AUEB NLP Group at ImageCLEFmed Caption 2020

Basil Karatzas[1], John Pavlopoulos[2,1*][0000−0001−9188−7425], Vasiliki Kougia[2,1][0000−0002−0172−6917], and Ion Androutsopoulos[1]

[1] Department of Informatics, Athens University of Economics and Business, Greece
[2] Department of Computer and Systems Sciences, Stockholm University, Sweden
karatzas.basil@gmail.com, {annis,kouyiav,ion}@aueb.gr

Abstract. This article concerns the participation of AUEB's NLP Group in the ImageCLEFmed Caption task of 2020. The goal of the task was to identify medical terms that best describe each image, in order to accelerate and improve the interpretation of medical images by experts and systems. The systems we implemented extend our previous work [7,8,9] on models that employ CNN image encoders combined with an image retrieval method or a feed-forward neural network. Our systems were ranked 1st, 2nd and 6th.

Keywords: Medical Images · Concept Detection · Image Retrieval · Image Captioning · Multi-label Classification · Multimodal · Ensemble · Convolutional Neural Network (CNN) · Machine Learning · Deep Learning

1 Introduction

ImageCLEF [4] is an evaluation campaign held annually since 2003 as part of CLEF[3], and revolves around image analysis and retrieval tasks. ImageCLEFmedical [11] is a collection of ImageCLEF tasks that are associated with the study of medical images. In 2020, it consisted of 3 tasks: VQA-Med, Caption and Tuberculosis.[4] The ImageCLEFmed Caption task concerns the automatic assignment of medical terms (called concepts) to medical images. The dataset of ImageCLEFmed Caption 2020 consisted of medical images, which were split to 7 categories according to their radiology modality (see Table 1). Writing a diagnostic report for a medical image is a demanding and very time-consuming task that needs to be handled by medical experts [2,14]. One of the main goals of the ImageCLEFmed Caption task is to assist the development of efficient, multi-label, medical-image tagging models, which could be used to assist the medical experts and reduce the time needed for the diagnosis as well as to reduce potential medical errors.

[*] Corresponding author.

[3] http://www.clef-initiative.eu/

[4] https://www.imageclef.org/2020/medical

Table 1. The names of the seven categories in the dataset (first column), as provided by the organisers, and the description of each category (second column), drawn from [10]. We also give the Concept Unique Identifier (CUI) of concepts that appear in every image of the respective category (third column) and their corresponding description from the UMLS Metathesaurus (fourth column).

Name	ROCO Description	CUI	UMLS Description
DRAN	ANGIOGRAPHY	C002978	Angiogram
DRCO	COMBINED MODALITIES	-	-
DRCT	COMPUTERIZED TOMOGRAPHY	C0040398 C0040405	Tomography X-Ray Computed Tomography
DRMR	MAGNETIC RESONANCE	C0024485	Magnetic Resonance Imaging
DRPE	PET	C0032743	Positron-Emission Tomography
DRUS	ULTRASOUND	C0041618	Ultrasonography
DRXR	X-RAY, 2D TOMOGRAPHY	C0043299	Diagnostic radiologic examination

In this paper we describe the medical image tagging systems of the AUEB NLP Group that were submitted to ImageCLEFmed Caption 2020. Following our last year's success [8], our 3 submissions were ranked 1st, 2nd and 6th.[5]

Overall, our submissions were based on two methods. The first method was based on the Mean@k-NN system of [9] that assigns concepts to each test image using the k nearest neighbors from the training dataset. The second method extends the ConceptCXN system of [9] and uses a DenseNet-121 CNN to encode any test image and a Feed Forward Neural Network classifier on top. The remaining of the paper describes the data, methods, submitted systems and our results, followed by conclusions and future directions.

2 Data

Initially, the ImageCLEFmed Caption datasets comprised a broad variety of clinical images, which were extracted from figures of scientific articles found in the open-access biomedical literature database PubMed Central.[6] Each image was assigned medical terms from the Unified Medical Language System (UMLS) [1]. These terms, called concepts, were extracted from the processed text of the respective figure caption. Since 2019, in order to discard compound or non-radiology images from the initial datasets, the organisers applied filters and also performed a manual revision of their data. A subset of the resulting dataset, which is called the extended Radiology Objects in COntext (ROCO) [10], was chosen to be used as the dataset of the competition this year (see Fig. 1). Additionally, this year, images were classified into 7 mutually exclusive categories, as shown in Table 1, depending on the type of the radiology exam.

The number of possible concepts was reduced compared to previous years, by removing concepts with few occurrences, since the large number of concepts in the previous years resulted in the task being difficult for models [15]. There were 111,156

[5] Our best performing system will become available in the bioCaption PyPi package.

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

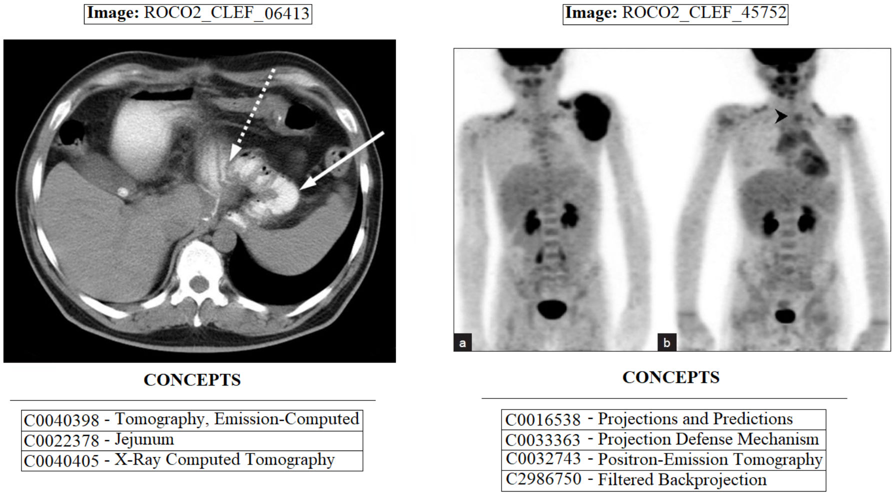

Image: ROCO2_CLEF_06413

CONCEPTS

C0040398	Tomography, Emission-Computed
C0022378	Jejunum
C0040405	X-Ray Computed Tomography

Image: ROCO2_CLEF_45752

CONCEPTS

C0016538	Projections and Predictions
C0033363	Projection Defense Mechanism
C0032743	Positron-Emission Tomography
C2986750	Filtered Backprojection

Fig. 1. Two images from ImageCLEFmed Caption 2020, with their gold CUIs. On the left is an image from a Computer Tomography (CT) and on the right is an image from a Positron Emission Tomography (PET).

possible concepts in 2018 [16], 5,528 in 2019 [8] and 3,047 in 2020. We also observed that there were concepts that appeared in every single image of a specific category, but rarely or never appeared in other categories. The Concept Unique Identifiers (CUIs) of these concepts are shown in Table 1.[7] These concepts rather describe the modality of the respective category. For example, DRPE images are always assigned with C0032743, whose UMLS term is POSITRON-EMISSION TOMOGRAPHY.

The dataset was split by the organisers to a training set of 64,753 images, a validation set of 15,970 images, and a test set of 3,534 images. For our experiments, we merged the provided training and validation sets and used 10% of the merged data as a development set. We will refer to the remaining 90% of the merged dataset as the training set for the rest of the paper.

3 Methods

This section describes the systems that were used in our submissions.

3.1 System 1: 2xCNN+FFNN

CNN+FNNN (a.k.a. ConceptCXN or DenseNet121+FFNN) [9,8] is the system that we submitted last year (and was ranked 1st) for the same task. It is a variation of CheXNet [13] that uses DenseNet-121 [3], which is a stack of 120 CNN layers, followed by

[7] We used UMLS Metathesaurus (`uts.nlm.nih.gov/home.html`) to map each CUI to its term.

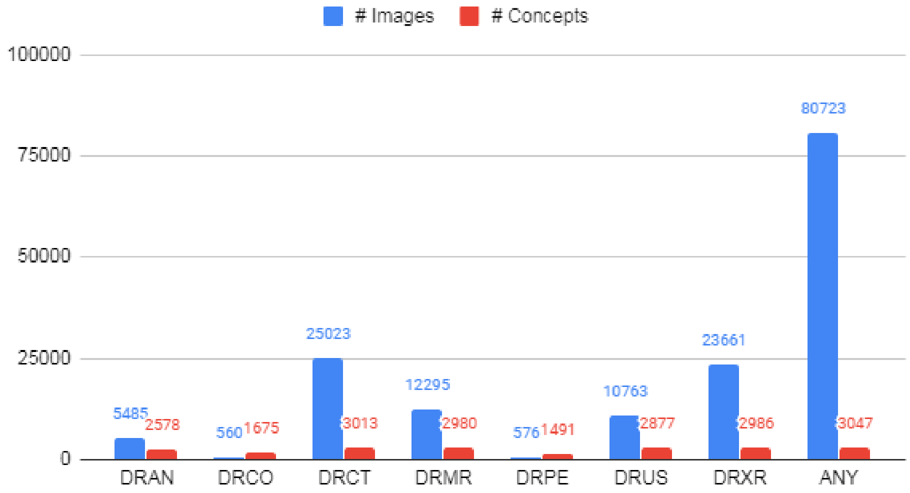

Fig. 2. Statistics regarding the number of images and concepts per category.

a feed-forward Neural Network (FFNN) that acts as a classifier layer on top. In the ImageCLEFmed Caption task of 2019, we changed the original FFNN to comprise 5,528 outputs (instead of 14), one per available concept. For the task of 2020, which also comprises different image categories (see Table 1), we followed the same approach per model (i.e., we employed one model per category). For example, the respective FFNN for the model of category C generates N_C outputs, which is the number of all possible concepts in category C.[8] The red bars of Fig. 2 depict the N_C values per category C, computed on the training set.

We trained the model by minimizing the binary cross entropy loss. We used Adam [6] as our optimizer and decreased the learning rate by a factor of 10 when the loss showed no improvement, following the work of [8]. We used a batch size of 16 and early stopping with a patience of 3 epochs. For each CNN+FFNN of a specific category (i.e., for each category, we fine-tuned a CNN and an FFNN on top), a classification threshold for all the concepts of the respective category was tuned by optimising the F1 score. Any concepts for which the respective output values exceeded that threshold were assigned to the corresponding image.

Two of our 2020 submissions consisted of ensembles of CNN+FFNN models, constructed in the following way. We trained 5 models per category, and kept the 2 best performing ones, according to their F1 score. We then created two ensembles, using the UNION and the INTERSECTION of the concepts returned by these two models. Hereafter, these two ensembles will be called 2xCNN+FFNN@U and 2xCNN+FFNN@I, respectively.

[8] We did not use image augmentation this year due to time restrictions, because of the large number of models we needed to train.

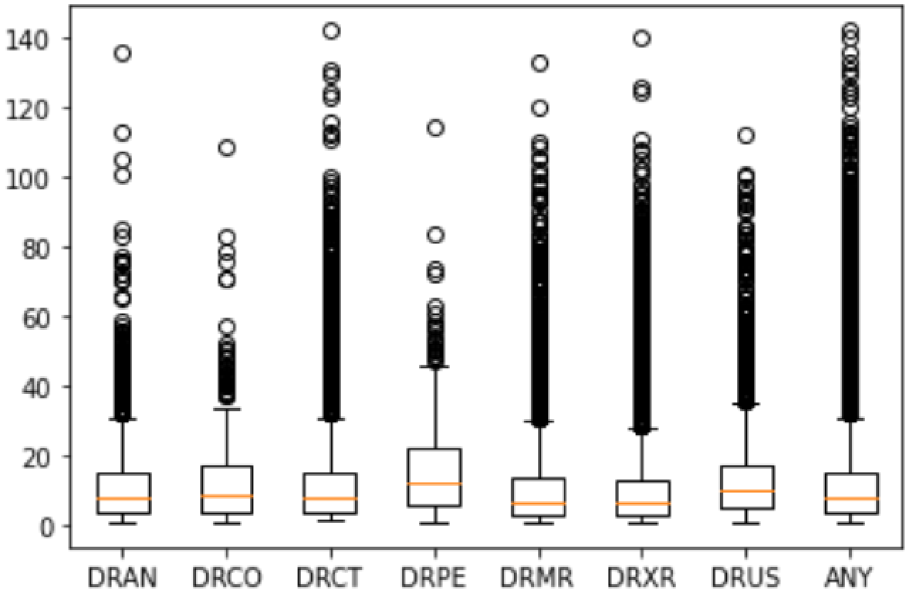

Fig. 3. Boxplots illustrating the number of gold concepts for the images of each category.

3.2 System 2: CNN+kNN

Following our previous work [8,7], the goal of our CNN+k-NN model for each test image was to retrieve similar images from the training set. The encoder of this model stemmed from our fine-tuned CNN+FFNN system, hence it is a CNN per category. We employed the output of the last average pooling layer of the CNN to represent each encoded image.[9] The encoded test image was then compared to all the images in the training set (encoded offline), using cosine similarity, and the k nearest images were returned.

After retrieving the k nearest images, CNN+k-NN returned the r concepts that were most frequently assigned to the k images. We tuned k and r for each category, investigating values from 1 to 200 for k, and values from 1 to 10 for r. During tuning, we also considered two other functions for r. First, we used the average number of concepts in the k images:

$$r = \frac{1}{k}\sum_{i=1}^{k} n_i \tag{1}$$

[9] Each image is rescaled to 224x224 and normalised with the mean and standard deviation of ImageNet.

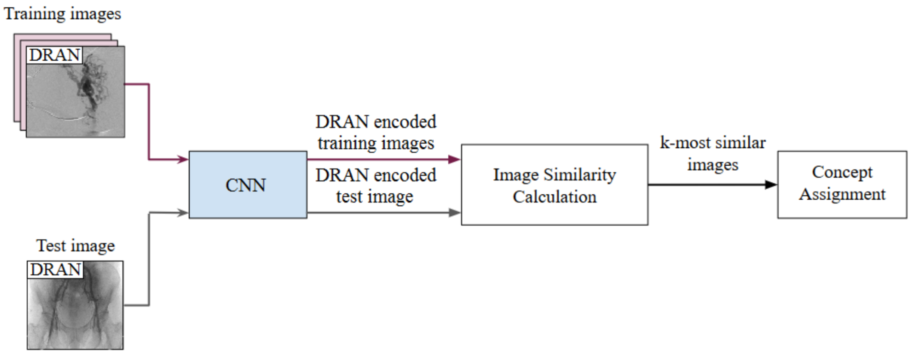

Fig. 4. Illustration of CNN+k-NN [8] for the DRAN category (we use DRAN as an example).

Second, we used a weighting based on cosine similarity to weigh the concepts:

$$r = \sum_{i=1}^{k} \frac{cos(g, g_i)}{\sum_{j=1}^{k} cos(g, g_j))} * n_i \tag{2}$$

where n_i is the number of concepts of the i-th retrieved image, g is the test image, g_i is the i-th closest training image, and $cos(g, g_i)$ the cosine similarity between g and g_i.

As with CNN+FFNN, our DenseNet-121 CNN was pretrained on ImageNet and fine-tuned on the ImageCLEFmed Caption dataset. However, we experimented also with adding an attention layer[10] [12] to our CNN. We call this model CNN+kNN@att. We also experimented with fine-tuning our CNN on a large dataset of radiography images called MIMIC-CXR [5] (before fine-tuning it further on the ImageCLEFmed Caption dataset), but we did not obtain any improvements.

4 Submissions and Results

In order to decide what models to use for the final submissions, we evaluated all models on our development set. Since images were separated into seven categories, each submission consisted of a model per category, resulting in seven models per submission. Two out of our three submissions employed different instances of the same system (2xCNN+FFNN@U & 2xCNN+FFNN@I), each for a different category, while the third one (called BEST@CATEGORY) combined results from different types of systems for each category.

The official measure of the competition was F1, macro-averaged over the images, without taking into account the different categories. To generate the predictions for the test set, we merged the training with the development set. We used a held-out set (20% of the merged data) to tune the hyper-parameters of the CNN+FFNN and CNN+k-NN models (see Table 4 and Table 4 for the final values). As shown in Fig. 5, the best score

[10] `https://bit.ly/323coXF`

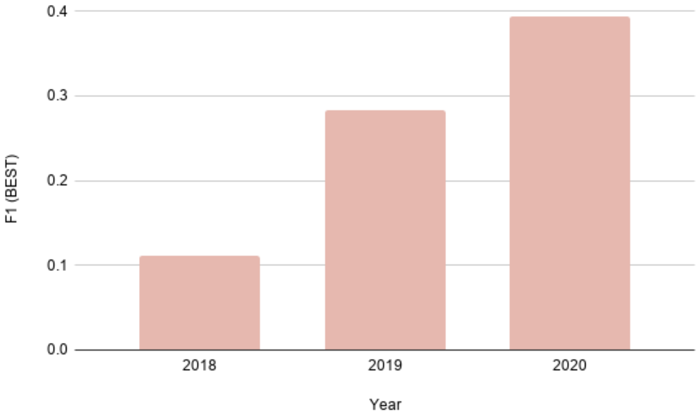

Fig. 5. The top F1 score achieved each year in the ImageCLEFmed Caption task.

Table 2. The F1 scores of our submitted models, measured on the development set.

	DRAN	**DRPE**	**DRCO**	**DRCT**	**DRMR**	**DRUS**	**DRXR**	**ANY**
BEST@CATEGORY	**0.3012**	**0.2485**	**0.1650**	**0.4456**	**0.3413**	**0.3093**	**0.3656**	**0.3715**
2xCNN+FFNN@U	0.3012	0.2485	0.1554	0.4455	0.3388	0.3063	0.3651	0.3704
2xCNN+FFNN@I	0.2995	0.2388	0.1560	0.4456	0.3400	0.3093	0.3656	0.3710

improves every year. This is probably because the task was indeed simplified each year (see Section 2).

Table 3. The results and rankings of our systems on the development and test set. The baseline of the last row only predicts the concepts that always appear in the images of the category.

Approach	**F1 Score**		**Ranking**
	Development	**Test**	
BEST@CATEGORY	**0.3715**	0.3933	2
2xCNN+FFNN@U	0.3704	0.3870	6
2xCNN+FFNN@I	0.3710	**0.3940**	1
BASELINE	0.3666	–	–

Table 4 presents the scores of our systems on the development and the official test set, along with the official rankings. 2xCNN+FFNN@I was the best. On the other hand, 2xCNN+FFNN@U, which returns the union (instead of the intersection) of the predicted concepts of the models in the ensemble, was ranked much lower. It is worth mentioning that a baseline, which simply returns the concepts always shown per category, achieves very high F1 on the development set. The submission that combined the best model per category (see Table 4) was ranked 2nd.

We noticed that models tended to predict only the concepts that always appear in each category, thus we also show statistics regarding the diversity of our submissions

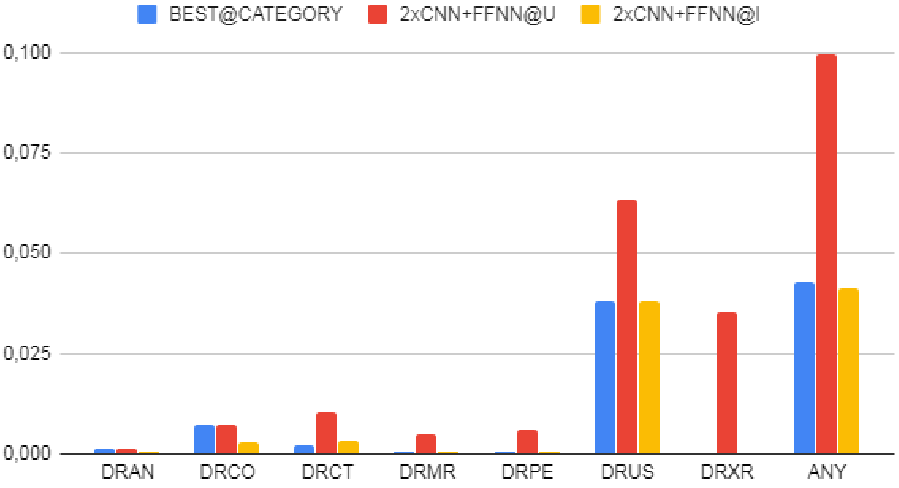

Fig. 6. The diversity (number of distinct concepts predicted / number of all possible concepts) for each of our submitted models per category.

for the test set (Fig. 6). We define diversity as the total number of distinct concepts the models predicted for a specific category divided by the total number of concepts found in the training set of that category. Observing the output of our models, we noticed that the ones with very low diversities only predicted the concepts that appeared in every image (as shown in Table 1) of the category they were trained on.

Table 4. The best models for each category according to the development scores, along with the hyper-parameter values used for the final submissions. t_1 and t_2 are the classification thresholds of the two models included in 2xCNN+FFNN, k is the number of the nearest images and r is the the number of concepts or the function that defines the number of concepts.

Category:	DRAN	DRPE	DRCO	DRCT	DRMR	DRUS	DRXR
Best Model:	2xCNN+FFNN@U		CNN+k-NN@att	CNN+k-NN		2xCNN+FFNN@I	
Hyper-parameters:	$t_1 = 0.34$ $t_2 = 0.7$	$t_1 = 0.3$ $t_2 = 0.23$	k = 91 r = average	k = 5 r = 2	k = 6 r = 1	$t_1 = 0.18$ $t_2 = 0.22$	$t_1 = 0.45$ $t_2 = 0.86$

Table 5. The thresholds used in our 2xCNN+FFNN ensembles, one for each CNN+FFNN. The two CNN+FFNN models and their corresponding thresholds are different for each category.

Category:	DRAN	DRPE	DRCO	DRCT	DRMR	DRUS	DRXR
Parameters:	$t_1 = 0.34$ $t_2 = 0.7$	$t_1 = 0.3$ $t_2 = 0.23$	$t_1 = 0.14$ $t_2 = 0.08$	$t_1 = 0.73$ $t_2 = 0.5$	$t_1 = 0.28$ $t_2 = 0.98$	$t_1 = 0.18$ $t_2 = 0.22$	$t_1 = 0.45$ $t_2 = 0.86$

5 Conclusions and Future Work

This article described the submissions of AUEB's NLP Group to the 2020 ImageCLEFmed Caption task. One of our submissions was ranked 1st, while the other two were ranked 2nd and 6th. All of our systems were based on a DenseNet-121 CNN [3] to encode images. A retrieval-based method achieved the best results in three out of seven categories. However, an ensemble of two CNN+FFNN multi-label classifiers was ranked 1st overall. Future work includes the assessment of our models on more datasets and improving retrieval-based methods, which are still under-explored.

References

1. Bodenreider, O.: The unified medical language system (umls): integrating biomedical terminology. Nucleic acids research **32**(Database issue), 4 (Jan 2004)
2. Chokshi, F.H., Hughes, D.R., Wang, J.M., Mullins, M.E., Hawkins, C.M., Jr, R.D.: Diagnostic radiology resident and fellow workloads: a 12-year longitudinal trend analysis using national medicare aggregate claims data. Journal of the American College of Radiology **12**(7), 664–669 (2015)
3. 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)
4. Ionescu, B., Müller, H., Péteri, R., Abacha, A.B., Datla, V., Hasan, S.A., Demner-Fushman, D., Kozlovski, S., Liauchuk, V., Cid, Y.D., Kovalev, V., Pelka, O., Friedrich, C.M., de Herrera, A.G.S., Ninh, V.T., Le, T.K., Zhou, L., Piras, L., Riegler, M., l Halvorsen, P., Tran, M.T., Lux, M., Gurrin, C., Dang-Nguyen, D.T., Chamberlain, J., Clark, A., Campello, A., Fichou, D., Berari, R., Brie, P., Dogariu, M., Ştefan, L.D., Constantin, M.G.: Overview of the ImageCLEF 2020: Multimedia retrieval in medical, lifelogging, nature, and internet applications. In: Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the 11th International Conference of the CLEF Association (CLEF 2020), vol. 12260. LNCS Lecture Notes in Computer Science, Springer, Thessaloniki, Greece (September 22-25 2020)
5. Johnson, A.E.W., Pollard, T.J., Greenbaum, N.R., Lungren, M.P., ying Deng, C., Peng, Y., Lu, Z., Mark, R.G., Berkowitz, S.J., Horng, S.: Mimic-cxr-jpg, a large publicly available database of labeled chest radiographs (2019)
6. Kingma, D.P., Ba, J.: Adam: A Method for Stochastic Optimization. arXiv:1412.6980 (2014)
7. 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)
8. Kougia, V., Pavlopoulos, J., Androutsopoulos, I.: AUEB NLP Group at ImageCLEFmed Caption 2019. In: CLEF2019 Working Notes. CEUR Workshop Proceedings. pp. 9–12. Lugano, Switzerland (2019)
9. Kougia, V., Pavlopoulos, J., Androutsopoulos, I.: Medical Image Tagging by Deep Learning and Retrieval. In: Experimental IR Meets Multilinguality, Multimodality, and Interaction Proceedings of the Eleventh International Conference of the CLEF Association (CLEF 2020). Thessaloniki, Greece (2020)
10. 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)

11. Pelka, O., Friedrich, C.M., García Seco de Herrera, A., Müller, H.: Overview of the ImageCLEFmed 2020 concept prediction task: Medical image understanding. In: CLEF2020 Working Notes. CEUR Workshop Proceedings, CEUR-WS.org $$, Thessaloniki, Greece (September 22-25 2020)
12. Raffel, C., Ellis, D.P.W.: Feed-forward networks with attention can solve some long-term memory problems (2015)
13. 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)
14. Rimmer, A.: Radiologist shortage leaves patient care at risk, warns royal college. British Medical Journal **359** (2017)
15. S. Singh, S. Karimi, K.H.S., Hamey, L.: Biomedical Concept Detection in Medical Images: MQ-CSIRO at 2019 ImageCLEFmed Caption Task. In: CLEF2019 Working Notes. CEUR Workshop Proceedings. p. 15. Lugano, Switzerland (2019)
16. 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)