ImageCLEF 2019: Multimedia Retrieval in Medicine, Lifelogging, Security and Nature

Bogdan Ionescu[1], Henning Müller[2], Renaud Péteri[3], Yashin Dicente Cid[2], Vitali Liauchuk[9], Vassili Kovalev[9], Dzmitri Klimuk[17], Aleh Tarasau[17], Asma Ben Abacha[11], Sadid A. Hasan[10], Vivek Datla[10], Joey Liu[10], Dina Demner-Fushman[11], Duc-Tien Dang-Nguyen[16], Luca Piras[5], Michael Riegler[6], Minh-Triet Tran[7], Mathias Lux[8], Cathal Gurrin[4], Obioma Pelka[12], Christoph M. Friedrich[12], Alba G. Seco de Herrera[13], Narciso Garcia[14], Ergina Kavallieratou[15], Carlos Roberto del Blanco[14], Carlos Cuevas[14], Nikos Vasillopoulos[15], Konstantinos Karampidis[15], Jon Chamberlain[13], Adrian Clark[13], and Antonio Campello[13,18]

[1] University Politehnica of Bucharest, Romania
bogdan.ionescu@upb.ro
[2] University of Applied Sciences Western Switzerland (HES-SO), Switzerland
[3] La Rochelle University, France
[4] Dublin City University, Ireland
[5] Pluribus One & University of Cagliari, Italy
[6] University of Oslo, Norway
[7] University of Science, Vietnam
[8] Klagenfurt University, Austria
[9] Institute for Informatics, Belarus
[10] Philips Research Cambridge, USA
[11] National Library of Medicine, USA
[12] University of Applied Sciences and Arts Dortmund, Germany
[13] University of Essex, UK
[14] E.T.S. Ingenieros Telecomunicación, Spain
[15] University of the Aegean, Greece
[16] University of Bergen, Norway
[17] Republican Research and Practical Centre for Pulmonology and TB, Belarus
[18] Filament, UK

Abstract. This paper presents an overview of the ImageCLEF 2019 lab, organized as part of the Conference and Labs of the Evaluation Forum - CLEF Labs 2019. ImageCLEF is an ongoing evaluation initiative (started 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 2019, the 17th edition of ImageCLEF runs four main tasks: (i) a *medical* task that groups three previous tasks (caption analysis, tuberculosis prediction, and medical visual question answering) with new data, (ii) a *lifelog* task (videos, images and other sources) about daily activities understanding, retrieval and summarization, (iii) a new *security* task addressing the problems of automatically identifying forged content and retrieve hidden information, and (iv) a new *coral* task about segmenting and labeling collections of coral images for 3D modeling. The strong participation, with 235 research groups registering, and 63 submitting over 359 runs, shows an important interest in this benchmark campaign.

Keywords: medical retrieval · life logging retrieval and summarization · file forgery detection · coral image segmentation and classification · ImageCLEF benchmarking · annotated data sets.

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 [9]. It increased its impact with the addition of medical tasks in 2004 [8], 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.

The tasks have changed much over the years but the general objective has always been the same, 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. A detailed analysis of several tasks and the creation of the data sets can be found in [29]. ImageCLEF has shown to have an important impact over the years, already detailed in 2010 [41, 42].

2 Overview of Tasks and Participation

ImageCLEF 2019 consists of four main tasks with the objective of covering a *diverse range* of multimedia retrieval applications, namely: *medicine*, *lifelogging*, *security*, and *nature*. Compared to 2018 [24], 2019 focused on a diversity of tasks [14, 2, 32, 7, 26, 10]. The visual question answering, caption and tuberculosis tasks from 2018 had a sequel and were organized as a specific medical track to foster collaboration. The life logging task also had a follow-up. New in 2019 are the coral and security tasks. Therefore, the 2019 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, 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:
  - *Tuberculosis*: This is the third 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

[19] http://www.imageclef.org/

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;
  - *Visual Question Answering*: This is the second edition of the 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[20] and with ImageCLEF [20, 2], we propose an enhanced and nicely curated enlarged data set. Like last year, given a medical image accompanied with a clinically relevant question, participating systems are tasked with answering the question based on the visual content;
  - *Caption*: This is the third edition of the task in this format, however, it is based on previous medical tasks. The proposed task is the first step towards automatic medical image captioning. Relevant UMLS (Unified Medical Language System®) concepts, that serve as building blocks from which captions can be composed, are to be automatically predicted. 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 [21, 13, 22], this year [32] the task focus on detecting UMLS® concepts in radiology images.
- **ImageCLEFlifelog**. This is the third edition of the task. It is now possible to record, capture, photograph and make a video almost in every moment of our life. Wearable devices have further expanded these possibilities and are able to keep track of all our vital functions: heart rate, burned calories, blood sugar and so on. All these data must be indexed, categorized and it must be possible to retrieve them easily for such applications to become usable. Hence, this task addresses the problems of lifelog data understanding, summarizing and retrieval.
- **ImageCLEFsecurity**. This is the first edition of the task. File Forgery Detection (FFD) is a serious problem concerning digital forensics examiners. Fraud or counterfeits are common causes for altering files. It is also common that anyone who wants to hide any kind of information in plain sight without

[20] https://visualqa.org/

Table 1: Key figures regarding participation in ImageCLEF 2019.

Task	Completed registrations	Groups that subm. results	Submitted runs	Submitted working notes
Tuberculosis	38	13	89	12
VQ Answering	60	17	80	12
Caption	49	11	60	8
Lifelog	17	10	67	10
Security	58	7	43	4
Coral	13	5	20	4
Overall	235	63	359	50

being perceived to use steganography. The objective of the specific task is first to examine if an image was forged, then if it could also hide a text message, and last to retrieve the potential hidden message from the forged stego images.

– **ImageCLEFcoral**. This is the first edition of the task. The increasing use of structure-from-motion photogrammetry for modelling large-scale environments from action cameras attached to drones has driven the next generation of visualisation techniques that can be used in augmented and virtual reality headsets. Advances in automatically annotating images for complexity and benthic composition have been promising. The task [7] aims to automatically identify areas of interest and to label them for monitoring coral reefs.

In order to participate in the evaluation campaign, the research groups had to register by following the instructions on the ImageCLEF 2019 web page[21]. To ease the overall management of the campaign, in 2019 the challenge was again organized through the crowdAI platform[22]. To actually get access to the data sets, the participants were required to submit a signed End User Agreement (EUA). Exception was the security task, for which no data usage agreement was required. Table 1 summarizes the participation in ImageCLEF 2019, 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. This is almost twice as many papers as in 2018. This is due to several factors: (i) in 2019 there were more tasks and sub-tasks and also a diversity of applications, attracting several different communities; (ii) the crowdAI platform facilitates an online registration which is easier than the previous registration system and much more accessible. It

[21] https://www.imageclef.org/2019

[22] https://www.crowdai.org/

provides visibility to a benchmark community outside of the classical CLEF, and ImageCLEF; (iii) the lab was promoted much more intensively, especially with online communities on social platforms such as Linkedin[23] and Facebook[24]. Overall, the success ratio, i.e., the number of teams completing the tasks reported to the number of teams completing the registration is more or less in the same range as in the previous editions, 27% for 2019 compared to 23% for 2018.

The following sections are dedicated to each of 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 [14, 2, 32, 7, 26, 10].

3 The Tuberculosis Task

Tuberculosis (TB) is a bacterial infection discovered about 130 years ago. The bacteria usually attack the lungs and the disease remains a persistent threat and an important cause of death worldwide [46]. Generally, TB can be cured with antibiotics. However, the different types of TB require different treatments and therefore the detection of the TB type and the evaluation of the severity stage are two important tasks. In the first and second editions of this task [13, 15] 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 the two editions it was concluded that the detection of MDR TB was not possible based in good quality only using the image. In the TBT subtask, there was a slight improvement in 2018 with respect to 2017 on the classification results. However, this was not strong considering the amount of additional data provided in the 2018 edition, both in terms of new images and meta-data. Most of the participants obtained good results in the severity scoring (SVR) subtask introduced in 2018. From a medical point of view, the 3 subtasks proposed previously had a limited utility. The MDR subtask was finally not feasible, and the TBT and SVR subtasks are tasks that expert radiologists can perform in a relatively short time. This encouraged us to add a new subtask based on providing an automatic report of the patient, an outcome that can have a major impact in the clinical routine.

3.1 Task Setup

Two subtasks were proposed in the ImageCLEF 2019 tuberculosis task [14]: (i) Severity score assessment (SVR subtask), (ii) Automatic CT report generation (CTR subtask).

The SVR subtask aims at assessing the TB severity score. The Severity score is a cumulative score of severity of a TB case assigned by a medical doctor (MD).

[23] https://www.linkedin.com/
[24] https://www.facebook.com/

Originally, the score varied from 1 ("critical/very bad") to 5 ("very good"). In the process of scoring, the MDs consider many factors like pattern of lesions, results of microbiological tests, duration of treatment, patient age and other data. The goal of this subtask is to assess the severity based on the CT image and additional meta-data, including disability, relapse, co-morbidity, bacillary and smoking history and a few more data items. The original severity score is included as training meta-data but the final score that participants have to assess is reduced to a binary category: "low" (scores 4 and 5) and "high" (scores 1, 2 and 3). In the case of the CTR subtask, the participants had to generate an automatic report based on the CT image. This report needed to include the following information in binary form (0 or 1): Left lung affected, right lung affected, presence of calcifications, presence of caverns, pleurisy, lung capacity decrease.

3.2 Data Set

Both subtasks (SVR and CTR) used the same data set containing 335 chest CT scans of TB patients along with a set of clinically relevant meta-data, divided into 218 patients for training and 117 for testing. The selected meta-data include the following binary measures: disability, relapse, symptoms of TB, comorbidity, bacillary, drug resistance, higher education, ex-prisoner, alcoholic, smoking history, and severity. For all patients we provided 3D CT images with an image size per slice of 512×512 pixels and number of slices varying from 50 to 400. For all patients we provided automatically extracted masks of the lungs obtained using the method described in [12].

3.3 Participating Groups and Submitted Runs

In 2019, 13 groups from 11 countries submitted at least one run to one of the two subtasks. There were 11 groups participating in the SVR subtask and 10 groups participating in the CTR subtask. Similar to previous editions, each group could submit up to 10 runs. 54 runs were submitted to the SVR subtask and 35 to the CTR subtask.

Similar to the previous edition, deep learning had a high presence in the submissions with 10 out of the 12 groups using convolutional neural networks (CNNs), at least in one of their attempts, for feature extraction or directly for patient classification. Five groups used 2D CNNs with pre-processed CT slices and four groups used 3D CNNs (three of them used partial CT volumes and only one used the entire CT scan). The remaining group used 2D CNN to classify feature maps derived from a graph model of the lungs. Despite the general use of CNNs, all these approaches differ in the pre-processing steps, using many techniques such as 2D projections, resizing, slice filtering or concatenations of multiple projections. In addition, one group considered the CT scans as a time sequence and applied optical flow. Another group modeled each CT scan with a set of random pixels and applied decision trees and weak classifiers. Finally, a

Table 2: Results obtained by the participants in the SVR subtask. Only the best run of each participant is reported here.

Group name	Run	AUC	Accuracy	Rank
UIIP_BioMed	SRV_run1_linear.txt	**0.7877**	0.7179	1
UIIP	subm_SVR_Severity	0.7754	0.7179	2
HHU	SVR_HHU_DBS2_run01.txt	0.7695	0.6923	3
CompElecEngCU	SVR_mlp-text.txt	0.7629	0.6581	6
SD VA HCS/UCSD	SVR_From_Meta_Report1c.csv	0.7214	0.6838	7
MedGIFT	SVR_SVM.txt	0.7196	0.6410	9
UniversityAlicante	SVR-SVM-axis-mode-4.txt	0.7013	0.7009	12
MostaganemFSEI	SVR_FSEI_run3_resnet_50_55.csv	0.6510	0.6154	22
SSN CoE	SVRtest-model1.txt	0.6264	0.6068	29
UoAP	SVRfree-text.txt	0.6111	0.6154	32
FIIAugt	SVRab.txt	0.5692	0.5556	38

Table 3: Results obtained by the participants in the CTR subtask. Only the best run of each participant is reported here.

Group name	Run	Mean AUC	Min AUC	Rank
UIIP_BioMed	CTR_run3_pleurisy_as_SegmDiff.txt	**0.7968**	**0.6860**	1
CompElecEngCU	CTRcnn.txt	0.7066	0.5739	4
MedGIFT	CTR_SVM.txt	0.6795	0.5626	5
SD VA HCS/UCSD	CTR_Cor_32_montage.txt	0.6631	0.5541	6
HHU	CTR_HHU_DBS2_run01.txt	0.6591	0.5159	7
UIIP	subm_CT_Report	0.6464	0.4099	10
MostaganemFSEI	CTR_FSEI_run1_lungnet_50_10slices.csv	0.6273	0.4877	14
UniversityAlicante	svm_axis_svm.txt	0.6190	0.5366	15
PwC	CTR_results_meta.txt	0.6002	0.4724	19
LIST	predictionCTReportSVC.txt	0.5523	0.4317	25

group applied a handcrafted technique for each CT finding in the CTR subtask based on image binarization and morphology.

3.4 Results

The SVR subtask was evaluated as a binary classification problem, including measures such as Area Under the ROC Curve (AUC) and accuracy. The ranking of the techniques is first based on the AUC and then on the accuracy. Similarly, the CTR subtask was considered a multi-binary classification problem (6 binary findings). Measures again include AUC and accuracy to evaluate the subtask. The ranking of this task is done first by average AUC and then by min AUC (both over the 6 CT findings). Tables 2 and 3 show the final results for each best run and its rank. More detailed results, including other performance measures, can be found in the overview article of the TB task [14].

3.5 Lessons Learned and Next Steps

The results obtained in the SVR subtask improved with respect to the 2018 edition. UIIP_BioMed obtained the highest rank in both editions, passing from

0.70 to 0.79 AUC. Most of the groups that participated in both editions present similar improvements. According to their reports, this improvement is mainly due to the integration of the new meta-data into their algorithms. In the case of the CTR subtask, also won by UIIP_BioMed, the results of this first edition are very promising with a mean AUC of 0.80. Most of the groups developed a single approach and applied it to detect each of the CT-findings in a multi-binary classification setup. However, UIIP_BioMed and a few other groups applied differing techniques for each finding, obtaining impressive results with somewhat trivial techniques in some of them, e.g., comparing the mask size of the right and left lungs to detect lung capacity decrease. This suggests that a delicate analysis of the images before applying any computer vision approach is essential. Thanks to the large amount of new meta-data offered it was possible to use a single data set for both subtasks. Having larger data sets without this information does not seem optimal and in future editions it is planned to focus on increasing the data set without reducing the amount of meta-data provided.

4 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) to support clinical decision making and improve patient engagement, opportunities to generate and leverage algorithms for automated medical image interpretation are currently being explored. Inspired by the success of visual question answering in the general domain, we conducted a pilot task (VQA-Med 2018) in ImageCLEF 2018 to focus on visual question answering in the medical domain [20]. Based on the success of the initial edition and the huge interest from both computer vision and medical informatics communities, we continued the task this year (VQA-Med 2019) [2] with enhanced focus on a well curated enlarged data set.

4.1 Task Setup

In the same way as in 2018, given a medical image accompanied by a clinically relevant question, participating systems in VQA-Med 2019 are tasked with answering the question based on the visual image content. In VQA-Med 2019, we specifically focused on radiology images and four main categories of questions: modality, plane, organ system, and abnormality. We mainly considered medical questions asking about one element only (e.g., "what is the organ principally shown in this MRI?", "in what plane is this mammograph taken?", "is this a t1 weighted, t2 weighted, or flair image?", "what is most alarming about this ultrasound?"), which can be answered from the image content without requiring additional medical knowledge or domain-specific inference.

Table 4: Participating groups in the VQA-Med 2019 task.

Team	Institution	# Runs
abhishekthanki	Manipal Institute of Technology (India)	8
AIOZ	AIOZ Pte Ltd (Singapore)	6
ChandanReddy	Virginia Tech (USA)	4
Dear stranger	School of Information Science and Engineering, Kunming (China)	6
deepak.gupta651	Indian Institute of Technology Patna (India)	1
Hanlin	Zhejiang University (China)	5
IBM Research AI	IBM Research, Almaden (USA)	4
IITISM@CLEF	Indian Institute of Technology Dhanbad (India)	3
JUST19	(Jordan) University of Science and Technology & University of Manchester (UK)	4
LIST	Faculty of Sciences and Technologies, Tangier (Morocco)	7
minhvu	Umeå University (Sweden) & University of Bern (Switzerland)	10
Team_Pwc_Med	Pricewaterhouse Coopers US Advisory (India)	5
Techno	Faculty of Technology Tlemcen (Algeria)	2
TUA1	Tokushima University (Japan)	1
Turner.JCE	Azrieli College of Engineering Jerusalem (Israel)	10
UMMS	Worcester Polytechnic Institute & University of Massachusetts Medical School (USA)	3
yan	Zhejiang University (China) & National Institute of Informatics (Japan)	1

4.2 Data Set

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[25] database with filters based on their captions, modalities, planes, localities, categories, 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, angiography, characteristic imaging appearance, radiographs, imaging features, ultrasound, and diagnostic radiology. Finally, we considered the most frequent question categories: Modality, Plane, Organ System, and Abnormality to create the data set, which included a training set of 3,200 medical images with 12,792 Question-Answer (QA) pairs (having 3 to 4 questions per image), a validation set of 500 medical images with 2,000 QA pairs, and a test set of 500 medical images with 500 questions. To further ensure the quality of the data, the test set was manually validated by two medical doctors. For more details, please refer to the task overview paper [2].

4.3 Participating Groups and Submitted Runs

Out of 104 online registrations, 61 participants submitted signed end user agreement forms. Finally, 17 groups submitted a total of 90 runs, indicating a notable interest in the VQA-Med 2019 task. Table 4 gives an overview of all participants and the number of submitted runs[26].

[25] https://medpix.nlm.nih.gov

[26] There was a limit of maximum 10 run submissions per team. The table includes only the valid runs that were graded (total# 80 out of 90 submissions).

4.4 Results

The evaluation of the participant systems of the VQA-Med 2019 task was conducted based on two primary metrics: accuracy and BLEU [2]. We use an adapted version of accuracy from the general domain VQA[27] task that strictly considers exact matching of a participant provided answer and the ground truth answer. We calculate the overall accuracy as well as the scores for each question category. To compensate for the strictness of the accuracy metric, BLEU [31] 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 task [20]. The overall results of the participating systems are presented in Table 5 and Table 6 for the two metrics in a descending order of the scores (the higher the better).

Table 5: Accuracy for several query aspects.

Team	Run ID	Modality	Plane	Organ	Abnormality	Overall
Hanlin	26889	0.202	0.192	0.184	0.046	0.624
Hanlin	26891	0.202	0.192	0.184	0.042	0.620
yan	26853	0.202	0.192	0.184	0.042	0.620
minhvu	26881	0.210	0.194	0.190	0.022	0.616
minhvu	27195	0.212	0.190	0.192	0.022	0.616
Hanlin	26890	0.202	0.192	0.184	0.038	0.616
minhvu	26862	0.206	0.192	0.194	0.022	0.614
minhvu	26863	0.208	0.194	0.188	0.024	0.614
minhvu	26879	0.206	0.194	0.192	0.022	0.614
minhvu	27197	0.204	0.194	0.194	0.022	0.614
Hanlin	26917	0.202	0.192	0.184	0.036	0.614
minhvu	26880	0.208	0.194	0.188	0.022	0.612
minhvu	27196	0.208	0.194	0.194	0.016	0.612
minhvu	27198	0.202	0.192	0.192	0.022	0.608
minhvu	26843	0.208	0.192	0.188	0.018	0.606
TUA1	26822	0.186	0.204	0.198	0.018	0.606
Hanlin	26922	0.202	0.192	0.184	0.020	0.598
UMMS	27306	0.168	0.190	0.184	0.024	0.566
AIOZ	26873	0.182	0.180	0.182	0.020	0.564
AIOZ	26833	0.188	0.174	0.182	0.018	0.562
IBM Research AI	27199	0.160	0.196	0.192	0.010	0.558
LIST	26908	0.180	0.184	0.178	0.014	0.556
IBM Research AI	27340	0.156	0.192	0.192	0.012	0.552
LIST	26906	0.166	0.178	0.182	0.012	0.538
Turner.JCE	26913	0.164	0.176	0.182	0.014	0.536
JUST19	27142	0.160	0.182	0.176	0.016	0.534
Turner.JCE	26882	0.174	0.176	0.170	0.014	0.534
Turner.JCE	26939	0.164	0.174	0.182	0.014	0.534
Turner.JCE	27187	0.176	0.174	0.164	0.016	0.530
JUST19	26870	0.160	0.182	0.176	0.010	0.528
JUST19	27143	0.160	0.182	0.176	0.010	0.528
JUST19	27293	0.160	0.182	0.176	0.010	0.528
AIOZ	26783	0.178	0.174	0.162	0.014	0.528
LIST	26900	0.156	0.178	0.180	0.012	0.526
IBM Research AI	27335	0.130	0.190	0.186	0.018	0.524
Turner.JCE	27083	0.176	0.174	0.164	0.010	0.524
AIOZ	26818	0.168	0.170	0.160	0.026	0.524
Turner.JCE	27001	0.152	0.174	0.182	0.014	0.522
AIOZ	26814	0.172	0.170	0.162	0.016	0.520
AIOZ	26819	0.172	0.170	0.162	0.016	0.520
Turner.JCE	26940	0.152	0.174	0.182	0.010	0.518
Turner.JCE	27002	0.152	0.174	0.164	0.014	0.504
Turner.JCE	26883	0.174	0.144	0.166	0.014	0.498
Turner.JCE	26781	0.156	0.176	0.164	0	0.496

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

Table 5: Accuracy for several query aspects.

Team	Run ID	Modality	Plane	Organ	Abnormality	Overall
Team_Pwc_Med	26941	0.148	0.150	0.168	0.022	0.488
Team_Pwc_Med	26955	0.148	0.150	0.168	0.022	0.488
Team_Pwc_Med	27295	0.148	0.150	0.168	0.018	0.484
Team_Pwc_Med	27296	0.148	0.150	0.168	0.018	0.484
UMMS	26931	0.156	0.168	0.152	0.004	0.480
Team_Pwc_Med	27297	0.148	0.150	0.168	0.010	0.476
IBM Research AI	26937	0.094	0.194	0.186	0	0.474
LIST	26829	0.154	0.162	0.138	0.012	0.466
Techno	27079	0.082	0.184	0.170	0.026	0.462
Techno	27100	0.082	0.184	0.170	0.026	0.462
LIST	26828	0.160	0.148	0.144	0.010	0.462
LIST	26831	0.142	0.148	0.138	0.010	0.438
LIST	26832	0.138	0.148	0.138	0.010	0.434
deepak.gupta651@gmail.com	27232	0.096	0.140	0.124	0.006	0.366
ChandanReddy	26884	0.094	0.126	0.064	0.010	0.294
ChandanReddy	26946	0.102	0.122	0.048	0.014	0.286
ChandanReddy	26947	0.094	0.126	0.058	0.008	0.286
Dear stranger	26895	0.062	0.140	0	0.008	0.210
Dear stranger	26894	0.078	0.114	0.002	0.006	0.200
Dear stranger	26919	0.076	0.086	0.004	0.012	0.178
Dear stranger	26920	0.076	0.086	0.004	0.012	0.178
abhishekthanki	27307	0.122	0	0.028	0.010	0.160
abhishekthanki	27298	0.122	0	0.026	0.010	0.158
abhishekthanki	26824	0.112	0	0.026	0.012	0.150
abhishekthanki	27315	0.114	0	0.026	0.010	0.150
abhishekthanki	27317	0.112	0	0.026	0.012	0.150
abhishekthanki	26936	0.104	0	0.024	0.014	0.142
abhishekthanki	26935	0.096	0	0.020	0.010	0.126
abhishekthanki	27316	0.086	0	0	0.012	0.098
IITISM@CLEF	26905	0.052	0.004	0.026	0.006	0.088
IITISM@CLEF	26953	0.052	0.004	0.026	0.006	0.088
Dear stranger	26927	0.054	0	0	0.010	0.064
Dear stranger	26928	0.054	0	0	0.010	0.064
ChandanReddy	26945	0	0.030	0.008	0	0.038
UMMS	26903	0.010	0	0	0.008	0.018
IITISM@CLEF	27304	0	0	0	0	0

Table 6: Results of the VQA task in terms of BLEU scores.

Team	Run ID	BLEU
Hanlin	26889	0.644
Hanlin	26891	0.640
yan	26853	0.640
Hanlin	26890	0.636
minhvu	26881	0.634
minhvu	27195	0.634
Hanlin	26917	0.634
minhvu	26862	0.633
minhvu	26863	0.633
TUA1	26822	0.633
minhvu	26879	0.632
minhvu	27196	0.632
minhvu	27197	0.632
minhvu	26880	0.631
minhvu	26843	0.623
minhvu	27198	0.622
Hanlin	26922	0.615
UMMS	27306	0.593
JUST19	27142	0.591
LIST	26908	0.583

Team	Run ID	BLEU
Turner.JCE	27002	0.538
Team_Pwc_Med	26955	0.534
Team_Pwc_Med	27295	0.531
Team_Pwc_Med	27296	0.531
Team_Pwc_Med	26941	0.530
AIOZ	26814	0.529
AIOZ	26819	0.529
Team_Pwc_Med	27297	0.521
Turner.JCE	26883	0.512
UMMS	26931	0.509
LIST	26828	0.495
LIST	26829	0.493
IBM Research AI	26937	0.486
Techno	27079	0.486
Techno	27100	0.486
abhishekthanki	26824	0.462
abhishekthanki	27317	0.462
LIST	26831	0.459
abhishekthanki	27298	0.455
abhishekthanki	26936	0.453

IBM Research AI	27199	0.582
AIOZ	26833	0.579
AIOZ	26873	0.576
Turner.JCE	26940	0.572
IBM Research AI	27340	0.569
Turner.JCE	26781	0.561
Turner.JCE	27187	0.558
LIST	26906	0.556
Turner.JCE	26939	0.554
Turner.JCE	27083	0.554
JUST19	26870	0.553
Turner.JCE	26913	0.552
Turner.JCE	27001	0.552
JUST19	27143	0.550
JUST19	27293	0.550
Turner.JCE	26882	0.547
LIST	26900	0.546
IBM Research AI	27335	0.542
AIOZ	26783	0.542
AIOZ	26818	0.540
abhishekthanki	27307	0.453
LIST	26832	0.451
abhishekthanki	27315	0.447
abhishekthanki	26935	0.433
Dear stranger	26895	0.393
deepak.gupta651@gmail.com	27232	0.389
ChandanReddy	26946	0.323
ChandanReddy	26884	0.318
Dear stranger	26894	0.310
ChandanReddy	26947	0.307
abhishekthanki	27316	0.301
Dear stranger	26919	0.270
Dear stranger	26920	0.270
ChandanReddy	26945	0.126
IITISM@CLEF	26905	0.096
IITISM@CLEF	26953	0.096
Dear stranger	26927	0.064
Dear stranger	26928	0.064
UMMS	26903	0.039
IITISM@CLEF	27304	0.025

4.5 Lessons Learned and Next Steps

Similar to last year, participants mainly used deep learning techniques to build their VQA-Med systems [2]. In particular, the best-performing systems leveraged deep convolutional neural networks (CNNs) like VGGNet or ResNet with a variety of pooling strategies e.g., global average pooling to encode image features and transformer-based architectures like BERT or recurrent neural networks (RNN) to extract question features. Then, various types of attention mechanisms are 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. Analyses of the question category-wise[28] accuracy in Table 5 suggest that in general, participating systems performed well to answer modality questions, followed by plane and organ questions because the possible types of answers for each of these question categories were finite [2]. However, for the abnormality type questions, systems did not perform good in terms of accuracy because of the underlying complexity of open-ended questions and possibly due to the strictness of the accuracy metric. To compensate for the strictness of the accuracy, we computed the BLEU scores to understand the similarity of the system generated answers and the ground-truth answers. The higher BLEU scores of the systems this year (0.631 best BLEU vs. 0.162 in 2018) further verify the effectiveness of the proposed deep learning-based models for the visual question answering task. Overall, the results obtained this year clearly denote the robustness of the provided data set compared to last year's task.

In this second edition of the VQA challenge, we focused on designing goal-oriented VQA data sets and therefore systems by selecting radiology images and clinically relevant questions and categories. We also targeted medical questions that can be answered from the image content without requiring additional medical knowledge or domain-specific inference. For example, we did not consider

[28] Note that the question category-wise accuracy scores are normalized (each divided by a factor of 4) so that the summation is equal to the overall accuracy.

questions such as: “Is this modality safe for pregnant women?”, “What is located immediately inferior to the right hemidiaphragm?”, “What can be typically visualized in this plane?”, “How would you measure the length of the kidneys?”. We would consider providing such kind of questions in the future editions of the challenge as well as context-sensitive questions, given the important role of context and background knowledge in medicine.

5 The Caption Task

The caption task was first proposed as part of the ImageCLEFmedical [22] in 2016. In 2017 and 2018 [13, 21] the ImageCLEFcaption task comprised two subtasks: concept detection and caption prediction. In 2019 [32], the task concentrates on extracting Unified Medical Language System® (UMLS) Concept Unique Identifiers (CUIs) [4] 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 [34].

5.1 Task Setup

The ImageCLEFmed Caption 2019 [32] follows the format of the concept detection subtask running as part of the ImageCLEFcaption task in 2017 [13] and 2018 [21]. As in the previous two editions, given a medical image, the participating teams are tasked with predicting concepts based on the visual image representation. In 2019, the focus is solely on radiology images. However, no single specific disease or anatomic structure is targeted, but several medical imaging modalities are addressed.

The balanced precision and recall trade-off in terms of F1-scores 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 subset of the Radiology Objects in COntext (ROCO) data set [33]. The training set include 56,629 images with 5,216 associated concepts. The number of related concepts to the validation set is 3,233 and contains 14,157 images. All images distributed are from biomedical journal articles extracted from the PubMed Central® (PMC)[29] repository [36].

For the concept detection evaluation, a test set containing 10,000 radiology images was distributed. The test set is not part of the ROCO data set but was extracted using the same procedures applied for the creation of ROCO. The maximum number of concepts per image varies between 34, 72 and 77 for the test, training and validation sets, respectively. All concepts in the ground truth, that were used for evaluation, are associated either to the training or validation.

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

Table 8: Performance of the participating teams in the ImageCLEF 2019 Concept Detection Task. The best run per team is selected. Teams with previous participation in 2018 are marked with an asterix.

Team	Institution	F1 Score
AUEB NLP Group	Department of Informatics Athens University of Economics and Business	0.2823094
damo	Beihang University, Beijing, China	0.2655099
ImageSem*	Institute of Medical Information Chinese Academy of Medical Sciences	0.2235690
UA.PT_Bioinformatics*	Biomedical Informatics Research Group Universidade de Aveiro, Portugal	0.2058640
richard_ycli	The Hong Kong University of Science and Technology, Kowloon Hong Kong	0.1952310
Sam Maksoud	The University of Queensland Brisbane, Australia	0.1749349
AI600	University of International Business and Economics, Beijing, China	0.1656261
MacUni-CSIRO	Macquarie University, North Ryde Sydney, Australia	0.1435435
pri2si17	Mentor Graphics LibreHealth Uttar Pradesh, India	0.0496821
AILAB*	University of the Aegean Samos, Greece	0.0202243
LIST	Faculty of Sciences and Techniques Abdelmalek Essadi University, Morocco	0.0013269

5.3 Participating Groups and Submitted Runs

In the third edition of the concept detection task [32], 49 teams signed the EUA out of 99 who downloaded it. 60 graded runs were submitted for evaluation by 11 teams from 7 countries. Each group was allowed 10 graded runs and 7 faulty runs altogether. 17 submitted runs were graded as faulty.

Three teams had taken part in the previous editions, while the majority were new to the task. As deep learning techniques have improved accuracy rates in many medical visual classification tasks in the past years [47], most submitted runs were based on these techniques. To optimize input for the predicting systems, several methods were used: image normalization, pre-classification based on body-parts, data augmentation regarding class imbalance, concept filtering and re-division. Transfer learning-based multi-label classification models and convolutional neural network (CNN) image encoders, as well as ensembles of adversarial auto-encoders and long short-term memory (LSTM) recurrent neural networks were the most frequently applied approaches.

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 was selected and is listed in Table 8. The complete list of submissions is presented in [32].

5.5 Lessons Learned and Next Steps

The results improved with respect to both previous editions, from 0.1583 in ImageCLEF 2017 and 0.1108 in ImageCLEF 2018 to 0.2823 this year in terms of F1-score. Three teams participated in the three editions. However, the majority were new to this task. The AUEB NLP Group [27] from Athens University of Economics and Business, who participated for the 1st time, achieved the highest ranked F1-score.

The decision to focus solely on radiology images proved to go into the right direction. Noisy concepts from a wide diversity of medical images were removed, reducing the number of concepts from 111,155 in the previous editions to 5,528 in ImageCLEF 2019, so an amount that is manageable. However, there is still an imbalance in the concept distribution over the images, which showed to be challenging for all teams.

The number of registered teams and submitted runs has increased over the three editions, showing the interest in this challenging task. In future work, better domain knowledge regarding the clinical relevance of the concepts in the development data should be explored. This will assist in creating efficient systems for automated medical data analysis.

6 The Lifelog Task

An increasingly wide range of personal devices, such as smartphones, video cameras as well as wearable devices that 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.

Despite the increasing number of successful related workshops and panels, lifelogging has seldom been the subject of a rigorous comparative benchmarking exercise. In this edition of this task we aim to bring the attention of lifelogging to an as wide as possible audience and to promote research into some of the key challenges of the coming years.

6.1 Task Setup

In 2019, the ImageCLEFlifelog task consists two sub-tasks: *Lifelog moment retrieval (LMRT)* This is the task used in 2018 with different topics. The participants are required to retrieve specific moments in a lifeloggers life. We define moments as semantic events, or activities that happened throughout the day. For example, they should return the relevant moments for the query “Find the moment(s) when the user1 is cooking in the kitchen.” Particular attention needs to be paid to the diversification of the selected moments with respect to the target scenario. The ground truth for this subtask was created using manual annotation; *Solve my life puzzle (Puzzle)* Given a set of lifelog images with associated metadata (e.g., biometrics, location, etc.), but no timestamps, the participants need to analyse these images and rearrange them in chronological order and predict the correct day (Monday or Sunday) and part of the day (morning, afternoon, or evening). The data set is arranged into 75% training and 25% test data.

Table 9: Statistics of the ImageCLEFlifelog 2019 data

Characters	**Size**
Number of Lifeloggers	2
Number of Days	43 days
Size of the Collection	14 GB
Number of Images	81,474 images
Number of Locations	61 semantic locations
Number of Puzzle Queries	20 queries
Number of LMRT Queries	20 queries

6.2 Data Set

The data consists of a medium-sized collection of multimodal lifelog data over 42 days by the two lifeloggers. The data consists of: *Multimedia Content* — Wearable camera images captured at a rate of about two images per minute and worn from breakfast to sleep. Accompanying this image data was a time-stamped record of music listening activities sourced from `Last.FM`[30] and an archive of all conventional (active-capture) digital photos taken by the lifelogger; *Biometrics Data* — Using the FitBit fitness trackers[31], the lifeloggers gathered 24×7 heart rate, calorie burn and steps. In addition, continuous blood glucose monitoring captured readings every 15 minutes using the Freestyle Libre wearable sensor[32]; *Human Activity Data* — The daily activities of the lifeloggers were captured in terms of the semantic locations visited, physical activities (e.g., walking, running, standing) from the Moves app[33], along with a time-stamped diet-log of all food and drink consumed; *Enhancements to the Data* — The wearable camera images were annotated with the outputs of a visual concept detector, which provided three types of outputs (Attributes, Categories and Concepts).

6.3 Participating Groups and Submitted Runs

In 2019, we received in total 50 valid submissions (46 official and 4 additional) for LMRT and 21 (all are official) for Puzzle, from 10 teams from 10 countries. Their submissions and the results are summarised in Tables 10 and 11.

[30] Last.FM Music Tracker and Recommender - https://www.last.fm/
[31] Fitbit Fitness Tracker (FitBit Versa) - https://www.fitbit.com
[32] Freestyle Libre wearable glucose monitor - https://www.freestylelibre.ie/
[33] Moves App for Android and iOS - http://www.moves-app.com/

Table 10: Official results of the ImageCLEFlifelog 2019 LMRT task.

Team	Run	P@10	CR@10	F1@10
Organiser [30]	RUN1*	0.41	0.31	0.29
	RUN2*	0.33	0.26	0.24
ATS [40]	RUN1	0.10	0.08	0.08
	RUN2	0.03	0.06	0.04
	RUN3	0.03	0.04	0.04
	RUN4	0.06	0.13	0.08
	RUN5	0.07	0.06	0.05
	RUN6	0.07	0.13	0.08
	RUN7	0.08	0.19	0.1
	RUN8	0.05	0.11	0.07
	RUN9	0.10	0.14	0.10
	RUN11	0.14	0.16	0.12
	RUN12	0.35	0.36	0.25
BIDAL [11]	RUN1	0.69	0.29	0.37
	RUN2	0.69	0.29	0.37
	RUN3	0.53	0.29	0.35
HCMUS [28]	RUN1	0.70	0.56	0.60
	RUN2	0.70	**0.57**	**0.61**
REGIM [1]	RUN1	0.28	0.16	0.19
	RUN2	0.25	0.14	0.17
	RUN3	0.25	0.10	0.14
	RUN4	0.09	0.05	0.06
	RUN5	0.07	0.09	0.06
	RUN6	0.07	0.08	0.06
UATP [35]	RUN1	0.03	0.01	0.02
	RUN2	0.08	0.02	0.03
	RUN3	0.09	0.02	0.03
	RUN4	0.1	0.02	0.03
	RUN5	0.1	0.02	0.04
	RUN6	0.06	0.06	0.06
UPB [16]	RUN1	0.17	0.22	0.13
ZJUTCVR [48]	RUN1	**0.71**	0.38	0.44
	RUN2†	0.74	0.34	0.43
	RUN3†	0.41	0.31	0.33
	RUN4†	0.48	0.35	0.36
	RUN5†	0.59	0.5	0.48
TUC_MI [39]	RUN1	0.02	0.10	0.03
	RUN2	0.04	0.08	0.04
	RUN3	0.03	0.06	0.03
	RUN4	0.10	0.11	0.09
	RUN5	0.08	0.13	0.09
	RUN6	0.00	0.00	0.00
	RUN7	0.04	0.06	0.05
	RUN8	0.04	0.01	0.02
	RUN9	0.02	0.01	0.01
	RUN10	0.15	0.15	0.12
	RUN11	0.03	0.07	0.04
	RUN12	0.06	0.11	0.06
	RUN13	0.01	0.01	0.01
	RUN14	0.06	0.21	0.09

Notes: * submissions from the organizer teams are just for reference.
† submissions submitted after the official competition.

6.4 Lessons Learned and Next Steps

We learned that all approaches are exploiting multi-modal instead of using only visual information. This trend was established from last year and now it is confirmed. We also confirmed the importance of deep neural networks in solving these challenges: all ten participants are using deep networks or exploiting the semantic concepts extracted by using some deep learning methods. Different from last year, we received more semi-automatic approaches, which combine human knowledge with state-of-the-art multi-modal information retrieval. Regarding the number of the signed-up teams and the submitted runs, the task keeps growing, with the highest number of registrations and participated teams. It is also a great success that team retention rate is high with two thirds of non-organiser teams from 2018 keeping participating in 2019. This confirms how interesting and challenging lifelogging is. As next steps, we do not plan to enrich the data set but rather provide richer and better concepts, improve the quality of the queries and narrow down the application of the challenges.

Table 11: Official Results of the ImageCLEFlifelog 2019 Puzzle Task.

Team	Run	Kendall's Tau	Part of Day	Final Score
Organiser [30]	RUN1*	0.06	0.31	0.18
	RUN2*	0.03	0.35	0.19
	RUN3*	0.03	0.34	0.18
	RUN4*	0.05	0.49	0.27
BIDAL [11]	RUN1	0.12	0.30	0.21
	RUN2	0.08	0.31	0.20
	RUN3	0.06	0.28	0.17
	RUN4	0.12	0.38	0.25
	RUN5	0.10	0.30	0.20
	RUN6	0.09	0.29	0.19
	RUN7	0.15	0.26	0.21
	RUN8	0.07	0.30	0.19
	RUN9	0.19	0.55	0.37
	RUN10	0.17	0.50	0.33
	RUN11	0.10	0.49	0.29
DAMILAB [23]	RUN6	0.02	0.40	0.21
	RUN7	0.02	0.47	0.25
HCMUS [28]	RUN03ME	**0.40**	**0.70**	**0.55**
	RUN3	**0.40**	0.66	0.53
	RUN04ME	**0.40**	**0.70**	**0.55**
	RUN4	**0.40**	0.66	0.53

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

7 The Security Task

File Forgery Detection (FFD) is a serious problem concerning digital forensics examiners. Fraud or counterfeits are common causes for altering files. Another example is a child predator who hides porn images by altering the image extension and in some cases by changing the image signature. Many proposals have been made to solve this problem and the most promising ones concentrate on the image content. It is also common that anyone who wants to hide any kind of information in plain sight without being perceived to use steganography. Steganography is the practice of concealing a file, message, image or video within another file, message, image, or video. The word steganography combines the Greek words steganos, meaning "covered" and graphein meaning "writing". The most usual cover medium for hiding data are images.

7.1 Task Setup

The objective of the specific task is first to examine if an image was forged, then if it could also hide a text message and finally to retrieve the potential hidden message from the forged stego images:

Competition Scenario. You are a professional digital forensic examiner collaborating with the police, who suspects that there is an ongoing fraud in the Central Bank. After obtaining a court order, police gain access to a suspects computer in the bank with the purpose to look for images proving the suspect guilty. However, police suspects that he has managed to change extension and signature of some images, so that they look like pdf files. Additionally, it is highly probable that the suspect has used steganography software to hide messages within

Table 12: Number of files in the data set

	Task 1	**Task 2**	**Task 3**
Training Set	2400	1000	1000
Test Set	1000	500	500

Table 13: Results of task 1 of the security task: identification of forged images

Participant	**runID**	**F-measure**	**Precision**	**Recall**	**Rank**
UA.PT_Bioinformatics	26850	**1.000**	**1.000**	**1.000**	1
nattochaduke	26738	**1.000**	**1.000**	**1.000**	2
agentili	26735	**1.000**	**1.000**	**1.000**	4
abcrowdai	26994	**0.748**	**0.798**	**0.703**	5

Table 14: Results of task 2 of the security task: identification of stego images

Participant	**runID**	**F-measure**	**Precision**	**Recall**	**Rank**
UA.PT_Bioinformatics	26934	**1.000**	**1.000**	**1.000**	1
agentili	26816	**0.888**	**0.908**	**0.868**	9
nattochaduke	26830	**0.660**	**0.508**	**0.944**	10
Yasser	26844	**0.626**	**0.524**	**0.776**	11
abcrowdai	26910	**0.525**	**0.467**	**0.600**	24
cen_amrita	27454	**0.438**	**0.422**	**0.456**	25

Table 15: Results of task 3 of the security task: retrieval of the messages.

Participant	**runID**	**Edit distance**	**Rank**
UA.PT_Bioinformatics	27447	**0.597828610**	1
João Rafael Almeida	26896	**0.563379028**	7

some images that could reveal valuable information of his collaborators. Police authorities asks you to: *Task 1: Identify Forged Images* — Perform detection of altered (forged) images (both extension and signature) and predict the actual type of the forged file; *Task 2: Identify Stego Images* — Identify the altered images that hide steganographic content; *Task 3: Retrieve the Message* — Retrieve the hidden messages (text) from the stego images.

7.2 Data Set

The data set contains 6,400 image and pdf files, divided into 3 sets. Each set is used for a specific task and the number of files contained in each one is shown in Table 12. All participants have access to the training data sets along with their respective ground truth. The test sets are distributed without the ground truth.

7.3 Participating Groups and Submitted Runs

Seven participating groups submitted at least one run to at least one of the tasks. Out of these 7 groups: 4 groups submitted 6 runs to the first task, 6 groups submitted 26 runs to the second task and 2 groups submitted 11 runs to the third task.

7.4 Results

Tables 13-15 summarize the evaluation scores per run and participant for tasks 1-3, refereeing just the best submission per participant. The runs of the first two tasks were compared according to their F-measure, precision and recall, while the ranking of the third task's runs was based on the Lenenshtein edit distance.

7.5 Lessons Learned and Next Steps

The security task was a new task in ImageCLEF 2019. The number of the registered teams/individuals and the submitted runs show that the security challenges receive much attention and can be interesting and challenging. Almost every participant signed to all three tasks although this was not mandatory. This highlights the importance of each task. The majority of the approaches exploits and combines deep learning techniques, achieving very good results. The most difficult task proved to be the third one, in which the participants had to retrieve hidden messages from the images. The third task results also show that there is room for improvement, as more advanced techniques need to be used for better results. The analysis of the specific task results indicates that the training set was small for the specific problem i.e., the extraction of the hidden messages. To leverage the power of advanced deep learning algorithms towards improving the state-of-the-art in steganalysis, we plan to increase the data set. We also plan to narrow down the application of the challenges, e.g., focus in steganalysis, probably in another domain.

8 The Coral Task

Although they represent only a small percentage of the sea floor, coral reefs are extremely important because they are the most bio-diverse marine environments — yet most coral reefs are in danger of being lost within the next 30 years, and with them the ecosystems they support [3]. This catastrophe will see the extinction of many marine species, such as shellfish, corals and many micro-organisms in the ocean. It also reduces reef fishery production, which is an important source of income and food [5, 37]. By monitoring changes in the structural complexity and composition of coral reefs we can help prioritize conservation efforts.

Autonomous Underwater Vehicles (AUV) can collect data for many hours at a time. However, the complexity of the images makes it impossible for human annotators to assess the contents of images on a large scale [6]. Advances in automatically annotating images for complexity and benthic composition have been promising [38, 19]; however, the type of images being collected using action cameras present a particular challenge. Following the success of the ImageCLEF annotation task running between 2012 and 2016 [18, 17, 44, 45, 43], the first edition of the ImageCLEFcoral task [7] aims to automatically annotate images with benthic substrates for monitoring coral reefs.

8.1 Task Setup

In the first edition of the ImageCLEFcoral task, the following two subtasks were proposed: *Coral reef image annotation and localisation* — This task is similar to the classic ImageCLEF annotation task. This subtask requires the participants to label the images with types of benthic substrate together with their bounding box; *Coral reef image pixel-wise parsing task* — This subtask requires the participants to label the images with types of benthic substrate together with a more detailed polygon bounding each substrate the images.

8.2 Data Set

The images used in 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. In particular, the data in the 2019 ImageCLEFcoral task was collected from several locations in the Wakatobi Marine Reserve in Sulawesi, Indonesia in July 2018. The images are part of a monitoring collection and therefore most have a tape measure running through a portion of the image.

The distributed collection data set comprises several sets of overlapping images taken in an area of underwater terrain. Each image was then labelled by experts with the following 13 types of benthic 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. The same set and annotations was provided for both subtasks. The training set contains contains 240 images with 6,670 substrates annotated and the test set contains 200 images with 5,370 substrates annotated.

8.3 Participating Groups and Submitted Runs

In the first edition of the ImageCLEFcoral task, there were 13 teams registered and 5 teams from 4 countries submitted 20 runs. Teams were limited to submit 10 runs per task.

8.4 Results

The task was evaluated using the PASCAL VOC style metric of intersection over union (IoU). The evaluation was carried out using the following 3 measures:

Table 16: Coral reef image annotation and localisation performance in terms of $MAP0.5IoU$, $R0.5IoU$, and $MAP0IoU$. The best run per team in terms of $MAP0.5IoU$ is selected.

Run id	**team**	$MAP0.5IoU$	$R0.5IoU$	$MAP0IoU$
27417	HHUD	0.2427	0.1309	0.4877
27349	VIT	0.14	0.0682	0.431
27497	ISEC	0.0006	0.0006	0.0006

Table 17: Pixel-wise coral reef parsing performance in terms of $MAP0.5IoU$, $R0.5IoU$, and $MAP0IoU$. The best run per team in terms of $MAP0.5IoU$ is selected.

Run id	**team**	$MAP0.5IoU$	$R0.5IoU$	$MAP0IoU$
27500	MTRU	0.0419	0.049	0.2398
27343	SOTON	0.0004	0.0015	0.0484
27505	HHUD	0.0	0.0	0.0

$MAP0.5IoU$ the localised mean average precision (MAP) for each submitted method for using the performance measure of IoU >=0.5 of the ground truth; $R0.5IoU$ the localised mean recall for each submitted method for using the performance measure of IoU >=0.5 of the ground truth; $MAP0IoU$ the image annotation average for each method with success if the concept is simply detected in the image without any localisation. Tables 16 and 17 present the best runs per team in terms of $MAP0.5IoU$. The complete overview of the results can be found in [7], including the accuracy per benthic substrate type.

8.5 Lessons Learned and Next Steps

In the first edition of the ImageCLEF coral task there were 5 groups participating in the 2 tasks, only one group participated in both tasks. The teams explored a variety of machine learning and deep learning approaches. The best run achieved 0.24 in terms of MAP50 score in the coral reef image annotation and localisation subtask and 0.05 in the coral reef image pixel-wise parsing subtask. Poor results were achieved in the coral reef image pixel-wise parsing subtask probably due to the submission of many self-intersecting polygons which were not taken into account in the evaluation.

This is a difficult task due to the complexity of the images including the morphology of the benthic organisms and the similarity between the growth forms. In 2020, we are planning to increase the amount of images provided and limited the submission of self-intersecting polygons.

9 Conclusions

This paper presents a general overview of the activities and outcomes of the ImageCLEF 2019 evaluation campaign. Four tasks were organised, covering challenges in the medical domain (caption analysis, tuberculosis prediction, and medical visual question answering), life logging (daily activities understanding, retrieval and summarization), security (automatically identifying forged content and retrieve hidden information), and nature (segmenting and labeling collections of coral images).

The participation increased in an important way with the diversification of the application domains, reaching more than 235 registrations and 63 teams submitting over 359 runs. Whereas several of the participants had participated in the past, there was also a large number of groups totally new to ImageCLEF and also collaborations of research groups in several tasks.

Most of the proposed solutions evolved around state-of-the-art deep neural networks architectures, also for the medical domain. In the tuberculosis task, results improved over last year and this improvement seems to be driven by the integration of the new meta-data. In the visual question answering task, deep learning techniques were predominant. Attention mechanisms proved to be very useful in improving the performance. In the caption task, results also improved compared to the previous editions. The use of radiology images for the decision, proved to be the best choice, as it focused the task. In the lifelog task, all approaches now exploited multi-modal techniques. Again, deep learning proved to be the state-of-the-art. Notably, semi-automatic approaches became more popular. In the security task, deep learning prevailed as well. Retrieving hidden messages from the images was the most difficult task. Results show that a larger amount of training data is desirable. Finally, the coral task was explored using general machine learning and also deep learning. The task seemed difficult and the lowest results were achieved in the coral reef image pixel-wise parsing.

ImageCLEF 2019 again brought together an interesting mix of tasks and approaches and we are looking forward to the fruitful discussions at the workshop.

Acknowledgements

The work of Bogdan Ionescu was partially supported by the Ministry of Innovation and Research, UEFISCDI, project SPIA-VA, agreement 2SOL/2017, grant PNIII-P2-2.1-SOL-2016-02-0002.

The data collection of the ImageCLEFcoral task was funded by an IAA grant with support from Professor David Smith and Operation Wallacea. The work of Antonio Campello was supported by Innovate UK, Knowledge Transfer Partnership project KTP010993, and hosted at Filament Consultancy Group Limited.

References

1. Abdallah, F.B., Feki, G., Ammar, A.B., , Amar, C.B.: Big Data For Lifelog Moments Retrieval Improvement. In: CLEF2019 Working Notes. CEUR Workshop Proceedings, CEUR-WS.org <http://ceur-ws.org>, Lugano, Switzerland (September 09-12 2019)

2. 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)
3. 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)
4. 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
5. 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)
6. Bullimore, R.D., Foster, N.L., Howell, K.L.: Coral-characterized benthic assemblages of the deep northeast atlantic: defining coral gardens to support future habitat mapping efforts. ICES Journal of Marine Science **70**(3), 511–522 (2013)
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. 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)
9. Clough, P., Sanderson, M.: The CLEF 2003 cross language image retrieval task. In: Proceedings of the Cross Language Evaluation Forum (CLEF 2003) (2004)
10. 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)
11. Dao, M.S., Vo, A.K., Phan, T.D., , Zettsu, K.: BIDAL@imageCLEFlifelog2019: The Role of Content and Context of Daily Activities in Insights from Lifelogs. In: CLEF2019 Working Notes. CEUR Workshop Proceedings, CEUR-WS.org <http://ceur-ws.org>, Lugano, Switzerland (September 09-12 2019)
12. 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)
13. 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)
14. 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)
15. 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)

16. Dogariu, M., Ionescu, B.: Multimedia Lab @ ImageCLEF 2019 Lifelog Moment Retrieval Task. In: CLEF2019 Working Notes. CEUR Workshop Proceedings, CEUR-WS.org <http://ceur-ws.org>, Lugano, Switzerland (September 09-12 2019)
17. Gilbert, A., Piras, L., Wang, J., Yan, F., Dellandrea, E., Gaizauskas, R.J., Villegas, M., Mikolajczyk, K.: Overview of the ImageCLEF 2015 scalable image annotation, localization and sentence generation task. In: CLEF Working Notes (2015)
18. Gilbert, A., Piras, L., Wang, J., Yan, F., Ramisa, A., Dellandrea, E., Gaizauskas, R.J., Villegas, M., Mikolajczyk, K.: Overview of the ImageCLEF 2016 scalable concept image annotation task. In: CLEF Working Notes. pp. 254–278 (2016)
19. Gonzalez-Rivero, M., Bongaerts, P., Beijbom, O., Pizarro, O., Friedman, A., Rodriguez-Ramirez, A., Upcroft, B., Laffoley, D., Kline, D., Bailhache, C., Vevers, R., Hoegh-Guldberg, O.: The catlin seaview surveykilometre-scale seascape assessment, and monitoring of coral reef ecosystems. Aquatic Conservation: Marine and Freshwater Ecosystems **24**, 184–198 (11 2014)
20. 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)
21. 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)
22. 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)
23. Hoang, T.H., Tran, M.K., Nguyen, V.T., Tran, M.T.: Solving Life Puzzle with Visual Context-based Clustering and Habit Reference. In: CLEF2019 Working Notes. CEUR Workshop Proceedings, CEUR-WS.org <http://ceur-ws.org>, Lugano, Switzerland (September 09-12 2019)
24. Ionescu, B., Müller, H., Villegas, M., de Herrera, A.G.S., Eickhoff, C., Andrearczyk, V., Cid, Y.D., Liauchuk, V., Kovalev, V., Hasan, S.A., Ling, Y., Farri, O., Liu, J., Lungren, M., Dang-Nguyen, D.T., Piras, L., Riegler, M., Zhou, L., Lux, M., Gurrin, C.: Overview of ImageCLEF 2018: Challenges, datasets and evaluation. In: Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Ninth International Conference of the CLEF Association (CLEF 2018), vol. 11018, pp. 309–334. LNCS Lecture Notes in Computer Science, Springer, Avignon, France (September 10-14 2018)
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. Karampidis, K., Vasillopoulos, N., Cuevas Rodrguez, C., del Blanco, C.R., Kavallieratou, E., Garcia, N.: Overview of the ImageCLEFsecurity 2019 task. In: CLEF2019 Working Notes. CEUR Workshop Proceedings, CEUR-WS.org, Lugano, Switzerland (September 09-12 2019)
27. Kougia, V., Pavlopoulos, J., Androusopoulos, I.: Aueb nlp group at imageclefmed caption 2019. In: CLEF2019 Working Notes. CEUR Workshop Proceed-

ings, CEUR-WS.org <http://ceur-ws.org>, Lugano, Switzerland (September 09-12 2019)

28. Le, N.K., Nguyen, D.H., Nguyen, V.T., Tran, M.T.: Lifelog Moment Retrieval with Advanced Semantic Extraction and Flexible Moment Visualization for Exploration. In: CLEF2019 Working Notes. CEUR Workshop Proceedings, CEUR-WS.org <http://ceur-ws.org>, Lugano, Switzerland (September 09-12 2019)
29. 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)
30. Ninh, V.T., Le, T.K., Zhou, L., Piras, L., Riegler, M., Lux, M., Tran, M.T., Gurrin, C., Dang-Nguyen, D.T.: LIFER 2.0: Discover Personal Lifelog Insight by Interactive Lifelog Retrieval System. In: CLEF2019 Working Notes. CEUR Workshop Proceedings, CEUR-WS.org <http://ceur-ws.org>, Lugano, Switzerland (September 09-12 2019)
31. 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)
32. 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)
33. 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)
34. 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
35. Ribeiro, R., Neves, A.J.R., Oliveira, J.L.: UAPTBioinformatics working notes at ImageCLEF 2019 Lifelog Moment Retrieval (LMRT) task. In: CLEF2019 Working Notes. CEUR Workshop Proceedings, CEUR-WS.org <http://ceur-ws.org>, Lugano, Switzerland (September 09-12 2019)
36. 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
37. 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)
38. Stokes, M., Deane, G.: Automated processing of coral reef benthic images. Limnology and Oceanography Methods **7**, 157–168 (2009)
39. Taubert, S., Kahl, S.: Automated Lifelog Moment Retrieval based on Image Segmentation and Similarity Scores. In: CLEF2019 Working Notes. CEUR Workshop Proceedings, CEUR-WS.org <http://ceur-ws.org>, Lugano, Switzerland (September 09-12 2019)

40. Tournadre, M., Dupont, G., Pauwels, V., Cheikh, B., Lmami, M., , Ginsca, A.L.: A Multimedia Modular Approach to Lifelog Moment Retrieval. In: CLEF2019 Working Notes. CEUR Workshop Proceedings, CEUR-WS.org <http://ceur-ws.org>, Lugano, Switzerland (September 09-12 2019)
41. 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)
42. 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)
43. Villegas, M., Paredes, R.: Overview of the ImageCLEF 2012 scalable web image annotation task. In: CLEF Working Notes (2012)
44. Villegas, M., Paredes, R.: Overview of the ImageCLEF 2014 scalable concept image annotation task. In: CLEF Working Notes. pp. 308–328. Citeseer (2014)
45. Villegas, M., Paredes, R., Thomee, B.: Overview of the ImageCLEF 2013 scalable concept image annotation subtask. In: CLEF Working Notes (2012)
46. World Health Organization, et al.: Global tuberculosis report 2016 (2016)
47. 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
48. Zhou, P., Bai, C., Xia, J.: ZJUTCVR Team at ImageCLEFlifelog2019 Lifelog Moment Retrieval Task. In: CLEF2019 Working Notes. CEUR Workshop Proceedings, CEUR-WS.org <http://ceur-ws.org>, Lugano, Switzerland (September 09-12 2019)