Overview of ImageCLEFmedical 2022 – Caption Prediction and Concept Detection

Johannes Rückert[1], Asma Ben Abacha[2], Alba G. Seco de Herrera[3], Louise Bloch[1,4], Raphael Brüngel[1,4], Ahmad Idrissi-Yaghir[1,4], Henning Schäfer[5], Henning Müller[6,7] and Christoph M. Friedrich[1,4]

[1]*Department of Computer Science, University of Applied Sciences and Arts Dortmund, Dortmund, Germany*

[2]*Microsoft, Redmond, Washington, USA*

[3]*University of Essex, Wivenhoe Park, Colchester CO4 3SQ, UK*

[4]*Institute for Medical Informatics, Biometry and Epidemiology (IMIBE), University Hospital Essen, Germany*

[5]*Institute for Transfusion Medicine, University Hospital Essen, Essen, Germany*

[6]*University of Applied Sciences Western Switzerland (HES-SO), Switzerland*

[7]*University of Geneva, Switzerland*

Abstract

The 2022 ImageCLEFmedical caption prediction and concept detection tasks follow similar challenges that were already run from 2017–2021. The objective is to extract Unified Medical Language System (UMLS) concept annotations and/or captions from the image data that are then compared against the original text captions of the images. The images used for both tasks are a subset of the extended Radiology Objects in COntext (ROCO) data set which was used in ImageCLEFmedical 2020. In the caption prediction task, lexical similarity with the original image captions is evaluated with the BiLingual Evaluation Understudy (BLEU) score. In the concept detection task, UMLS terms are extracted from the original text captions, combined with manually curated concepts for image modality and anatomy, and compared against the predicted concepts in a multi-label way. The F1-score was used to assess the performance. The task attracted a strong participation with 20 registered teams. In the end, 12 teams submitted 157 graded runs for the two subtasks. Results show that there is a variety of techniques that can lead to good prediction results for the two tasks. Participants used image retrieval systems for both tasks, while multi-label classification systems were used mainly for the concept detection, and Transformer-based architectures primarily for the caption prediction subtask.

Keywords

Concept Detection, Computer Vision, ImageCLEF 2022, Image Understanding, Image Modality, Radiology, Caption Prediction

1. Introduction

The caption task was first proposed as part of the ImageCLEFmedical [1] in 2016. In 2017 and 2018 [2, 3] the ImageCLEFmedical caption task comprised two subtasks: concept detection and caption prediction. In 2019 [4] and 2020 [5], the task concentrated on extracting Unified Medical Language System® (UMLS) Concept Unique Identifiers (CUIs) [6] from radiology images.

In 2021 [7], both subtasks, concept detection and caption prediction, were running again due to participants demands. The focus in 2021 was on making the task more realistic by using fewer images which were all manually annotated by medical doctors. As additional data of similar quality is hard to acquire, the 2022 ImageCLEFmedical caption task continues with both subtasks albeit with an extended version of the Radiology Objects in COntext (ROCO) [8] data set used for both subtasks, which was already used in 2020 and 2019.

This paper sets forth the approaches for the caption task: automated cross-referencing of medical images and captions into predicted coherent captions implying UMLS concept detection in radiology images as a first step. This task is a part of the ImageCLEF benchmarking campaign, which has proposed medical image understanding tasks since 2003; a new suite of tasks is generated each subsequent year. Further information on the other proposed tasks at ImageCLEF 2022 can be found in Ionescu et al. [9].

This is the 6th edition of the ImageCLEFmedical caption task. Just like in 2016 [1], 2017 [2], 2018 [3], and 2021 [7], both subtasks of concept detection and caption prediction are included in ImageCLEFmedical Caption 2022. Like in 2020, an extended subset of the ROCO [8] data set is used to provide a much larger data set compared to 2021.

Manual generation of the knowledge of medical images is a time-consuming process prone to human error. As this process requires assistance for the better and easier diagnoses of diseases that are susceptible to radiology screening, it is important that we better understand and refine automatic systems that aid in the broad task of radiology-image metadata generation. The purpose of the ImageCLEFmedical 2022 caption prediction and concept detection tasks is the continued evaluation of such systems. Concept detection and caption prediction information is applicable to unlabelled and unstructured data sets and medical data sets that do not have textual metadata. The ImageCLEFmedical caption task focuses on the medical image understanding in the biomedical literature and specifically on concept extraction and caption prediction based on the visual perception of the medical images and medical text data such as medical caption or UMLS CUIs paired with each image (see Figure 1).

For the development data, an extended subset of the ROCO [8] data set from 2020 was used, with new images from the same source added for the validation and test sets.

This paper presents an overview of the ImageCLEFmedical caption task 2022 including the task and participation in Section 2, the data creation in Section 3, and the evaluation methodology in Section 4. The results are described in Section 5, followed by conclusion in Sections 6.

2. Task and Participation

In 2022, the ImageCLEFmedical caption task consisted of two subtasks: concept detection and caption prediction.

The concept detection subtask follows the same format proposed since the start of the task in 2017. Participants are asked to predict a set of concepts defined by the UMLS CUIs [6] based on the visual information provided by the radiology images.

The caption prediction subtask follows the original format of the subtask used between 2017 and 2018. The task is running again since 2021 because of participant demand. This subtask aims to automatically generate captions for the radiology images provided.

In 2022, 20 teams registered and signed the End-User-Agreement that is needed to download the development data. 12 teams submitted 157 runs for evaluation (all 12 teams submitted working notes) attracting more attention than in 2021. Each of the groups was allowed a maximum of 10 graded runs per subtask.

Table 1 shows all the teams who participated in the task and their submitted runs. 11 teams participated in the concept detection subtask this year, 3 of those teams also participated in 2021. 10 teams submitted runs to the caption prediction subtask, 4 of those teams also participated in 2021. Overall, 9 teams participated in both subtasks, two teams participated only in the concept detection subtask and one team participated only in the caption prediction subtask.

3. Data Creation

Figure 1 shows an example from the data set provided by the task.

UMLS CUI	UMLS Meaning
C1306645	Plain x-ray
C0030797	Pelvis
C0332466	Fused structure
C0034014	Bone structure of pubis
C0205094	Anterior
C0005976	Bone Transplantation
C0021102	Implants

Caption: Anteroposterior pelvic radiograph of a 30-year-old female diagnosed with Ehlers-Danlos Syndrome demonstrating fusion of pubic symphysis and both sacroiliac joints (anterior plating, bone grafting and sacroiliac screw insertion)

Figure 1: Example of a radiology image with the corresponding UMLS® CUIs and caption extracted from the 2022's ImageCLEFmedical caption task. CC-BY [Ali et al. (2020)] [23]

In the previous edition, in an attempt to make the task more realistic, the data set contained a smaller number of real radiology images annotated by medical doctors which resulted in high-quality concepts.

Additional data of similar quality is hard to acquire and so it was decided to return to the data set already used in 2020 and 2019, which originates from biomedical articles of the PMC

Table 1
Participating groups in the ImageCLEFmedical 2022 caption task and their graded runs submitted to both subtasks: T1-Concept Detection and T2-Caption Prediction. Teams with previous participation in 2021 are marked with an asterisk (*).

Team	Institution	Runs T1	Runs T2
AUEB-NLP-Group* [10]	Department of Informatics, Athens University of Economics and Business, Athens, Greece	6	9
CSIRO* [11]	Australian e-Health Research Centre, Commonwealth Scientific and Industrial Research Organisation, Herston, Queensland, Australia and CSIRO Data61, Imaging and Computer Vision Group, Pullenvale, Queensland, Australia and Queensland University of Technology, Brisbane, Queensland, Australia	10	9
eecs-kth [12]	KTH Royal Institute of Technology, Stockholm, Sweden	10	10
CMRE-UoG (fdallaserra) [13]	Canon Medical Research Europe, Edinburgh, UK and University of Glasgow, Glasgow, UK	5	6
IUST_NLPLAB [14]	School of Computer Engineering, Iran University of Science and Technology, Tehran, Islamic Republic Of Iran	10	10
kdelab* [15] [16]	KDE Laboratory, Department of Computer Science and Engineering, Toyohashi University of Technology, Aichi, Japan	10	8
MAI_ImageSem* [17]	Institute of Medical Information and Library, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing, China	–	2
Morgan_CS [18]	Morgan State University, Baltimore, MD, USA	8	4
PoliMi-ImageClef [19]	Politecnico di Milano, Milan, Italy	10	–
SDVA-UCSD [20]	San Diego VA HCS, San Diego, CA, USA	1	–
SSNSheerinKavitha [21]	Department of CSE, Sri Sivasubramaniya Nadar College of Engineering, India	6	7
vcmi [22]	University of Porto, Porto, Portugal and INESC TEC, Porto, Portugal	9	7

Open Access Subset[1] [24] and was extended with new images added since the last time the data set was updated.

All captions were pre-processed by removing punctuation, numbers and words containing

---

[1]https://www.ncbi.nlm.nih.gov/pmc/tools/openftlist/ [last accessed: 28.06.2022]

numbers. Additionally, lemmatization was applied using spaCy[2] and the pre-trained model *en_core_web_lg*. Finally, all captions were converted to lower-case.

From the resulting captions, UMLS concepts were generated using a reduced subset of the UMLS 2020 AB release[3], which includes the sections (restriction levels) 0, 1, 2, and 9. To improve the feasibility of recognizing concepts from the images, concepts were filtered based on their semantic type. Concepts with very low frequency were also removed, based on suggestions from previous years.

Additional concepts were assigned to all images addressing their image modality. Six modality concepts were covered: x-ray, computer tomography (CT), magnetic resonance imaging (MRI), ultrasound, and positron emission tomography (PET) as well as modality combinations (e.g., PET/CT) as standalone concept. For images of the x-ray modality further concepts on the represented anatomy were assigned, covering specific anatomical body regions of the Image Retrieval in Medical Application (IRMA) [25] classification: cranium, spine, upper extremity/arm, chest, breast/mamma, abdomen, pelvis, and lower extremity/leg. Both of the described concept extensions were created performing a two-stage process, each. In the first stage predictions via classification models were created and assigned as annotations. For modality prediction for all images a model trained on the ROCO dataset [8], and for anatomy prediction for x-ray modality images a model trained on an existing IRMA-annotated image dataset [26] was used. In the second stage, these annotations underwent manual quality control measures, involving correction of faulty predictions and filtering of images that did not represent one of the minded modality or anatomy concepts. Three annotators were involved. Each individual modality concept was processed by a single annotator due to the low complexity of this task part. Anatomy concepts of x-ray modality images were each, too, processed by a single annotator per concept. However, due to the complexity/ambiguity of this task, the one annotator most-experienced in anatomy classification re-evaluated the assessments of the other two. This re-evaluation resulted in very few adjustments, indicating high agreement between annotators.

The following subsets were distributed to the participants where each image has one caption and multiple concepts (UMLS-CUI):

- *Training set* including 83,275 radiology images and associated captions and concepts.
- *Validation set* including 7,645 radiology images and associated captions and concepts.
- *Test set* including 7,645 radiology images.

4. Evaluation Methodology

In this year's edition, the performance evaluation is carried out in the same way as last year, with both subtasks being evaluated separately.

For the concept detection subtask, the balanced precision and recall trade-off were measured in terms of F1-scores. In addition, a secondary F1-score was introduced in this edition, where the score is computed using a subset of concepts that was manually curated and only contains x-ray anatomy and image modality concepts.

---

[2]https://spacy.io/api/lemmatizer/ [last accessed: 28.06.2022]
[3]https://www.nlm.nih.gov/pubs/techbull/nd20/nd20_umls_release.html [last accessed: 28.06.2022]

Caption prediction performance is evaluated based on the BiLingual Evaluation Understudy (BLEU) scores [27], which a geometric mean of n-gram scores from 1 to 4. As a preprocessing step for the evaluation, all captions were lowercased and stripped of all punctuation and English stop words. Additionally, to increase coverage, lemmatization was applied using spaCy and the pre-trained model *en_core_web_lg*. BLEU values are then computed for each test image, treating the entire caption as one sentence, even though it may contain multiple sentences. The average of the BLEU values for all images is reported as the primary ranking score. Since evaluating generated text and image captioning is very challenging and should be based on a single metric, additional evaluation metrics were explored in this year's edition in order to find the metric that correlate well with human judgements for this task. First, the Recall-Oriented Understudy for Gisting Evaluation (ROUGE) [28] score was adopted as a secondary metric that counts the number of overlapping units such as n-grams, word sequences, and word pairs between the generated text and the reference. Specifically, the ROUGE-1 (F-measure) score was calculated, which measures the number of matching unigrams between the model-generated text and a reference. All individual scores for each caption are then summed and averaged over the number of captions, resulting in the final score. In addition to ROUGE, the Metric for Evaluation of Translation with Explicit ORdering (METEOR) [29] was explored, which is a metric that evaluates the generated text by aligning it to reference and calculating a sentence-level similarity score. Furthermore, the Consensus-based Image Description Evaluation (CIDEr) [30] metric was also adopted. CIDEr is an automatic evaluation metric that calculates the weights of n-grams in the generated text and the reference text based on term frequency and inverse document frequency (TF-IDF), and then compares them based on cosine similarity. Another used metric is the Semantic Propositional Image Caption Evaluation (SPICE) [31], which maps the reference and generated captions to semantic scene graphs through dependency parse trees and measures the similarity between the scene graphs for the evaluation. Finally, BERTScore [32] was used, which is a metric that computes a similarity score for each token in the generated text with each token in the reference text. It leverages the pre-trained contextual embeddings from BERT-based models and matches words by cosine similarity. In this work, the pre-trained model *microsoft/deberta-xlarge-mnli*[4] was utilized, since it is the model that correlates best with human evaluation according to the authors[5].

5. Results

For the concept detection and caption prediction subtasks, Tables 2 and 3 show the best results from each of the participating teams. The results will be discussed in this section.

5.1. Results for the Concept Detection subtask

In 2022, 11 teams participated in the concept prediction subtask, submitting 85 runs. Table 2 presents the results achieved in the submissions.

---

[4]https://huggingface.co/microsoft/deberta-xlarge-mnli [last accessed: 28.06.2022]
[5]https://github.com/Tiiiger/bert_score [last accessed: 28.06.2022]

Table 2

Performance of the participating teams in the ImageCLEFmedical 2022 Concept Detection subtask. Only the best run based on the achieved F1-score is listed for each team, together with the corresponding secondary F1-score based on manual annotations as well as the team rankings based on the primary and secondary F1-score

Group Name	Best Run	F1	Secondary F1	Rank (secondary)
AUEB-NLP-Group	182358	**0.4511**	0.7907	1 (6)
fdallaserra	182324	0.4505	0.8222	2 (4)
CSIRO	182343	0.4471	0.7936	3 (5)
eecs-kth	181750	0.4360	0.8546	4 (2)
vcmi	182097	0.4329	**0.8634**	5 (1)
PoliMi-ImageClef	182296	0.4320	0.8512	6 (3)
SSNSheerinKavitha	181995	0.4184	0.6544	7 (8)
IUST_NLPLAB	182307	0.3981	0.6732	8 (7)
Morgan_CS	182150	0.3520	0.6281	9 (9)
kdelab	182346	0.3104	0.4120	10 (11)
SDVA-UCSD	181691	0.3079	0.5524	11 (10)

AUEB-NLP-Group Like in previous years, the AUEB-NLP-Group submitted the best performing result with a primary F1-score of 0.4511 [10] and a secondary F1-score of 0.7907. The winning approach was an ensemble of two EfficientNetV2-B0 backbones followed by a single classification layer where the union of predicted concepts was used to form the ensemble. This solution outperformed their retrieval-based system which won last year's concept detection subtask [33].

fdallaserra The second best system, with an only slightly worse primary F1-score of 0.4505 and a better secondary F1-score of 0.8222 [13] was proposed by CMRE-UoG (fdallaserra). Their best approach consisted of an image retrieval system which used an ensemble of five DenseNet-201, each of which retrieves 100 different images. Then CUIs appearing in at least 30% of the images are taken, and finally a union of each model's predicted CUIs is assigned to each image.

CSIRO The CSIRO group reached a primary F1-score of 0.4471 [11] and a secondary F1-score of 0.7936. They experimented with a range of different backbones for multi-label classification system, and their best approach is an ensemble of 43 DenseNet-161 with top-1% threshold optimisation.

eecs-kth The eecs-kth team reached a primary F1 score of 0.4360 [12] and a secondary F1 score of 0.8546. Their best approach utilized a multi-label classification system based on DenseNet161 with a single classification layer.

vcmi The VCMI (vcmi) team reached a primary F1-score of 0.4329 [22] and the best overall secondary F1-score of 0.8634. They combined a multi-label classification system based on DenseNet-121 with an information retrieval approach for their best approach, where the retrieval system is used if the classification did not assign any labels.

PoliMi The PoliMi team reached a primary F1-score of 0.4320 [19] and a secondary F1-score of 0.8512. They used a ResNext50-based multi-label classification system.

SSNSheerinKavitha The SSN MLRG (SSNSheerinKavitha) team reached a primary F1-score of 0.4184 [21] and a secondary F1-score of 0.6544. They employed DenseNet for multi-label classification and an information retrieval system.

IUST_NLPLAB The IUST_NLPLAB team reached a primary F1-score of 0.3981 [14] and a secondary F1-score of 0.6732. They used a multi-label classification model based on ResNet for their best results.

Morgan_CS The CS_Morgan (Morgan_CS) team from Morgan State University (USA) reached a primary F1-score of 0.3520 [18] and a secondary F1-score of 0.6280. They used a fusion of Vision Transformers for their best approach, which outperformed their multi-label classification systems.

Kdelab The Kdelab team reached a primary F1-score of 0.3104 [15] and a secondary F1-score of 0.4120. They exclusively experimented with image retrieval systems and their best approach consisted of an ensemble of different backbone networks (DenseNet, EfficientNet, ResNet) using simple majority voting.

SDVA-UCSD The SDVA-UCSD team reached a primary F1-score of 0.3079 [20] and a secondary F1-score of 0.5524. They used a multi-label classification system with ResNet and DenseNet backbones.

To summarize, in the concept detection subtasks, the groups used primarily multi-label classification systems and image retrieval systems, much like in the 2021 challenge. Multi-label classification systems outperformed retrieval-based systems for most of the teams who experimented with both, and while the winner was a multi-label classification approach, the second placing team with an F1-score only 0.0006 less than the winning team, used a retrieval-based system for which they took last year's winning approach and tuned it to include more CUIs by reducing the threshold for the percentage of retrieved images in which the CUI had to appear from 50% to 30% [13].

This year's models for concept detection do not show an increased F1-score compared to last year, however due to the much larger data set and number of concepts used in this year's challenge, this is not surprising. Comparing it to the 2020 results, where a data set of similar size was used, the F1-scores show a clear improvement. There are no radically new approaches used in this year's concept detection subtask, but the teams experimented with, optimised and re-combined many different existing techniques and created competitive solutions using both multi-label classification systems and image retrieval systems.

5.2. Results for the Caption Prediction subtask

In this sixth edition, the caption prediction subtask attracted 10 teams which submitted 72 runs. Table 3 presents the results of the submissions.

Table 3
Performance of the participating teams in the ImageCLEF 2022 Caption Prediction subtask. Only the best run based on the achieved BLEU score is listed for each team, together with the corresponding secondary ROUGE score as well as the team rankings based on the primary BLEU and secondary ROUGE score. The best results are highlighted.

Group Name	Best Run	BLEU	Secondary ROUGE	Rank (secondary)
IUST_NLPLAB	182275	**0.4828**	0.1422	1 (8)
AUEB-NLP-Group	181853	0.3222	0.1665	2 (5)
CSIRO	182268	0.3114	0.1974	3 (2)
vcmi	182325	0.3058	0.1738	4 (4)
eecs-kth	182337	0.2917	0.1157	5 (9)
fdallaserra	182342	0.2913	**0.2012**	6 (1)
kdelab	182351	0.2783	0.1584	7 (6)
Morgan_CS	182238	0.2549	0.1441	8 (7)
MAI_ImageSem	182105	0.2211	0.1847	9 (3)
SSNSheerinKavitha	182248	0.1595	0.0425	10 (10)

Table 4
Performance of the participating teams in the ImageCLEF 2022 Caption Prediction subtask for additional metrics METEOR, CIDEr, SPICE, and BERTScore. These correspond to the best F1 score-based runs of each team, listed in Table 3. The best results are highlighted.

Group Name	Best Run	METEOR	CIDEr	SPICE	BERTScore
IUST_NLPLAB	182275	**0.0928**	0.0304	0.0072	0.5612
AUEB-NLP-Group	181853	0.0737	0.1902	0.0313	0.5989
CSIRO	182268	0.0841	0.2693	0.0462	**0.6234**
vcmi	182325	0.0746	0.2047	0.0358	0.6044
eecs-kth	182337	0.0624	0.1317	0.0218	0.5728
fdallaserra	182342	0.0819	0.2564	0.0464	0.6101
kdelab	182351	0.0735	**0.4114**	**0.0512**	0.6003
Morgan_CS	182238	0.0559	0.1481	0.0232	0.5835
MAI_ImageSem	182105	0.0675	0.2513	0.0393	0.6059
SSNSheerinKavitha	182248	0.0226	0.0169	0.0072	0.5451

IUST_NLPLAB The IUST_NLPLAB team presented the best model for the caption prediction subtask. They reached a BLEU score of 0.4828, outperforming the competition by a large margin, and a ROUGE score of 0.1422 [14]. Additionally, they reached the overall best METEOR score of 0.0928. For their best run, they employed a multi-label classification system based on ResNet50 which treats every word as a label and assigns 26 words in the order of their probability to each image.

AUEB-NLP-Group The AUEB-NLP-Group submitted the second best performing result with a BLEU score of 0.3222 [10] and a ROUGE score of 0.1664. Their best approach utilizes the Show & Tell model [34] consisting of a CNN-RNN encoder-decoder with an EfficientNetB0 backbone. While they were clearly behind the BLEU score of the winners, they outscore

them in most of the other scores.

CSIRO The CSIRO group reached a BLEU score of 0.3114 [11] and a ROUGE score of 0.1974. Additionally, they reached the overall best BERTScore of 0.6234. They experimented with different encoder-to-decoder models and achieved their best scores with CvT-21 as the encoder and DistilGPT2 as the decoder, warm-started with a MIMIC-CXR checkpoint with a penalty for n-grams of size 3 that are repeated.

vcmi The VCMI (vcmi) team reached a BLEU score of 0.3058 [22] and a ROUGE score of 0.1738. They used a vision encoder-to-decoder system for the best results.

eecs-kth The eecs-kth team reached a BLEU score of 0.2917 [12] and a ROUGE score of 0.1157. They employed an information retrieval system based on AlexNet which summarizes the captions of a number of similar images using Pegasus.

fdallaserra The CMRE-UoG (fdallaserra) group reached a BLEU score of 0.2913 [13] and the overall best ROUGE score of 0.2012. They used a CNN Transformer approach with multi-modal (image + CUIs) input for their best results.

Kdelab The Kdelab team reached a BLEU score of 0.2782 [16] and a ROUGE score of 0.1584. Additionally, they reached the overall best CIDEr score of 0.4114 and overall best SPICE score of 0.0512. They used an image retrieval approach with an ensemble of different backbone networks for their best submission results.

Morgan_CS The CS_Morgan (Morgan_CS) team reached a BLEU score of 0.2549 [18] and a ROUGE score of 0.1441. They used a very similar approach as for the concept detection, namely a fusion of Vision Transformers.

MAI_ImageSem The MAI_ImageSem team reached a BLEU score of 0.2211 [17] and a ROUGE score of 0.1847. For the best results, they use pre-trained BLIP (Bootstrapping Language-Image Pre-training), a pre-training framework for vision-language understanding consisting of a multi-modal encoder-decoder and a captioning and filtering module.

SSNSheerinKavitha The SSN MLRG (SSNSheerinKavitha) team reached a BLEU score of 0.1595 [21] and a ROUGE score of 0.0425. For their best run, they employed a Sparse Auto Encoder (SAE) with a Multi-Layer Perceptron (MLP) and a Gated Recurrent Unit (GRU).

To summarize, in the caption prediction subtask most teams experimented with Transformer-based architectures and image retrieval systems. Only one team used a multi-label classification approach, and it achieved by far the best BLEU score. However, it did not score as well on most of the other employed metrics, with the second placing team outscoring the winners in all but the BLEU and METEOR metrics, which highlights the difficulty of evaluating caption similarity. One metric to highlight especially is SPICE, which is specifically designed for the evaluation of image captions. The winners scored a value of 0.0072 in this metric with the rest of the field (except the last placing team) scoring between 0.0218 and 0.0512.

Transfer Learning has frequently been used for pre-training, from a variety of different data sets. As in the previous years, simpler architectures ended up yielding better results compared to more complex ones in many instances.

Similar to the concept detection, the BLEU scores in the caption prediction subtask are overall lower compared to last year, which can be explained by the larger and more complex data set and more varied captions. Since there was no caption prediction subtask running in 2020, no comparable scores for a similar data set exist.

6. Conclusion

This year's caption task of ImageCLEFmedical once again ran with both subtasks, concept detection and caption prediction. It returned to a larger, ROCO-based data set for both challenges after a smaller, manually annotated data set was used last year. It attracted 12 teams who submitted 157 runs overall, a stronger participation compared to last year. For the concept detection subtask, a secondary F1-score was introduced to distinguish manually curated concepts from automatically generated ones. For the caption prediction, a number of additional scores were added to better illustrate the difficulty of evaluating the quality of predicted captions. All but one team participated in the concept detection subtask, with only two teams choosing not to participate in the caption prediction subtask as well. Only one team used the generated concepts as the input for the caption prediction model, most teams approached the subtasks with separate systems. For the concept detection challenge, most teams employed multi-label classification systems or image retrieval systems, while the caption prediction challenge was predominantly approached using Transformer-based architectures and image retrieval systems, with only the winning team using a multi-label classification system.

The scores for both subtasks have not improved compared to the 2021 edition. However, the larger and more complex ROCO-based data set with more concepts and more varied captions make the scores difficult to compare. Looking at the 2020 edition, which used a similar data set, the concept detection scores have clearly increased (there was no caption prediction subtask).

For next year's ImageCLEFmedical Caption challenge, some possible improvements include adding more manually validated concepts like increased anatomical coverage and directionality information, reducing recurring captions, more fine-grained CUI filters, improving the caption pre-processing, and using a different primary score for the caption prediction challenge, since the BLEU score has some disadvantages which were highlighted by this year's caption prediction results.

What should also be addressed is how to deal with models that were pre-trained on PMC data, because strictly speaking they have seen the real captions and can have an advantage when some of these images appear in test data.

Acknowledgments

References

[1] 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, pp. 219–232.

[2] C. Eickhoff, I. Schwall, A. G. S. de Herrera, H. Müller, Overview of ImageCLEFcaption 2017 - Image Caption Prediction and Concept Detection for Biomedical Images, in: Working Notes of CLEF 2017 - Conference and Labs of the Evaluation Forum, Dublin, Ireland, September 11-14, 2017., 2017. URL: http://ceur-ws.org/Vol-1866/invited_paper_7.pdf.

[3] A. G. S. de Herrera, C. Eickhoff, V. Andrearczyk, H. Müller, Overview of the ImageCLEF 2018 Caption Prediction Tasks, in: Working Notes of CLEF 2018 - Conference and Labs of the Evaluation Forum, Avignon, France, September 10-14, 2018., 2018. URL: http://ceur-ws.org/Vol-2125/invited_paper_4.pdf.

[4] O. Pelka, C. M. Friedrich, A. G. S. de Herrera, H. Müller, Overview of the ImageCLEFmed 2019 Concept Detection Task, in: L. Cappellato, N. Ferro, D. E. Losada, H. Müller (Eds.), Working Notes of CLEF 2019 - Conference and Labs of the Evaluation Forum, Lugano, Switzerland, September 9-12, 2019, volume 2380 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2019. URL: http://ceur-ws.org/Vol-2380/paper_245.pdf.

[5] O. Pelka, C. M. Friedrich, A. García Seco de Herrera, H. Müller, Overview of the Image-CLEFmed 2020 concept prediction task: Medical image understanding, in: CLEF2020 Working Notes, volume 1166 of *CEUR Workshop Proceedings*, CEUR-WS.org, Thessaloniki, Greece, 2020.

[6] O. Bodenreider, The Unified Medical Language System (UMLS): integrating biomedical terminology, Nucleic Acids Research 32 (2004) 267–270. doi:10.1093/nar/gkh061.

[7] O. Pelka, A. Ben Abacha, A. García Seco de Herrera, J. Jacutprakart, C. M. Friedrich, H. Müller, Overview of the ImageCLEFmed 2021 concept & caption prediction task, in: CLEF2021 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, Bucharest, Romania, 2021, pp. 1101–1112.

[8] 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, 2018, pp. 180–189. doi:10.1007/978-3-030-01364-6_20.

[9] B. Ionescu, H. Müller, R. Péteri, 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.

[10] F. Charalampakos, G. Zachariadis, J. Pavlopoulos, V. Karatzas, C. Trakas, I. Androutsopoulos, AUEB NLP group at ImageCLEFmed caption 2022, in: CLEF2022 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, Bologna, Italy, 2022.

[11] L. Lebrat, A. Nicolson, R. S. Cruz, G. Belous, B. Koopman, J. Dowling, CSIRO at ImageCLEFmed caption 2022, in: CLEF2022 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, Bologna, Italy, 2022.

[12] G. M. Moschovis, E. Fransén, Neuraldynamicslab at ImageCLEF medical 2022, in: CLEF2022 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, Bologna, Italy, 2022.

[13] F. D. Serra1, F. Deligianni, J. Dalton, A. Q. O'Neil, CMRE-UoG team at ImageCLEFmed caption 2022 task: Concept detection and image captioning, in: CLEF2022 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, Bologna, Italy, 2022.

[14] M. Hajihosseini, Y. Lotfollahi, M. Nobakhtian, M. M. Javid, F. Omidi, S. Eetemadi, IUST_NLPLAB at ImageCLEFmed caption tasks 2022, in: CLEF2022 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, Bologna, Italy, 2022.

[15] R. Tsuneda, T. Asakawa, K. Shimizu, T. Komoda, M. Aono, Kdelab at ImageCLEF 2022: Medical concept detection with image retrieval and code ensemble, in: CLEF2022 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, Bologna, Italy, 2022.

[16] R. Tsuneda, T. Asakawa, K. Shimizu, T. Komoda, M. Aono, Kdelab at ImageCLEF2022 medical caption prediction task, in: CLEF2022 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, Bologna, Italy, 2022.

[17] X. Wang, J. Li, ImageSem Group at ImageCLEFmed Caption 2022 Task: Generating Medical Image Descriptions based on Visual- Language Pre-training, in: CLEF2022 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, Bologna, Italy, 2022.

[18] M. M. Rahman, O. Layode, CS_Morgan at ImageCLEFmed caption 2022: Deep learning based multilabel classification and transformers for concept detection & caption prediction, in: CLEF2022 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, Bologna, Italy, 2022.

[19] S. A. M. Ghayyomnia, K. de Gast, M. J. Carmana, Polimi-imageclef group at ImageCLEFmed caption 2022, in: CLEF2022 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, Bologna, Italy, 2022.

[20] A. Gentili, ImageCLEFmed concept detection, finding duplicates, in: CLEF2022 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, Bologna, Italy, 2022.

[21] N. M. S. Sitara, S. Kavitha, SSN MLRG at ImageCLEF 2022: Medical concept detection and caption prediction using transfer learning and transformer based learning approaches, in: CLEF2022 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, Bologna, Italy, 2022.

[22] I. Rio-Torto, C. Patrício, H. Montenegro, T. Gonçalves, Detecting Concepts and Generating Captions from Medical Images: Contributions of the VCMI Team to ImageCLEFmed Caption 2022, in: CLEF2022 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, Bologna, Italy, 2022.

[23] A. Ali, P. Andrzejowski, N. K. Kanakaris, P. V. Giannoudis, Pelvic Girdle Pain, Hypermo-

bility Spectrum Disorder and Hypermobility-Type Ehlers-Danlos Syndrome: A Narrative Literature Review, Journal of Clinical Medicine 9 (2020) 3992. doi:`10.3390/jcm9123992`.

[24] R. J. Roberts, PubMed Central: The GenBank of the published literature, Proceedings of the National Academy of Sciences of the United States of America 98 (2001) 381–382. doi:`10.1073/pnas.98.2.381`.

[25] T. M. Lehmann, H. Schubert, D. Keysers, M. Kohnen, B. B. Wein, The IRMA code for unique classification of medical images, in: H. K. Huang, O. M. Ratib (Eds.), Medical Imaging 2003: PACS and Integrated Medical Information Systems: Design and Evaluation, SPIE, 2003. doi:`10.1117/12.480677`.

[26] T. Deserno, B. Ott, 15.363 IRMA Bilder in 193 Kategorien für ImageCLEFmed 2009, 2009. URL: https://publications.rwth-aachen.de/record/667225. doi:`10.18154/RWTH-2016-06143`.

[27] 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.

[28] C.-Y. Lin, ROUGE: A Package for Automatic Evaluation of Summaries, in: Text Summarization Branches Out, Association for Computational Linguistics, 2004, pp. 74–81. URL: https://aclanthology.org/W04-1013.

[29] M. Denkowski, A. Lavie, Meteor Universal: Language Specific Translation Evaluation for Any Target Language, in: Proceedings of the Ninth Workshop on Statistical Machine Translation, Association for Computational Linguistics, 2014, pp. 376–380. URL: http://aclweb.org/anthology/W14-3348. doi:`10.3115/v1/W14-3348`.

[30] R. Vedantam, C. L. Zitnick, D. Parikh, CIDEr: Consensus-based image description evaluation, in: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2015, pp. 4566–4575. URL: http://ieeexplore.ieee.org/document/7299087/. doi:`10.1109/CVPR.2015.7299087`.

[31] P. Anderson, B. Fernando, M. Johnson, S. Gould, SPICE: Semantic Propositional Image Caption Evaluation, in: B. Leibe, J. Matas, N. Sebe, M. Welling (Eds.), Computer Vision – ECCV 2016, Lecture Notes in Computer Science, Springer International Publishing, 2016, pp. 382–398. doi:`10.1007/978-3-319-46454-1_24`.

[32] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, Y. Artzi, Bertscore: Evaluating text generation with BERT, in: 8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020, 2020. URL: https://openreview.net/forum?id=SkeHuCVFDr.

[33] 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, pp. 1184–1200.

[34] K. Xu, J. Ba, R. Kiros, K. Cho, A. C. Courville, R. Salakhutdinov, R. S. Zemel, Y. Bengio, Show, attend and tell: Neural image caption generation with visual attention, in: F. R. Bach, D. M. Blei (Eds.), Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015, volume 37 of *JMLR Workshop and Conference Proceedings*, JMLR.org, 2015, pp. 2048–2057. URL: http://proceedings.mlr.press/v37/xuc15.html.