CMRE-UoG team at ImageCLEFmedical Caption 2022: Concept Detection and Image Captioning

Francesco Dalla Serra[1,2], Fani Deligianni[2], Jeffrey Dalton[2] and Alison Q O'Neil[1,3]

[1]Canon Medical Research Europe, Edinburgh, UK
[2]University of Glasgow, Glasgow, UK
[3]University of Edinburgh, Edinburgh, UK

Abstract

This work presents the proposed solutions of our team for the ImageCLEFmedical Caption 2022 task [1]. This task is structured as two subtasks: (1) *Concept Detection* subtask – which consists of detecting Concept Unique Identifiers (CUIs) from the Unified Medical Language System (UMLS) [2] attributed to each image; and (2) the *Caption Prediction* subtask – which involves generating an accurate description of the content of the image, based on the concepts detected in the first subtask. For both subtasks, the dataset corresponds to a subset of the Radiology Objects in the COntext (ROCO) dataset [3].

In the **Concept Detection** subtask, we experiment with two different strategies: a) *supervised learning* – we train a Convolutional Neural Network (CNN) [4, 5] to classify the full set of CUIs; b) *image retrieval* – we retrieve the top K most "similar" images from the training set based on the cosine similarity score between the image representations (extracted from the last average pooling layer), and combine the associated CUIs using a soft majority voting approach, similar to the ImageCLEFmed Caption 2021 winning approach [6]. Our best submission consists of the second image retrieval approach, for which we used an ensemble of five different CNNs. This approach ranked 2nd with an F1 score equal to 0.451, with a margin of approximately 5×10^{-4} from the 1st position.

In the **Caption Prediction** subtask, we adopt an image encoder-decoder Transformer model [7], which takes as input the image representation – generated using a CNN image encoder – and generates a text caption describing the image. Furthermore, we considered a multimodal encoder-decoder Transformer model, which differs from the previous by taking as an additional input the CUIs extracted from the previous subtask alongside an image representation. Our multimodal approach ranked 6th, with a BLEU score [8] of 0.291, and ranked 1st place in terms of ROUGE [9] (the secondary metric for this subtask), with a score of 0.201.

Keywords

ImageCLEF, Concept Detection, Image Captioning, Image Retrieval, Medical Imaging, UMLS, Convolutional Neural Network, Transformer

1. Introduction

The ImageCLEFmedical Caption 2022 challenge [1] is the 6th edition of the Caption Task, organized by the CLEF initiative[1] and part of ImageCLEF 2022 [10]. This challenge aims at

CLEF 2022: Conference and Labs of the Evaluation Forum, September 5–8, 2022, Bologna, Italy
francesco.dallaserra@mre.medical.canon (F. Dalla Serra); fani.deligianni@glasgow.ac.uk (F. Deligianni); jeff.dalton@glasgow.ac.uk (J. Dalton); alison.oneil@mre.medical.canon (A. Q. O'Neil)
0000-0001-8863-1533 (F. Dalla Serra); 0000-0003-1306-5017 (F. Deligianni); 0000-0003-2422-8651 (J. Dalton); 0000-0001-8371-0603 (A. Q. O'Neil)

[1]http://www.clef-initiative.eu/ [last accessed: 30.06.2022]

promoting research in the field of medical image captioning, which consists of describing the content of a medical image in the form of free text. This is normally a time-consuming task performed by highly specialized experts. Therefore, this could aid radiologists by speeding up the diagnosis and treatment workflow, while reducing human errors, due to extensive working hours or distractions.

This year's challenge was divided into two subtasks: *Concept Detection* and *Caption Prediction*. Concept Detection consists in detecting the Concept Unique Identifiers (CUIs) from the Unified Medical Language System (UMLS) [2] relevant to each image. This is considered to be the first step towards the Caption Prediction task, which consists in generating a textual description of the image based on the detected CUIs.

In these working notes, we present the different methods applied by our team in both subtasks. For the Concept Detection task, we considered a supervised learning approach, comparing two different Convolutional Neural Network (CNN) architectures [4, 5]; and an image retrieval approach, similar to last year's winning solution [6], where the predicted CUIs are selected from the top K most similar images. For the Caption Prediction subtask, we applied a Transformer encoder-decoder architecture [7], using the images alone or the image-CUI pairs as input. Our best approaches ranked 2nd out of 11 participating teams for the Concept Detection subtask (with a small margin in terms of F1 score from the 1st team) and 6th out of 10 participating teams for the Caption Prediction subtask.

2. Dataset

The data released for the ImageCLEFmedical Caption 2022 task correspond to a subset of the Radiology Objects in the COntext (ROCO) dataset [3]. This multimodal dataset contains image-caption pairs collected from open access biomedical journal articles in PubMedCentral[2]. This dataset comprises several different medical imaging modalities, including Computed Tomography, Ultrasound, X-Ray, Fluoroscopy, Positron Emission Tomography, Mammography, Magnetic Resonance Imaging, Angiography and PET-CT. However, in this edition of the Caption Task, the modality information was not available to participants.

The dataset used for this challenge consists of 83,275 radiology images in the training set, 7,645 radiology images in the validation set, and 7,601 radiology images in the test set. Each image is paired with the associated caption and the set of extracted UMLS CUIs. The original split of the dataset is used throughout all our experiments and no additional sources are considered for both subtasks.

For the *Concept Detection* subtask, the set of CUIs assigned to each radiology image was extracted from the caption texts, based on the UMLS 2020 AB release. The organizers filtered them based on their semantic type and removed those with very low occurrences. This results in a total of 8,347 CUIs.

For the *Caption Prediction* subtask, the original captions of each image were pre-processed by the organizers as follows: (1) removing numbers and words containing numbers; (2) removing the punctuation; (3) applying lemmatization; and (4) converting the text to lower-case.

[2] https://www.ncbi.nlm.nih.gov/pmc/ [last accessed: 30.06.2022]

3. Methods

In this section we describe the methods implemented by our team for both ImageCLEFmedical Caption 2022 subtasks: *Concept Detection* and *Caption Prediction*.

3.1. Concept Detection

We participated in the Concept Detection task with four different submissions. These can be categorized into two main approaches, as described below.

3.1.1. Supervised Learning

Two different CNN architectures were considered and fine-tuned on the full set of CUIs contained in the training set. Namely, ResNet-152 [4] and DenseNet-201 [5] are the two CNN considered for this task. The task was framed as a multi-label classification task, where each label corresponds to a different CUI, resulting in a total of $8,374$ classes.

3.1.2. Image Retrieval

Following last year's winning method [6], we considered an image retrieval approach, where a set of CUIs were assigned to each query image (corresponding to an image in the validation and test set) based on the top N most similar images from the training set and their associated CUIs.

More specifically, this was achieved by considering the following three steps. (1) A single DenseNet-201 was fine-tuned with supervised learning, as described in 3.1.1. (2) After discarding the final fully-connected layer, the CNN was used as image encoder, and the top N most similar images to a query image were retrieved based on the cosine similarity between image embeddings. (3) An aggregation step was performed to select the set of CUIs to associate to each query image. This consisted in a *soft majority voting*, similar to the method proposed in [6], where a CUI is attributed to the query image if it appears in at least 30% of the retrieved images. The proposed pipeline is shown in Figure 1. Differently from [6], which considers only the CUIs appearing in at least 50% of the retrieved images, we aim to assign also less frequent CUIs by taking those that appear in at least 30% of retrieved images. Another difference from [6] and our method, is that (for simplicity) we only consider a single DenseNet-201 model to retrieve N different images.

Alternatively, an ensemble of five DenseNet-201 models was considered, each fine-tuned using a different seed. The three steps described above were performed individually for each model. Finally, for each image, we assigned the union of each model's predicted CUIs.

3.2. Caption Prediction

Our team participated to the Caption Prediction task with six submissions, all based on a CNN-Transformer approach, similar to previous works in radiology report generation [13, 14]. The major difference among our submissions consists in a) the input modalities – image-only *vs.* multimodal (image & CUIs expressed as text), b) the text pre-processing – where we filter out

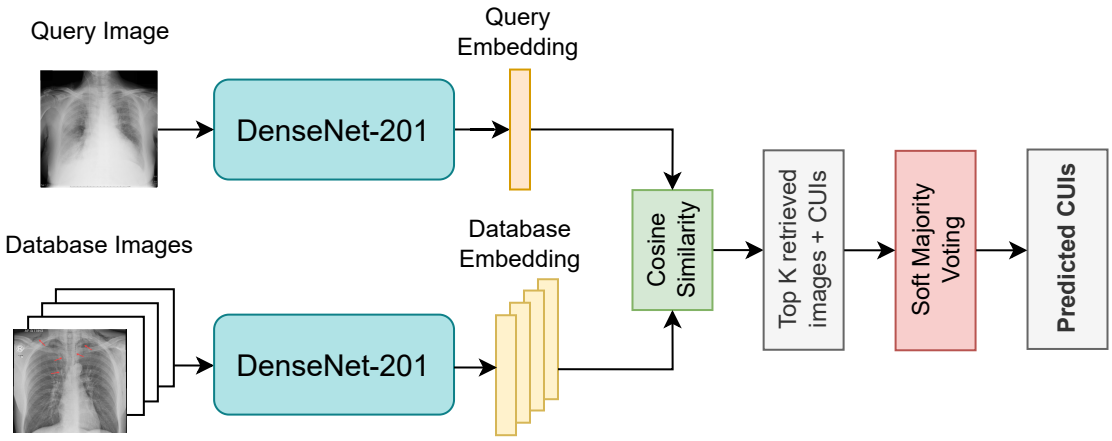

Figure 1: Diagram of the proposed image retrieval pipeline, using a single DenseNet-201 model, which was first fine-tuned on the full set of CUIs. The datatbase images correspond to the full ImageCLEFmedical Caption 2022 training set. The images are taken from the ImageCLEFmedical Caption 2022 dataset (*top image*: CC BY [Shimoyama et al. (2017)] [11]; *bottom image*: CC BY [Alnofal et al. (2021)] [12]).

the low frequency terms from the vocabulary, and c) the text post-processing – which consists of removing repeating words.

The chosen architecture consists of a ResNet-101 image encoder followed by a vanilla Transformer encoder-decoder [7], composed of 3 attention layers (for both the encoder and the decoder), 8 heads, 512 hidden units and no pre-trained weights initialization, similar to the [13] baseline method. The image encoder extracts a 49×2048 feature map from each image. Each of the $N = 49$ latent vectors is considered to be an image token which is inputted to the Transformer. In the multimodal setup, where the CUIs extracted in the Concept Detection task are considered as additional input, the extracted CUIs are concatenated into a single string and tokenized into M text tokens (one for each CUI). The textual tokens are based on a custom-built vocabulary, where each token corresponds to a word or a CUI appearing in the training set. A $[SEP]$ token is used to separate the two input modalities. Both the textual (CUIs) and the visual inputs are projected into the input embedding space and summed with the related positional and segment embedding; the segment embedding is then available to the model to allow distinction of textual and visual inputs. The model is treated as a language model, following the sequence-to-sequence paradigm. The multimodal Encoder-Decoder Transformer architecture is shown in Figure 2.

4. Experiments & Results

In this section, we describe the implementation details of the four submitted solutions for the *Concept Detection* subtask, and the six submissions for the *Caption Prediction* subtask. For each subtask, we highlight our best performing approach, and we show how it compares with other participating teams' solutions by presenting the final ranking.

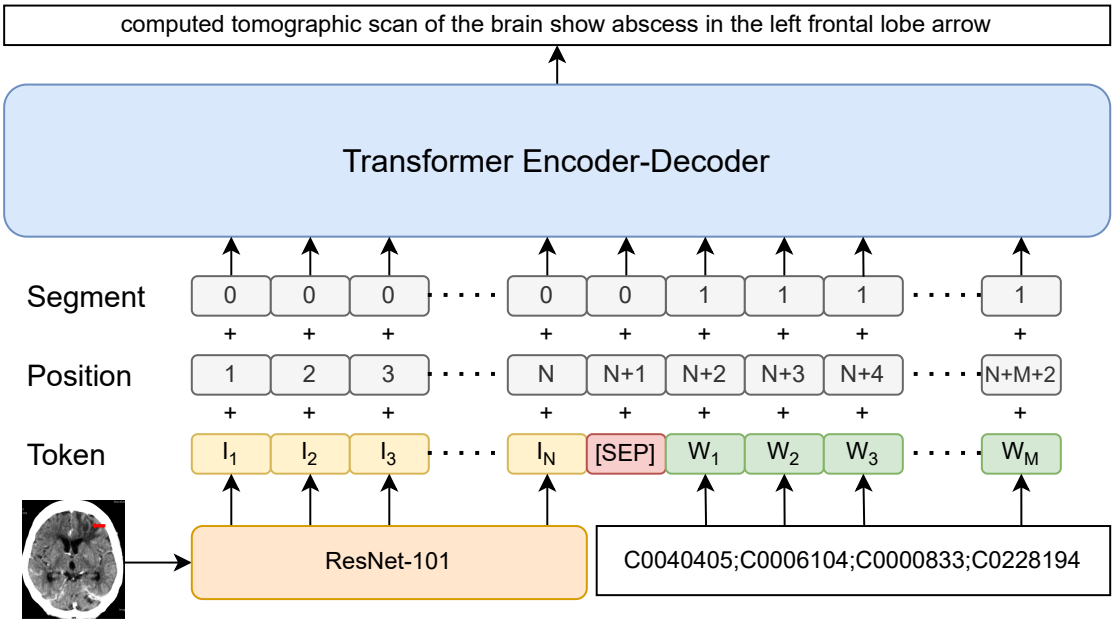

Figure 2: Multimodal Encoder-Decoder Transformer. The image, CUIs and caption are taken from the ImageCLEFmedical Caption 2022 dataset (CC BY [Wang et al. (2021)] [15]). The four CUIs in this example correspond to "X-Ray Computed Tomography", "Brain", "Abscess" and "Left frontal lobe structure", respectively.

4.1. Concept Detection

The first two submitted solutions – *#182230* and *#182232* – consist of a DenseNet-201 and a ResNet-152 architecture, respectively. These are initialized with ImageNet [16] pre-trained weights and fine-tuned using a binary cross entropy loss, on the $8,374$ CUIs. The same set of hyperparameters was used for both submissions. The two models were trained for 20 epochs, and a batch size of 64 on a single NVIDIA RTX A5000 GPU. The initial learning rate was set to 10^{-3} and was reduced by a factor of 0.5 when the F1 score, computed on the validation set, was not improving for 3 consecutive epochs. The input images are resized by matching the smaller edge to 224 pixels and maintaining the original aspect ratio. We applied the following data augmentation techniques: random horizontal flipping; and random crop of 224×224 pixels.

The third and fourth submissions – *#182260* and *#182324* – apply the image retrieval approach. Submission *#182260* is a single DenseNet-201 fine-tuned with the same set of hyperparameters described above, discarding the final fully-connected layer. We then computed the cosine similarity between the image embedding in the test set and the training set (the training set is considered as our database), extracted from the last average pooling layer of DenseNet-201. The CUIs associated with the 50 images in the database with the highest similarity score were retrieved. Finally, an aggregation step consists of selecting only the CUIs which appeared in at least 30% of the retrieved images, which we named *soft majority voting*. Following these steps, we predicted which CUIs to attribute to each image in the test set. Similarly, submission *#182324* consists of an ensemble of five different DenseNet-201, fine-tuned using different seeds.

Table 1
F1 scores on the test set, for each of our submitted solutions.

Run	Approach	Network	F1 Score
#182230	Supervised Learning	DenseNet-201	0.443
#182232	Supervised Learning	ResNet-152	0.440
#182260	Image Retrieval	DenseNet-201	0.446
#182324	Image Retrieval	5× DenseNet-201	**0.451**

Table 2
F1 and Secondary F1 scores on the test set for each team's best solution. The ranking is based on the F1 Score. For both metrics, we highlight in **bold** the best score and underline the second best score.

Team	Run	F1 Score	Secondary F1	Rank
AUEB-NLP-Group	#182358	**0.451**	0.791	1
Ours	#182324	0.451	0.822	2
CSIRO	#182343	0.447	0.794	3
eecs-kth	#181750	0.436	0.855	4
vcmi	#182097	0.433	**0.863**	5
PoliMi-ImageClef	#182296	0.432	0.851	6
SSNSheerinKavitha	#181995	0.418	0.654	7
IUST_NLPLAB	#182307	0.398	0.673	8
Morgan_CS	#182150	0.352	0.628	9
kdelab	#182346	0.310	0.412	10
SDVA-UCSD	#181691	0.308	0.552	11

Differently to *#182260*, we retrieved 100 different images for each of the five networks. Next, the *soft majority voting* step was applied to each of them. Finally, we assigned the union of each model's predicted CUIs to each image.

The results on the test set are shown in Table 1. We can notice that our best submission was an image retrieval approach, by ensembling five different DenseNet-201 models. This follows the trend of last year's winning solution [6].

We compare our top submission with other teams' submissions, considering also a *Secondary F1* score, computed on a manually curated concept subset, corresponding to anatomy and image modality only. The results are shown in Table 2. Our best submission ranked 2nd in this challenge, with a very small margin from the 1st position, in terms of F1 score[3]. Furthermore, we note that our best method achieved the 4th best Secondary F1 score (the best among the top 3 ranked solution), meaning that it effectively managed to predict the anatomy and image modality corresponding CUIs along with other types of CUIs.

[3]All the scores shown in our tables are rounded to the third decimal place, therefore, two submissions may appear to have the same score. A more detailed list of the best scores of each team can be seen on the challenge web page: https://www.imageclef.org/2022/medical/caption [last accessed: 30.06.2022].

Table 3
BLEU scores on the test set, for each our submitted solutions.

Run	Input	Pre-Processing	Post-Processing	BLEU
#182349	Image			0.271
#182255	Image	✓		0.280
#182273	Image	✓	✓	0.276
#182350	Image + CUIs			0.278
#182342	Image + CUIs	✓		**0.291**
#182344	Image + CUIs	✓	✓	0.285

4.2. Caption Prediction

We participated to the Caption Prediction task with six different submissions. These can be grouped into *image-only* (#182349, #182255 and #182273) and *image + CUIs* (#182350, #182342 and #182344), as described in section 3.2. For all models we considered the same set of hyper-parameters. Specifically, we trained each model for 15 epochs using a batch size equal to 16 and Adam optimizer with weight decay [17], with the initial learning rate set to 5×10^{-5} for the visual encoder (ResNet-101) and 10^{-4} for the remaining parameters. The learning rates were reduced by a factor of 0.8 when the BLEU-1 score [8], computed on the validation set, did not improve for 3 consecutive epochs. Apart from the visual encoder, which was initialized using ImageNet pre-trained weights, the remaining parameters were initialized from a random uniform distribution.

Another difference among the submissions is found in the text *pre-processing stage* of the target captions. This matches the post-processing steps performed on the predicted caption before computing the different caption metrics:

- the caption text is converted to lower-case (this step is applied to all submissions);
- all punctuation is removed;
- stopwords are removed;
- lemmatization is applied.

Finally, we applied a *post-processing stage* by noticing that our model suffers a commonly reported repetition problem – where a language model tends to unnecessarily repeat chunks of text (*e.g. "... connect center femoral head center femoral head center femoral head..."*). This is simply addressed by removing repeating words in the predicted captions.

The BLEU scores for each of our submissions are shown in Table 3. It can be seen that the multimodal approach (image - CUIs) generally yields higher BLEU scores. Moreover, the model benefits from the pre-processing stage but not from removing repeating words. This last observation can be attributed to the fact that there are captions where the same word is repeated multiple times in the ground-truth. Therefore, a less naive approach should be considered to overcome the repetition problem. Overall, our best submission is #182342, which is a multimodal architecture where the pre-processing stage is applied but no post-processing.

Table 4 shows the final ranking of the Caption Prediction task. This is based on the BLEU score, and our best submission ranked 6th. However, by looking at the other caption metrics

Table 4
Caption scores computed on the test set for each teams' best solution. The ranking is based on the BLEU score. For all the metrics, we highlight in **bold** the best score and underline the second best score.

Team	Run	BLEU	ROUGE	METEOR	CIDEr	SPICE	BERTScore	Rank
IUST_NLPLAB	#182275	**0.483**	0.142	**0.093**	0.030	0.007	0.561	1
AUEB-NLP-Group	#181853	0.322	0.166	0.074	0.190	0.031	0.599	2
CSIRO	#182268	0.311	0.197	0.084	0.269	0.046	**0.623**	3
vcmi	#182325	0.306	0.174	0.075	0.205	0.036	0.604	4
eecs-kth	#182337	0.292	0.116	0.062	0.132	0.022	0.573	5
Ours	#182342	0.291	**0.201**	0.082	0.256	0.046	0.610	6
kdelab	#182351	0.278	0.158	0.074	**0.411**	**0.051**	0.600	7
Morgan_CS	#182238	0.255	0.144	0.056	0.148	0.023	0.583	8
MAI_ImageSem	#182105	0.221	0.185	0.067	0.251	0.039	0.606	9
SSNSheerinKavitha	#182248	0.160	0.043	0.027	0.017	0.007	0.545	10

we notice that we ranked 1st in terms of ROUGE score [9] and 2nd when considering SPICE [18] and the BERTScore [19]. This shows how problematic it is to evaluate captioning methods, and it is usually good practice to keep track of multiple metrics.

5. Conclusion

This manuscript presents our proposed solutions for the ImageCLEFmedical Caption 2022 task. In particular, we applied well known techniques in the field of image classification (supervised learning and image retrieval) and image captioning, and showed how they yield promising results in the medical domain. Our best submission ranked 2nd in the Concept Detection subtask, showing a good balance between detecting the full set of CUIs (determined using the F1 score) and the CUIs associated with the image modality and the anatomy (determined from the Secondary F1 score). Furthermore, we ranked 6th in the Caption Prediction subtask, based on the BLEU score, but we highlighted how the ranking can vary considerably depending on the metric.

References

[1] J. Rückert, A. Ben Abacha, A. García Seco de Herrera, L. Bloch, R. Brüngel, A. Idrissi-Yaghir, H. Schäfer, H. Müller, C. M. Friedrich, Overview of ImageCLEFmedical 2022 – Caption Prediction and Concept Detection, in: CLEF2022 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, Bologna, Italy, 2022.

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

[3] 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, Springer, 2018, pp. 180–189.
[4] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.
[5] G. Huang, Z. Liu, L. Van Der Maaten, K. Q. Weinberger, Densely connected convolutional networks, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 4700–4708.
[6] F. Charalampakos, V. Karatzas, V. Kougia, J. Pavlopoulos, I. Androutsopoulos, AUEB NLP group at ImageCLEFmed caption tasks 2021, in: CLEF2021 Working Notes, CEUR Workshop Proceedings, CEUR-WS. org, Bucharest, Romania, 2021.
[7] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, Advances in neural information processing systems 30 (2017).
[8] 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.
[9] C.-Y. Lin, ROUGE: A package for automatic evaluation of summaries, in: Text summarization branches out, 2004, pp. 74–81.
[10] 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.
[11] Y. Shimoyama, O. Umegaki, Y. Ooi, T. Agui, N. Kadono, T. Minami, Bacillus cereus pneumonia in an immunocompetent patient: a case report, JA Clinical Reports 3 (2017) 1–5.
[12] W. Y. Alnofal, M. R. Alshadely, M. A. Khatib, Spontaneous Subcutaneous Emphysema and Pneumomediastinum Associated With Influenza B Virus in a Young Male Adult: A Case Report, Cureus 13 (2021).
[13] Z. Chen, Y. Song, T.-H. Chang, X. Wan, Generating radiology reports via memory-driven transformer, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, Online, 2020, pp. 1439–1449.
[14] F. Liu, X. Wu, S. Ge, W. Fan, Y. Zou, Exploring and distilling posterior and prior knowledge for radiology report generation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 13753–13762.
[15] V. V. Wang, C. Y. Chang, A. P. Radhakrishnan, Invasive Aspergillus rhinosinusitis complicated with cerebral abscess, Revista da Sociedade Brasileira de Medicina Tropical 54 (2021).
[16] 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.
[17] D. P. Kingma, J. Ba, Adam: A method for stochastic optimization, in: ICLR (Poster), 2015.
[18] P. Anderson, B. Fernando, M. Johnson, S. Gould, SPICE: Semantic propositional image caption evaluation, in: European conference on computer vision, Springer, 2016, pp. 382–398.
[19] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, Y. Artzi, BERTScore: Evaluating text generation with BERT, in: International Conference on Learning Representations, 2020.