IUST_NLPLAB at ImageCLEFmedical Caption Tasks 2023

Notebook for the ImageCLEF Lab at CLEF 2023

Yasaman Lotfollahi[1,†], Melika Nobakhtian[1,†], Malihe Hajihosseini[1,*] and Sauleh Eetemadi[2]

[1] *Student at School of Computer Engineering, Iran University of Science and Technology, Tehran, Islamic Republic Of Iran.*
[2] *Assistant Professor of Computer Science, School of Computer Engineering, Iran University of Science and Technology, Tehran, Islamic Republic Of Iran.*

Abstract

We present models implemented by the IUST_NLPLAB group for ImageCLEFmedical Caption Task 2023. This task contains two subtasks: Concept Detection and Caption Prediction. Under the first subtask, the model should extract medical concepts contained in radiology images. These concepts can be used for context-based image and information retrieval. Under the second subtask, the model predicts the caption for a medical image. This can be used for improving the diagnosis and treatment of diseases by saving time, money and helping physicians. This was our second experience to participate in this competition. We used diffrent models for both subtasks. We were able to get the 4th rank in the concepts detection subtask with a score of 0.49. Also, in the caption prediction subtask, we were able to get the 12th rank based on the BERTScore evaluation metric. This is despite the fact that our model has won the first rank based on ROUGE, BLEU and METEOR. From this, it can be concluded that the type of evaluation metric determined has an important effect on the results of this subtask.

Keywords

Medical Image Captioning, Concept Detection, Caption Prediction, Deep Learning

1. Introduction

ImageCLEF[1] is part of CLEF[1]. ImageCLEF was launched in 2003 and added a medical task in 2004. Although it started with four participants, in 2020 was able to attract more than one hundred and ten participants from all around the world to participate in the competition. ImageCLEF includes various sections that retrieve and classify visual information using textual and visual data and their combinations.

CLEF 2023: Conference and Labs of the Evaluation Forum, September 18–21, 2023, Thessaloniki, Greece

[*] Corresponding author.

[†] These authors contributed equally.

y_lotfollahi@comp.iust.ac.ir (Y. Lotfollahi); m_nobakhtian@comp.iust.ac.ir (M. Nobakhtian); m_hajihosseini@comp.iust.ac.ir (M. Hajihosseini); sauleh@iust.ac.ir (S. Eetemadi)
http://sauleh.github.io/ (S. Eetemadi)
0000-0002-4808-5141 (Y. Lotfollahi); 0000-0001-9788-2764 (M. Nobakhtian); 0000-0002-0067-1184 (M. Hajihosseini); 0000-0003-1376-2023 (S. Eetemadi)

CEUR Workshop Proceedings (CEUR-WS.org)

[1] Conference and Labs of the Evaluation Forum

In 2022, imageclef used the AIcrowd[2] platform to publish contest data and receive submissions from participating groups[2]. In that platform, groups could see the score earned after each submission and plan to improve their models'score. However, the score obtained by other groups could not be seen.

In ImageCLEF 2023[3], the contest data was made available to participating groups via a private GitHub link. Also, sciebo[3] system was used to receive the results sent by the groups. Participating groups could have a maximum of 10 successful submissions in each subtask. In each run, in addition to the test data results in csv format, a txt file containing a brief description of the method should be attached. Unfortunately, unlike the AIcrowd platform, in the sciebo system, the scores obtained after each submission were not presented, and it was not possible to improve the models and analyze them by comparing the obtained results.

In ImageCLEFmedical 2023, four tasks were proposed

1. Image Captioning.
2. Controlling the Quality of Synthetic Medical Images created via GANs.
3. Visual Question Answering for Colonoscopy Images.
4. Medical Dialogue Summarization.

We selected the Image Captioning task from the ImageCLEFmedical section to participate in the competition. ImageCLEF medical Image Captioning task in 2023[4], like last year, contained two subtasks: Concepts Detection and Caption Prediction. Each group could participate in one or both subtasks. In this paper, we present the methods our group, IUST_NLPLAB, from the Iran University of Science and Technology[4], School of Computer Engineering[5], Natural Language Pocessing Laboratory[6] used in both subtasks. This is our second time participating in the ImageCLEF competition. We participated in both subtasks and registered 7 successful submissions in the concept detection subtask and 10 successful submissions in the caption prediction subtasks.

In the concept detection subtask, we were able to win the fourth place in the competition with a gap of about 2 percent in F1-score from the first ranked group. Also, in the subtask of caption prediction, we were able to get the 12th rank of the competition based on BERTScore[5], which was the main evaluation metric of the competition. But based on other evaluation metrics such as ROUGE[6], BLEU[7] and METEOR[8], our group was able to win the first rank among other participating groups.

In the following sections, we will describe the task, datasets, models developed and the results we achieved in detail.

2. Task description

This year the ImageCLEF evaluation campaign hosted the 7th edition of the medical image caption task. Unlike some of the previous editions which only contained the caption prediction

[2] https://www.aicrowd.com/ (last accessed: 2023-07-05)
[3] https://hochschulcloud.nrw/en/ (last accessed: 2023-07-05)
[4] http://www.iust.ac.ir/en (last accessed: 2023-07-05)
[5] http://ce-inter.iust.ac.ir/ (last accessed: 2023-07-05)
[6] https://nlplab.iust.ac.ir (last accessed: 2023-07-05)

Table 1
Most frequent concepts in the training data

UMLS CUI	UMLS Meaning	frequency
C0040405	X-Ray Computed Tomography	20955
C1306645	Plain x-ray	17108
C0024485	Magnetic Resonance Imaging	10062
C0041618	Ultrasonography	8390
C0817096	Chest	6805
C1999039	Anterior-Posterior	5907
C0449900	Contrast used	4945
C0002978	angiogram	4194
C0037303	Bone structure of cranium	3058
C1996865	Postero-Anterior	2911
C0039985	Plain chest X-ray	2884
C0000726	Abdomen	2824
C0030797	Pelvis	2590
C0023216	Lower Extremity	1989
C0205129	Sagittal	1930

task (e.g., 2016[9]) or only the concept detection task (e.g., 2019[10]), the 7th edition, as like as last year, contained both subtasks as described below.

2.1. Concept Detection

In this subtask, the goal is to extract medical concepts in medical images. These concepts are selected from UMLS[7][11] Concept Unique Identifiers (CUIs) specified in the dataset. The extraction of these concepts can be used for image retrieval and context-based information purposes.

The 2023 dataset contains 2,125 medical concepts, which has decreased compared to last year's dataset. Table 1 shows a list of the 15 most frequent concepts in the training collection based on their frequency. According to the table published in the [2], most of the most frequent concepts of the 2023 dataset are in common with the most frequent concepts of the 2022 dataset, but their frequency has decreased compared to last year. The lowest rate frequency of concepts in the training set is related to six concepts, each of them was repeated only 2 times.

2.2. Caption Prediction

In this subtask, the goal is to generate a suitable caption for the input medical images. Extracting medical concepts can help in producing a more appropriate captions. This subtask consists of a combination of text and image processing and is more complicated than the previous subtask.

[7] Unified Medical Language System®

3. Data

The dataset introduced for the ImageCLEFmedical Caption 2023 is a subset of the Radiology Objects in COntext (ROCO)[12] dataset. The dataset published in 2023 was structurally similar to the dataset of 2022. In this year's dataset, there were 60,918 training data, which was reduced compared to last year's dataset, but the validation and testing datasets included 10,437 and 10,473 data, respectively, which increased compared to last year. For each image in the training and validation dataset, the concepts in the image and a suitable caption of it were provided.

In the following, more details of the data of each subtask are provided.

3.1. Image Concepts

In this subtask, each image in the dataset has several related concepts. These concepts have originated from the Unified Medical Language System (UMLS)[11] Concept Unique Identifiers (CUIs). The generated concepts are based on a reduced subset of the UMLS 2022 AB release[8] this year. Filtering images according to their semantic type was performed to reach a higher possibility of recognizing concepts in images. Concepts with a low occurrence were removed based on recommendations from previous years.

Each image has a different number of concepts. The overall number of concepts is 2125. An image has at least one related concept and at most 24 concepts. Most images in the dataset have three concepts.

3.2. Image Captions

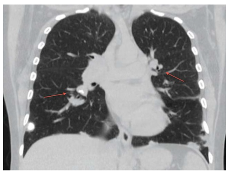
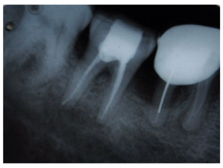
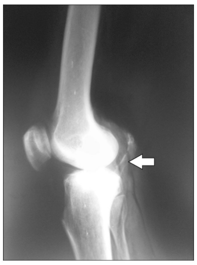
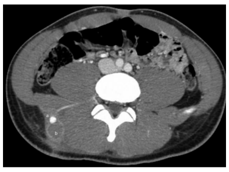
A caption is provided for each image in the training and validation sets in this subtask. Last year, the provided captions were pre-processed in four stages, but according to the explanations provided by the organizers, in this year only one pre-processing step, removal of links from the captions, was done on the captions.

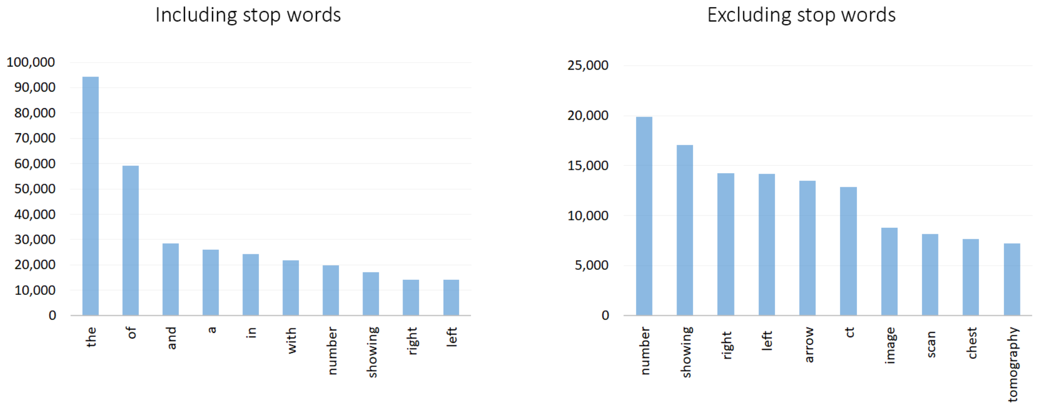
Based on the analysis performed on the training dataset, 63 images have one-word captions, which is the shortest caption length in the dataset. The maximum length of the caption is 410 words, which is related to one image. Also, the average number of words in captions is 20 words. We also calculated the TTR[9] for this annotation dataset. TTR is obtained by dividing the number of unique words by the text size and is a simple measure of lexical diversity[13]. Considering the stop words, the TTR value in this dataset is 0.07 and without considering the stop words, it is 0.05, both of them have increased compared to last year. Figure 1 displays the frequently recurring words in the captions of the training set, along with their frequency including and excluding stop words.

[8] https://www.nlm.nih.gov/pubs/techbull/nd22/nd22_umls_2022ab_release_available.html (last accessed: 2023-07-05)

[9] Type-token ratio

Table 2
Sample images from the training set along with their concepts and captions[14, 15, 16, 17].

Image	Concepts	Caption
 CC BY-NC [Unterstell et al. (2013)]	• C0040405 (X-Ray Computed Tomography) • C0817096 (Chest) • C0497156 (Lymphadenopathy)	• Thoracic CT scan showing perihilar pulmonary lymphadenomegaly.
 CC BY [Paul et al. (2015)]	• C1306645 (Plain x-ray) • C0037303 (Bone structure of cranium) • C0001168 (Complete obstruction)	• Obturation.
 CC BY-NC [Ananthi Kumarasamy et al. (2014)]	• C1306645 (Plain x-ray) • C0023216 (Lower Extremity) • C0205129 (Sagittal) • C0449900 (Contrast used) • C0017067 (Ganglia) • C0206207 (Joint Capsule)	• Contrast X-ray of right knee. Lateral view showing the stalk of the ganglion not communicating with the joint capsule.
 CC BY-NC [Counihan et al. (2015)]	• C0040405 (X-Ray Computed Tomography) • C0449900 (Contrast used) • C0000726 (Abdomen) • C0030797 (Pelvis) • C1510412 (Pseudoaneurysm) • C0205417 (Lobular) • C0278403 (Subcutaneous Tissue) • C0223651 (Iliac crest structure) • C0018944 (Hematoma)	• Contrast-enhanced CT scan of the lower abdomen and pelvis showing a single lobe of a presumed, bilobed pseudoaneurysm (a) as well as a 3.5 × 5.5 × 6 cm rim-enhancing, lobular collection of the superior right gluteal subcutaneous tissues, just superior to the right iliac crest and lateral to the paraspinal musculature, consistent with a hematoma (b)



(a) Ten most frequent words in the training set.

(b) Ten most frequent words in the training set without stop words.

Figure 1: Ten most frequent words in the training set

4. Methods

In this section, we present our methods for concept detection and caption prediction subtask.

4.1. Concept Detection

In concept detection subtask, we used different image preprocessing methods. When we used CLIP[18] and PubMedCLIP[19] models, we used their preprocess method. When used other pretrained models such as Resnet[20] and Efficientnet[21], we used CLAHE[22]. CLAHE is the one of ways to increase quality of image.

In the following, we explain our developed models for concept detection subtask. We used two methods: Ensemble Models as v1 and Multi-Label Classification Method as v2. Table 3 shows the details of the all developed models in concept detection subtask.

4.1.1. Ensemble Models

One of the systems that we designed was based on ensemble systems. We adopted this method according to the winner of last year, the AUEB-NLP group[23]. We utilized two instances of EfficientNetV2B0[24] for this model. All layers of base models were frozen until the last convolutional layer during the training process. A dense layer was added to each model to predict concepts. Models were trained for 50 epochs with a batch size of 256. We considered different thresholds for each concept to find out if a concept is related to an image or not. We tried certain thresholds on validation data and found the best one regarding F1-score. Every five-epoch model weights and best thresholds were saved. After training, the best weights and thresholds were chosen for each model. To predict concepts, if both models assigned a

Table 3
Concept detection submissions' details of IUST_NLPLAB group.

Run ID	Type	Base model	Best Epochs	Batch size	Best Threshold	F1-score
7	v2	PubMedCLIP ViT-B/32	20	256	0.2	0.495
5	v2	Resnet50	5	256	0.3	0.485
6	v2	CLIP ViT-B/32	30	256	0.3	0.440
3	v1	EfficientnetB0	50	256	-	0.435
4	v2	Resnet50	5	128	0.3	0.433
8	v2	PubMedCLIP RN50	30	256	0.2	0.421
2	v2	PubMedCLIP RN50x4	20	256	0.3	0.380

concept to an image, we concluded that this image has this concept, in other words, we used an intersection of concepts.

4.1.2. Multi-Label Classification Method

In this approach, we built a multi-label classification model to predict the correct concepts for each input image. We used CNNs with pre-trained weights from ImageNet[25]. These networks were modified by removing their final layer and adding a classification layer. Then, they were fine-tuned on the target dataset. We tried fine-tuning different pre-trained models and applied various thresholds to find the best results. In this year, we also used the vision-language models of CLIP[18] and its medical version, PubMedCLIP[19], which had achieved good results in many tasks.

4.2. Caption Prediction

In the caption prediction subtask, we utilized the approach we developed for the previous year's challenge, where we achieved first place. This methodology treats each word in a caption as a label corresponding to the associated image. We trained a multi-label classification model to predict the words that will ultimately form a caption for the given image.

To extract image features for the subtask, we used a pre-trained CNN on ImageNet[25] and fine-tuned it on the training set. During fine-tuning, we excluded the last layer of the CNN and added a dropout layer and a dense layer. We tried various CNN models to explore different possibilities. Similar to the concept detection subtask, in this subtask, we used the vision-language models of CLIP[18] and its medical version, PubMedCLIP[19].

The model generates captions for images by predicting the corresponding words. The probability of each word is computed in the output layer using the sigmoid activation function. Two methods are used to select the candidate words:

1. The top N words with the highest probability are chosen. N is a hyper-parameter that will define the length of captions.
2. A threshold is applied to the model output. Words with probabilities higher than the threshold are chosen to create the caption.

Table 4
Caption prediction submissions' details of IUST_NLPLAB group.

Run ID	Base model	Best Epochs	Word limit	Threshold	Sorted
6	PubMedCLIP RN50x4	90	20	-	Yes
2	ResNet50	20	19	-	Yes
4	CLIP RN50x4	90	20	-	Yes
10	PubMedCLIP RN50x4	95	-	0.1	Yes
8	CLIP RN50x4	95	-	0.1	Yes
5	PubMedCLIP RN50x4	90	20	-	No
1	ResNet50	20	19	-	No
3	CLIP RN50x4	90	20	-	No
9	PubMedCLIP RN50x4	95	-	0.1	No
7	CLIP RN50x4	95	-	0.1	No

Table 5
Caption prediction submissions' results of IUST_NLPLAB group.

Run ID	BERTScore	ROUGE	BLEURT	BLEU	METEOR	CIDEr	CLIPScore
6	0.567	0.290	0.223	0.268	0.104	0.177	0.807
2	0.565	0.271	0.209	0.241	0.089	0.159	0.805
4	0.561	0.280	0.210	0.259	0.095	0.162	0.806
10	0.556	0.275	0.212	0.264	0.096	0.142	0.801
8	0.553	0.269	0.203	0.264	0.096	0.134	0.803
5	0.549	0.290	0.201	0.268	0.100	0.174	0.804
1	0.546	0.271	0.189	0.241	0.089	0.156	0.803
3	0.544	0.280	0.186	0.259	0.094	0.158	0.803
9	0.539	0.275	0.177	0.264	0.096	0.142	0.797
7	0.537	0.269	0.168	0.264	0.095	0.133	0.797

After extracting the correct words, we need to sort them to create the full caption. Two methods are used to arrange the words:

1. Words are arranged from highest to lowest probability.
2. Words are ordered based on their statistical occurrence within the training set. Each word is assigned to its most common position in the caption.

Different values of N and threshold were applied to the output. Although BERTScore[5] is the primary score in this year's competition, we were not able to use this metric to evaluate our models because of our resource limits. Therefore, we used the BLEU[7] score to evaluate our models and find the best hyperparameters. Details of each submission and their results are described in table 4 and 5.

5. Results

In the previous parts, the details of the models implemented by our group and the results obtained by each one were explained.

In the concept detection subtask, two metrics, F1-Score and F1-Score Manual, which was calculated using a subset of manually validated concepts, were used to evaluate the models, but the results of the competition was based on the F1-Score metric. In 2023, 9 groups from all over the world participated in the concept detection subtask and managed to register successful submissions. The details of the results announced by the organizers of the competition in this subtask are presented in Table 6. Among the submissions of our group, Run ID 7, which used the basic model of PubMedCLIP ViT-B/32[19] to extract the features of the images, was able to get the best result with a difference of about 2 percent from the first group and achived 4th rank in this competition, which has increased four ranks compared to last year's results.

In the caption prediction subtask, seven metrics: BERTScore[5], ROUGE[6], BLEURT[26], BLEU[7], METEOR[8], CIDEr[27] and CLIPScore[28], were used to evaluate the models, but the results of the competition was based on the BERTScore metric. In 2023, 13 groups from all over the world participated in the caption prediction subtask and managed to register successful submissions. The details of the results announced by the organizers of the competition in this subtask are presented in Table 7. Among the submissions of our group, Run ID 6, which used the basic model of PubMedCLIP RN50x4[19] to extract the features of the images, was able to get the best result with a difference of about 7 Percent from the first group and achived 12th rank in this competition.

The noteworthy point is that based on three metrics ROUGE, BLEU and METEOR, our group has been able to get the first rank among the participating groups.

Differences in rankings based on different metrics can show challenges in evaluating generated captions. This is due to the differences in how these metrics evaluate the quality of generated captions. For example, the BLEU score measures n-gram overlap between the generated and reference captions, rewarding precision, and the presence of matching n-grams. In contrast, BERTScore used contextualized embeddings from BERT to capture semantic similarity, taking into account both word order and correctness. Consequently, a result with higher n-gram overlap but potential issues in word order or overall fluency could receive a better BLEU score and a lower BERTScore.

While our models can generate relevant words to describe the input image, they struggle to shape them into actual sentences which are semantically similar to the original caption. This issue can be one of the reasons why our results have low BERTScores, while having high BLEU scores.

6. Conclusion

This paper describes the participation of IUST_NLPLAB at Iran University of Science and Technology at ImageCLEFmedical caption 2023 task.

In the concept detection subtask, we ranked 4 among 9 participating teams. We used MLC and ensemble models in this subtask. Our MLC methods with PubMedCLIP ViT-B/32 as a base model had better overall score.

In the caption prediction subtask, last year, our group won the first place in the competition based on the BLEU evaluation metric, but this year it won the 12th place in the competition based on the BERTScore metric. Based on the published results, our group was able to win

Table 6
Results of ImageCLEFmedical 2023 concept detection subtask.

Team Name	Best Run ID	F1-Score	F1-Score Manual
AUEB-NLP-Group	4	0.522272	0.925842
KDE-Lab_Med	10	0.507414	0.932091
VCMI	8	0.499812	0.916184
IUST_NLPLAB	7	0.495863	0.880381
Clef-CSE-GAN-Team	1	0.495730	0.910585
CS_Morgan	2	0.483401	0.890151
SSNSheerinKavitha	1	0.464894	0.860296
closeAI2023	5	0.448105	0.856928
SSN_MLRG	3	0.017250	0.112211

Table 7
Results of ImageCLEFmedical 2023 Caption prediction subtask.

Team Name	BERTScore	ROUGE	BLEURT	BLEU	METEOR	CIDEr	CLIPScore
closeAI2023	0.628106	0.240061	0.320915	0.184624	0.087254	0.237704	0.807454
AUEB-NLP-Group	0.617034	0.213014	0.295011	0.169212	0.071982	0.146601	0.803888
PCLmed	0.615190	0.252756	0.316561	0.217150	0.092063	0.231535	0.802123
VCMI	0.614736	0.217545	0.308386	0.165322	0.073449	0.172042	0.808184
KDE-Lab_Med	0.614538	0.222341	0.301391	0.156465	0.072441	0.181853	0.806207
SSN_MLRG	0.601933	0.211177	0.277434	0.141797	0.061514	0.128443	0.775915
DLNU_CCSE	0.600546	0.202888	0.262998	0.105948	0.055716	0.133207	0.772518
CS_Morgan	0.581949	0.156419	0.224238	0.056632	0.043649	0.083982	0.759258
Clef-CSE-GAN-Team	0.581625	0.218103	0.269043	0.145035	0.070155	0.173664	0.789327
Bluefield-2023	0.577966	0.153448	0.271642	0.154316	0.060069	0.100910	0.783725
IUST_NLPLAB	0.566886	0.289774	0.222957	0.268452	0.100354	0.177266	0.806763
SSNSheerinKavitha	0.544106	0.086648	0.215170	0.074905	0.025768	0.014313	0.687312

first place in all 10 of its submissions based on the three metric ROUGE, BLEU and METEOR. Based on these results, it can be concluded that the selection of evaluation metric in the analysis of the models presented for this subtask is very important and the results based on different evaluation metric can have significant differences from each other.

This year was our second experience of participating in this competition and we hope to be able to participate in these competitions in the coming years and gain new experiences.

Acknowledgments

This work has been supported by School of Computer Engineering of Iran University of Science and Technology and we are very grateful for their support.

References

[1] B. Ionescu, H. Müller, R. Peteri, J. Rückert, A. Ben Abacha, A. G. S. de Herrera, C. M. Friedrich, L. Bloch, R. Brüngel, A. Idrissi-Yaghir, H. Schäfer, S. Kozlovski, Y. D. Cid, V. Kovalev, L.-D. Ştefan, M. G. Constantin, M. Dogariu, A. Popescu, J. Deshayes-Chossart, H. Schindler, J. Chamberlain, A. Campello, A. Clark, Overview of the ImageCLEF 2022: Multimedia retrieval in medical, social media and nature applications, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction, Proceedings of the 13th International Conference of the CLEF Association (CLEF 2022), LNCS Lecture Notes in Computer Science, Springer, Bologna, Italy, 2022.

[2] M. Hajihosseini, Y. Lotfollahi, M. Nobakhtian, M. M. Javid, F. Omidi, S. Eetemadi, Iust_nlplab at imageclefmedical caption tasks (2022).

[3] B. Ionescu, H. Müller, A. Drăgulinescu, W. Yim, A. Ben Abacha, N. Snider, G. Adams, M. Yetisgen, J. Rückert, A. Garcıa Seco de Herrera, C. M. Friedrich, L. Bloch, R. Brüngel, A. Idrissi-Yaghir, H. Schäfer, S. A. Hicks, M. A. Riegler, V. Thambawita, A. Storås, P. Halvorsen, N. Papachrysos, J. Schöler, D. Jha, A. Andrei, A. Radzhabov, I. Coman, V. Kovalev, A. Stan, G. Ioannidis, H. Manguinhas, L. Ştefan, M. G. Constantin, M. Dogariu, J. Deshayes, A. Popescu, Overview of ImageCLEF 2023: Multimedia retrieval in medical, socialmedia and recommender systems applications, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction, Proceedings of the 14th International Conference of the CLEF Association (CLEF 2023), Springer Lecture Notes in Computer Science LNCS, Thessaloniki, Greece, 2023.

[4] J. Rückert, A. Ben Abacha, A. G. Seco de Herrera, L. Bloch, R. Brüngel, A. Idrissi-Yaghir, H. Schäfer, H. Müller, C. M. Friedrich, Overview of ImageCLEFmedical 2023 – Caption Prediction and Concept Detection, in: CLEF2023 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, Thessaloniki, Greece, 2023.

[5] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, Y. Artzi, Bertscore: Evaluating text generation with bert, arXiv preprint arXiv:1904.09675 (2019).

[6] C.-Y. Lin, Rouge: A package for automatic evaluation of summaries, in: Text summarization branches out, 2004, pp. 74–81.

[7] K. Papineni, S. Roukos, T. Ward, W.-J. Zhu, Bleu: a method for automatic evaluation of machine translation, in: Proceedings of the 40th annual meeting of the Association for Computational Linguistics, 2002, pp. 311–318.

[8] M. Denkowski, A. Lavie, Meteor universal: Language specific translation evaluation for any target language, in: Proceedings of the ninth workshop on statistical machine translation, 2014, pp. 376–380.

[9] A. García Seco de Herrera, R. Schaer, S. Bromuri, H. Müller, Overview of the ImageCLEF 2016 medical task, in: Working Notes of CLEF 2016 (Cross Language Evaluation Forum), 2016.

[10] B. Ionescu, H. Müller, R. Péteri, Y. D. Cid, V. Liauchuk, V. Kovalev, D. Klimuk, A. Tarasau, A. B. Abacha, S. A. Hasan, V. Datla, J. Liu, D. Demner-Fushman, D.-T. Dang-Nguyen, L. Piras, M. Riegler, M.-T. Tran, M. Lux, C. Gurrin, O. Pelka, C. M. Friedrich, A. G. S. de Herrera, N. Garcia, E. Kavallieratou, C. R. del Blanco, C. C. Rodríguez, N. Vasillopoulos, K. Karampidis, J. Chamberlain, A. Clark, A. Campello, ImageCLEF 2019: Multimedia retrieval in

medicine, lifelogging, security and nature, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction, Proceedings of the Tenth International Conference of the CLEF Association (CLEF 2019), LNCS Lecture Notes in Computer Science, Springer, Lugano, Switzerland, 2019.

[11] O. Bodenreider, The unified medical language system (umls): integrating biomedical terminology, Nucleic acids research 32 (2004) D267–D270.

[12] O. Pelka, S. Koitka, J. Rückert, F. Nensa, C. M. Friedrich, Radiology objects in context (roco): a multimodal image dataset, in: Intravascular Imaging and Computer Assisted Stenting and Large-Scale Annotation of Biomedical Data and Expert Label Synthesis: 7th Joint International Workshop, CVII-STENT 2018 and Third International Workshop, LABELS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16, 2018, Proceedings 3, Springer, 2018, pp. 180–189.

[13] K. Kettunen, Can type-token ratio be used to show morphological complexity of languages?, Journal of Quantitative Linguistics 21 (2014) 223–245.

[14] N. Unterstell, A. L. Bressan, L. A. Serpa, P. P. d. Fonseca e Castro, A. C. Gripp, Systemic sarcoidosis induced by etanercept: first brazilian case report, An. Bras. Dermatol. 88 (2013) 197–199.

[15] A. Zbiciak, T. Markiewicz, A new extraordinary means of appeal in the polish criminal procedure: the basic principles of a fair trial and a complaint against a cassatory judgment, Access to Justice in Eastern Europe 6 (2023) 1–18.

[16] S. Ananthi Kumarasamy, B. S. Kannadath, S. Soundamourthy, A. Subramanian, S. P. Sinhasan, R. V. Bhat, Semimembranosus ganglion cyst, Anat. Cell Biol. 47 (2014) 207–209.

[17] M. Counihan, M. E. Pontell, B. Selvan, A. Trebelev, A. Nunez, Delayed presentation of a lumbar artery pseudoaneurysm resulting from isolated penetrating trauma, J. Surg. Case Rep. 2015 (2015) rjv083.

[18] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al., Learning transferable visual models from natural language supervision, in: International conference on machine learning, PMLR, 2021, pp. 8748–8763.

[19] S. Eslami, G. de Melo, C. Meinel, Does clip benefit visual question answering in the medical domain as much as it does in the general domain?, arXiv preprint arXiv:2112.13906 (2021).

[20] M. Talo, Convolutional neural networks for multi-class histopathology image classification, ArXiv abs/1903.10035 (2019).

[21] M. Tan, Q. Le, Efficientnet: Rethinking model scaling for convolutional neural networks, in: International conference on machine learning, PMLR, 2019, pp. 6105–6114.

[22] K. Zuiderveld, Contrast limited adaptive histogram equalization, Graphics gems (1994) 474–485.

[23] V. Kougia, J. Pavlopoulos, I. Androutsopoulos, Aueb nlp group at imageclefmed caption 2022, in: Conference and Labs of the Evaluation Forum, 2019.

[24] M. Tan, Q. V. Le, Efficientnetv2: Smaller models and faster training, 2021. arXiv:2104.00298.

[25] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, Imagenet: A large-scale hierarchical image database, in: 2009 IEEE conference on computer vision and pattern recognition, Ieee, 2009, pp. 248–255.

[26] T. Sellam, D. Das, A. P. Parikh, Bleurt: Learning robust metrics for text generation, arXiv

preprint arXiv:2004.04696 (2020).

[27] R. Vedantam, C. Lawrence Zitnick, D. Parikh, Cider: Consensus-based image description evaluation, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 4566–4575.

[28] J. Hessel, A. Holtzman, M. Forbes, R. L. Bras, Y. Choi, Clipscore: A reference-free evaluation metric for image captioning, arXiv preprint arXiv:2104.08718 (2021).