Pubilched in Experimental IR Meets Multilinguality, Multimodality, and Interaction : 12th International Conference of the CLEF Association, CLEF 2021, Virtual Event, September 21–24, 2021, Proceedings, which should be cited to refer to this work.
DOI:10.1007/978-3-030-85251-1_23

Overview of the ImageCLEF 2021: Multimedia Retrieval in Medical, Nature, Internet and Social Media Applications

Bogdan Ionescu[1], Henning Müller[2], Renaud Péteri[3], Asma Ben Abacha[4], Mourad Sarrouti[4], Dina Demner-Fushman[4], Sadid A. Hasan[5], Serge Kozlovski[6], Vitali Liauchuk[6], Yashin Dicente Cid[7], Vassili Kovalev[6], Obioma Pelka[8], Alba García Seco de Herrera[9], Janadhip Jacutprakart[9], Christoph M. Friedrich[8], Raul Berari[10], Andrei Tauteanu[10], Dimitri Fichou[10], Paul Brie[10], Mihai Dogariu[1], Liviu Daniel Ştefan[1], Mihai Gabriel Constantin[1], Jon Chamberlain[9], Antonio Campello[11], Adrian Clark[9], Thomas A. Oliver[12], Hassan Moustahfid[12], Adrian Popescu[13], and Jérôme Deshayes-Chossart[13]

[1] University Politehnica of Bucharest, Bucharest, Romania
bogdan.ionescu@upb.ro
[2] University of Applied Sciences Western Switzerland (HES-SO), Delémont, Switzerland
[3] University of La Rochelle, La Rochelle, France
[4] National Library of Medicine, Bethesda, USA
[5] CVS Health, Wellesley, MA, USA
[6] United Institute of Informatics Problems, Minsk, Belarus
[7] University of Warwick, Coventry, UK
[8] University of Applied Sciences and Arts Dortmund, Dortmund, Germany
[9] University of Essex, Colchester, UK
[10] teleportHQ, Cluj-Napoca, Romania
[11] Wellcome Trust, London, UK
[12] Pacific Islands Fisheries Science Center, Silver Spring, USA
[13] Université Paris-Saclay, CEA, List, Palaiseau, France

Abstract. This paper presents an overview of the ImageCLEF 2021 lab that was organized as part of the Conference and Labs of the Evaluation Forum – CLEF Labs 2021. ImageCLEF is an ongoing evaluation initiative (first run in 2003) that promotes the evaluation of technologies for annotation, indexing and retrieval of visual data with the aim of providing information access to large collections of images in various usage scenarios and domains. In 2021, the 19th edition of ImageCLEF runs four main tasks: (i) a *medical* task that groups three previous tasks, i.e., caption analysis, tuberculosis prediction, and medical visual question answering and question generation, (ii) a *nature* coral task about segmenting and labeling collections of coral reef images, (iii) an *Internet* task addressing the problems of identifying hand-drawn and digital user interface components, and (iv) a new *social media* aware task on estimating potential real-life effects of online image sharing. Despite the current pandemic situation, the benchmark campaign received a strong participation with over 38 groups submitting more than 250 runs.

Keywords: Visual question answering and generation · medical image classification · coral image segmentation and classification · recognition of website user interface components · prediction of effects of online image sharing · ImageCLEF lab

1 Introduction

ImageCLEF[14] is the image retrieval and classification lab of the CLEF (Conference and Labs of the Evaluation Forum) conference. ImageCLEF has started in 2003 with only four participants [12]. It increased its impact with the addition of medical tasks in 2004 [11], attracting over 20 participants already in the second year. An overview of ten years of the medical tasks can be found in [25]. It continued the ascending trend, reaching over 200 participants in 2019 and over 110 in 2020 despite the outbreak of the covid-19 pandemic. The tasks have changed much over the years but the general objective has always been the same, i.e., *to combine text and visual data to retrieve and classify visual information*. Tasks have evolved from more general object classification and retrieval to many specific application domains, e.g., nature, security, medical, Internet. A detailed analysis of several tasks and the creation of the data sets can be found in [31]. ImageCLEF has shown to have an important impact over the years, already detailed in 2010 [44, 45].

Since 2018, ImageCLEF uses the crowdAI platform, now migrated to AIcrowd[15] from 2020, to distribute the data and receive the submitted results. The system allows having an online leader board and gives the possibility to keep data sets accessible beyond competition, including a continuous submission of runs and addition to the leader board. Over the years, ImageCLEF and also CLEF have shown a strong scholarly impact that was analyzed in [44, 45]. For instance, the term "ImageCLEF" returns on Google Scholar[16] over 5,800 article results (search on June 11th, 2021). This underlines the importance of evaluation campaigns for disseminating best scientific practices. We introduce here the four tasks that were run in the 2021 edition[17], namely: ImageCLEFmedical, ImageCLEFcoral ImageCLEFdrawnUI, and the new ImageCLEFaware.

2 Overview of Tasks and Participation

ImageCLEF 2021 consists of four main tasks with the objective of covering a *diverse range* of multimedia retrieval applications, namely: *medicine*, *nature*, *Internet*, and *social media* applications. It followed the 2019 tradition [24] of diversifying the use cases [4, 9, 5, 33, 26, 37]. The 2021 tasks are presented as follows:

[14] http://www.imageclef.org/
[15] https://www.aicrowd.com/
[16] https://scholar.google.com/
[17] https://www.imageclef.org/2021/

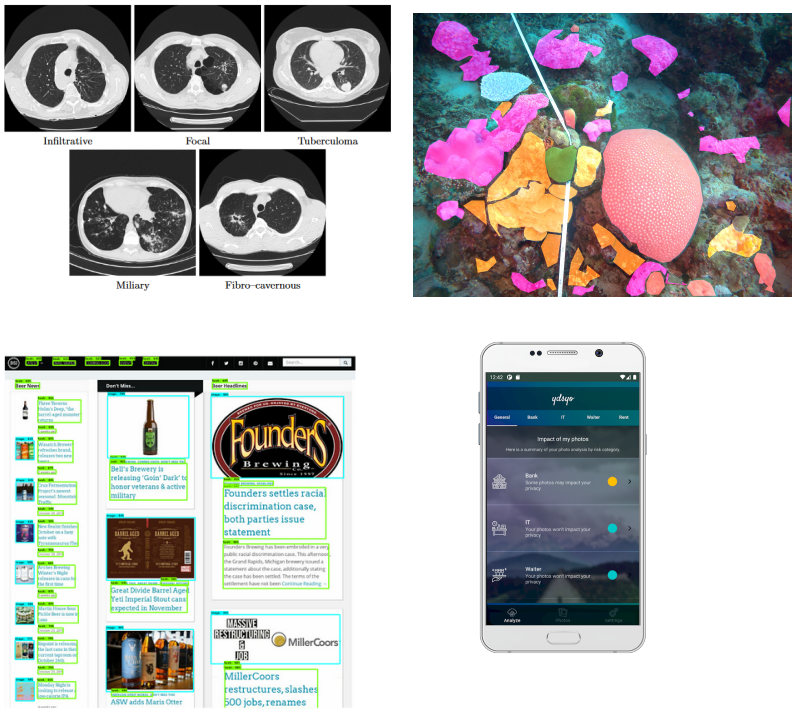

Fig. 1: Sample images from (left to right, top to bottom): ImageCLEFmedical tuberculosis prediction, ImageCLEFcoral with segmenting and labeling collections of coral reef images, ImageCLEFdrawnUI with recognition of website UIs, and ImageCLEFaware with estimating potential real-life effects of online image sharing.

– **ImageCLEFmedical**. Medical tasks have been part of ImageCLEF every year since 2004. In 2018, all but one task were medical, but little interaction happened between the medical tasks. For this reason, starting with 2019, the medical tasks were focused towards one specific problem but combined as a single task with several subtasks. This allows exploring synergies between the domains:

  - *Visual Question Answering*: This is the fourth edition of the VQA-Med task. With the increasing interest in artificial intelligence (AI) to support clinical decision making and improve patient engagement, opportunities to generate and leverage algorithms for automated medical image interpretation are currently being explored. In view of this and inspired by the success of the previous VQA-Med editions [21, 3, 2], we propose this year two tasks on visual question answering (VQA) and visual Question Generation (VQG) [4]. For the VQA task, given a radiology image accompanied with a relevant question, participating systems are tasked

with answering the question based on the visual image content, while for the VQG task, given a radiology image, participating systems are tasked with generating relevant questions based on the visual content of the medical image;

- *Tuberculosis*: This is the fifth edition of the task. The main objective is to provide an automatic CT-based evaluation of tuberculosis (TB) patients. This is done by detecting visual TB-related findings and by assessing the TB type based on the automatic analysis of lung CT scans. Being able to generate this automatic analysis from the image data allows to have a preliminary assessment of the medical case and limit laboratory analyses to determine the TB type. This can lead to quicker decisions on the best treatment strategy, reduced use of antibiotics, and lower impact on the patient. In this year edition, participants need to directly classify one of the five TB types: Infiltrative, Focal, Tuberculoma, Miliary, Fibrocavernous [26].
- *Caption*: This is the fifth edition of the task in this format, however, it is based on previous medical tasks. Based on the lessons learned in previous years [22, 15, 23, 34, 35], this year [33] we brought back the "caption prediction" subtask which focuses on composing coherent captions for the entirety of a radiology image. This year we continue with the "concept detection" subtask which focuses on identifying the presence and location of relevant concepts in the same corpus of radiology images. In the 2021 edition, the dataset is the same as the dataset of the ImageCLEF-VQAMed 2021 task. This encourages teams to participate in both tasks, as detected concepts can be used as building blocks for the VQA tasks. But also generated questions and answers can be used to evaluate the concept detection models.

– **ImageCLEFdrawnUI**. Traditionally, user interfaces (UI) are drawn by designers before being translated into code by developers. As this process is error prone and time consuming, the use of deep learning to automatize it and help UI professionals is gaining traction. In this second edition of the task [5], participants need to develop a machine learning system able to detect the position and type of UI elements in images. The task is separated into two subtasks. The wireframe subtask takes, as in the last edition, hand drawn wireframes as input. Issues from last year, such as class imbalance have been addressed by adding new images. The new screenshot subtask takes digital images as input and is a more difficult challenge due to the ambiguous way the images can be analyzed.
– **ImageCLEFcoral**. This is the third edition of the task. As in previous years [7, 8], the task addresses the problem of automatically segmenting and labeling a collection of images that can be used in combination to create 3D models for the monitoring of coral reefs. The task is separated into two subtasks which aim to label the images with types of benthic substrate. The first subtask uses bounding boxes to annotate the images while the second subtask segment the images pixel-wise using polygons. This year [9], the

Table 1: Key figures regarding participation in ImageCLEF 2021.

Task	Completed registrations	Groups that subm. results	Submitted runs	Submitted working notes
VQ Answering	33	13	75	8
. Tuberculosis	29	11	64	9
Caption	23	10	75	8
Coral	3	3	8	3
DrawnUI	8	3	28	2
Aware	7	2	6	0
Overall	103	42	256	30

training and test data form the complete set of images required to form a 3D reconstruction of the environment.

– **ImageCLEFaware**. This was the first edition of the task [26]. The disclosure of personal data is done in a particular context and users are often unaware that their data can be reused in other contexts. It is thus important to give feedback to users about the effects of personal data sharing. The objective was to automatically provide a rating of a visual user profile in different real-life situations. A new dataset was created specifically for this task and will be shared publicly in the following months. Data were sampled from YFCC100 and were further anonymized in order to comply with GPDR.

To participate in the evaluation campaign, the research groups had to register by following the instructions on the ImageCLEF 2021 web page[18]. To ease the overall management of the campaign, in 2021 the challenge was organized through the AIcrowd platform[19]. To actually get access to the data sets, the participants were required to submit a signed End User Agreement (EUA). Table 1 summarizes the participation in ImageCLEF 2021, including the number of completed registrations, indicated both per task and for the overall lab. The table also shows the number of groups that submitted runs and the ones that submitted a working notes paper describing the techniques used. Teams were allowed to register for participating in several different tasks.

After a decrease in participation in 2016, the participation increased in 2017 and 2018, and increased again in 2019. In 2018, 31 teams completed the tasks and 28 working notes papers were received. In 2019, 63 teams completed the tasks and 50 working notes papers were retrieved. In 2020, 40 teams completed the tasks and submitted working notes papers. In 2021, 42 teams completed the tasks and we received 30 working notes papers. Although there is a slight increase in the number of teams succeeding to conclude the tasks, we can clearly see a drop in participation compared to 2019. We expect that this is mostly due to the current pandemic situation which caught us for the second time during the organizing of the lab. Nevertheless, we still received a hefty number

[18] `https://www.imageclef.org/2021/`

[19] `https://www.aicrowd.com/`

of systems, i.e., 256 runs, which allow for an effective comparison of the results of the proposed solutions.

In the following sections, we present the tasks. Only a short overview is reported, including general objectives, description of the tasks and data sets, and a short summary of the results. A detailed review of the received submissions for each task is provided with the task overview working notes: ImageCLEFmedical VQA [4], Tuberculosis [26], and Caption [33], ImageCLEFcoral [9], ImageCLEFdrawnUI [5], and ImageCLEFaware [37].

3 The Visual Question Answering Task

Visual Question Answering is an exciting problem that combines natural language processing (NLP) and computer vision (CV) techniques. With the increasing interest in artificial intelligence (AI) technologies to support clinical decision making and improve patient engagement, opportunities to generate and leverage algorithms for automated medical image interpretation are being explored at a faster pace. To offer more training data and evaluation benchmarks, we organized the first visual question answering (VQA) task in the medical domain in 2018 [21], and continued the task in 2019 [3] and 2020 [2]. Following the strong engagement from the research community in the previous editions of VQA in the medical domain (VQA-Med) and the ongoing interests from both computer vision and medical informatics communities, we continued the task this year (VQA-Med 2021) [4] with an enhanced focus on (i) answering medical questions about abnormalities and (ii) generating relevant natural language questions about radiology images based on their visual content[20].

3.1 Task Setup

Two subtasks were proposed:

- Visual question answering (VQA) task: given a radiology image accompanied by a relevant question, participating systems in VQA-Med 2021 were tasked with answering the question based on the visual image content.
- Visual question generation (VQG) task: given a radiology image, participating systems were tasked with generating relevant natural language questions about the abnormality present in the image.

3.2 Data Set

For the visual question answering task, we automatically constructed the training, validation, and test sets by: (i) applying several filters to select relevant images and associated annotations, and, (ii) creating patterns to generate the questions and their answers. We selected relevant medical images from the MedPix[21]

[20] `https://www.imageclef.org/2021/medical/vqa`
[21] `https://medpix.nlm.nih.gov/`

database with filters based on their captions, localities, and diagnosis methods. We selected only the cases where the diagnosis was made based on the image. Finally, we considered the most frequent abnormality question categories to create the data set, which included a training set of 4,500 radiology images with 4,500 question-answer (QA) pairs (the same dataset used in 2020), a new validation set of 500 radiology images with 500 QA pairs, and a new test set of 500 radiology images with 500 questions about Abnormality. To further ensure the quality of the data, the reference answers of the test set were manually validated by a medical doctor.

For the visual question generation task, we automatically constructed the validation and test sets by using a collection of radiology images and their associated captions. We automatically generated questions from the images and their captions using two different approaches. To generate questions from the images, we used a variational autoencoder-based model called VQGR [40] trained on the VQA-RAD dataset (A CNN was used to encode the images and an LSTM to decode the questions). The second approach used a T5-based model fine-tuned on the SQuAD and MS MARCO datasets to generate questions from the image captions. Then, a medical doctor curated the list of created questions. The final curated corpus for the VQG task was comprised of 85 radiology images with 200 questions for validation, and 100 radiology images with 302 reference questions for the test set. For more details, please refer to the VQA-Med 2021 overview paper [4].

3.3 Participating Groups and Submitted Runs

Out of 48 online registrations, 33 participants submitted signed end user agreement forms. Finally, 13 teams submitted a total of 75 successful runs; 68 runs for the VQA task and 7 runs for the VQG task, indicating a notable interest in the VQA-Med challenge. Table 2 gives an overview of all participating teams and the number of submitted runs (please note that were allowed only 10 runs per team).

3.4 Results

Similar to the evaluation setup of the VQA-Med 2020 challenge [2], the evaluation of the participant systems for the VQA task in VQA-Med 2021 is also conducted based on two primary metrics: accuracy and BLEU. We used an adapted version of accuracy from the general domain VQA[22] task that strictly considers exact matching of a participant provided answer and the ground truth answer. To compensate for the strictness of the accuracy metric, BLEU [32] is used to capture the word overlap-based similarity between a system-generated answer and the ground truth answer. The overall methodology and resources for the BLEU metric are essentially similar to last year's VQA task. The BLEU

[22] `https://visualqa.org/evaluation.html`

Table 2: Participating groups in the VQA-Med 2021 tasks.

Team	*Institution*	# *Valid Runs*
Zhao_Ling_Ling	Yunnan University (China)	10
Zhao_Shi_	School of Information Science and Engineering, Yunnan University (China)	2
dua_dua	School of Computer Science and Engineering, Sun Yat-sen University (China)	10
Li_Yong_	South China Normal University (China)	10
TeamS	D4L data4life gGmbH&Hasso Plattner Institute (Germany)	10
sheerin	Siva Subramaniya Nadar College of Engineering (India)	5
IALab_PUC	IALab group of the Pontifical Catholic University (Chile)	5
Chabbiimen	Research Groups in Intelligent Machines& Higher Institute of Informatics and Communication Technologies (Tunisia)	5

Table 3: Maximum Accuracy and Maximum BLEU Scores for the VQA Task (out of each team's submitted runs).

Team	*Accuracy*	*BLEU*
dua_dua	0.382	0.416
Zhao_Ling_Ling	0.362	0.402
TeamS	0.348	0.391
zhao_shi_	0.316	0.352
IALab_PUC	0.236	0.276
Li_Yong_	0.222	0.255
sheerin	0.196	0.227
Baseline 1	0.288	0.326
Baseline 2	0.134	0.156

Table 4: Maximum Average BLEU Scores for the VQG Task (out of each team's submitted runs).

Team	*Average BLEU*
Chabbiimen	0.383
Baseline	0.274

metric is also used to evaluate the submissions for the VQG task, where we essentially compute the word overlap-based average similarity score between the system-generated questions and the ground truth question for each given test image [23]. The overall results of the participating systems are presented in Table 3 and Table 4 in a descending order of the accuracy and average BLEU scores respectively (the higher the better).

[23] `https://github.com/abachaa/VQA-Med-2021/tree/main/EvaluationCode`

3.5 Lessons Learned

Similar to last three years, participants continued to use state-of-the-art deep learning techniques to build their VQA-Med systems for both VQA and VQG tasks [21, 3, 2]. In particular, most systems leveraged encoder-decoder architectures with, e.g., deep convolutional neural networks (CNNs) like VGGNet or ResNet. A variety of pooling strategies were explored, e.g., global average pooling to encode image features and transformer-based architectures like BERT or recurrent neural networks (RNN) to extract question features (for the VQA task). Various types of attention mechanisms are also used coupled with different pooling strategies such as multimodal factorized bilinear (MFB) pooling or multi-modal factorized high-order pooling (MFH) in order to combine multimodal features followed by bilinear transformations to finally predict the possible answers in the VQA task and generate possible question words in the VQG task.

Analyses of the results in Table 3 suggest that in general, participating systems performed well for the VQA task. For the VQG task, results in Table 4 suggest that the task was comparatively challenging than the VQA task, but participating systems achieved better BLEU scores compared to last year's VQG results [2].

4 The Tuberculosis Task

Tuberculosis (TB) is a bacterial infection caused by a germ called Mycobacterium tuberculosis. About 130 years after its discovery, the disease remains a persistent threat and one of the top 10 causes of death worldwide according to the WHO [47]. The bacteria usually attack the lungs and generally TB can be cured with antibiotics. However, the different types of TB require different treatments, and therefore detection of the specific case characteristics is an important real-world task.

In the previous editions of this task, the setup evolved from year to year. In the first two editions [15, 17] participants had to detect Multi-drug resistant patients (MDR subtask) and to classify the TB type (TBT subtask) both based only on the CT image. After 2 editions it was concluded to drop the MDR subtask because it seemed impossible to solve based only on the image, and the TBT subtask was also suspended because of a very little improvement in the results between the 1st and the 2nd editions. At the same time, most of the participants obtained good results in the severity scoring (SVR) subtask introduced in 2018. In the 3d edition Tuberculosis task [16] was restructured to allow usage of the uniform dataset, and included two subtasks - continued Severity Score (SVR) prediction subtask and a new subtask based on providing an automatic report (CT Report) on the TB case. In the 4th edition [27], the SVR subtask was dropped and the automated CT report generation task was modified to be lung-based rather than CT-based.

Because of the fairly high results achieved by the participants in the CTR task last year, we decided to discontinue the CTR task at the moment and switch

to the task which was not yet solved with high quality. So in this year's edition, it was decided to bring back to life the Tuberculosis Type classification task from the 1st and 2nd ImageCLEFmed Tuberculosis editions. The task dataset was updated, extended in size, and some additional information was added for part of the CT scans.

We hoped that utilizing the newest deep learning approaches together with available at the moment pre-trained models and additional data sets will allow the participants to achieve better results for the TB Type classification compared to the early editions of the task.

4.1 Task Setup

In this task, participants had to automatically categorize each TB case into one of the following five types: (1) Infiltrative, (2) Focal, (3) Tuberculoma, (4) Miliary, (5) Fibro-cavernous. So the task is a multi-label classification problem.

4.2 Data Set

In this edition, the data set containing chest CT scans of 1,338 TB patients was used: 917 images for the training (development) data set and 421 for the test set. Some of the scans were accompanied by additional meta-information, depending on data available for different cases. Each CT image corresponded to only one TB type and to one unique patient. For all patients, we provided 3D CT images with a slice size of 512×512 pixels and a variable number of slices (the median number was 128).

Same as in the previous year, for all patients we provided two versions of automatically extracted masks of the lungs obtained using the methods described in [14, 29].

4.3 Participating Groups and Submitted Runs

In 2021, 11 groups from 9 countries submitted at least one run. Similar to the previous editions, each group could submit up to 10 runs. 64 scored runs were submitted in total. All groups used CNNs in some way, and two groups used a combination of CNN and RNN. Several groups tried a few different methods during their experiments, all reported approaches are listed below.

The majority of participants (seven groups) used 2D CNN to analyze either selected projections of CT images or all slices. Two of these groups further used per-slice features output of 2D CNN to train RNN in order to extract inter-slice information. Four groups tried to utilize 3D CNNs for whole CT analysis. Different neural network architectures and model training tweaks were used by participants, the majority of participants also used transfer learning techniques. All participants used some approaches for artificial data set enlargement and a few pre-processing steps, such as resizing, normalization, slice filtering etc.

Table 5: Results obtained by the participants of the task. Only the best run of each participant is reported here.

Group name	*Run ID*	*Kappa*	*Accuracy*	*Run rank*
SenticLab.UAIC	135715	0.221	0.466	1
hasibzunair	135720	0.200	0.423	4
SDVA-UCSD	135721	0.190	0.371	8
Emad_Aghajanzadeh	135689	0.181	0.404	11
MIDL-NCAI-CUI	134939	0.140	0.333	23
uaic2021	135708	0.129	0.333	28
IALab_PUC	134688	0.120	0.401	30
KDE-lab	133407	0.117	0.382	31
JBTTM	134791	0.038	0.221	42
Zhao_Shi_	133103	0.015	0.380	47
YNUZHOU	133288	-0.008	0.385	55

4.4 Results

The task was evaluated as a multi-label classification problem and scored using unweighted Cohen's Kappa and accuracy metrics. The ranking of this task is done first by Kappa and then by accuracy. Table 5 shows the final results for each group's best run and includes the run rank. More detailed results, including other performance measures, are presented in the overview article [26].

4.5 Lessons Learned and Next Steps

The results obtained in the task should be compared to the same TBT subtask presented in the 2018 edition. Before comparison, we should note, that although the task setup is the same in both editions, the data set was significantly changed, which means participants needed to deal with different images and labels distribution, so the scores can't be compared directly.

Top scores in the 2018 and 2021 editions are pretty close. The best result of 2021 achieved by SenticLab.UAIC group is slightly worse than the best result of 2018 - 0.221 vs 0.231 (-0.01 drop). On the other hand, four groups overcome 2nd best result from 2018. We should also mention that the group SDVA-UCSD participated in both editions and was able to improve Kappa score from 0.15 to 0.19. The best performer, SenticLab.UAIC group used per-slice analysis, which combined selection of relevant slices and their analysis by EfficientNet-B4 network. The 2nd-ranked hasibzunair group developed a hybrid CNN-RNN model and used pre-training on human action videos. The 3rd ranked SDVA-UCSD group used 3D ResNet34 with convolutional block attention.

Results analysis shows, that while the best result was not improved this year compared to the similar 2018 subtask, overall the top-5 scores of 2021 look better than in the 2018 edition, and the group which participated in both editions was able to improve its result. Analyzing participants working notes we observed the variability of participants approaches (top-3 groups used very different methods)

and usage of modern machine learning techniques and methods. As a result, we can conclude that the task is successful and its outcome is informative and useful.

Possible updates for future editions of TBT task should consider: (i) extending the additional meta-information for CT scans; (ii) including some kind of lesion location information to the data set.

5 The Caption Task

The caption task was first proposed as part of the ImageCLEFmedical [23] in 2016. In 2017 and 2018 [15, 22] the ImageCLEFcaption task comprised two subtasks: concept detection and caption prediction. In 2019 [34] and 2020 [35], the task concentrated on extracting Unified Medical Language System® (UMLS) Concept Unique Identifiers (CUIs) [6] from radiology images.

In 2021 [33], both subtasks, concept detection and caption prediction, were running again due to participants demands. To make the task more realistic, the focus in ImageCLEF 2021 lies in using real radiology images annotated by medical doctors in contrast to earlier years where images have been extracted from medical publications. Since this task can be considered as a first step of the Visual Question Answering Task 3, this year both tasks used the same dataset.

5.1 Task Setup

The ImageCLEFmed Caption 2021 [33] follows the format of the ImageCLEFmed caption previous tasks. In 2021, the overall task comprises two sub-tasks: "Concept Detection" and "Caption prediction". The concept detection subtask focuses on predicting Unified Medical Language System® (UMLS) Concept Unique Identifiers (CUIs) [6] based on the visual image representation in a given image. The caption prediction subtask focuses composing coherent captions for the entirety of the images.

The detected concepts are evaluated using the balanced precision and recall trade-off in terms of F1-scores, as in previous years. The predicted captions are evaluated using the BLEU score independent from the first subtask and designed to be robust to variability in style and wording.

5.2 Data Set

In 2021, the dataset is the same as the ImageCLEFVQA task [4] (see details in Section 4.2). The VQA-Med collection of radiology images and their annotations were used as a basis for the extraction of the concepts and captions. Semi-automatic text preprocessing was then applied to improve the quality of the annotations.

Following this approach, we provided new training, validation, and test sets for both tasks:

– The Caption and Concept training sets contain 2,756 radiology images and associated captions and concepts.

– The validation sets contain 500 radiology images and associated captions and concepts.
– The test sets contain 500 radiology images and associated reference captions and concepts.

We have also validated all the captions manually and checked the coherence of the generated concepts in the training, validation, and test sets.

As an additional source for training machine learning systems, the ROCO dataset [36], that has been used in the preceding years could be used by the participants.

Table 6: Performance of the participating teams in the ImageCLEF 2021 Concept Detection Task. The best run per team is selected. Teams with previous participation in 2020 are marked with an asterix.

Team	*Institution*	*F1 Score*
AUEB NLP Group*	Information Processing Laboratory, Department of Informatics, Athens University of Economics and Business, Athens, Greece	0.505
NLIP-Essex*-ITESM	School of Computer Science and Electronic Engineering, University of Essex, Colchester, UK and Instituto Tecnologico y de Estudios Superiores de Monterrey, Monterrey, Mexico	0.469
ImageSem	Institute of Medical Information and Library, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing, China	0.419
IALab PUC	Department of Computer Science, Pontificia Universidad Católica de Chile, Región Metropolitana, Chile	0.360
RomiBed	The Center for machine vision and signal analysis, University of Oulu, Oulu, Finland	0.143

5.3 Participating Groups and Submitted Runs

In the fifth edition of the ImageCLEFcaption task, 23 teams registered and signed the End-User-Agreement license, needed to download the development data. 75 graded runs were submitted for evaluation by 10 teams (8 submitted working notes) attracting more attention than last year. Each of the group was allowed 10 graded runs per subtask. In the concept detection task 5 teams participated and 2 teams also took part in the 2020 challenge. The caption prediction task raised interest of 8 teams, that submitted their results, 2 teams decided not to submit working notes.

In the concept detection subtask, the groups typically used deep learning models trained as multi-label classificators or more Information Retrieval oriented solutions. For the IR solutions, image embeddings from deep learning

Table 7: Performance of the participating teams in the ImageCLEF 2021 Caption Prediction Task. The best run per team is selected.

Team	*Institution*	*BLEU Score*
IALab PUC	Department of Computer Science, Pontificia Universidad Católica de Chile, Región Metropolitana, Chile	0.5098
AUEB NLP Group	Information Processing Laboratory, Department of Informatics, Athens University of Economics and Business, Athens, Greece	0.4610
AEHRC-CSIRO	Australian e-Health Research Centre, Commonwealth Scientific and Industrial Research Organisation, Herston, Australia	0.4319
kdelab	Department of Computer Science and Engineering, Toyohashi University of Technology, Aichi, Japan	0.3616
jeanbenoit_delbrouck	Laboratory of Quantitative Imaging and Artificial Intelligent, Department of Biomedical Data Science, Stanford University, Stanford, United States	0.2850
ImageSem	Institute of Medical Information and Library, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing, China	0.2565
RomiBed	Center for machine vision and signal analysis, University of Oulu, Finland	0.2427
ayushnanda14	Department of Computer Science and Engineering, Siva Subramaniya Nadar College of Engineering, Kalavakkam, India	0.1029

models are typically used. In this year, more modern deep learning architectures like EfficientNets [41] and Visual Transformers (ViT) [18] have been proposed for the solutions. In the caption prediction task, several teams used variations of the Show, Attend and Tell model. New has been the occurrence of Transformer based architectures and general language models like GPT-2 [38]. Transfer Learning has frequently been used and some teams in both subtasks tried to pretrain with more medically oriented datasets like ROCO or CheXpert.

To get a better overview of the submitted runs, the best results for each team are presented in Tables 6 and 7.

5.4 Results

This years models for concept detection show again increased F1-scores in comparison to earlier years. This could partly be explained by a smaller number of potential concepts in the images. More modern architectures have been used and show improvements. Transformer based architectures and solutions arrived at both tasks. For concept detection in this year machine learning based methods and information retrieval oriented solutions have been used more equally by

all groups. In former years the majority of proposed solutions used multi-label approaches. Some participants noticed, that less complex solutions showed the best results. An in-depth analysis is presented in [33].

5.5 Lessons Learned and Next Steps

The participants appreciated, that more realistic medical images have been used in contrast to the publication based images from last years. On the other hand the size of the training and testing datasets is small in comparison to other datasets. This leads to simpler solutions as less concepts are present and the captions show less variation. One expectable next step would be to increase the number of concepts and variation of image descriptions further by increasing the dataset size. The use of the ROCO dataset as a pretraining solutions showed no improvement for the groups that used it. It can be assumed, that the descriptions/captions of the VQA-task images have a different focus in comparison to the ROCO images.

6 The DrawnUI Task

Creating high quality User Interfaces (UI) is a complex process involving several actors such as designers and developers. As more companies push to increase their online presence, the automatization of this process is gaining interest. Pix2Code [1] and UI2Code [10] were proposed in 2018 to tackle this challenge, those solutions took as input a screenshot and output a domain specific language representing the UI.

The first edition of the ImageCLEFdrawnUI task [20] took place in 2020 with a data set of 3,000 wireframe. Participants were tasked to create a computer vision system to localize and identify different UI elements in the drawings. Two of the three participating teams obtained results exceeding the baseline using various object detection algorithm combined with data preprocessing, cleaning and augmentation.

6.1 Task Setup

The 2021 ImageCLEFdrawnUI task (see the detailed overview paper [5]) is the second edition of the task and consist of two challenges. Given hand drawn (wireframes) and digital (screenshots) images of user interfaces, participants must develop a machine learning models to predict the bounding boxes coordinates and type of each UI elements in the images. For each task, the data sets are separated in 75% for training and 25% for validation. The $MAP0.5IoU$ and $R0.5IoU$ [19] were used to evaluate the submissions.

6.2 Data Set

For the wireframe task, the data set contained 4,291 hand-drawn wireframe images. Each images was drawn based on actual screenshots of mobile and web

UIs. Images from the RICO data set [13] were used for the mobile UI while a custom parser was used to obtain the web pages UIs. For the drawing itself, three persons were involved and had to use a predefined dictionary of 21 shapes and were instructed to focus on an unambiguous drawing instead of fidelity to the original screenshot to facilitate the following annotation step and thereafter the computer vision task. The VOTT software[24] was used for annotation by two different annotators and verified by a single person afterward. In the previous edition, there was a large class imbalance in the dataset, to overcome this, new images containing a larger proportion of the rare class were introduce and the class distribution was more carefully monitored during the creation of both train and validation set.

Table 8: Participation in the DrawUI 2021 task, wireframe subtask: the best score from all runs for each team.

Team	*#Runs*	*MAP@0.5*	*R@0.5*
vyskocj	10	0.900	0.934
pwc	10	0.836	0.865
AIMultimediaLab	1	0.216	0.319

Table 9: Participation in the DrawUI 2021 task, screenshot subtask: the best score from all runs.

Team	*#Runs*	*MAP@0.5*	*R@0.5*
vyskocj	7	0.628	0.83

For the screenshot task, the data set consisted of 9,276 screenshots. A custom parser was used to obtain the images, In addition to the screenshot, the parser also screened the Document Oriented Model to extract the position and type of each HTML element of the webpages. Those UI elements were then attributed when applicable to one of the 6 elements of the retained dictionary (TEXT, IMAGE, HEADING, BUTTON, INPUT, LINK).

6.3 Participating Groups and Submitted Runs

8 teams registered for both tasks. For the wireframe task, 3 teams from 3 countries submitted 21 runs. For the screenshot task, 1 team submitted 7 run. Teams were limited to submit 10 runs.

[24] `https://github.com/microsoft/VoTT`

6.4 Results

The $MAP0.5IoU$ and $R0.5IoU$ scores have been compiled using the Python API of COCO[25]. For both subtasks, the participants used recent object detection model architectures such as YOLOv5 and Faster R-CNN supplemented by a Feature Pyramid Network. Data augmentation methods were also employed, ranging from color and contrast normalization, to random cutting out of objects and relative resizing of the images. In the screenshot subtask, low-quality data points were filtered out based on color similarity checking. Overall, these experiments brought the mAP score to 0.900 for the wireframe subtask and 0.628 for the screenshot one, representing a promising improvement compared to the 2020 edition.

6.5 Lessons Learned and Next Steps

Based on the high scores obtained when tackling it, the wireframe challenge is nearing full completion. For the screenshot subtask, it was also demonstrated that a smaller sized model converged faster to an adequate level, indicating that large resource allocation is not a necessity for satisfactory results. Although the participation rate was very low, our baseline scores were still surpassed and the contestants proposed uniquely adapted modifications of the data set and the models for solving the subtasks.

For the next editions of the task, the further development and extension the two data sets remains a priority. We will stress making them more challenging from a technical perspective, as well as showcasing them to the UI-based communities, attracting more participants interested in the ML-facilitated development of user interfaces.

7 The Coral Task

There is a crucial need to implement effective monitoring techniques to protect coral reefs immediately and in the long term [46]. This monitoring process can be made by collecting 3D visual data using autonomous underwater vehicles which will provide useful information for both annotation and further study of the coral. The ImageCLEFcoral task organisers have developed a novel multi-camera system that allows large amounts of imagery to be captured by a SCUBA diver or autonomous underwater vehicle in a single dive.

In its 3rd edition, the ImageCLEFcoral data form the complete set of images required to form a 3D reconstruction of the environment. This allows the participants to explore novel probabilistic computer vision techniques based around image overlap and transposition of data points.

[25] `https://github.com/cocodataset/cocoapi`

7.1 Task Setup

Following the format of previous editions of the ImageCLEFcoral task [7, 8], in 2021 participants were again asked to devise and implement algorithms for automatically annotating regions in a collection of images containing several types of benthic substrate, such as hard coral or sponge. As in previous editions, the overall task comprises two sub-tasks: "Coral reef image annotation and localisation" and "Coral reef image pixel-wise parsing" subtasks. The "Coral reef image annotation and localisation" subtask uses bounding boxes for the annotation, with sides parallel to the edges of the image, around identified features. The "Coral reef image pixel-wise parsing" subtasks uses a series of boundary image coordinates which form a single polygon around each identified feature; this has been dubbed *pixel-wise parsing* (these polygons should not have self-intersections). Participants were invited to make submissions for either or both tasks.

Algorithmic performance is evaluated on the unseen test data using the popular intersection over union metric from the PASCAL VOC[26] exercise. This computes the area of intersection of the output of an algorithm and the corresponding ground truth, normalising that by the area of their union to ensure its maximum value is bounded.

7.2 Data Set

As in previous editions, the data for this ImageCLEFcoral task originates from a growing, large-scale collection of images taken from coral reefs around the world as part of a coral reef monitoring project with the Marine Technology Research Unit at the University of Essex. The images contain annotations of the following 13 types of substrates: Hard Coral – Branching, Hard Coral – Submassive, Hard Coral – Boulder, Hard Coral – Encrusting, Hard Coral – Table, Hard Coral – Foliose, Hard Coral – Mushroom, Soft Coral, Soft Coral – Gorgonian, Sponge, Sponge – Barrel, Fire Coral – Millepora and Algae - Macro or Leaves.

In 2021, the training and test data form the complete set of images required to form a 3D reconstruction of the environment. The training dataset contains images from 6 subsets from 4 locations. 1 subset is complete (containing all the images to build the 3D model) and 5 subsets contain a partial collection. The test data contains the images required to complete 4 of the partial image sets from each of the 4 locations (the final partial subset is not used for testing, only training).

In addition, participants are encouraged to use the publicly available NOAA NCEI data[27] and/or CoralNet[28] to train their approaches.

[26] `http://host.robots.ox.ac.uk/pascal/VOC/`
[27] `https://www.ncei.noaa.gov/`
[28] `https://coralnet.ucsd.edu/`

7.3 Participating Groups and Submitted Runs

In this third edition of the ImageCLEFcoral task, 8 teams registered, of which 3 teams submitted 8 runs. Teams were limited to submit 10 runs per subtask. To get a better overview of the submitted runs, the best results for each team are presented in Tables 10 and 11. An in-depth analysis is presented in [9].

Table 10: Coral reef image annotation and localisation performance in terms of $MAP0.5IoU$. The best run per team is selected.

Run id	*Team*	$MAP0.5IoU$
139118	UAlbany	0.457
138115	University of West Bohemia	0.121

Table 11: Pixel-wise coral reef parsing performance in terms of $MAP0.5IoU$. The best run per team is selected.

Run id	*Team*	$MAP0.5IoU$
139084	University of West Bohemia	0.075
138389	MTRU	0.021

7.4 Results

The results from both tasks showed lower performance than has been achieved in previous years. More detailed analysis of the results is presented in [9], where pixel accuracy per class is investigated. This gives us a better indication as to which classes are difficult to train for and identify. Previous years' tasks used only training data from a single location, so the reason for obtaining good performance when testing with a dataset from the same area is clear. By contrast, this year both the training and test datasets were from multiple locations. In addition, some participants included large-scale training datasets from a fifth location.

7.5 Lessons Learned and Next Steps

The varied morphology and distribution of substrates across different datasets and locations suggest that trying to develop a single generic algorithm to detect coral reef substrate type will be challenging. This proved to be the case for the datasets used in this task, even with the incorporation of considerably larger datasets from other sources as training corpora. The next steps for this work are to leverage the image overlap of the data to develop probabilistic labelled models in 3D and develop cross-compatibility in large datasets for use in this task.

8 The Aware Task

Social networks engage the users to share their personal data in order to interact with other users. The context of the sharing is chosen by the users but they do not have control on further data use. These data are automatically aggregated into profiles which are exploited by social networks to propose personalized advertising/services to users. Depending on their visibility, data can be also consulted by other entities to make decisions which have a high impact on the user's life. It is thus important to give users feedback about the potential real-life effects of their personal data sharing.

We designed a task focused on the automatic rating of visual user profile in four impactful situations. Each profile includes 100 photos and its appeal is manually evaluated via crowdsourcing. Participants are asked to provide automatic visual profile ratings obtained by using a training set which includes visual- and situation-related information. These ratings are then ranked and compared to manual ones in order to assess the feasibility of providing automatic feedback related to the effects of personal photos sharing. Two teams submitted results for this first edition of the task.

8.1 Task Setup

This is the first edition of the task and consists of one challenge. Participants are provided with automatic object detections for the images and with object ratings per situation. Then, the objective is to propose a ranking of user profiles which is as close as possible to the crowdsourced one. Data were split into 360/40/100 profiles for training/validation and test. The Pearson correlation coefficient between manual and automatic profile rankings was used to evaluate the quality of proposed runs. The final scores were calculated by averaging correlations obtained for individual situations.

8.2 Data Set

A data set of 500 user profiles with 100 photos per profile was created and annotated with an "appeal" score for four real-life situations via crowdsourcing. The modeled situations are demands for: a bank credit, an accommodation, a job as an IT engineer, a job as a waiter. Participants to the experiment were asked to provide a global rating of each profile in each situation modeled using a 7-points Likert scale ranging from "strongly unappealing" to "strongly appealing". The averaged "appeal" score was used to create a ground truth composed of ranked users in each modeled situation. User profiles are created by repurposing a subset of the YFCC100M dataset [43].

Situations are modeled by crowdsourcing visual objects ratings. Similar to profile crowdsourcing, object ratings are collected for each situation using a 7-points Likert scale with ratings between -3 (strongly negative influence) to +3 (strongly positive influence). The averaged rating is computed and provided to participants. A Faster R-CNN object detector was trained in order to detect

objects in images. The detection dataset combines objects from OpenImages [28], ImageNet [39] and COCO [30]. Only objects with at least one non-zero situation rating were kept. All objects detected in the 100 images of a profile were provided to participants, along with the detection probability and the associated bounding box. Given a situation, the combination of the ratings of objects and of their automatic detection enables the automatic computation of a profile score.

Given the personal nature of the included profiles, the dataset was anonymized in order to comply with GDPR. Participants did not have access to the images, and the user IDs and the object names were hashed.

8.3 Participating Groups and Submitted Runs

We received in total 6 valid submissions from 2 teams. SIP_Team was from the University of Paris, France. v18nguye is an independent researcher. None of the two participants provided details about their participation.

Table 12: Results of the Aware 2021 task.

Team	*# Runs*	*Pearson*
SIP_Team	3	0.597
v18nguye	3	0.388

8.4 Lessons Learned and Next Steps

While no details were provided about the implemented methods, the scores reported in Table 12 give a good correlation between automatic and manual profile rankings. This means that automatic methods for computing visual profile ratings are effective.

These initial results encourage us to pursue the task next year. We plan to: (1) enrich the dataset with new objects which have a strong influence in at least one of the modeled situations, (2) use more recent object detectors, such as EfficientDet [42], which should boost results via an improved photo analysis and (3) increase the number of user profiles in order to have a more representative training set.

9 Conclusion

This paper presents a general overview of the activities and outcomes of the ImageCLEF 2021 evaluation campaign. Four tasks were organised, covering challenges in the medical domain (visual question answering and visual question generation, tuberculosis prediction, and caption analysis), nature (segmenting and labeling collections of coral images), Internet (identifying website user interface

components), and social networks (analysis of the real-life effects of personal data sharing). Despite the outbreak of the COVID-19 pandemic and lock-down during the benchmark, 103 teams registered, 42 teams completed the tasks and submitted over 256 runs.

As anticipated already, most of the proposed solutions evolved around state-of-the-art deep neural network architectures. In the VQA task most systems leveraged encoder-decoder architectures with, e.g., deep convolutional neural networks (CNNs) like VGGNet or ResNet. Systems were able to solve the VQA task with good performance. The VQG task proved to be more challenging, however, results improved compared to the last year's edition. In the tuberculosis task, the best result was not improved this year compared to the similar 2018 task. However, overall, the top-5 scores of 2021 look better than in the 2018 edition, and the group which participated in both editions was able to improve its result. The methods employed a variety of different approaches. In the caption task, the more realistic medical images were closer to a real-world use case scenario. On the other hand, the size of the training and testing datasets is smaller. This led to simpler solutions as less concepts are present and the captions show less variation.

In the drawnUI task, the wireframe challenge achieved close to perfect solutions. For the screenshot task, it was also demonstrated that a smaller sized model converged faster to an adequate level, indicating that large resource allocation is not a necessity for satisfactory results. In the coral task, the varied morphology and distribution of substrates across different datasets and locations suggest that trying to develop a single generic algorithm to detect coral reef substrate type will be challenging. This was also visible from the results which are still low for an on-the-field application. The aware task is a new concept and was in its first edition this year. Despite the incipient participation, achieved results prove the feasibility of the concept.

ImageCLEF 2021 brought again together an interesting mix of tasks and approaches and we are looking forward to the fruitful discussions at the CLEF 2021 workshop.

Acknowledgements

Data collection for the Tuberculosis task was supported by the National Institute of Allergy and Infectious Diseases, National Institutes of Health, US Department of Health and Human Services, CRDF project DAA9-19-65987-1.

The aware task was fully supported and the drawnUI was partially supported under project AI4Media, A European Excellence Centre for Media, Society and Democracy, H2020 ICT-48-2020, grant #951911.

References

1. Beltramelli, T.: pix2code : Generating Code from a Graphical User Interface Screenshot. Proceedings of the ACM SIGCHI Symposium on Engineering Interactive Computing Systems pp. 1–9 (2018)

2. Ben Abacha, A., Datla, V.V., Hasan, S.A., Demner-Fushman, D., Müller, H.: Overview of the vqa-med task at imageclef 2020: Visual question answering and generation in the medical domain. In: CLEF 2020 Working Notes. CEUR Workshop Proceedings, CEUR-WS.org, Thessaloniki, Greece (September 22-25 2020)
3. Ben Abacha, A., Hasan, S.A., Datla, V.V., Liu, J., Demner-Fushman, D., Müller, H.: VQA-Med: Overview of the medical visual question answering task at imageclef 2019. In: CLEF2019 Working Notes. CEUR Workshop Proceedings, CEUR-WS.org <http://ceur-ws.org>, Lugano, Switzerland (September 09-12 2019)
4. Ben Abacha, A., Sarrouti, M., Demner-Fushman, D., Hasan, S.A., Müller, H.: Overview of the vqa-med task at imageclef 2021: Visual question answering and generation in the medical domain. In: CLEF 2021 Working Notes. CEUR Workshop Proceedings, CEUR-WS.org, Bucharest, Romania (September 21-24 2021)
5. Berari, R., Tauteanu, A., Fichou, D., Brie, P., Dogariu, M., Ştefan, L.D., Constantin, M.G., Ionescu, B.: Overview of ImageCLEFdrawnUI 2021: The detection and recognition of hand drawn and digital website uis task. In: CLEF2021 Working Notes. CEUR Workshop Proceedings, CEUR-WS.org <http://ceur-ws.org>, Bucharest, Romania (September 21-24 2021)
6. Bodenreider, O.: The Unified Medical Language System (UMLS): integrating biomedical terminology. Nucleic Acids Research **32**(Database-Issue), 267–270 (2004). https://doi.org/10.1093/nar/gkh061
7. Chamberlain, J., Campello, A., Wright, J.P., Clift, L.G., Clark, A., García Seco de Herrera, A.: Overview of ImageCLEFcoral 2019 task. In: CLEF2019 Working Notes. CEUR Workshop Proceedings, CEUR-WS.org (2019)
8. Chamberlain, J., Campello, A., Wright, J.P., Clift, L.G., Clark, A., García Seco de Herrera, A.: Overview of the ImageCLEFcoral 2020 task: Automated coral reef image annotation. In: CLEF2020 Working Notes. CEUR Workshop Proceedings, CEUR-WS.org (2020)
9. Chamberlain, J., García Seco de Herrera, A., Campello, A., Clark, A., Oliver, T.A., Moustahfid, H.: Overview of the ImageCLEFcoral 2021 task: Coral reef image annotation of a 3d environment. In: CLEF2021 Working Notes. CEUR Workshop Proceedings, CEUR-WS.org, Bucharest, Romania (September 21-24 2021)
10. Chen, C., Su, T., Meng, G., Xing, Z., Liu, Y.: From UI Design Image to GUI Skeleton : A Neural Machine Translator to Bootstrap Mobile GUI Implementation. International Conference on Software Engineering **6** (2018)
11. Clough, P., Müller, H., Sanderson, M.: The CLEF 2004 cross–language image retrieval track. In: Peters, C., Clough, P., Gonzalo, J., Jones, G.J.F., Kluck, M., Magnini, B. (eds.) Multilingual Information Access for Text, Speech and Images: Result of the fifth CLEF evaluation campaign. Lecture Notes in Computer Science (LNCS), vol. 3491, pp. 597–613. Springer, Bath, UK (2005)
12. Clough, P., Sanderson, M.: The CLEF 2003 cross language image retrieval task. In: Proceedings of the Cross Language Evaluation Forum (CLEF 2003) (2004)
13. Deka, B., Huang, Z., Franzen, C., Hibschman, J., Afergan, D., Li, Y., Nichols, J., Kumar, R.: Rico: A mobile app dataset for building data-driven design applications. In: UIST 2017 - Proceedings of the 30th Annual ACM Symposium on User Interface Software and Technology. pp. 845–854 (2017). https://doi.org/10.1145/3126594.3126651
14. Dicente Cid, Y., Jimenez-del-Toro, O., Depeursinge, A., Müller, H.: Efficient and fully automatic segmentation of the lungs in CT volumes. In: Orcun Goksel, Jimenez-del-Toro, O., Foncubierta-Rodriguez, A., Müller, H. (eds.) Proceedings of the VISCERAL Challenge at ISBI. pp. 31–35. No. 1390 in CEUR Workshop Proceedings (Apr 2015)

15. Dicente Cid, Y., Kalinovsky, A., Liauchuk, V., Kovalev, V., , Müller, H.: Overview of ImageCLEFtuberculosis 2017 - predicting tuberculosis type and drug resistances. In: CLEF2017 Working Notes. CEUR Workshop Proceedings, CEUR-WS.org <http://ceur-ws.org>, Dublin, Ireland (September 11-14 2017)
16. Dicente Cid, Y., Liauchuk, V., Klimuk, D., Tarasau, A., Kovalev, V., Müller, H.: Overview of ImageCLEFtuberculosis 2019 - automatic ct-based report generation and tuberculosis severity assessment. In: CLEF2019 Working Notes. CEUR Workshop Proceedings, CEUR-WS.org <http://ceur-ws.org>, Lugano, Switzerland (September 9-12 2019)
17. Dicente Cid, Y., Liauchuk, V., Kovalev, V., , Müller, H.: Overview of ImageCLEFtuberculosis 2018 - detecting multi-drug resistance, classifying tuberculosis type, and assessing severity score. In: CLEF2018 Working Notes. CEUR Workshop Proceedings, CEUR-WS.org <http://ceur-ws.org>, Avignon, France (September 10-14 2018)
18. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In: Proceedings of the 9th International Conference on Learning Representations (ICLR 2021): 03-07.05.2021; Online Event (2021), https://openreview.net/forum?id=YicbFdNTTy
19. Everingham, M., Gool, L.V., Williams, C.K.I., Winn, J., Zisserman, A.: The PASCAL Visual Object Classes (VOC) Challenge. Int J Comput Vis **88**, 303–338 (2010). https://doi.org/10.1007/s11263-009-0275-4
20. Fichou, D., Berari, R., Brie, P., Dogariu, M., Ştefan, L.D., Constantin, M.G., Ionescu, B.: Overview of ImageCLEFdrawnUI 2020: The Detection and Recognition of Hand Drawn Website UIs Task. In: CLEF2020 Working Notes. CEUR Workshop Proceedings, CEUR-WS.org <http://ceur-ws.org>, Thessaloniki, Greece (September 22-25 2020)
21. Hasan, S.A., Ling, Y., Farri, O., Liu, J., Lungren, M., Müller, H.: Overview of the ImageCLEF 2018 medical domain visual question answering task. In: CLEF2018 Working Notes. CEUR Workshop Proceedings, CEUR-WS.org <http://ceur-ws.org>, Avignon, France (September 10-14 2018)
22. García Seco de Herrera, A., Eickhoff, C., Andrearczyk, V., , Müller, H.: Overview of the ImageCLEF 2018 caption prediction tasks. In: CLEF2018 Working Notes. CEUR Workshop Proceedings, CEUR-WS.org <http://ceur-ws.org>, Avignon, France (September 10-14 2018)
23. García Seco de Herrera, A., Schaer, R., Bromuri, S., Müller, H.: Overview of the ImageCLEF 2016 medical task. In: Working Notes of CLEF 2016 (Cross Language Evaluation Forum) (September 2016)
24. Ionescu, B., Müller, H., Péteri, R., Cid, Y.D., Liauchuk, V., Kovalev, V., Klimuk, D., Tarasau, A., Abacha, A.B., Hasan, S.A., Datla, V., Liu, J., Demner-Fushman, D., Dang-Nguyen, D.T., Piras, L., Riegler, M., Tran, M.T., Lux, M., Gurrin, C., Pelka, O., Friedrich, C.M., de Herrera, A.G.S., Garcia, N., Kavallieratou, E., del Blanco, C.R., Rodríguez, C.C., Vasillopoulos, N., Karampidis, K., Chamberlain, J., Clark, A., Campello, A.: ImageCLEF 2019: Multimedia retrieval in medicine, lifelogging, security and nature. In: Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the 10th International Conference of the CLEF Association (CLEF 2019), vol. 11438. LNCS Lecture Notes in Computer Science, Springer, Lugano, Switzerland (September 9-12 2019)

25. Kalpathy-Cramer, J., García Seco de Herrera, A., Demner-Fushman, D., Antani, S., Bedrick, S., Müller, H.: Evaluating performance of biomedical image retrieval systems: Overview of the medical image retrieval task at ImageCLEF 2004–2014. Computerized Medical Imaging and Graphics **39**(0), 55 – 61 (2015)
26. Kozlovski, S., Liauchuk, V., Dicente Cid, Y., Kovalev, V., Müller, H.: Overview of ImageCLEFtuberculosis 2021 - CT-based tuberculosis type classification. In: CLEF 2021 Working Notes. CEUR Workshop Proceedings, CEUR-WS.org <http://ceur-ws.org>, Bucharest, Romania (September 21-24 2021)
27. Kozlovski, S., Liauchuk, V., Dicente Cid, Y., Tarasau, A., Kovalev, V., Müller, H.: Overview of ImageCLEFtuberculosis 2020 - automatic CT-based report generation. In: CLEF2020 Working Notes. CEUR Workshop Proceedings, CEUR-WS.org <http://ceur-ws.org>, Thessaloniki, Greece (September 22-25 2020)
28. Kuznetsova, A., Rom, H., Alldrin, N., Uijlings, J.R.R., Krasin, I., Pont-Tuset, J., Kamali, S., Popov, S., Malloci, M., Duerig, T., Ferrari, V.: The open images dataset V4: unified image classification, object detection, and visual relationship detection at scale. CoRR **abs/1811.00982** (2018), `http://arxiv.org/abs/1811.00982`
29. Liauchuk, V., Kovalev, V.: Imageclef 2017: Supervoxels and co-occurrence for tuberculosis CT image classification. In: CLEF2017 Working Notes. CEUR Workshop Proceedings, CEUR-WS.org <http://ceur-ws.org>, Dublin, Ireland (September 11-14 2017)
30. Lin, T., Maire, M., Belongie, S.J., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft COCO: common objects in context. In: Fleet, D.J., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V. Lecture Notes in Computer Science, vol. 8693, pp. 740–755. Springer (2014). https://doi.org/10.1007/978-3-319-10602-1_48, `https://doi.org/10.1007/978-3-319-10602-1\_48`
31. Müller, H., Clough, P., Deselaers, T., Caputo, B. (eds.): ImageCLEF – Experimental Evaluation in Visual Information Retrieval, The Springer International Series On Information Retrieval, vol. 32. Springer, Berlin Heidelberg (2010)
32. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: BLEU: a method for automatic evaluation of machine translation. In: Proceedings of the 40th annual meeting on association for computational linguistics. pp. 311–318. Association for Computational Linguistics (2002)
33. Pelka, O., Ben Abacha, A., García Seco de Herrera, A., Jacutprakart, J., Friedrich, C.M., Müller, H.: Overview of the ImageCLEFmed 2021 concept & caption prediction task. In: CLEF2021 Working Notes. CEUR Workshop Proceedings, CEUR-WS.org, Bucharest, Romania (September 21-24 2021)
34. Pelka, O., Friedrich, C.M., García Seco de Herrera, A., Müller, H.: Overview of the ImageCLEFmed 2019 concept prediction task. In: CLEF2019 Working Notes. CEUR Workshop Proceedings, CEUR-WS.org <http://ceur-ws.org>, Lugano, Switzerland (September 09-12 2019)
35. Pelka, O., Friedrich, C.M., García Seco de Herrera, A., Müller, H.: Overview of the ImageCLEFmed 2020 concept prediction task: Medical image understanding. In: CLEF2020 Working Notes. CEUR Workshop Proceedings, CEUR-WS.org, Thessaloniki, Greece (September 22-25 2020)
36. Pelka, O., Koitka, S., Rückert, J., Nensa, F., Friedrich, C.M.: Radiology Objects in COntext (ROCO): a multimodal image dataset. In: Proceedings of the Third International Workshop on Large-Scale Annotation of Biomedical Data and Expert Label Synthesis (LABELS 2018), Held in Conjunction with MICCAI 2018. vol.

11043, pp. 180–189. LNCS Lecture Notes in Computer Science, Springer, Granada, Spain (September 16 2018)
37. Popescu, A., Deshayes-Chossart, J., Ionescu, B.: Overview of ImageCLEFaware 2021: Estimating potential real-life effects of online image sharing task. In: CLEF 2021 Working Notes. CEUR Workshop Proceedings, CEUR-WS.org <http://ceur-ws.org>, Bucharest, Romania (September 21-24 2021)
38. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I.: Language-models are unsupervised multitask learners. Tech. rep., Open-AI (2019)
39. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M.S., Berg, A.C., Li, F.: Imagenet large scale visual recognition challenge. International Journal of Computer Vision **115**(3), 211–252 (2015). https://doi.org/10.1007/s11263-015-0816-y, https://doi.org/10.1007/s11263-015-0816-y
40. Sarrouti, M., Ben Abacha, A., Demner-Fushman, D.: Visual question generation from radiology images. In: Proceedings of the First Workshop on Advances in Language and Vision Research. pp. 12–18. Association for Computational Linguistics, Online (Jul 2020), https://www.aclweb.org/anthology/2020.alvr-1.3
41. Tan, M., Le, Q.: EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. In: Chaudhuri, K., Salakhutdinov, R. (eds.) Proceedings of the 36th International Conference on Machine Learning (ICML 2019): 10-15.06.2019; Long Beach, California, US. vol. 97, pp. 6105–6114. Long Beach, California, USA (06 2019), http://proceedings.mlr.press/v97/tan19a.html
42. Tan, M., Pang, R., Le, Q.V.: Efficientdet: Scalable and efficient object detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10781–10790 (2020)
43. Thomee, B., Shamma, D.A., Friedland, G., Elizalde, B., Ni, K., Poland, D., Borth, D., Li, L.: YFCC100M: the new data in multimedia research. Commun. ACM **59**(2), 64–73 (2016)
44. Tsikrika, T., García Seco de Herrera, A., Müller, H.: Assessing the scholarly impact of ImageCLEF. In: CLEF 2011. pp. 95–106. Springer Lecture Notes in Computer Science (LNCS) (sep 2011)
45. Tsikrika, T., Larsen, B., Müller, H., Endrullis, S., Rahm, E.: The scholarly impact of CLEF (2000–2009). In: Information Access Evaluation. Multilinguality, Multimodality, and Visualization, pp. 1–12. Springer (2013)
46. Wilkins, K.W., Rosa-Marín, A., Cziesielski, M., Hughes, H., Love, C., Nowakowski, C.: Short and long-term visions for protecting coral reefs. Limnol. Oceanogr. Bull **30** (2021)
47. World Health Organization, et al.: Global tuberculosis report 2019 (2019)