Overview of the ImageCLEF 2020: Multimedia Retrieval in Medical, Lifelogging, Nature, and Internet Applications

Bogdan Ionescu[1], Henning Müller[2], Renaud Péteri[3], Asma Ben Abacha[4], Vivek Datla[5], Sadid A. Hasan[6], Dina Demner-Fushman[4], Serge Kozlovski[7], Vitali Liauchuk[7], Yashin Dicente Cid[8], Vassili Kovalev[7], Obioma Pelka[9], Christoph M. Friedrich[9], Alba García Seco de Herrera[10], Van-Tu Ninh[11], Tu-Khiem Le[11], Liting Zhou[11], Luca Piras[12], Michael Riegler[13], Pål Halvorsen[13], Minh-Triet Tran[14], Mathias Lux[15], Cathal Gurrin[11], Duc-Tien Dang-Nguyen[16], Jon Chamberlain[10], Adrian Clark[10], Antonio Campello[17], Dimitri Fichou[18], Raul Berari[18], Paul Brie[18], Mihai Dogariu[1], Liviu Daniel Ştefan[1], and Mihai Gabriel Constantin[1]

[1] University Politehnica of Bucharest, Romania bogdan.ionescu@upb.ro
[2] University of Applied Sciences Western Switzerland (HES-SO), Switzerland
[3] University of La Rochelle, France
[4] National Library of Medicine, USA
[5] Philips Research Cambridge, USA
[6] CVS Health, USA
[7] United Institute of Informatics Problems, Belarus
[8] University of Warwick, UK
[9] University of Applied Sciences and Arts Dortmund, Germany
[10] University of Essex, UK
[11] Dublin City University, Ireland
[12] Pluribus One & University of Cagliari, Italy
[13] University of Oslo, Norway
[14] University of Science, Vietnam
[15] Klagenfurt University, Austria
[16] University of Bergen, Norway
[17] Wellcome Trust, UK
[18] teleportHQ, Romania

Abstract. This paper presents an overview of the ImageCLEF 2020 lab that was organized as part of the Conference and Labs of the Evaluation Forum - CLEF Labs 2020. 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 2020, the 18th 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 *lifelog* task (videos, images and other sources) about daily activity understanding, retrieval and summarization, (iii) a *coral* task about segmenting and labeling collections

of coral reef images, and (iv) a new *Internet* task addressing the problems of identifying hand-drawn user interface components. Despite the current pandemic situation, the benchmark campaign received a strong participation with over 40 groups submitting more than 295 runs.

Keywords: visual question answering · visual question generation · lifelogging retrieval and summarization · medical image classification · coral image segmentation and classification · recognition of hand-drawn website user interface components · ImageCLEF benchmark · annotated data · common evaluation framework

1 Introduction

ImageCLEF[19] 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 [11]. It increased its impact with the addition of medical tasks in 2004 [10], attracting over 20 participants already in the second year. An overview of ten years of the medical tasks can be found in [29]. It continued the ascending trend, reaching over 200 participants in 2019. 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 [34]. ImageCLEF has shown to have an important impact over the years, already detailed in 2010 [47, 48].

Since 2018, ImageCLEF uses the crowdAI platform, now migrated to AIcrowd[20] 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 [47, 48]. For instance, the term "ImageCLEF" returns on Google Scholar[21] over 5,300 article results (search on July 3rd, 2020). This underlines the importance of evaluation campaigns for disseminating best scientific practices. We introduce here the four tasks that were run in the 2020 edition[22], namely: ImageCLEFmedical, ImageCLEFlifelog, ImageCLEFcoral, and the new ImageCLEFdrawnUI.

2 Overview of Tasks and Participation

ImageCLEF 2020 consists of four main tasks with the objective of covering a *diverse range* of multimedia retrieval applications, namely: *medicine*, *lifelogging*,

[19] `http://www.imageclef.org/`
[20] `https://www.aicrowd.com/`
[21] `https://scholar.google.com/`
[22] `https://www.imageclef.org/2020/`

nature, and *Internet applications*. It followed the 2019 tradition [28] of diversifying the use cases [2, 8, 19, 38, 35, 31]. The 2020 tasks are presented as follows:

- **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 third 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. The clinicians' confidence in interpreting complex medical images can be enhanced by a "second opinion" provided by an automated system. Since patients may now access structured and unstructured data related to their health via patient portals, such access motivates the need to help them better understand their conditions regarding their available data, including medical images. In view of this and inspired by the success of visual question answering in the general domain[23] and the previous VQA-Med editions [23, 3], we propose this year two tasks on visual question answering (VQA) and visual Question Generation (VQG) [2]. For the VQA task, given a radiology image accompanied with a clinically relevant question, participating systems are tasked with answering the question based on the visual content, while for the VQG task, given a radiology image, participating systems are tasked with generating relevant questions based on the visual content;
  - *Tuberculosis*: This is the fourth 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 a TB severity score based on the automatic analysis of lung CT scans and clinically relevant meta-data. Being able to generate this automatic analysis from the image data allows to limit laboratory analyses to determine the TB stage. 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, we decided to concentrate on the automated CT lung-based report generation task and labels include presence of TB lesions in general, presence of pleurisy and caverns in particular [31];
  - *Caption*: This is the fourth edition of the task in this format, however, it is based on previous medical tasks. The proposed task is to automatically predict UMLS (Unified Medical Language System®) concepts, which is the first step towards automatic medical image semantic tagging. These relevant UMLS® concepts can be further adopted for several medical

[23] `https://visualqa.org/`

imaging tasks such as image captioning, multi-modal image classification and image retrieval. There is a considerable need for automatic mapping of visual information to textual content, as the interpretation of knowledge from medical images is time-consuming. In view of better-structured medical reports, the more information and image characteristics known, the more efficient are the radiologist regarding interpretation. Based on the lessons learned in previous years [26, 15, 27, 37], this year [38] the task focuses on detecting UMLS® concepts in radiology images including a more diverse wealth of imaging modality information.

- **ImageCLEFlifelog**. This is the fourth edition of the task. The increasingly wide range of personal devices, such as smartphones, video cameras as well as wearable devices allow capturing pictures, videos, and audio clips for every moment of our lives are becoming available. Considering the huge volume of data created, there is a need for systems that can automatically analyse the data in order to categorize, summarize and also query to retrieve the information the user may need. This year edition of the task comes with new, enriched data, focused on daily living activities and the chronological order of the moments. Two tasks are proposed: lifelog moment retrieval (LMRT) requiring participants to retrieve a number of specific predefined activities in a lifelogger's life, and sport performance lifelog (SPLL) requiring participants to predict the expected performance (e.g., estimated finishing time) for an athlete who trained for a sport event [35].
- **ImageCLEFcoral**. The increasing use of structure-from-motion photogrammetry for modelling large-scale environments from action cameras has driven the next generation of visualization techniques. 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. Last year was the first time a coral annotation task formed part of ImageCLEF [7]. Participants' entries showed that some level of automatically annotating corals and benthic substrates was possible, despite this being a difficult task due to the variation of colour, texture and morphology between and within classification types. This year [8], the volume of training data has been increased and there are four subsets of test data ranging in geographical similarity and ecological connectedness to the training data. The intention is to explore how well systems trained on one area of data will perform on data from other geographical regions.
- **ImageCLEFdrawnUI**. This task is new for 2020. Building websites requires a very specific set of skills. Currently, the two main ways to achieve this is either by using a visual website builder or by programming. Both approaches have a steep learning curve. Enabling people to create websites by drawing them on a whiteboard or on a piece of paper would make the webpage building process more accessible. In this context, the detection and recognition of hand drawn website UIs task addresses the problem of automatically recognizing the hand drawn objects representing website UIs, which are further used to be translated automatically into website code.

Table 1: Key figures regarding participation in ImageCLEF 2020.

Task	Completed registrations	Groups that subm. results	Submitted runs	Submitted working notes
VQ Answering	30	11	62	11
Tuberculosis	21	9	67	9
Caption	23	7	47	7
Lifelog	12	6	48	6
Coral	15	4	53	4
DrawnUI	14	3	18	3
Overall	115	40	295	40

To participate in the evaluation campaign, the research groups had to register by following the instructions on the ImageCLEF 2020 web page[24]. To ease the overall management of the campaign, in 2020 the challenge was organized through the AIcrowd platform[25]. 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 2020, 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. Given the previous ascending trend, we estimate that this drop is mostly due to the outbreak of the COVID-19 pandemic and lock-down, started during the registration time and continued till the end of the challenge. This triggered a significant perturbation of the tasks. Although additional time was granted, the final participation is lower. Nevertheless, we see a significant improvement in the involvement of the teams and success ratio, which is more important that the sole high participation. The number of teams registering is less than half of as in 2019, however, the number of groups submitting results was not proportionally reduced, and the success ratio, i.e., the number of teams completing the tasks reported to the number of teams completing the registration, is higher, i.e., 35%, compared to 27% for 2019, and 23% for 2018.

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

[24] `https://www.imageclef.org/2020/`
[25] `https://www.aicrowd.com/`

VQA [2], Tuberculosis [31], and Caption [38], ImageCLEFlifelog [35], ImageCLEFcoral [8], and ImageCLEFdrawnUI [19].

3 The Visual Question Answering Task

Visual Question Answering is an exciting problem that combines natural language processing and computer vision 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 [23], and continued the task in 2019 [3]. Following the strong engagement from the research community in both 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 2020) [2] with an enhanced focus on answering questions about abnormalities from the visual content of associated radiology images. Furthermore, we introduced an additional task this year, visual question generation (VQG), consisting in generating relevant natural language questions about radiology images based on their visual content.

3.1 Task Setup

For the visual question answering task, similar to 2019, given a radiology medical image accompanied by a clinically relevant question, participating systems in VQA-Med 2020 were tasked with answering the question based on the visual image content. In VQA-Med 2020, we specifically focused on questions about abnormality (e.g., "what is most alarming about this ultrasound image?"), which can be answered from the image content without requiring additional medical knowledge or domain-specific inference. Additionally, the visual question generation (VQG) task was introduced for the first time in this third edition of the VQA-Med challenge. This task required participants to generate relevant natural language questions about radiology images using their visual content.

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[26] 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. Examples of the selected diagnosis methods include: CT/MRI imaging,

[26] `https://medpix.nlm.nih.gov/`

angiography, characteristic imaging appearance, radiographs, imaging features, ultrasound, and diagnostic radiology. Finally, we considered the most frequent abnormality question categories to create the data set, which included a training set of 4,000 radiology images with 4,000 Question-Answer (QA) pairs, a validation set of 500 radiology images with 500 QA pairs, and a test set of 500 radiology images with 500 questions. To further ensure the quality of the data, the test set was manually validated by a medical doctor. The participants were also encouraged to utilize VQA-Med-2019 data set as additional training data.

For the visual question generation task, we automatically constructed the training, validation, and test sets in a similar fashion by using a separate collection of radiology images and their associated captions. We semi-automatically generated questions from the image captions first by using a rule-based sentence-to-question generation approach[27], and then, three annotators manually curated the list of question-answer pairs by removing or editing the noises related to grammatical inconsistencies. The final curated corpus for the VQG task was comprised of 780 radiology images with 2,156 associated questions (and answers) for training, 141 radiology images with 164 questions for validation, and 80 radiology images for testing. For more details, please refer to [2].

3.3 Participating Groups and Submitted Runs

Out of 47 online registrations, 30 participants submitted signed end user agreement forms. Finally, 11 groups submitted a total of 49 successful runs for the VQA task, while 3 groups submitted a total of 13 successful runs for the VQG task, indicating a notable interest in the VQA-Med 2020 challenge. Table 2 and Table 3 give an overview of all participants and the number of submitted runs (please note that were allowed only 5 runs per team).

3.4 Results

Similar to the evaluation setup of the VQA-Med 2019 challenge [3], the evaluation of the participant systems for the VQA task in the VQA-Med 2020 challenge is also conducted based on two primary metrics: accuracy and BLEU. We used an adapted version of accuracy from the general domain VQA[28] 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 [36] 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 [3]. The BLEU 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. The overall results of the participating systems are presented in Table 4

[27] http://www.cs.cmu.edu/~ark/mheilman/questions/

[28] https://visualqa.org/evaluation.html

Table 2: Participating groups in the VQA-Med 2020 VQA task.

Team	*Institution*	# *Valid Runs*
bumjun_jung	Machine Intelligence Lab, University of Tokyo (Japan)	5
dhruv_sharma	Virginia Tech (USA)	1
going	Sun Yat-Sen University (China)	5
harendrakv	Vadict Innovation Solutions (India)	5
kdevqa	Toyohashi University of Technology (Japan)	4
NLM	National Library of Medicine (USA)	5
sheerin	individual participation (India)	5
Shengyan	Yunnan University (China)	5
TheInceptionTeam	Jordan University of Science and Technology (Jordan)	5
umassmednlp	University of Massachusetts Medical School (USA)	4
z_liao	The Australian Institute for Machine Learning, The University of Adelaide (Australia)	5

Table 3: Participating groups in the VQA-Med 2020 VQG task.

Team	*Institution*	# *Valid Runs*
NLM	National Library of Medicine (USA)	3
TheInceptionTeam	Jordan University of Science and Technology (Jordan)	5
z_liao	The Australian Institute for Machine Learning, The University of Adelaide (Australia)	5

Table 4: Maximum Accuracy and Maximum BLEU Scores for VQA Task (out of each team's submitted runs).

Team	*Accuracy*	*BLEU*
z_liao	0.496	0.542
TheInceptionTeam	0.480	0.511
bumjun_jung	0.466	0.502
going	0.426	0.462
NLM	0.400	0.441
harendrakv	0.378	0.439
Shengyan	0.376	0.412
kdevqa	0.314	0.350
sheerin	0.282	0.330
umassmednlp	0.220	0.340
dhruv_sharma	0.142	0.177

and Table 5 in a descending order of the accuracy and average BLEU scores respectively (the higher the better).

Table 5: Maximum Average BLEU Scores for VQG Task (out of each team's submitted runs).

Team	*Average BLEU*
z_liao	0.348
TheInceptionTeam	0.339
NLM	0.116

3.5 Lessons Learned and Next Steps

Similar to last two years, participants continued to use state-of-the-art deep learning techniques to build their VQA-Med systems for both VQA and VQG tasks [23, 3]. 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. Additionally, the top performing systems first classified the questions into two types: yes/no, and abnormality, then added another multi-class classification framework for abnormality-related question answering, while using the same backbone architecture along with utilizing additional training data, leading to better results.

Analyses of the results in Table 4 suggest that in general, participating systems performed well for the VQA task and achieved better accuracy results relatively compared to last year's results for answering abnormality-related questions [3]. They obtained slightly lower BLEU scores as we focused on only abnormality questions this year that are generally complex than modality, plane, or organ category questions given in the last year. Overall, the VQA task results obtained this year entail the robustness of the provided data set compared to last year's task due to the enhanced focus on the abnormality-related questions for corpus creation. For the VQG task, results in Table 5 suggest that the task was comparatively challenging than the VQA task as the systems achieved lower BLEU scores. As BLEU is not the ideal metric to semantically compare the generated questions with the ground-truth questions, this could also urge the necessity of an embedding-based similarity metric to be explored in the future edition of this task. We would like to also expand the VQG corpus with more images and questions to enable effective development of learning models.

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 [49]. 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 important. In particular, detection of the TB type and presence of different lesion types are important real-world tasks.

In the previous editions of this task, setup evolved from year to year. In the first two editions of this task [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 third edition, SVR subtask was included again for the updated data set, and a new subtask based on providing an automatic report (CT Report) for the TB case was added [16].

In this year's edition, we decided to skip the SVR subtask and concentrate on the automated CT report generation task, since it has an important outcome that can have a major impact in the real-world clinical routines. To make the task both more attractive for participants and practically valuable, this year's report generation was lung-based rather than CT-based, which means the labels for left and right lungs were provided independently. The set of target labels in the CT Report was updated in accordance with the opinion of medical experts.

4.1 Task Setup

In this task, participants had to generate automatic lung-wise reports based on the CT image data. Each report should include the probability scores (ranging from 0 to 1) for each of the three labels and for each of the lungs. Two labels indicated the presence of a specific lesion in the lung - caverns and pleurisy, the third label indicated that the lung is affected by any lesion (not limited to the mentioned two).

The resulting list of entries for each CT included six entries: "left lung affected", "right lung affected", "caverns in the left lung", "caverns in the right lung", "pleurisy in the left lung", "pleurisy in the right lung".

4.2 Data Set

In this edition, the data set containing chest CT scans of 403 TB patients was used, divided into 283 patients for training and 120 for testing. For all patients, we provided 3D CT images with an image size per slice of 512×512 pixels and a variable number of slices (the median number was 128).

For all patients, we provided two versions of automatically extracted masks of the lungs obtained using methods described in [14, 32]. The first version of segmentation was retrieved using the same technique as the previous years and

Table 6: Results obtained by the participants of the task. Only the best run of each participant is reported here.

Group name	*Run ID*	*Mean AUC*	*Min AUC*	*Run rank*
SenticLab.UAIC	68148	0.924	0.885	1
SDVA-UCSD	67950	0.875	0.811	6
chejiao	68118	0.791	0.682	16
CompElecEngCU	67732	0.767	0.733	21
KDE-lab	60707	0.753	0.698	28
FAST_NU_DS	67947	0.705	0.644	37
uaic2020	68081	0.659	0.562	40
JBTTM	67681	0.601	0.432	49
sztaki_dsd	68061	0.595	0.546	50

provides accurate masks, but it tends to miss large abnormal regions of lungs in the most severe TB cases. The second version of segmentation was retrieved using a non-rigid image registration scheme, which on the contrary provides more rough bounds, but behaves more stable in terms of including lesion areas.

4.3 Participating Groups and Submitted Runs

In 2020, 9 groups from 8 countries submitted at least one run. Similar to the previous editions, each group could submit up to 10 runs. 67 runs were submitted in total. The trend toward using convolutional neural networks (CNNs) is stronger again. Last year, 10 out of the 12 groups used CNNs at least in one of their attempts, and this year all groups used CNNs in some way. Several groups tried a few different methods during their experiments, all reported approaches are listed below.

The majority of participants (six groups) used variations of the projection-based approach. These groups extracted axial, coronal, and sagittal projections from the CT image and executed further analysis using 2D CNNs. Different CNN architectures and model training tweaks were used. Two groups also used conventional methods like SVM or handcrafted features in addition to 2D CNNs for projection analysis. Four groups tried 3D CNN for direct analysis of the CT volumetric data. Two groups used per slice analysis, and one of the groups performed additional manual adaptation of lung-based labeling to slice-based labeling. All participants used different techniques for artificial data set enlargement and a few pre-processing steps, such as resizing, normalization, slice filtering or concatenations etc.

4.4 Results

The task was evaluated as a multi-binary classification problem and measured using Area Under the ROC Curve (AUC) metric. AUC was calculated over the 3 target labels ("caverns", "pleurisy", "affected") in a lung-wise manner. The ranking of this task is done first by average AUC and then by min AUC. Table 6

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 [31].

4.5 Lessons Learned and Next Steps

The results obtained in the task improved with respect to the similar CTR subtask presented in the 2019 edition. SenticLab.UAIC group achieved 0.92 mean AUC, which is a significant improvement compared to 0.80 achieved last year by UIIP_BioMed. The group used per-slice analysis, which required some manual pre-processing of training data to utilize per-slice affection labeling. The second-ranked, SDVA-UCSD group, also overcome last year's top result with a score of 0.88 achieved using 3D CNN. Groups that participated in both editions demonstrated improvements over last year results. Only one group applied differing techniques for each finding, the others used a single approach to detect each of the CT-findings in a multi-binary classification setup.

Overall improvement of results, appearing of new more efficient approaches, variability in network architectures and training schemes, suggests that future development and extension of the proposed task is reasonable and may introduce new valuable results. Possible updates for future editions should consider: (i) extending the number of lesion classes; (ii) inclusion of lesion location information, up to switching from binary classification to a detection task.

5 The Caption Task

A large amount of data found in hospital information systems, including radiology reports are stored as free-text. This poses certain problems, as some of these medical narratives are written differently with respect to grammar, acronyms, abbreviations, transcription errors and misspellings. The virtuosity to search through such unstructured database systems and retrieve relevant information is demanding and labour-intensive, hence developing standardized semantic tagging for such stored data is crucial.

The caption task was first proposed as part of the ImageCLEFmedical [27] in 2016. In 2017 and 2018 [15, 26] the ImageCLEFcaption task comprised two subtasks: concept detection and caption prediction. In 2019 [37], the task concentrated on extracting Unified Medical Language System® (UMLS) Concept Unique Identifiers (CUIs) [5] from radiology images. These automatically predicted concepts enable perceivable order for unlabeled and unstructured radiology images and for data sets lacking text information, as multi-modal approaches prove to obtain better results regarding image classification [40].

In 2020, additional label information is included. For each images in the data set, the imaging modality technique is distributed. This extra information can be adopted for pre-filtering and fine-tuning approaches.

Table 7: Explorative analysis on data distribution ImageCLEFmed 2020 Concept Detection Task [38].

Imaging technique	*Train*	*Validation*	*Test*	*Sum*
Angiography	4,713	1,132	325	6,170
Combined Modalities	487	73	49	609
Computer Tomography	20,031	4,992	1,140	26,163
Magnetic Resonance	11,447	2,848	562	14,857
PET	502	74	38	614
Ultrasound	8629	2,134	502	11,265
X-Ray	18,944	4,717	918	24,579
Sum	65,753	15,970	3534	84,257

5.1 Task Setup

The ImageCLEFmed Caption 2020 [38] follows the format of the ImageCLEFmed caption 2019 [37], as well as the concept detection subtask running as part of the ImageCLEFcaption task in 2017 [15] and 2018 [26]. As in all three previous editions, participating teams are tasked with predicting Unified Medical Language System® (UMLS) Concept Unique Identifiers (CUIs) [5] based on the visual image representation in a given image.

In 2017 and 2018, all images commonly found in biomedical literature, were distributed. However in 2019, the focus was reduced to solely radiology images, without targeting any specific disease or anatomic structure. This led to more focused semantic scope of UMLS Concepts that were to be predicted. In 2020, the focus is still on radiology images. Additional information regarding the imaging modality was included. This extra label knowledge can be adopted for certain pre-processing steps, as well as for fine-tuning the models.

The performance of the participating teams was evaluated using the balanced precision and recall trade-off in terms of F1-scores, as in the three previous years. This was measured per image and averaged across all test images and computed with the default implementation of the Python scikit-learn (v0.17.1-2) library.

5.2 Data Set

The training and validation sets distributed are a extension of the Radiology Objects in COntext (ROCO) data set [39]. The training set include 64,753 images and the validation set has 15,970 images. Both sets are associated with 3,047 concepts. All images distributed originate from biomedical journal articles extracted from the PubMed Central® (PMC)[29] repository [43].

For the concept detection evaluation, the test set containing 3,534 images was distributed. This test set does not originate from the ROCO data set and was created using the same procedures applied for the creation of ROCO. It has images from PubMed Central® articles archived between 02.2019 - 02.2020,

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

hence containing no overlap with previous editions. The maximum number of concepts per image varies between 140, 142 and 95 for the training, validation and training sets, respectively. The original imaging technique used for acquiring each image is added as extra label information and the distribution across the training, validation and test set is displayed in Table 7. All concepts in the ground truth that were used for evaluation, as well as in the validation set, are associated and exist in the training set.

5.3 Participating Groups and Submitted Runs

In the fourth edition of the concept detection task, 23 teams registered and signed the End-User-Agreement license, needed to download the development data. 57 graded runs were submitted for evaluation by 7 teams from the following countries: Germany, Great Britain, India, Greece and United States of America. Each of the group was allowed 10 graded runs and 5 faulty runs altogether. 10 of the submitted runs were faulty and were not used for the official evaluation.

Majority of the participating teams were new to the task. Only one team, the AUEB Natural Language Processing Group, participated for the second time. Similar to 2019 [37], deep learning techniques were broadly adopted for training the concept detection models, as improved accuracy rates have been published in the past year [50]. Many teams incorporated the addition modality information for pre-processing steps, fine-tuning of the models, filtering of concepts and late fusion ensemble approaches. The commonly used approaches adopted by most participating teams are: transfer learning with pre-trained deep learning models such as CheXNet [41] and ImageNet [44] on multi-label classification models, image encoding using convolutional neural networks (CNNs), adversarial auto-encoders and long short-term memory (LSTM) recurrent neural networks (RNNs).

5.4 Results

The binary ground truth vector is compared to the predicted UMLS CUIs. To get a better overview of the submitted runs, the best results for each team are presented in Table 8. An in-depth analysis is presented in [38].

5.5 Lessons Learned and Next Steps

The F1-score improved with respect to the previous three editions, from 0.1583 in ImageCLEF 2017, 0.1108 in ImageCLEF 2018 and 0.2823 in ImageCLEF 2019, to 0.3940 this year. The majority of the participating teams this year were new to the task. The AUEB NLP Group [30] from Athens University of Economics and Business, the only teams with previous participation, achieved the highest ranked F1-score.

The decision made for the ImageCLEFmed Caption 2019 to focus on radiology images proved to go into the right direction. By doing so, noisy concepts

Table 8: Performance of the participating teams in the ImageCLEF 2020 Concept Detection Task. The best run per team is selected. Teams with previous participation in 2019 are marked with an asterix.

Team	*Institution*	*F1 Score*
AUEB NLP Group*	Department of Informatics, Athens University of Economics and Business	0.3940
PwC_Healthcare	PRICEWATERHOUSECOOPERS Service Delivery Center PVT. LTD. India	0.3924
Essex	School of computer Science and Electronic Engineering, University of Essex, United Kingdom	0.3808
IML	Interactive Machine Learning Group, German Research Center for Artificial Intelligence (DFKI)	0.3745
TUC_MC	Technische Unversität Chemnitz	0.3512
Morgan_CS	Morgan State University	0.1673
CSE_SSN	Department of Computer Science and Engineering, SSN College of Engineering, Chennai, India	0.1347

were removed, as the biomedical content contained a wide diversity scope. This led to the reduction in the number of concepts from 111,155 in the previous editions to 5,528 in ImageCLEF 2019, and to 3,047 this year, making the amount manageable. The inclusion of imaging modality was adopted by all teams at several model creating steps, which shows to be supportive towards improving the prediction models. Challenging for all teams however, is the imbalance in the concept distribution and imaging modality over the images.

For future improvements, as the UMLS CUIs were extracted from the original PubMed figure captions, it is intended to manually evaluate the clinical relevance content. The natural language captions contain some parts that have important context relation to the published article and not necessarily medical semantic information. By manually screening the extracted CUIs, a data set with expressive and suitable content will be generated, leading to robust concept prediction models that can be incorporated in clinical routine.

6 ImageCLEFlifelog

The goal of the ImageCLEFlifelog 2020 is to continue to promote research in lifelogging as an application supporting human memory and well-being. This year, the ImageCLEFlifelog task is again divided into two sub-tasks: Lifelog Moment Retrieval (LMRT) and Sport Performance Lifelog (SPLL). The core task of Lifelog is LMRT, which has the same format as of previous editions but with a large-scaled data set and different test topics to measure the retrieval performance of participants' system. Again, the LMRT task mainly focuses on images which means that participants need to retrieve photos as the evidence

of relevant moments for some predefined queries. The evaluation metrics are unchanged, which use precision, cluster recall and f1-score for top-10 retrieved results. These metrics require participants to diversify their results while still retrieving the correct moments. The data used in LMRT task is the merging version of three previous NTCIR challenges in three years: 2016 [20], 2017 [21], and 2019 [22]. It was collected using many wearable devices to capture daily life activities, moments, well-being status and current locations of the lifelogger passively and continuously for years. The data contains five main types: multimedia contents, biometrics data, location and GPS, visual concepts and annotations, and human activity information.

The Sport Performance Lifelog (SPLL) is a new task in 2020. The data is totally different from and independent of the data used in the LMRT task. The aim of SPLL is to monitor the change of both well-being status and improvement during the training process of 16 people for a sport event. In particular, participants are required to predict the expected performance of these people in different measurements after the training. This yields three subtasks as follows:

- *Subtask 1*: Predict the change in running speed given by the change in seconds used per km (kilometer speed) from the initial run to the run at the end of the reporting period.
- *Subtask 2*: Predict the change in weight since the beginning of the reporting period to the end of the reporting period in kilos.
- *Subtask 3*: Predict the change in weight from the beginning of February to the end of the reporting period in kilos using the images.

6.1 Task Setup

The ImageCLEFlifelog 2020 proposes two tasks which are *Lifelog Moment Retrieval (LMRT)* and *Sport Performance Lifelog (SPLL)*. The LMRT task has the same requirements and evaluation methodology as the ones of three previous editions but with brand-new topics and different data set structure. Particularly, in this task, participants are required to retrieve moments which are relevant to a predefined topic. The moments are defined as "semantic events or activities that happened through out the day" [12]. For instance, participants should find the images of the relevant moments for the topic "Find the moments that the lifelogger was looking at items in a toy shop". To achieve full-score of each query, participants need to pay attention not only on the precision of the top-10 retrieved results but they should also re-arrange them to increase the diversification of the selected moments with respect to the narrative of each topic.

The ground-truth of this task was manually created. The SPLL task is a new task with the aim of predicting the expected performance (weight change, running speed improvement) of 16 people who trained for a sport event. For this task, there are two evaluation metrics to rank the submissions of participants which are accuracy of the change (primary score) and absolute difference between the actual change and the predicted one (secondary score). While the primary score is ranked in descending order, the secondary score is arranged in ascending

Table 9: Statistics of the ImageCLEFlifelog 2020 LMRT data

Characters	*Size*
Number of Lifeloggers	1
Number of Days	114 days
Size of the Collection	37.1 GB
Number of Images	191,439 images
Number of Locations	166 semantic locations
Number of LMRT Dev Queries	10 queries
Number of LMRT Test Queries	10 queries

order. If there is a draw in the primary score, the secondary score is considered to rank the teams.

6.2 Data Set

LMRT Task — The data is a large-scaled collection of multimodal lifelog data gathered from 114 days of three different years in one lifelogger's life. It was a merging data from three previous NTCIR challenges: NTCIR-12, NTCIR-13, and NTCIR-14. The statistics of the LMRT 2020 dataset is demonstrated in Table 9. In general, the data can be divided into five main types with some similar features as in previous editions including:

- *Multimedia Content* — Non-annotated egocentric images captured passively from OMG Autographer and Narrative Clip worn by the lifelogger for 16-18 hours a day. The total number of images per day ranges from 1,500 to 2,500.
- *Biometrics Data* — Using the FitBit fitness trackers[30], the lifeloggers gathered 24×7 heart rate, calorie burn and steps.
- *Semantic Locations and GPS* — GPS data with 166 semantic locations are captured using Moves app and smartphones. In addition, time zones are inferred using the GPS data, which is essential to convert the time in different wearable devices into the same format and time zone.
- *Human Activity Data* — The daily activities of the lifeloggers were captured in terms of physical activities (e.g., walking, running, transporting) from the Moves app[31].
- *Visual Concepts and Annotations* — The wearable camera images were annotated with the outputs of a visual concept detector, which provided three types of outputs (Attributes, Categories and Concepts). Two visual concepts which include attributes and categories of the place in the image are extracted using PlacesCNN [51]. The remaining one is the detected object category and its bounding box extracted by using Mask R-CNN [25] trained on MSCOCO data set [33].

[30] Fitbit Fitness Tracker (FitBit Versa) - `https://www.fitbit.com/`
[31] Moves App for Android and iOS - `http://www.moves-app.com/`

Table 10: Official results of the ImageCLEFlifelog 2020 LMRT task.

Team	*Run*	*P@10*	*CR@10*	*F1@10*	*Team*	*Run*	*P@10*	*CR@10*	*F1@10*
Organiser	RUN1*	0.19	0.31	0.21	HCMUS	RUN1	0.79	0.73	0.72
	RUN2*	0.23	**0.44**	0.27		RUN2	0.78	0.73	0.72
	RUN3*	**0.36**	0.38	**0.32**		RUN3	0.79	0.69	0.71
REGIM	RUN1	0.04	0.08	0.05		RUN4	0.80	0.74	0.74
	RUN2	0.16	0.22	0.17		RUN5	0.81	0.77	0.75
	RUN3	0.17	**0.24**	**0.19**		RUN6	0.81	0.79	0.77
	RUN4	0.00	0.00	0.00		RUN7	0.82	**0.81**	0.79
	RUN5	**0.19**	0.16	0.16		RUN8	0.77	0.76	0.74
	RUN6	0.03	0.05	0.04		RUN9	0.85	**0.81**	**0.81**
	RUN7	0.17	**0.24**	**0.19**		RUN10	**0.86**	**0.81**	**0.81**
UATP	RUN1	0.02	0.07	0.03	BIDAL	RUN1	0.69	0.68	0.65
	RUN2	0.02	0.07	0.03		RUN2	0.68	0.63	0.58
	RUN3	**0.50**	**0.58**	**0.52**		RUN3	0.68	0.69	0.65
DCU-DDTeam	RUN1	0.07	0.13	0.09		RUN4	0.70	0.69	0.66
	RUN2	0.22	0.39	0.25		RUN5	0.72	0.69	0.66
	RUN3	0.44	**0.63**	0.41		RUN6	0.73	0.69	0.67
	RUN4	**0.58**	0.53	**0.48**		RUN7	0.75	0.65	0.64
	RUN5	0.16	0.36	0.21		RUN8	0.73	0.69	0.67
						RUN9	0.73	**0.70**	**0.69**
						RUN10	**0.74**	**0.70**	**0.69**

Notes: *submissions from the organizer teams are just for reference.

Table 11: Official Results of the ImageCLEFlifelog 2020 SPLL Task.

Team	*Run*	*Primary score*	*Secondary score*
Organiser	RUN1*	0.47	313.30
	RUN2*	0.41	203.10
BIDAL	RUN1	0.77	306.90
	RUN2	0.52	309.10
	RUN3	0.59	254.70
	RUN4	0.59	372.60
	RUN5	0.53	375.20
	RUN6	0.65	319.60
	RUN7	0.71	250.20
	RUN8	**0.82**	245.60
	RUN9	**0.82**	128.00
	RUN10	0.65	**112.00**

Notes: *submissions from the organizer teams are just for reference.

SPLL Task — The data is collected from 16 people during their training for a 5 km run. Fitbit Versa 2 sport watch is used to capture the heart rate and calories information while the PMSYS system is employed to collect information about subjective wellness, training load, and injury data. Moreover, information such as meals, drinks, medication, etc. is also collected via Google Forms. The data contain information about daily sleeping patterns, daily heart rate, sport activities, logs of food consumed during the training period from at least 2 participants and self reported data like mode, stress, fatigue, readiness to train and other measurements also used for professional soccer teams [46]. For this

task, we have the data approved by the Norwegian Center for Research Data with proper copyright and ethical approval to release.

6.3 Participating Groups and Submitted Runs

We received in total 50 valid submissions from 6 teams. These include 38 valid submissions for LMRT and 12 valid ones for SPLL. Their submissions and the results are summarised in Tables 10 and 11. A detailed analysis of the results is presented in the task overview paper [35].

6.4 Lessons Learned and Next Steps

For the LMRT task, we learned that most of the approaches are building interactive systems using multi-modal data and extended visual concepts to retrieve the relevant moments. One team tried to implement an automatic retrieval system but the results are not as competitive as the interactive ones. We also confirm that visual concepts extracted automatically from different deep networks are extremely useful when creating the indexing system for retrieval. If visual concepts and annotations of visual images are enriched, the interactive retrieval systems can be improved in precision and diversification, significantly. The ImageCLEFlifelog 2020 results are competitive with great improvements compared to previous systems. In this year's challenge, only 6 teams participated in the LMRT task, including an organizer team. We received 50 valid submissions. Each team was allowed to submit up to 10 runs. For the LMRT task, among five teams which participated in ImageCLEFlifelog 2019 (including the organizer team), four teams managed to obtain better results with the highest F1-score up to 0.81. The mean (SD) increase of final F1-score from these five teams is 0.25 (0.18). The new team from Dublin City University also managed to achieve the 4th rank with a 0.48 F1-score. For the SPLL task, as the task is new, only one team from The Big Data Analytics Laboratory submitted 10 runs. Their best submission achieves an accuracy of performance change and the absolute difference between the prediction and actual change are 0.82 and 128 respectively, which is a good result.

For the next edition of the LMRT task, we plan to provide better concepts and descriptions of the egocentric images including activities, locations, and visual objects, while still expanding the data set. This year, the submitted results are better, with competitive scores. For the SPLL task, although the number of non-organizer teams participating in the task is only one, results show that the task has potential and should be improved in the next run.

7 The Coral Task

Coral reefs are some of the most biodiverse regions of the ocean, yet are undergoing unprecedented decline through a combination of factors such as climate change, ocean acidification, fertiliser run-off from land and unsustainable fishing

practices [4]. Marine biologists and ecologists want to find ways for those living in the vicinity of reefs to maintain their food supplies [6, 45] without destroying the very reefs on which they depend. It is therefore crucial that they are able to monitor the health of reefs and the classes of structure they contain — but currently, they have to do this manually.

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. These images can be used within a structure-from-motion framework to reconstruct 3D point clouds of large regions of reef; and while these point clouds produce information of interest to marine biologists and ecologists on reef complexity, determining benthic substrate 3D point clouds is a significantly more difficult task than from the 2D images. That is why ImageCLEFcoral task encourages vision researchers to develop automatic ways of performing the annotation, yielding information that helps the marine researchers monitor coral reefs.

7.1 Task Setup

Following the success of the first edition of the ImageCLEFcoral task [7], in 2020 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. The images were captured using an underwater multi-camera system developed at the Marine Technology Research Unit at the University of Essex (MTRU), UK[32].

The ground truth annotations of the training and test sets were made by a combination of marine biology MSc students at Essex and experienced researchers. All annotations were double checked by an experienced coral reef researcher. The annotations were performed using a web-based tool, initially developed in a collaborative project with London-based company Filament Ltd and subsequently extended by one of the organisers. This tool was designed to be simple to learn, quick to use and, almost uniquely, allowing many people to work concurrently [7].

The overall task comprises two sub-tasks. In the first, the annotation is a bounding box, with sides parallel to the edges of the image, around identified features. In the second, participants submit 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.

As in the first edition, algorithmic performance is evaluated on the unseen test data using the popular intersection over union metric from the PASCAL VOC[33] exercise. This computes the area of intersection of the output of an algorithm and the corresponding ground truth, normalizing that by the area of their union to ensure its maximum value is bounded.

[32] `https://essexnlip.uk/marine-technology-research-unit/`
[33] `http://host.robots.ox.ac.uk/pascal/VOC/`

7.2 Data Set

The images used in both editions of the 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 (MTRU) at the University of Essex.

The data set comprises 440 human-annotated training images, with 12,082 substrates, from the Wakatobi Marine Reserve, Indonesia; this is the complete training and test sets used in the ImageCLEFcoral 2019 task. The test set comprises a further 400 test images, with 8,640 substrates annotated, from four geographical regions, 100 images per subset:

1. Wakatobi Marine Reserve, Indonesia – the same location as the training images;
2. Spermonde archipelago, Indonesia – geographically similar location to the training set with a similar benthic composition;
3. Seychelles, Indian Ocean – geographically distinct but ecologically connected coral reef;
4. Dominica, Caribbean – geographically and ecologically distinct rocky reef.

The images are part of a monitoring collection and therefore many have a tape measure running through a portion of the image. As in 2019, the data set comprises an area of underwater terrain. Many images contain the same ground features captured from different viewpoints. Each image contains some of the same thirteen types of benthic substrates as in 2019, namely hard coral — branching, submassive, boulder, encrusting, table, foliose, mushroom; soft coral — gorgonian; sponge — barrel; fire coral — millepora; algae — macro or leaves.

The test set from the same area as the training set will give an indication as to how well a submitted algorithm can localise and classify marine substrate, i.e., the maximum performance. We hypothesise that performance will deteriorate with other test subsets as the composition, morphology and identifying features of the substrate change and exhibit less similarity with the training data.

7.3 Participating Groups and Submitted Runs

In this second edition of the ImageCLEFcoral task, 15 teams registered, of which 4 teams submitted 53 runs. Teams were limited to submit 10 runs per task. The majority of submissions use deep neural networks, generally convolutional ones. For example, some of the submissions were performed using a R-CNN with ResNet 101 backbone, with 30 epochs of training on the full training data set. Data augmentation (using flips, random crops and contrast, hue, saturation and brightness adjustments) was employed, then averaging over the top five models. Others used different types of networks, so there is a good comparison of different approaches. However, at least one submission is based on k-nearest neighbours, perhaps one of the longest-standing clustering techniques, with statistical features. It is also interesting that most training seemed to use sub-sampled images, though the image size varied from group to group and run to run.

Table 12: Coral reef image annotation and localisation performance in terms of $MAP0.5IoU$, and $MAP0IoU$. The best run per team in terms of $MAP0.5IoU$ is selected.

Run id	*Team*	$MAP0.5IoU$	$MAP0IoU$
68143	FAV ZČU PiVa	0.582	0.853
67539	FAV ZČU CV	0.49	0.822
68181	FHD	0.457	0.775
68201	HHU	0.392	0.806

Table 13: Pixel-wise coral reef parsing performance in terms of $MAP0.5IoU$, and $MAP0IoU$. The best run per team in terms of $MAP0.5IoU$ is selected.

Run id	*Team*	$MAP0.5IoU$	$MAP0IoU$
67864	FAV ZČU PiVa	0.678	0.845
68190	FHD	0.474	0.715
67620	FAV ZČU CV	0.304	0.602

7.4 Results

As in 2019, the task was evaluated using the PASCAL VOC style metric of intersection over union (IoU), as discussed above. The evaluation was carried out using two measures: MAP 0.5 IoU — the localised mean average precision (MAP) for each submitted method for using the performance measure of IoU >=0.5 of the ground truth; and MAP 0 IoU — the image annotation average for each method with success if the concept is simply detected in the image without any localisation. Tables 12 and 13 present the best runs per team in terms of $MAP0.5IoU$. The complete overview of the results can be found in [8], including the results on each of the geographical locations in the test set and the accuracy per benthic substrate type.

The $MAP0.5IoU$ score from FAV ZČU PiVa of 0.582 over the entire test set is excellent, bearing in mind both the difficulty of the problem and the number of classes involved. There is a significant margin before the best run from the second-placed team, FAV ZČU CV, and the other teams' best submissions, which are quite closely spaced. FAV ZČU PiVa also made the best-ranked submission for $MAP0IoU$ but the other teams' best-scoring submissions are much closer to this. However, when one compares the accuracy obtained by these runs, the best-scoring one for $MAP0.5IoU$ does not yield the highest accuracy of all the submissions. Clearly then, there is some inconsistency in the evaluation measures employed — and this is perhaps more of an indication that performance evaluation should be revisited.

It is interesting to review the scores obtained from the four categories of test data. For the first three geographic regions, performance is quite similar, which is good, but performance drops off for other geographic regions. Although not at all unexpected, this shows how difficult it will be to develop a system for marine biologists who can take it to any part of the world and preserve its accuracy.

The results of the pixel-wise parsing task, in which teams attempt to identify the boundaries of features rather than their bounding boxes are shown in Table 13. The $MAP0.5IoU$ score of the best-placed team, FAV ZČU PiVa, is actually higher than for the first task, showing that their approach is able to identify the boundaries of the image features somewhat better than those of the other teams. This makes the performance gap between first- and second-placed teams somewhat larger than for the first task. Again, the best-scoring run in terms of $MAP0.5IoU$ is not the best in terms of accuracy.

7.5 Lessons Learned and Next Steps

The results of the 2020 coral exercise are interesting and demonstrate how well modern deep neural networks in particular are at a range of problems. For the coral exercise, the authors regard a performance approaching 70% for a 13-class problem as excellent. The results show that the best pixel-wise parsing technique outperformed the best bounding box one, suggesting that future exercises should concentrate on pixel-wise parsing. There are always difficulties with overlapping bounding boxes and other types of feature in the background of bounding boxes which together reduce the value of that type of annotation.

An in-depth analysis of the test results is not presented here but it is clear that there are genuine performance differences between the four geographical categories of test images described above. This is an immensely important practical problem for coral annotation, and also for vision systems in general. We anticipate future coral annotation tasks will explore ways to overcome this difficulty. Close examination of the ground truth annotations for the pixel-parsing task shows that annotators tend to place the bounding polygons just outside the boundaries of the features being annotated. We are considering producing other annotations that lie within feature boundaries and encourage teams in a future exercise to train the same architecture with both, then see which works best. That would give us the opportunity to learn something about how annotations should be produced.

The fact that different measures rank-order the different runs differently does not come as a surprise but does show how difficult it is to devise a simple measure that encapsulates performance well. There is clearly research to be done in this regard. Although there are performance differences between the runs, there is no indication as to whether they are statistically significant or not. This analysis can be done however, and we shall explore this as future work. Bearing in mind the point made about performance measures in the previous paragraph, it will be especially interesting to ascertain whether different performance measure yield statistically-significant but inconsistent results.

8 The DrawnUI Task

User interfaces (UIs) represent the medium where interactions between humans and computers occur. The increasing dependence on web and mobile applications has led many enterprises to prioritize the development of UIs in an effort to

improve the overall user experience. Currently, the performance of any modern digital product is strongly correlated to the quality and usability of its user interface. However, building one poses a complex problem, requiring the interaction of multiple specialists, each with their own domain-specific knowledge. The process becomes increasingly error prone as the number of workers increases. Moreover, UI experts are in limited supply too, with 22 million developers in the whole world[34], among which only 10 million are estimated to also be JavaScript UI developers[35].

Recently, the use of machine learning to facilitate the creation of UIs has been demonstrated as a viable solution. In 2018, pix2code proposed an open-source, machine-learning based approach to generate low fidelity, domain specific languages from screenshots [1]. In the same year, Chen Chunyang et al. [9] created their own data set based on Android applications, providing 185,277 pairs of UI images and GUI skeletons. The data set and code were open-sourced as well.

8.1 Task Setup

The 2020 ImageCLEF DrawnUI task is at its first edition and consists of a single task. The participants are required to develop a computer vision model to predict the type and position (bounding box) of different UI elements in hand-drawn wireframes. The data set is split approximately 75% for training and 25% for testing. During the competition, the submissions were evaluated using the overall precision. In addition, $MAP0.5IoU$ and $R0.5IoU$ were computed after the competition [18].

8.2 Data Set

The task data set consists of 3,000 hand-drawn wireframe images based on 1,000 different templates of mobile and web UIs. Mobile UI templates were manually selected from the RICO data set [13] while web pages UIs were parsed using a custom web parser. Three people were involved in this drawing step, which involved the use of a predefined shape dictionary with 21 different UI elements. This shape dictionary was focused on unambiguous drawing instead of fidelity to the original screenshot in order to facilitate the annotation step. Finally, a last check was performed by a master annotator to ensure consistency.

8.3 Participating Groups and Submitted Runs

14 teams registered and 3 teams from 2 countries submitted 18 runs. Teams were limited to submit 10 runs.

[34] `http://evansdata.com/`
[35] `http://appdevelopermagazine.com/`

Table 14: Participation in the DrawUI 2020 task: the best score from all runs for each team.

Team	*# Runs*	*Overall precision*	*MAP*0.5*IoU*	*R*0.5*IoU*
zip	7	**0.970**	0.755	0.555
CudaMemError1	8	0.950	**0.793**	**0.598**
OG_SouL	3	0.940	0.641	0.501

Table 15: Overall precision, $MAP0.5IoU$, and $R0.5IoU$ for each run. Organizers baseline is marked with an asterix.

Team	*Run ID*	*Method description*	*Overall precision*	*MAP*0.5*IoU*	*R*0.5*IoU*
zip	67816	resnet50 Faster R-CNN, full-size, grayscale	**0.970**	0.582	0.445
zip	68014	inception resnet v2 Faster R-CNN, full-size, merging	0.956	0.693	0.519
zip	68003	inception resnet v2 Faster R-CNN, full-size, grayscale	0.956	0.694	0.520
zip	67814	resnet50 Faster R-CNN, 12MP, grayscale	0.955	0.675	0.517
CudaMemError1	67814	fusiont-3	0.950	0.715	0.556
CudaMemError1	67833	obj wise 2	0.950	0.681	0.533
CudaMemError1	67710	resnet101	0.949	0.649	0.505
dimitri.fichou*	67413	baseline: Faster R-CNN, data augmentation	0.947	0.572	0.403
zip	67991	resnet50 Faster R-CNN, full-size, all data	0.944	0.647	0.472
zip	68015	inception resnet v2 Faster R-CNN, full-size, merging	0.941	0.755	0.555
OG_SouL	67391	Transfer Learning using Mask R-CNN pre-trained with COCO	0.940	0.573	0.417
zip	67733	-	0.939	0.687	0.536
CudaMemError1	67722	resnet101	0.934	0.723	0.585
CudaMemError1	67706	-	0.934	**0.793**	**0.598**
CudaMemError1	67829	obj fusion	0.932	0.738	0.556
CudaMemError1	67707	-	0.931	0.792	0.594
CudaMemError1	67831	image wise fusion	0.929	0.791	0.600
OG_SouL	67699	Mask R-CNN, multi-pass inference, grayscale	0.918	0.637	0.501
OG_SouL	67712	Mask R-CNN, multi-pass inference	0.917	0.641	0.496

8.4 Results

The $MAP0.5IoU$ and $R0.5IoU$ scores have been compiled using an adapted version of the COCO data set evaluator[36]. All submissions fared better than expected on this challenge, confirming our assumptions regarding the usage of machine learning in streamlining the process of wireframing. While transferring paper information into its digital counterpart is only one part of the design and implementation process, the high accuracy of the results clearly indicates potential for further extending this challenge to other areas, such as predicting directly the nested UI structure.

[36] `https://github.com/philferriere/cocoapi/`

8.5 Lessons Learned and Next Steps

Each submission used object detection algorithms such as Faster R-CNN [42] or Mask R-CNN [24] with different types of data augmentation and pre-processing. Two teams obtained scores superior to our baseline according to the overall precision. All submissions were superior to our baseline according to the $MAP0.5IoU$ and $R0.5IoU$. Although overall precision is as high as 0.97 and may show the task as more or less solved, this is not the best metric in terms of localization as it does not take into account a high number of false negatives or poor results on the rare classes of the data set. Mean Average Precision and Recall are more appropriate metrics in this case. In this case, best results are significantly lower, e.g., 0.79 for MAP, meaning that there is still room for improvement.

As future challenges, for the next edition of this task, we plan to tackle two different problems: (i) predicting the nested structure of the UI based on either the wireframe or the bounding boxes. The current task was focused on absolute positioning but the final UI is built using relative positioning, to handle responsiveness. This task is particularly challenging and could be solved with a mix of computer vision and natural language processing; (ii) object detection from screenshots instead of drawings. Mockups are often used by designers as a medium to hand off their designs to the developers. It is possible to parse the web to obtain a similar data set to the one from DrawnUI 2020 by analysing the DOM trees and capturing screenshots. However, due to the nature of the world wide web, compiling a clean data set will represent a challenge. Instead, we propose to only manually clean the test set and let the participant train using a large, raw data set. The challenge here will be close to real life data set, where the data contains numerous errors.

9 Conclusions

This paper presents a general overview of the activities and outcomes of the ImageCLEF 2020 evaluation campaign. Four tasks were organised, covering challenges in the medical domain (visual question answering and visual question generation, tuberculosis prediction, and caption analysis), lifelogging (daily activity understanding, retrieval and summarization), nature (segmenting and labeling collections of coral images), and Internet (identifying hand-drawn website user interface components). Despite the outbreak of the COVID-19 pandemic and lock-down during the benchmark, 115 teams registered, 40 teams completed the tasks and submitted over 295 runs. Although the number of registrations was lower than in 2019, the success rate of the participants increased with over 8 percentage points.

Most of the proposed solutions evolved around state-of-the-art deep neural network architectures, also for the medical domain. For the visual question answering, most systems leveraged encoder-decoder architectures with various pooling strategies and attention mechanisms. There was a visible improvement in performance compared to previous editions. The visual question generation,

on the other hand, proved to be more challenging. For the tuberculosis prediction task, results also improved compared to previous editions. Best runners employed per-slice analysis involving some manual pre-processing of the training data. Classification is achieved with deep neural networks. For the caption analysis task, all participants embraced the imaging modality in their prediction deep models. A challenge was posed by the imbalance in the concept distribution. However, results improved compared to last year. For the lifelog task, most approaches built interactive systems using multi-modal information and visual concepts for the retrieval. Automated retrieval systems proved to be less competitive. The most reliable information were the visual concepts extracted automatically from the data. The sport performance subtask, although newly introduced, lead to good results. Overall, results also improved compared to last year. For the coral task, pixel-wise parsing outperformed bounding boxing. Also, geographical position of the corals influenced significantly the results. Finally, for the drawn UI task, even in the first edition, systems were able to achieve very high performance in terms of precision (up to 97%) with variation of R-CNNs. The detection problem seems to be solved, however the precise UI localization is not yet that accurate and leaves room for improvement.

ImageCLEF 2020 brought again together an interesting mix of tasks and approaches and we are looking forward to the fruitful discussions at the CLEF 2020 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.

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. Birkeland, C.: Global status of coral reefs: In combination, disturbances and stressors become ratchets. In: World Seas: an Environmental Evaluation, pp. 35–56. Elsevier (2019)
5. 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

6. Brander, L.M., Rehdanz, K., Tol, R.S., Van Beukering, P.J.: The economic impact of ocean acidification on coral reefs. Climate Change Economics **3**(01), 1250002 (2012)
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. 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)
10. 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)
11. Clough, P., Sanderson, M.: The CLEF 2003 cross language image retrieval task. In: Proceedings of the Cross Language Evaluation Forum (CLEF 2003) (2004)
12. Dang-Nguyen, D.T., Piras, L., Riegler, M., Tran, M.T., Zhou, L., Lux, M., Le, T.K., Ninh, V.T., Gurrin, C.: Overview of ImageCLEFlifelog 2019: Solve my life puzzle and Lifelog Moment Retrieval. In: CLEF2019 Working Notes. CEUR Workshop Proceedings, CEUR-WS.org, Lugano, Switzerland (September 09-12 2019)
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. 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

19. 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)
20. Gurrin, C., Joho, H., Hopfgartner, F., Zhou, L., Albatal, R.: Overview of ntcir-12 lifelog task. In: NTCIR (2016)
21. Gurrin, C., Joho, H., Hopfgartner, F., Zhou, L., Gupta, R., Albatal, R., Dang-Nguyen, D.T.: Overview of ntcir-13 lifelog-2 task (2017)
22. Gurrin, C., Joho, H., Hopfgartner, F., Zhou, L., Ninh, V.T., Le, T.K., Albatal, R., Dang-Nguyen, D.T., Healy, G.: Overview of the ntcir-14 lifelog-3 task (2019)
23. 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)
24. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask R-CNN. IEEE Transactions on Pattern Analysis and Machine Intelligence **42**(2), 386–397 (2020). https://doi.org/10.1109/TPAMI.2018.2844175
25. He, K., Gkioxari, G., Dollár, P., Girshick, R.B.: Mask r-cnn. 2017 IEEE International Conference on Computer Vision (ICCV) pp. 2980–2988 (2017)
26. 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)
27. 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)
28. 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)
29. 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)
30. Kougia, V., Pavlopoulos, J., Androusopoulos, I.: Aueb nlp group at imageclefmed caption 2019. In: CLEF2019 Working Notes. CEUR Workshop Proceedings, CEUR-WS.org <http://ceur-ws.org>, Lugano, Switzerland (September 09-12 2019)
31. 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)

32. 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)
33. Lin, T.Y., Maire, M., Belongie, S.J., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. ArXiv **abs/1405.0312** (2014)
34. 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)
35. Ninh, V.T., Le, T.K., Zhou, L., Piras, L., Riegler, M., l Halvorsen, P., Tran, M.T., Lux, M., Gurrin, C., Dang-Nguyen, D.T.: Overview of ImageCLEF Lifelog 2020: Lifelog Moment Retrieval and Sport Performance Lifelog. In: CLEF2020 Working Notes. CEUR Workshop Proceedings, CEUR-WS.org <http://ceur-ws.org>, Thessaloniki, Greece (September 22-25 2020)
36. 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)
37. 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)
38. 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)
39. 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)
40. Pelka, O., Nensa, F., Friedrich, C.M.: Adopting semantic information of grayscale radiographs for image classification and retrieval. In: Proceedings of the 11th International Joint Conference on Biomedical Engineering Systems and Technologies (BIOSTEC 2018) - Volume 2: BIOIMAGING, Funchal, Madeira, Portugal, January 19-21, 2018. pp. 179–187 (2018). https://doi.org/10.5220/0006732301790187
41. Rajpurkar, P., Irvin, J., Zhu, K., Yang, B., Mehta, H., Duan, T., Ding, D.Y., Bagul, A., Langlotz, C., Shpanskaya, K.S., Lungren, M.P., Ng, A.Y.: Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning. CoRR **abs/1711.05225** (2017), `http://arxiv.org/abs/1711.05225`
42. Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. IEEE Transactions on Pattern Analysis and Machine Intelligence **39**(6), 1137–1149 (2017). https://doi.org/10.1109/TPAMI.2016.2577031
43. Roberts, R.J.: PubMed Central: The GenBank of the published literature. Proceedings of the National Academy of Sciences of the United States of America **98**(2), 381–382 (Jan 2001). https://doi.org/10.1073/pnas.98.2.381

44. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A.C., Fei-Fei, L.: Imagenet large scale visual recognition challenge. International Journal of Computer Vision **115**(3), 211–252 (Dec 2015). https://doi.org/10.1007/s11263-015-0816-y
45. Speers, A.E., Besedin, E.Y., Palardy, J.E., Moore, C.: Impacts of climate change and ocean acidification on coral reef fisheries: An integrated ecological–economic model. Ecological economics **128**, 33–43 (2016)
46. Thambawita, V., Hicks, S., Borgli, H., Pettersen, S., Johansen, D., Johansen, H., Kupka, T., Stensland, H., Jha, D., Grønli, T.M., Fredriksen, P.M., Eg, R., Hansen, K., Fagernes, S., Biorn-Hansen, A., Dang Nguyen, D.T., Hammer, H., Jain, R., Riegler, M., Halvorsen, P.: Pmdata: A sports logging dataset (02 2020). https://doi.org/10.31219/osf.io/k2apb
47. 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)
48. 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)
49. World Health Organization, et al.: Global tuberculosis report 2019 (2019)
50. Xu, Y., Mo, T., Feng, Q., Zhong, P., Lai, M., Chang, E.I.: Deep learning of feature representation with multiple instance learning for medical image analysis. In: IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2014, Florence, Italy, May 4-9, 2014. pp. 1626–1630 (2014). https://doi.org/10.1109/ICASSP.2014.6853873
51. Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., Torralba, A.: Places: A 10 million image database for scene recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence (2017)