noise of the retrieval results, the concept documents that is irrelevant to the test image should be filtered out; on the other hand, the topic with the highest probability may not be the sole correct choice, and external semantic information can be used to select useful topics.

4.3. Results of fusion strategies

Table 7 shows the results of different fusion strategies. "F1_500" is the combination of concepts from "TL_F500_gft" and "RT", and "F1_1000" is the combination of concepts from "TL_F1000_gft" and "RT", removing duplicated CUIs. It can be observed that "F1_500" and "F1_1000" recalled more relevant concepts than a single model (best recall of 0.1711), while the overall accuracy was reduced by introducing too much noise. "F2_500" and "F2_1000" separately used concepts predicted by "TL_F500_gft" and "TL_F1000_gft" as a hint for choosing candidate topics in the image retrieval-based topic models. This strategy improved the precision of topic model (precision of 0.1976 for "F2_500" and 0.2002 for "F2_1000") significantly by selecting useful topics, but it also neglected many low-frequency concepts and reduced the recall heavily. "F3_500" and "F3_1000" filtered some irrelevant CUIs documents based on concepts predicted by "TL_F500_gft" and "TL_F1000_gft". Compared with former methods, the topic model recalled more useful concepts (recall of 0.1180).

Table 7. Results of concepts detection by fusion strategies on the Rec-test set.

Methods	P	R	F1
F1_500	0.0551	0.1644	0.0763
F1_1000	0.0569	**0.1711**	**0.0789**
F2_500	0.1976	0.0380	0.0557
F2_1000	**0.2002**	0.0398	0.0578
F3_500	0.0393	0.1153	0.0518
F3_1000	0.0403	0.1180	0.0532

4.4. Quality and error analysis

4.4.1. Impact of data quality

As mentioned in section 1 and section 2.1.1, our ImageSem group participated the ImageCLEFcaption 2018 task and applied similar methods on the ImageCLEFcaption 2018 collection, and our transfer learning method achieved second-best F1 score of 0.0928 in the concept detection task. Compared with results on the reconstructed ImageSem collection in table 5, in which the best overall result was 0.1298, the robustness of transfer learning methods across different datasets was verified. The retrieval-based method achieved 0.0907 on the ImageCLEFcaption data, but declined to 0.0515 on the ImageSem data. One possible reason is that in the case of the same retrieved images, topic models are very sensitive to the variation of concepts distribution. The other reason is that a large number of high-frequency concepts in the ImageCLEFcaption 2018 collection were easier to be captured, but not necessarily meaningful.

4.4.2. Case analysis

Despite the low scores on statistical evaluation, we think there is still some useful information learned from the large scale multimodal collection. Figure 2 shows a sample of test images. It is observed that medical concepts annotated by MetaMap in the image caption ranged variously on

frequency distribution. Concepts with higher frequency, such as 'C1704922 image' and 'C0205123 Coronal' are more likely to be detected. The deep transfer learning methods are good at predicting high-frequency concepts of limited scope, but cannot recognize low-frequency concepts in the training set or out of vocabulary concepts. The image retrieval-based topic models can reveal the high-frequency concepts and low-frequency concepts at the same time, but dependent heavily on the quality of the retrieved images. The higher similarity between the query image and the retrieved images, the more related concepts can be recalled, otherwise, a lot of noise words would be brought in. However, images retrieved by LIRE were often similar with query images in lower level, such as color, grayscale, contour, texture, etc. As shown in figure 3, the given query figure was a magnetic resonance image of a coronal section. Obviously, besides the very first image in the red frame, most of retrieved images were irrelevant with the query image, differ in either image types or body parts, which brought in plenty of irrelevant concepts. The fusion strategy, to some extent, may balance the results of the two methods and release the influence of data heterogeneity.

Figure ID: 10.1177_1176935116684824-fig1

Caption: T1-weighted magnetic resonance image of a coronal section through the brain. Gray matter and white matter are indicated, as well as the ventricles and the corpus callosum. Adapted with permission from Fieremans.40

Concepts:

CUIs;	UMLS Terms;	Frequency in Rec-training set
C0010090	Corpus Callosum	86
C0007799	Cerebral Ventricles	82
C0018220	Gray Matter	106
C0682708	White matter	247
C0028580	Nuclear Magnetic Resonance	47
C0006104	Brain	985
C1444656	Indicated	2062
C2937289	Adapt (substance)	319
C1524004	Authorization documentation	425
C1704922	***Image***	***19121***
C1413227	CD5 gene	2373
C0205123	***Coronal***	***4731***
C1305866	Weighing patient	2497
C1522472	section sample	4878

Concepts identified by transfer learning model:

C1547282;C0336721;***C1704922***;C0449911;C0030705;C0205091;C0205090;C0040405;C0205131;C0523207

Concepts identified by retrieval-based topic model:

C0001861;C0032718;C2826234;C0227614;C3810846;C0932489;C1704788;C0024485;C0282046;C0205097;C0205448;C0412620;C2349975;C1706765;C0205143;C0449911;C1441672;C0332173;C0205129;C1962958;C0009924;C0728985;***C1704922***;C0441889;***C0205123***;C1301820

Concepts Identified by fusion strategy:

C0449911;***C1704922***;C0040405;C0336721;C0523207;C0030705;C1547282;C0205090;C0205091;C0205131;C0932489;C0205448;C0412620;C2349975;C0282046;C1301820;C0441889;C0009924;***C0205123***;C1441672;C1962958;C0227614

Figure 2. A medical image with its caption and concepts annotated by the MetaMap. The lower part shows the typical concepts identified by the transfer learning model, the retrieval-based topic model and a fusion strategy.

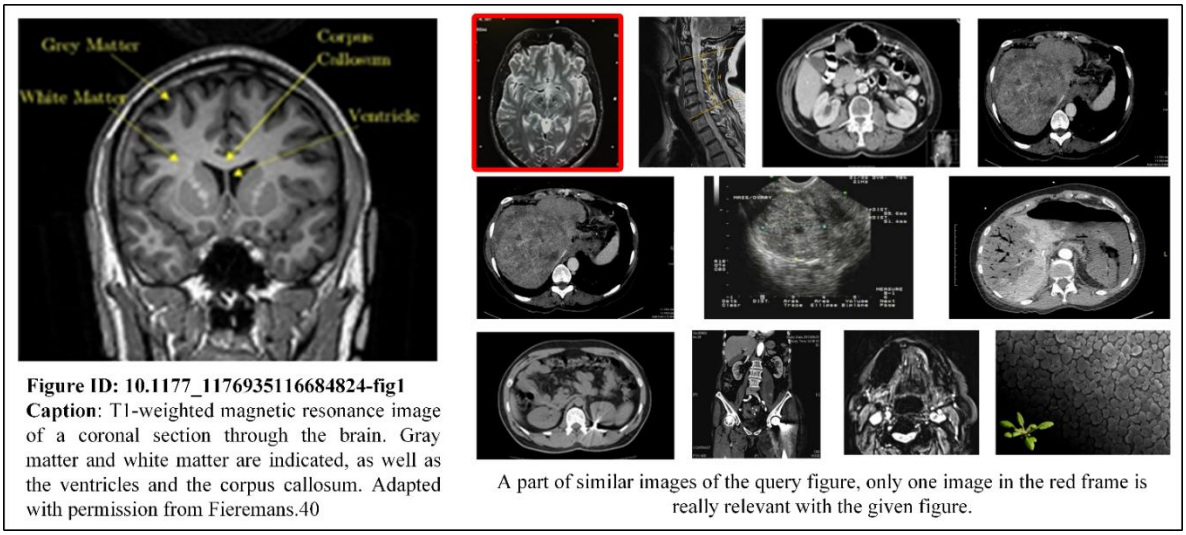

Figure 3. A medical image and its similar images retrieved from the Rec-training set.

5. Conclusion and further work

This study applied the deep transfer learning model, the image retrieval-based topic model as well as fusion strategies of both methods to identify concepts from medical images. The experiments showed the preferable performance of deep transfer learning models on predicting high-frequency concepts for medical images, the best F1 score of 0.1298 verified the effectiveness of the CNNs model on multi-label classification. The image retrieval-based topic model recalled high and low-frequency concepts simultaneously, but depended heavily on the retrieval results and brought noises with the overall accuracy reduced. Due to the variety and diversity of the medical images as well as the massive quantity of medical concepts, the work of semantic concept detection of large-scale open medical images still needs further research and improvement.

In future work, we will perform deeper data processing on the basis of the ImageSem collection, by adding more available image-text pairs, clustering the images into different groups based on the image type, the anatomy part, etc., and creating high quality label sets respectively. In addition, we will separately train deep models for different category of images, and seek more useful semantic clues from the external data.

Acknowledgments

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); the Fundamental Research Funds for the Central Universities (Grant No. 3332018153) and the CAMS Innovation Fund for Medical Sciences (CIFMS) (Grant No.2017-I2M-B&R-10).

Conflict of interest

All authors declare no conflicts of interest in this paper.

References

1. G. J. Litjens, T. Kooi and B. E. Bejnordi, et al., A survey on deep learning in medical image analysis, *Med. Image Anal.*, **42**(2017), 60–88.
2. X. Kong, T. Tan and L. Bao, et al., Classification of breast mass in 3D ultrasound images with annotations based on convolutional neural networks. *Chin. J. Biomed. Eng.*, **37**(2018), 414–422.
3. S. J. Pan and Q. Yang, A survey on transfer learning, *IEEE T. Knowl. Data. En.*, **22**(2010), 1345–1359.
4. A. Esteva, B. Kuprel and R. A. Novoa, et al., Dermatologist-level classification of skin cancer with deep neural networks, *Nature*, **542(**2017), 115–118.
5. Y. Yu, H. Lin and J. Meng, et al., Classification modeling and recognition for cross modal and multi-label biomedical image. *J. Image Graph.*, **23**(2018), 917–927.
6. C. Eickhoff, I. Schwall and A. G. Seco de Herrera, et al., Overview of ImageCLEFcaption 2017–image caption prediction and concept detection for biomedical images. In: G. J. F. Jones, S. Lawless and J. Gonzalo, et al., editors. *Lect. Notes. Comput. SC.*: Experimental IR meets multilinguality, multimodality, and interaction. 8th International Conference of the CLEF Association (CLEF 2017); September 11–14, 2017; Dublin, Ireland. Cham: Springer; 2017 Aug 17. **10456**(2017), 315–337.
7. ImageCLEFcaption 2018: ImageCLEF/LifeCLEF–Multimedia Retrieval in CLEF [Internet]. Avignon, France: the CLEF initiative labs. 2018-[cited 2019 Feb 26]. Available from: http://www.imageclef.org/2018/caption.
8. UMLS: Unified Medical Language System [Internet]. Bethesda, Maryland: U.S. National Library of Medicine. 1986-[cited 2019 Feb 26]. Available from: https://www.nlm.nih.gov/research/umls/.
9. O. Bodenreider, The unified medical language system (umls): integrating biomedical terminology. *Nucleic Acids Res.*, **32**(2004), 267–270.
10. A. G. Seco de Herrera, C. Eickhoff and V. Andrearczyk, et al., Overview of the ImageCLEF 2018 caption prediction tasks. Paper presented at: CLEF 2018. Working Notes of CLEF 2018-Conference and Labs of the Evaluation Forum, CEUR Workshop Proceedings; 2018 Sep 10–14; Avignon, France.
11. E. Pinho, J. F. Silva and J. M. Silva, Towards representation learning for biomedical concept detection in medical images: UA.PT bioinformatics in ImageCLEF 2017. Paper presented at: CLEF 2017. Working notes of CLEF 2017-Conference and Labs of the Evaluation Forum, CEUR Workshop Proceedings; 2017 Sep 11–14; Dublin, Ireland.
12. D. Katsios and E. Kavallieratou, Concept detection on medical images using deep residual learning network. Paper presented at: CLEF 2017. Working notes of CLEF 2017-Conference and Labs of the Evaluation Forum, CEUR Workshop Proceedings; 2017 Sep 11–14; Dublin, Ireland.
13. N. N. Hoavy, J. Mothe and M.I. Randrianarivony, IRIT & MISA at ImageCLEF 2017-multi label classification. Paper presented at: CLEF 2017. Working notes of CLEF 2017-Conference and Labs of the Evaluation Forum, CEUR Workshop Proceedings; 2017 Sep 11–14; Dublin, Ireland.
14. L. Valavanis and T. Kalamboukis, IPL at ImageCLEF 2018: a KNN-based concept detection approach. Paper presented at: CLEF 2018. Working notes of CLEF 2018-Conference and Labs of the Evaluation Forum, CEUR Workshop Proceedings; 2018 Sep 10–14; Avignon, France.
15. M. M. Rahman, T. Lagree and M. Taylor, A cross-modal concept detection and caption prediction approach in ImageCLEFcaption track of ImageCLEF 2017. Paper presented at: CLEF

2017. Working notes of CLEF 2017-Conference and Labs of the Evaluation Forum, CEUR Workshop Proceedings; 2017 Sep 11–14; Dublin, Ireland.
16. Y. Zhang, X. Wang and Z. Guo, et al., ImageSem at ImageCLEF 2018 caption task: image retrieval and transfer learning, Paper presented at: CLEF 2018. Working notes of CLEF 2018-Conference and Labs of the Evaluation Forum, CEUR Workshop Proceedings; 2018 Sep 10–14; Avignon, France.
17. E. Pinho and C. Costa, Feature learning with adversarial networks for concept detection in medical images: UA.PT bioinformatics at ImageCLEF 2018. Paper presented at: CLEF 2018. Working notes of CLEF 2018-Conference and Labs of the Evaluation Forum, CEUR Workshop Proceedings; 2018 Sep 10–14; Avignon, France.
18. PMC: PubMed Central [Internet]. Bethesda, Maryland: National Center for Biotechnology Information (NCBI), U.S. National Institutes of Health's, National Library of Medicine. 2000-[cited 2019 Feb 26]. Available from: https://www.ncbi.nlm.nih.gov/pmc/.
19. QuickUMLS: System for Medical Concept Extraction [Internet]. Georgetown University, Washington: Luca Soldaini and Nazli Goharian. 2016-[cited 2019 Feb 26]. Available from: https://github.com/Georgetown-IR-Lab/QuickUMLS
20. MetaMap: A Tool for Recognizing UMLS Concepts in Text [Internet]. Bethesda, Maryland: U.S. National Institutes of Health's, National Library of Medicine. 1996-[cited 2019 Feb 26]. Available from: https://metamap.nlm.nih.gov/.
21. LIRE: Lucene Image Retrieval [Internet]. Klagenfurt University, AT: Mathias Lux. 2015-[cited 2019 Feb 26]. Available from: http://www.lire-project.net/.
22. R. Gan and J. Yin, Using LIRe to implement image retrieval system based on multi-feature descriptor. Proceedings of the Third International Conference on Digital Manufacturing & Automation; 2012 Jul 31–Aug 2; Guilin, China. IEEE; 2012 Sep 13. 1014–1017p.
23. C. Szegedy, V. Vanhoucke and S. Ioffe, et al., Rethinking the inception architecture for computer vision. Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition; 2016 Jun 27–30; La Vegas, NV, USA. IEEE; 2016 Dec 12. 2818–2826p.
24. O. Russakovsky, J. Deng and H. Su, et al., ImageNet large scale visual recognition challenge. *Int. J. Comput. Vis.*, **115**(2015), 211–252.
25. D.M. Blei, A.Y. Ng and M. I. Jordan, Latent dirichlet allocation. *J. Mach. Learn. Res.*, **3** (2003), 993–1022.
26. Gensim, topic modelling for humans [Internet]. Masaryk University, Czech: Radim Řehůřek. 2009-[cited 2019 Feb 26]. Available from: https://radimrehurek.com/gensim/.

AIMS Press