AUEB NLP Group at ImageCLEFmed Caption 2019

Vasiliki Kougia, John Pavlopoulos, and Ion Androutsopoulos

Department of Informatics, Athens University of Economics and Business, Greece
{kouyiav,annis,ion}@aueb.gr

Abstract. We present the systems that AUEB's NLP Group used to participate in the ImageCLEFmed 2019 Caption task. The goal of this task is to automatically select medical concepts related to each image, as a first step towards generating image captions, medical reports, or to help in medical diagnosis. We participated with four systems, all using CNN image encoders. The encoder of each system is combined with an image retrieval method or a feed-forward neural network to predict concepts. Our systems were ranked 1st, 2nd, 3rd, and 5th.

Keywords: Medical Images · Concept Detection · Image Retrieval · Multi-label Classification · Image Captioning · Machine Learning · Deep Learning

1 Introduction

Deep learning methods are being developed to automatically interpret biomedical images in order to help clinicians who examine large numbers of images daily [10]. The ImageCLEFmed Caption task [12] is part of ImageCLEF 2019 [6].[1] Image CLEF is a campaign that suggests novel challenges and develops benchmarking resources for the evaluation of systems operating on images. The ImageCLEFmed Caption Task ran for the 3rd year in 2019. It included a Concept Detection sub-task, where the goal was to perform multi-label classification of medical images by automatically selecting medical concepts that should be assigned to each image. The concepts come from the Unified Medical Language System (UMLS).[2] Selecting the appropriate concepts per image can be a first step towards automatically generating image captions, longer medical reports, and can also assist, more generally, in computer-assisted diagnosis [9]. In the two previous years, ImageCLEFmed also included a Caption Prediction (generation) sub-task [2, 4], which was not included this year.

This paper presents the four Concept Detection systems that AUEB's NLP Group used to participate in ImageCLEFmed 2019 Caption. The systems were ranked 1st, 2nd, 3rd, and 5th. The system that was ranked 3rd consists of a DenseNet-121 [5] Convolutional Neural Network (CNN) image encoder and a k-Nearest Neighbors (k-NN) retrieval component that uses the encoding of the image being classified to retrieve similar training images with known concepts; these are then used to assign concepts to the new

[1] https://www.imageclef.org/2019/medical/caption/

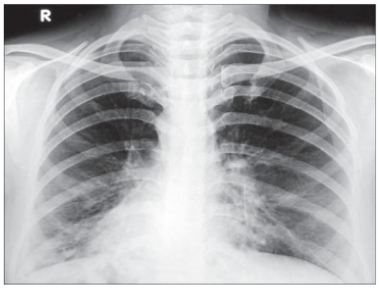
[2] https://www.nlm.nih.gov/research/umls/

image. The top-ranked system is a re-implementation of CheXNet [14], with modifications for ImageCLEFmed Caption 2019. CheXnet also uses the DenseNet-121 encoder [5], combined with a feed-forward neural network (FFNN) that performs multi-label classification. The second-best system is an ensemble combining concept probability scores obtained from the CheXNet-based system and image similarity scores produced by k-NN retrieval of similar training images. The system ranked 5th uses the VGG-19 image encoder [15], which was also used by Jing et al. [7], combined with a FFNN for multi-label classification.

2 Data

The ImageCLEFmed Caption 2019 dataset is a subset of the Radiology Objects in COntext (ROCO) dataset [13]. It consists of medical images extracted from open access biomedical journal articles of PubMed Central.[3] Each image was extracted along with its caption. The caption was processed using QuickUMLS [16] to produce the gold UMLS concept unique identifiers (CUIs). An image can be associated with multiple CUIs (Figure 1). Each CUI is accompanied by its corresponding UMLS term.

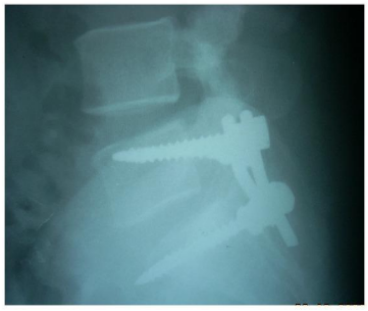
Image: ROCO_CLEF_00055



Concepts

CUI	UMLS Term
C0043299	x-ray procedure
C1548003	radiograph
C0817096	thoracics
C1962945	radiogr

Image: ROCO_CLEF_00051



Concepts

CUI	UMLS Term
C0267716	hernia
C0520904	postop nausea
C0867390	postoperative stroke
C0032786	care postop
C0241311	after surgery

Fig. 1. Two images from ImageCLEFmed Caption 2019, with their gold CUIs and UMLS terms.

In ImageCLEFmed Caption 2017 [2] and 2018 [4], the datasets were noisy. They included generic and compound images, covering a wide diversity of medical images;

[3] https://www.ncbi.nlm.nih.gov/pmc/

there was also a large total number of concepts (111,155) and some of them were too generic and did not appropriately describe the images [18]. In the ROCO dataset, compound and non-radiology images were filtered out using a CNN model. This led to 80,786 radiology images in total, of which 56,629 images were provided as the training set, 14,157 as the validation set, and the remaining 10,000 images were used for testing. In ImageCLEFmed Caption 2019, the total number of UMLS concepts was reduced to 5,528, with 6 concepts assigned to each training image on average. The minimum number of concepts per training image is 1, and the maximum is 72. Table 1 shows the 6 most frequent concepts of the training set and how many training images they were assigned to, according to the gold annotations. We note that 312 of the 5,528 total concepts are not assigned to any training image; and 1,530 concepts are assigned to only one training image.

CUI	UMLS term	Images
C0441633	diagnostic scanning	6,733
C0043299	x-ray procedure	6,321
C1962945	radiogr	6,318
C0040395	tomogr	6,235
C0034579	pantomogr	6,127
C0817096	thoracics	5,981

Table 1. The 6 most frequent concepts (CUIs) in the training set of ImageCLEFmed Caption 2019 and how many training images they are assigned to, according to the gold annotations.

We randomly selected 20% of the training images and used them as our development set (11,326 images along with their gold concepts). The models we used to produce the submitted results were trained on the entire training set. The validation set was used for hyper-parameter tuning and early stopping.

3 Methods

This section describes the four methods we developed for ImageCLEFmed Caption 2019.

3.1 System 1: DenseNet-121 Encoder + k-NN Image Retrieval (Ranked 3rd)

In this system, we followed a retrieval approach, extending the 1-NN baseline of our previous work on biomedical image captioning [9]. Given a test image, the previous 1-NN baseline returned the caption of the most similar training image, using a CNN encoder to map each image to a dense vector. For ImageCLEFmed Caption 2019, we retrieve the k-most similar training images and use their concepts, as described below.

We use the DenseNet-121 [5] image encoder, a CNN with 121 layers, where all layers are directly connected to each other improving information flow and avoiding vanishing gradients. We started with DenseNet-121 pre-trained on ImageNet [1] and fine-tuned it on ImageCLEFmed Caption 2019 training images.[4] The fine-tuning was

[4]We used the implementation of `https://keras.io/applications/#densenet`.

performed as when training DenseNet-121 in System 2, including data augmentation (Section 3.2). Without fine-tuning, the performance of the pre-trained encoder was worse. ImageCLEFmed Caption 2019 images were rescaled to 224×224 and normalized with the mean and standard deviation of ImageNet to match the requirements of DenseNet-121 and how it was pre-trained on ImageNet. Having fine-tuned DenseNet-121, we used it to obtain dense vector encodings, called *image embeddings*, of all training images. The image embeddings are extracted from the last average pooling layer of DenseNet-121. Given a test image (Fig. 2), we again use the fine-tuned DensNet-121 to obtain the image's embedding. We then retrieve the k training images with the highest cosine similarity (computed on image embeddings) to the test image, and return the r concepts that are most frequent among the concepts of the k images. We set r to the average number of concepts per image of the particular k retrieved images. We tuned the value of k in the range from 1 to 200 using the validation set, which led to $k = 199$. Further fine-tuning may improve performance further. This system ranked 3rd.

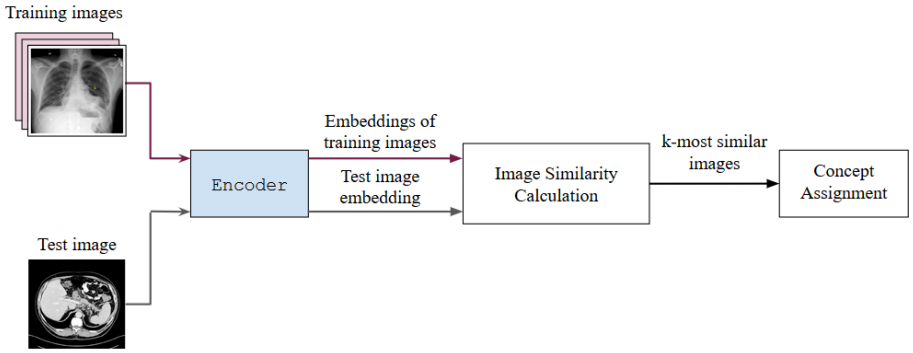


Fig. 2. Illustration of how System 1 (DenseNet-121 and k-NN image retrieval) works at test time.

3.2 System 2: CheXNet-based, DenseNet-121 Encoder + FFNN (Ranked 1st)

This system, which is based on CheXNet [14], achieved the best results in ImageCLEFmed Caption 2019. In its original form, CheXNet maps X-rays of the ChestX-ray 14 dataset [17] to 14 labels. It uses DenseNet-121 [5] to encode images, adding a FFNN to assign one or more of the 14 labels (classes) to each image.

We re-implemented CheXNet in Keras[5] and extended it for the many more labels (5,528 vs. 14) of ImageCLEFmed Caption 2019. The images of ImageCLEFmed Caption 2019 were again rescaled to 224×224 and normalized using the mean and standard deviation values of ImageNet. Also the training images of ImageCLEFmed Caption 2019 were augmented by applying random horizontal flip. Image embeddings are again extracted from the last average pooling layer of DenseNet-121. In this system, however, the image embeddings are then passed through a dense layer with 5,528 outputs and sigmoid activations to produce a probability per label. We trained the model by minimizing

[5] `https://keras.io/`

binary cross entropy loss. We used Adam [8] with its default hyper-parameters, early stopping on the validation set, and patience of 3 epochs. We also decayed the learning rate by a factor of 10 when the validation loss stopped improving.

At test time, we predict the concepts for each test image using their probabilities, as estimated by the trained model. For each concept (label), we assign it to the test image if the corresponding predicted probability exceeds a threshold t. We use the same t value for all 5,528 concepts. We tuned t on the validation set, which led to $t = 0.16$.

3.3 System 3: Based on Jing et al., VGG-19 Encoder + FFNN (Ranked 5th)

This system is based on the work of Jing et al. [7], who presented an encoder-decoder model to generate tags and medical reports from medical images. Roughly speaking, the full model of Jing et al. uses a VGG-19 [15] image encoder, a multi-label classifier to produce tags (describing concepts) from the images, and a hierarchical LSTM that generates texts by attending on both image and tag embeddings; the top level of the LSTM generates sentence embeddings, and the bottom level generates the words of each sentence. We implemented in Keras a simplified version of the first part of Jing et al.'s model, the part that performs multi-label image classification.

Again, we rescale the ImageCLEFmed Caption 2019 images to 224×224 and normalize them using the mean and standard deviation of ImageNet. We feed the resulting images to the VGG-19 CNN, which has 19 layers and uses small kernels of size 3×3. We used VGG-19 pre-trained on ImageNet.[6] We feed whole images to VGG-19, unlike Jing et al. [7], who divide each image into regions and encode each region separately. The output of the last fully connected layer of VGG-19 is then given as input to a dense layer with a softmax activation to obtain a probability distribution over the concepts. The model is trained using categorical cross entropy, which is calculated as:

$$E = -\sum_{i=1}^{|C|} y_{true,i} \log_2(y_{pred,i}) \tag{1}$$

where C is the set of $|C| = 5,528$ concepts, y_{true} is the ground truth binary vector of a training image, and y_{pred} is the predicted softmax probability distribution over the concepts C for the training image. Categorical cross entropy sums loss terms only for the gold concepts of the image, which have a value of 1 in y_{true}. When using softmax and categorical cross-entropy, usually y_{true} is a one-hot vector and the classes are mutually exclusive (single-label classification). To use softmax with categorical cross entropy for multi-label classification, where y_{true} is binary but not necessarily one-hot, the loss is divided by the number of gold labels (true concepts) [3, 11]. Jing et al. [7] achieve this by dividing the ground truth binary vector y_{true} by its L1 norm, which equals the number of gold labels. Hence, the categorical cross-entropy loss is computed as follows:

$$E = -\sum_{i=1}^{|C|} \frac{y_{true,i}}{\| y_{true} \|_1} \log_2(y_{pred,i}) = -\frac{1}{M} \sum_{j=1}^{M} \log_2(y_{pred,j}) \tag{2}$$

[6] https://keras.io/applications/#vgg19

where M is the number of gold labels (true concepts) of the training image, which is different per training image. In this model, the loss of Eq. 2 achieved better results on the development set, compared to binary cross entropy with a sigmoid activation per concept. We used the Adam optimizer with initial learning rate 1e-5 and early stopping on the validation set with patience 3 epochs. Given a test image, we return the six concepts with the highest probability scores, since the average number of gold concepts per training image is 6.

3.4 System 4: Ensemble, k-NN Image Retrieval + CheXNet (Ranked 2nd)

This method is an ensemble of System 1 (DenseNet-121 + k-NN Image Retrieval) and System 2 (CheXNet-based), where System 1 is modified to produce a score for each returned concept.

Given a test image g, we use System 1 (Fig. 2) to retrieve the k most similar training images $g_1, \ldots, g_k$, their gold concepts, and the cosine similarities $s(g, g_1), \ldots, s(g, g_k)$ between the test image g and each one of the k retrieved images. Let C be again the set of $|C| = 5,528$ concepts. For each concept $c_j \in C$, the modified System 1 assigns to c_i the following score:

$$v_1(c_j, g) = \sum_{i=1}^{k} s(g, g_i)\ \delta(c_j, g_i) \tag{3}$$

where $\delta(c_j, g_i) = 1$ if c_j is a gold concept of the retrieved training image g_i, and $\delta(c_j, g_i) = 0$ otherwise. In other words, the score of each concept c_j is the sum of the cosine similarities of the retrieved documents where c_j is a gold concept.

For the same test image g, we also obtain concept probabilities from System 2, i.e., a vector of 5,528 probabilities. Let $v_2(c_j, g)$ be the probability of concept c_j being correct for test image g according to System 2. For each $c_j \in C$, the ensemble's score $v(c_j, g)$ of c_j is simply the average of $v_1(c_j, g)$ and $v_2(c_j, g)$. The ensemble returns the six concepts with the highest $v(c_j, g)$ scores, as in System 3, on the grounds that the average number of gold concepts per training image is 6.

4 Results

Systems were evaluated in ImageCLEFmed Caption 2019 by computing their F1 scores on each test image (in effect comparing the binary ground truth vector y_{true} to the predicted concept probabilities y_{pred}) and then averaging over all test images [6]. Table 2 reports the evaluation results of our four systems on the development and test data, as well as their ranking among the approximately 60 systems that participated in the task. The ensemble (System 4) had the best results on development data, but the CheXNet-based system (System 2) had the best results on the test set.

System	Description	F1 Score		Ranking
		Dev	Test	
S1	DenseNet [5] + k-NN	0.2575244	0.2740204	3
S2 (CheXNet-based [14])	DenseNet [5] + FFNN	0.2599914	**0.2823094**	1
S3 (based on Jing et al. [7])	VGG-19 [15] + FFNN	0.2497768	0.2639952	5
S4 (ensemble)	Combination of S1, S2	**0.2644322**	0.2792511	2

Table 2. Results of our four systems on development and test data.

5 Conclusions and Future Work

We described the four systems that AUEB's NLP Group used to participate in ImageCLEFmed 2019 Caption. The four systems were ranked 1st, 2nd, 3rd, and 5th. Our top system was a re-implementation of CheXNet [14], with modifications to handle the much larger label set of ImageCLEFmed 2019 Caption and data augmentation. The system that was ranked 3rd used DenseNet [5] to encode images and k-NN retrieval to return the concepts of the most similar training images. Our second-best system was an ensemble of the previous two (CheXNet-based and k-NN based), indicating that the two approaches are complementary. Our weakest system, which nevertheless was ranked 5th, was based on the multi-label classification part of the system of Jing et al. [7], which aims to generate draft medical reports using an encoder-decoder approach.

In future work, we aim to experiment with, combine, and improve upon additional methods and datasets for medical image captioning. Towards that direction, we recently published a survey on medical image to text methods [9], which we also plan to extend.

References

1. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: ImageNet: A Large-Scale Hierarchical Image Database. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 248–255. Miami Beach, FL, USA (2009)
2. Eickhoff, C., Schwall, I., de Herrera, A.G.S., Müller, H.: Overview of ImageCLEFcaption 2017 - the Image Caption Prediction and Concept Extraction Tasks to Understand Biomedical Images. In: CLEF2017 Working Notes. CEUR Workshop Proceedings, CEUR-WS.org <http://ceur-ws.org>, Dublin, Ireland (September 11-14 2017)
3. Gong, Y., Jia, Y., Leung, T., Toshev, A., Ioffe, S.: Deep Convolutional Ranking for Multilabel Image Annotation. In: International Conference on Learning Representations (2014)
4. de Herrera, A.G.S., Eickhoff, C., Andrearczyk, V., Müller, H.: Overview of the ImageCLEF 2018 Caption Prediction Tasks. In: CLEF2018 Working Notes. CEUR Workshop Proceedings, CEUR-WS.org <http://ceur-ws.org>, Avignon, France (September 10-14 2018)
5. Huang, G., Liu, Z., van der Maaten, L., Weinberger, K.Q.: Densely Connected Convolutional Networks. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. pp. 4700–4708. Honolulu, HI, USA (2017)
6. Ionescu, B., Müller, H., Péteri, R., Cid, Y.D., Liauchuk, V., Kovalev, V., Klimuk, D., Tarasau, A., Abacha, A.B., Hasan, S.A., Datla, V., Liu, J., Demner-Fushman, D., Dang-Nguyen, D.T., Piras, L., Riegler, M., Tran, M.T., Lux, M., Gurrin, C., Pelka, O., Friedrich, C.M., de Herrera, A.G.S., Garcia, N., Kavallieratou, E., del Blanco, C.R., Rodríguez, C.C., Vasillopoulos, N., Karampidis, K., Chamberlain, J., Clark, A., Campello, A.: ImageCLEF 2019: Multimedia

Retrieval in Medicine, Lifelogging, Security and Nature. In: Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the 10th International Conference of the CLEF Association (CLEF 2019), LNCS Lecture Notes in Computer Science, Springer, Lugano, Switzerland (September 9-12 2019)
7. Jing, B., Xie, P., Xing, E.: On the Automatic Generation of Medical Imaging Reports. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Long Papers). pp. 2577–2586. Melbourne, Australia (2018)
8. Kingma, D.P., Ba, J.: Adam: A Method for Stochastic Optimization. arXiv:1412.6980 (2014)
9. Kougia, V., Pavlopoulos, J., Androutsopoulos, I.: A Survey on Biomedical Image Captioning. In: Workshop on Shortcomings in Vision and Language of the Annual Conference of the North American Chapter of the Association for Computational Linguistics. pp. 26–36. Minneapolis, MN, USA (2019)
10. Litjens, G., Kooi, T., Bejnordi, B.E., Setio, A.A.A., Ciompi, F., Ghafoorian, M., Laak, J.A.V.D., Ginneken, B.V., Sánchez, C.I.: A Survey on Deep Learning in Medical Image Analysis. Medical Image Analysis **42**, 60–88 (2017)
11. Mahajan, D., Girshick, R., Ramanathan, V., He, K., Paluri, M., Li, Y., Bharambe, A., van der Maaten, L.: Exploring the Limits of Weakly Supervised Pretraining. In: European Conference on Computer Vision. pp. 181–196. Munich, Germany (2018)
12. Pelka, O., Friedrich, C.M., de Herrera, A.G.S., Müller, H.: Overview of the ImageCLEFmed 2019 Concept Prediction Task. In: CLEF2019 Working Notes. CEUR Workshop Proceedings, vol. ISSN 1613-0073. CEUR-WS.org <http://ceur-ws.org/Vol-2380/>, Lugano, Switzerland (September 09-12 2019)
13. Pelka, O., Koitka, S., Rückert, J., Nensa, F., Friedrich, C.M.: Radiology Objects in COntext (ROCO): A Multimodal Image Dataset. In: MICCAI Workshop on Large-scale Annotation of Biomedical data and Expert Label Synthesis. pp. 180–189. Granada, Spain (2018)
14. Rajpurkar, P., Irvin, J., Zhu, K., Yang, B., Mehta, H., et al.: CheXNet: Radiologist-Level Pneumonia Detection on Chest X-rays with Deep Learning. arXiv:1711.05225 (2017)
15. Simonyan, K., Zisserman, A.: Very Deep Convolutional Networks for Large-Scale Image Recognition. arXiv:1409.1556 (2014)
16. Soldaini, L., Goharian, N.: QuickUMLS: A Fast, Unsupervised Approach for Medical Concept Extraction. In: MedIR workshop (2016)
17. Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M., Summers, R.M.: ChestX-ray8: Hospital-scale Chest X-ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. pp. 2097–2106. Honolulu, HI, USA (2017)
18. Zhang, Y., Wang, X., Guo, Z., Li, J.: ImageSem at ImageCLEF 2018 Caption Task: Image Retrieval and Transfer Learning. In: CLEF2018 Working Notes. CEUR Workshop Proceedings. Avignon, France (2018)