Overview of ImageCLEF 2018: Challenges, Datasets and Evaluation

Bogdan Ionescu[1(✉)], Henning Müller[2], Mauricio Villegas[3], Alba García Seco de Herrera[4], Carsten Eickhoff[5], Vincent Andrearczyk[2], Yashin Dicente Cid[2], Vitali Liauchuk[6], Vassili Kovalev[6], Sadid A. Hasan[7], Yuan Ling[7], Oladimeji Farri[7], Joey Liu[7], Matthew Lungren[8], Duc-Tien Dang-Nguyen[9], Luca Piras[10], Michael Riegler[11,12], Liting Zhou[9], Mathias Lux[13], and Cathal Gurrin[9]

[1] University Politehnica of Bucharest, Bucharest, Romania
bionescu@alpha.imag.pub.ro
[2] University of Applied Sciences Western Switzerland (HES-SO), Sierre, Switzerland
[3] omni:us, Berlin, Germany
[4] University of Essex, Colchester, UK
[5] Brown University, Providence RI, USA
[6] United Institute of Informatics Problems, Minsk, Belarus
[7] Artificial Intelligence Lab, Philips Research North America, Cambridge, MA, USA
[8] Department of Radiology, Stanford University, Stanford, CA, USA
[9] Dublin City University, Dublin, Ireland
[10] University of Cagliari & Pluribus One, Cagliari, Italy
[11] University of Oslo, Oslo, Norway
[12] Simula Metropolitan Center for Digital Engineering, Oslo, Norway
[13] Klagenfurt University, Klagenfurt, Austria

Abstract. This paper presents an overview of the ImageCLEF 2018 evaluation campaign, an event that was organized as part of the CLEF (Conference and Labs of the Evaluation Forum) Labs 2018. ImageCLEF is an ongoing initiative (it started in 2003) that promotes the evaluation of technologies for annotation, indexing and retrieval with the aim of providing information access to collections of images in various usage scenarios and domains. In 2018, the 16th edition of ImageCLEF ran three main tasks and a pilot task: 1) a *caption prediction* task that aims at predicting the caption of a figure from the biomedical literature based only on the figure image; 2) a *tuberculosis* task that aims at detecting the tuberculosis type, severity and drug resistance from CT (Computed Tomography) volumes of the lung; 3) a *LifeLog* task (videos, images and other sources) about daily activities understanding and moment retrieval, and 4) a pilot task on *visual question answering* where systems are tasked with answering medical questions. The strong participation, with over 100 research groups registering and 31 submitting results for the tasks, shows an increasing interest in this benchmarking campaign.

1 Introduction

One or two decades ago getting access to large visual data sets for research was a problem and open data collections that could be used to compare algorithms of researchers were rare. Now, it is getting easier to access data collections but it is still hard to obtain annotated data with a clear evaluation scenario and strong baselines to compare against. Motivated by this, ImageCLEF has for 16 years been an initiative that aims at evaluating multilingual or language independent annotation and retrieval of images [21,39,23,25,5]. The main goal of ImageCLEF is to support the advancement of the field of visual media analysis, classification, annotation, indexing and retrieval. It proposes novel challenges and develops the necessary infrastructure for the evaluation of visual systems operating in different contexts and providing reusable resources for benchmarking. It is also linked to initiatives such as Evaluation-as-a-Service (EaaS) [17,18].

Many research groups have participated over the years in these evaluation campaigns and even more have acquired its datasets for experimentation. The impact of ImageCLEF can also be seen by its significant scholarly impact indicated by the substantial numbers of its publications and their received citations [36].

There are other evaluation initiatives that have had a close relation with ImageCLEF. LifeCLEF [22] was formerly an ImageCLEF task. However, due to the need to assess technologies for automated identification and understanding of living organisms using data not only restricted to images, but also videos and sound, it was decided to be organised independently from ImageCLEF. Other CLEF labs linked to ImageCLEF, in particular the medical task, are: CLEFeHealth [14] that deals with processing methods and resources to enrich difficult-to-understand eHealth text and the BioASQ [4] tasks from the Question Answering lab that targets biomedical semantic indexing and question answering but is now not a lab anymore. Due to their medical orientation, the organisation is coordinated in close collaboration with the medical tasks in ImageCLEF. In 2017, ImageCLEF explored synergies with the MediaEval Benchmarking Initiative for Multimedia Evaluation [15], which focuses on exploring the "multi" in multimedia: speech, audio, visual content, tags, users, context. MediaEval was founded in 2008 as VideoCLEF, a track in the CLEF Campaign.

This paper presents a general overview of the ImageCLEF 2018 evaluation campaign[1], which as usual was an event organised as part of the CLEF labs[2].

The remainder of the paper is organized as follows. Section 2 presents a general description of the 2018 edition of ImageCLEF, commenting about the overall organisation and participation in the lab. Followed by this are sections dedicated to the four tasks that were organised this year: Section 3 for the Caption Task, Section 4 for the Tuberculosis Task, Section 5 for the Visual Question Answering Task, and Section 6 for the Lifelog Task. For the full details and complete results on the participating teams, the reader should refer to the

[1] http://imageclef.org/2018/
[2] http://clef2018.clef-initiative.eu/

corresponding task overview papers [20,11,19,7]. The final section concludes the paper by giving an overall discussion, and pointing towards the challenges ahead and possible new directions for future research.

2 Overview of Tasks and Participation

ImageCLEF 2018 consisted of three main tasks and a pilot task that covered challenges in diverse fields and usage scenarios. In 2017 [21] the proposed challenges were almost all new in comparison to 2016 [40], the only exception being Caption Prediction that was a subtask already attempted in 2016, but for which no participant submitted results. After such a big change, for 2018 the objective was to continue most of the tasks from 2017. The only change was that the 2017 Remote Sensing pilot task was replaced by a novel one on Visual Question Answering. The 2018 tasks are the following:

- **ImageCLEFcaption:** Interpreting and summarizing the insights gained from medical images such as radiology output is a time-consuming task that involves highly trained experts and often represents a bottleneck in clinical diagnosis pipelines. Consequently, there is a considerable need for automatic methods that can approximate this mapping from visual information to condensed textual descriptions. The task addresses the problem of bio-medical image concept detection and caption prediction from large amounts of training data.
- **ImageCLEFtuberculosis:** The main objective of the task is to provide a tuberculosis severity score based on the automatic analysis of lung CT images of patients. Being able to extract this information from the image data alone allows to limit lung washing and laboratory analyses to determine the tuberculosis type and drug resistances. This can lead to quicker decisions on the best treatment strategy, reduced use of antibiotics and lower impact on the patient.
- **ImageCLEFlifelog:** An increasingly wide range of personal devices, such as smart phones, video cameras as well as wearable devices that allow capturing pictures, videos, and audio clips of every moment of life 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 to retrieve query-information that the user may desire. Hence, this task addresses the problems of lifelog data understanding, summarization and retrieval.
- **ImageCLEF-VQA-Med (pilot task):** Visual Question Answering is a new and exciting problem that combines natural language processing and computer vision techniques. With the ongoing drive for improved patient engagement and access to the electronic medical records via patient portals, patients can now review structured and unstructured data from labs and images to text reports associated with their healthcare utilization. Such access can help them better understand their conditions in line with the details received from their healthcare provider. Given a medical image accompanied

with a set of clinically relevant questions, participating systems are tasked with answering the questions based on the visual image content.

In order to participate in the evaluation campaign, the research groups first had to register by following the instructions on the ImageCLEF 2018 web page. To ease the overall management of the campaign, this year the challenge was organized through the crowdAI platform[3]. To get access to the datasets, the participants were required to submit a signed End User Agreement (EUA) form. Table 1 summarizes the participation in ImageCLEF 2018, including the number of registrations (counting only the ones that downloaded the EUA) and the number of signed EUAs, indicated both per task and for the overall Lab. The table also shows the number of groups that submitted results (runs) and the ones that submitted a working notes paper describing the techniques used.

The number of registrations could be interpreted as the initial interest that the community has for the evaluation. However, it is a bit misleading because several persons from the same institution might register, even though in the end they count as a single group participation. The EUA explicitly requires all groups that get access to the data to participate, even though this is not enforced. Unfortunately, the percentage of groups that submit results is often limited. Nevertheless, as observed in studies of scholarly impact [36,37], in subsequent years the datasets and challenges provided by ImageCLEF often get used, in part due to the researchers that for some reason (e.g. alack of time, or other priorities) were unable to participate in the original event or did not complete the tasks by the deadlines.

After a decrease in participation in 2016, the participation again increased in 2017 and for 2018 it increased further. The number of signed EUAs is considerably higher, mostly due to the fact that this time each task had an independent EUA. Also, due to the change to crowdAI, the online registration became easier and attracted other research groups than usual, which made the registration-to-participation ratio lower than in previous years. Nevertheless, in the end, 31 groups participated and 28 working notes papers were submitted, which is a slight increase with respect to 2017. The following four sections are dedicated to each of the tasks. Only a short overview is reported, including general objectives, description of the tasks and datasets and a short summary of the results.

3 The Caption Task

This task studies algorithmic approaches to medical image understanding. As a testbed for doing so, teams were tasked with automatically "guessing" fitting keywords or free-text captions that best describe an image from a collection of images published in the biomedical literature.

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

Table 1: Key figures of participation in ImageCLEF 2018.

Task	Registered & downloaded EUA	Signed EUA	Groups that subm. results	Submitted working notes
Caption	84	46	8	6
Tuberculosis	85	33	11	11
VQA-Med	58	28	5	5
Lifelog	38	25	7	7
Overall	265*	132*	31	29

* Total for all tasks, not unique groups/emails.

3.1 Task Setup

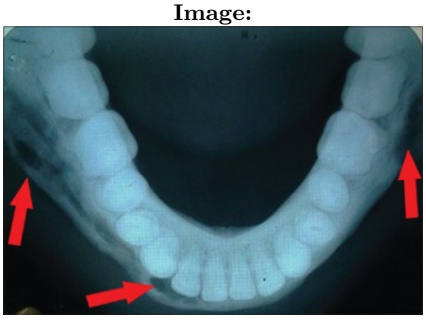
Following the structure of the 2017 edition, two sub tasks were proposed. The first task, concept detection, aims to extract the main biomedical concepts represented in an image based only on its visual content. These concepts are UMLS (Unified Medical Language System®) Concept Unique Identifiers (CUIs). The second task, caption prediction, aims to compose coherent free-text captions describing the image based only on the visual information. Participants were, of course, allowed to use the UMLS CUIs extracted in the first task to compose captions from individual concepts. Figure 1 shows an example of the information available in the training set. An image is accompanied by a set of UMLS CUIs and a free-text caption. Compared to 2017 the data sets was modified strongly to respond to some of the difficulties with the task in the past [13].

3.2 Dataset

The dataset used in this task is derived from figures and their corresponding captions extracted from biomedical articles on PubMed Central® (PMC)[4]. This data set was changed strongly compared to the same task run in 2017 to reduce the diversity on the data and limit the number of compound figures. A subset of clinical figures was automatically obtained from the overall set of 5.8 million PMC figures using a deep multimodal fusion of Convolutional Neural Networks (CNN), described in [2]. In total, the dataset is comprised of 232,305 image–caption pairs split into disjoint training (222,305 pairs) and test (10,000 pairs) sets. For the Concept Detection subtask, concepts present in the caption text were extracted using the QuickUMLS library [30]. After having observed a strong breadth of concepts and image types in the 2017 edition of the task, this year's continuation focused on radiology artifacts, introducing a greater topical focus to the collection.

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

Image:



Concept detection:

- C0024689: Mandibular Diseases
- C2964201: Mandibular arc
- C0447303: Occlusal surface of tooth
- C0307097: Occlusal-HP
- C0591895: Occlusal - Brand Name
- C1962945: Radiographic imaging procedure
- C1548003: Diagnostic Service Section ID - Radiograph
- C0227123: Gum of mandible
- C0024687: Mandible
- C1306645: Plain x-ray
- C0024693: Mandibular Injuries
- C0043299: Diagnostic radiologic examination
- C1550309: Submandibular
- C1548009: Cineradiograph

Caption prediction: *Mandibular* true *occlusal radiograph.*

Fig. 1: Example of an image and the information provided in the training set in the form of the original caption and the extracted UMLS concepts.

3.3 Participating Groups and Submitted Runs

In 2018, 46 groups registered for the caption task compared with the 37 groups registered in 2017. 8 groups submitted runs, one less than in 2017. 28 runs were submitted to the concept detection subtask and 16 to the caption prediction task. Although the caption prediction task appears like an extension of the concept detection task, only two groups participated in both, and 4 groups participated only in the caption prediction task.

3.4 Results

The submitted runs are summarized in Tables 2 and 3, respectively. Similar to 2017, there were two main approaches used on the concept detection subtask: multi-modal classification and retrieval.

ImageSem [41] was the only group applying a retrieval approach this year achieving 0.0928 in terms of mean F1 scores. They retrieved similar images

Table 2: Concept detection performance in terms of F_1 scores.

Team	Run	$MeanF_1$
UA.PT_Bioinformatics	aae-500-o0-2018-04-30_1217	0.1108
UA.PT_Bioinformatics	aae-2500-merge-2018-04-30_1812	0.1082
UA.PT_Bioinformatics	lin-orb-500-o0-2018-04-30_1142	0.0978
ImageSem	run10extended_results_concept_1000_steps_25000_learningrate_0.03_batch_20	0.0928
ImageSem	run02extended_results-testdata	0.0909
ImageSem	run4more1000	0.0907
ImageSem	run01candidate_image_test_0.005	0.0894
ImageSem	run05extended_results_concept_1000_top20	0.0828
UA.PT_Bioinformatics	faae-500-o0-2018-04-27_1744	0.0825
ImageSem	run06top2000_extended_results	0.0661
UA.PT_Bioinformatics	knn-ip-aae-train-2018-04-27_1259	0.0569
UA.PT_Bioinformatics	knn-aae-all-2018-04-26_1233	0.0559
IPL	DET_IPL_CLEF2018_w_300_annot_70_gboc_200	0.0509
UMass	result_concept_new	0.0418
AILAB	results_v3	0.0415
IPL	DET_IPL_CLEF2018_w_300_annot_40_gboc_200	0.0406
AILAB	results	0.0405
IPL	DET_IPL_CLEF2018_w_300_annot_30_gboc_200	0.0351
UA.PT_Bioinformatics	knn-orb-all-2018-04-24_1620	0.0314
IPL	DET_IPL_CLEF2018_w_200_annot_30_gboc_200	0.0307
UA.PT_Bioinformatics	knn-ip-faae-all-2018-04-27_1512	0.0280
UA.PT_Bioinformatics	knn-ip-faae-all-2018-04-27_1512	0.0272
IPL	DET_IPL_CLEF2018_w_200_annot_20_gboc_200	0.0244
IPL	DET_IPL_CLEF2018_w_200_annot_15_gboc_200	0.0202
IPL	DET_IPL_CLEF2018_w_100_annot_20_gboc_100	0.0161
AILAB	results_v3	0.0151
IPL	DET_IPL_CLEF2018_w_200_annot_5_gboc_200	0.0080
ImageSem	run03candidate_image_test_0.005douhao	0.0001

from the training set and clustered concepts of those images. The multi–modal classification approach was more popular [28,27,38]. Best results were achieved by UA.PT Bioinformatics [27] using a traditional bag-of-visual-words algorithm. They experimented with logistic regression and k–Nearest Neighbors (k–NN) for the classification step. Morgan State University [28] used a deep learning based approach by using both image and text (caption) features of the training set for modeling. However, instead of using the full 220K-image collection, they relied on a subset of 4K images, applying the Keras[5] framework to generate deep learning based features. IPL [38] used and encoder of the ARAE [44] model creating a textual representation for all captions. In addition, the images were mapped to continuous representation space with a CNN.

In the Caption Prediction subtask, ImageSem [41] achieved the best results using an image retrieval strategy and tuning the parameters such as the most similar images and the number of candidate concepts. The other 4 groups used different deep learning approaches in very interesting ways from generating captions word by word or in sequences of words. Morgan State University [28] and WHU used a long short-term memory (LSTM) network while UMass [33] and KU Leuven [32] applied different CCNs.

[5] https://keras.io/

Table 3: Caption prediction performance in terms of BLEU scores.

Team	Run	Mean BLEU
ImageSem	run04Captionstraining	0.2501
ImageSem	run09Captionstraining	0.2343
ImageSem	run13Captionstraining	0.2278
ImageSem	run19Captionstraining	0.2271
ImageSem	run03Captionstraining	0.2244
ImageSem	run07Captionstraining	0.2228
ImageSem	run08Captionstraining	0.2221
ImageSem	run06Captionstraining	0.1963
UMMS	test_captions_output4_13_epoch	0.1799
UMMS	test_captions_output2_12_epoch	0.1763
Morgan	result_caption	0.1725
UMMS	test_captions_output1	0.1696
UMMS	test_captions_output5_13_epoch	0.1597
UMMS	test_captions_output3_13_epoch	0.1428
KU Leuven	23_test_valres_0.134779058389_out_file_greedy	0.1376
WHU	CaptionPredictionTesting-Results-zgb	0.0446

After discussions in the 2017 submissions where groups used external data and possibly included part of the test data, no group augmented the training set in 2018. It is further noticeable that, despite the dataset being less noisy than in 2018, the achieved results were slightly lower than observed in the previous year, in both tasks.

3.5 Lessons Learned and Next Steps

Interestingly and despite this year's focus on radiology modalities, a large number of target concepts was extracted in the training set. Such settings with hundreds of thousands of classes are extremely challenging and fall into the realm of extreme classification methods. In future editions of the task, we plan to focus on detecting only the most commonly used UMLS concepts and truncate the concept distribution in order to shift the intellectual challenge away from extreme or one-shot classification settings that were not originally meant to be the key challenge in this task.

The new filtering for finding images with lower variability and fewer combined figures helped to make the task more realistic and considering the difficulty of the task the results are actually fairly good.

Most techniques used relied on deep learning but best results were often obtained also with other techniques, such as using retrieval and handcrafted features. This may be due to the large number of concepts and in this case limited amount of training data. As PMC is increasing in size very quickly it should be easy to find more data for future contests.

4 The Tuberculosis Task

Tuberculosis (TB) remains a persistent threat and a leading cause of death worldwide also in recent years with multiple new strains appearing worldwide. Recent studies report a rapid increase of drug-resistant cases [29] meaning that the TB organisms become resistant to two or more of the standard drugs. One of the most dangerous forms of drug-resistant TB is so-called multi-drug resistant (MDR) tuberculosis that is simultaneously resistant to several of the most powerful antibiotics. Recent published reports show statistically significant links between drug resistance and multiple thick-walled caverns [42]. However, the discovered links are not sufficient for a reliable early recognition of MDR TB. Therefore, assessing the feasibility of MDR detection based on Computed Tomography (CT) imaging remains an important but very challenging task. Other tasks proposed in the ImageCLEF 2018 tuberculosis challenge are automatic classification of TB types and TB severity scoring using CT volumes.

4.1 Task Setup

Three subtasks were proposed in the ImageCLEF 2018 tuberculosis task [11]:

- Multi-drug resistance detection (MDR subtask);
- Tuberculosis type classification (TBT subtask);
- Tuberculosis severity scoring (SVR subtask).

The goal of the MDR subtask is to assess the probability of a TB patient having a resistant form of tuberculosis based on the analysis of a chest CT. Compared to 2017, datasets for the MDR detection subtask were extended by means of adding several cases with extensively drug-resistant tuberculosis (XDR TB), which is a rare and the most severe subtype of MDR TB.

The goal of the TBT subtask is to automatically categorize each TB case into one of the following five types: Infiltrative, Focal, Tuberculoma, Miliary, and Fibro-cavernous. The SVR subtask is dedicated to assess the TB severity based on a single CT image of a patient. The severity score is the results of a cumulative score of TB severity assigned by a medical doctor.

4.2 Dataset

For all three subtasks 3D CT volumes were provided with a size of 512×512 pixels and number of slices varying from 50 to 400. All CT images were stored in the NIFTI file format with .nii.gz file extension (g-zipped .nii files). This file format stores raw voxel intensities in Hounsfield Units (HU) as well as the corresponding image metadata such as image dimensions, voxel size in physical units, slice thickness, etc. For all patients automatically extracted masks of the lungs were provided. The details of the lung segmentation used can be found in [9].

Tables 4, 5 and 6 present for each of the subtasks the division of the datasets between training and test sets (columns), and the corresponding ground truth

Table 4: Dataset for the MDR subtask.

# Patients	Train	Test
DS	134	99
MDR	125	137
Total patients	259	236

Table 5: Dataset for the TBT subtask.

# Patients (# CTs)	Train	Test
Type 1 – Infiltrative	228 (376)	89 (176)
Type 2 – Focal	210 (273)	80 (115)
Type 3 – Tuberculoma	100 (154)	60 (86)
Type 4 – Miliary	79 (106)	50 (71)
Type 5 – Fibro-cavernous	60 (99)	38 (57)
Total patients (CTs)	677 (1008)	317 (505)

Table 6: Dataset for the SVR subtask.

# Patients	Train	Test
Low severity	90	62
High severity	80	47
Total patients	170	109

labels (rows). The dataset for the MDR subtask was composed of 262 MDR and 233 Drug-Sensitive (DS) patients, as shown in Table 4. In addition to CT image data, age and gender for each patient were provided for this subtask. The TBT task contained in total 1,513 CT scans of 994 unique patients divided as shown in Table 5. Patient metadata includes only age. The dataset for the SVR subtask was represented by a total number of 279 patients with a TB severity score assigned for each case by medical doctors. The scores were presented as numbers from 1 to 5, so for a regression task. In addition, for the 2-class prediction task the severity labels were binarized so that scores from 1 to 3 corresponded to "high severity" and 4-5 corresponded to "low severity" (see Table 6).

4.3 Participating Groups and Submitted Runs

In the second year of the task, 11 groups from 9 countries submitted at least one run to one of the subtasks. There were 7 groups participating in the MDR subtask, 8 in the TBT subtask, and 7 groups participating in the SVR subtask. Each group could submit up to 10 runs. Finally, 39 runs were submitted by the groups in the MDR subtask, 39 in the TBT and 36 in the SVR subtasks. Several Deep Learning approaches were employed by 8 out of the 11 participating groups. The approaches were based on using 2D and 3D Convolutional Neural Networks (CNNs) for both classification and feature extraction, transfer learning and a few other techniques. In addition, one group used texture-based graph models of

the lungs, one group used texture-based features combined with classifiers and one group used features based on image binarization and morphology.

4.4 Results

The MDR subtask is designed as a 2-class problem. The participants submitted for each patient in the test set the probability of belonging to the MDR group. The Area Under the ROC Curve (AUC) was chosen as the measure to rank the results. The accuracy was provided as well. For the TBT subtask, the participants had to submit the tuberculosis type. Since the 5-class problem was not balanced, Cohen's Kappa[6] coefficient was used to compare the methods. Again, the accuracy was provided for this subtask. Finally, the SVR subtask was considered in two ways: as a regression problem with scores from 1 to 5, and as a 2-class classification problem (low/high severity). The regression problem was evaluated using Root Mean Square Error (RMSE), and AUC was used to evaluate the classification approaches. Tables 7, 8 and 9 show the final results for each run and its rank.

4.5 Lessons Learned and Next Steps

Similarly to 2017 [10], in the MDR task all participants achieved a relatively low performance, which is only slightly higher than the performance of a random classifier. The best accuracy achieved by participants was 0.6144, and the best reached AUC was 0.6178. These results are better than in the previous years but still remain unsatisfactory for clinical use. The overall increase of performance compared to 2017 may be partly explained by the introduction of patient age and gender, and also by adding more severe cases with XDR TB. For the TBT subtask, the results are slightly worse compared to 2017 in terms of Cohen's Kappa with the best run scoring a 0.2312 Kappa value (0.2438 in 2017) and slightly better with respect to the best accuracy of 0.4227 (0.4067 in 2017). It is worth to notice that none of the groups achieving best performance in the 2017 edition participated in 2018. The group obtaining best results in this task this year (the UIIP group) obtained a 0.1956 Kappa value and 0.3900 accuracy in the 2017 edition. This shows a strong improvement, possibly linked to the increased size of the dataset. The newly-introduced SVR subtask demonstrated good performance in both regression and classification problems. The best result in terms of regression achieved a 0.7840 RMSE, which is less than 1 grade of error in a 5-grade scoring system. The best classification run demonstrated a 0.7708 AUC. These results are promising taking into consideration the fact that TB severity was scored by doctors using not only CT images but also additional clinical data. The good participation also highlights the importance of the task.

[6] https://en.wikipedia.org/wiki/Cohen's_kappa

Table 7: Results for the MDR subtask.

Group Name	Run	AUC	Rank AUC	Acc	Rank Acc
VISTA@UEvora	MDR-Run-06-Mohan-SL-F3-Personal.txt	**0.6178**	1	0.5593	8
San Diego VA HCS/UCSD	MDSTest1a.csv	0.6114	2	**0.6144**	1
VISTA@UEvora	MDR-Run-08-Mohan-voteLdaSmoF7-Personal.txt	0.6065	3	0.5424	17
VISTA@UEvora	MDR-Run-09-Sk-SL-F10-Personal.txt	0.5921	4	0.5763	3
VISTA@UEvora	MDR-Run-10-Mix-voteLdaSl-F7-Personal.txt	0.5824	5	0.5593	9
HHU-DBS	MDR_FlattenCNN_DTree.txt	0.5810	6	0.5720	4
HHU-DBS	MDR_FlattenCNN2_DTree.txt	0.5810	7	0.5720	5
HHU-DBS	MDR_Conv68adam_fl.txt	0.5768	8	0.5593	10
VISTA@UEvora	MDR-Run-07-Sk-LDA-F7-Personal.txt	0.5730	9	0.5424	18
UniversityAlicante	MDRBaseline0.csv	0.5669	10	0.4873	32
HHU-DBS	MDR_Conv48sgd.txt	0.5640	11	0.5466	16
HHU-DBS	MDR_Flatten.txt	0.5637	12	0.5678	7
HHU-DBS	MDR_Flatten3.txt	0.5575	13	0.5593	11
UIIP_BioMed	MDR_run_TBdescs2_zparts3_thrprob50_rf150.csv	0.5558	14	0.4576	36
UniversityAlicante	testSVM_SMOTE.csv	0.5509	15	0.5339	20
UniversityAlicante	testOpticalFlowwFrequencyNormalized.csv	0.5473	16	0.5127	24
HHU-DBS	MDR_Conv48sgd_fl.txt	0.5424	17	0.5508	15
HHU-DBS	MDR_CustomCNN_DTree.txt	0.5346	18	0.5085	26
HHU-DBS	MDR_FlattenX.txt	0.5322	19	0.5127	25
HHU-DBS	MDR_MultiInputCNN.txt	0.5274	20	0.5551	13
VISTA@UEvora	MDR-Run-01-sk-LDA.txt	0.5260	21	0.5042	28
MedGIFT	MDR_Riesz_std_correlation_TST.csv	0.5237	22	0.5593	12
MedGIFT	MDR_HOG_std_euclidean_TST.csv	0.5205	23	0.5932	2
VISTA@UEvora	MDR-Run-05-Mohan-RF-F3I650.txt	0.5116	24	0.4958	30
MedGIFT	MDR_AllFeats_std_correlation_TST.csv	0.5095	25	0.4873	33
UniversityAlicante	DecisionTree25v2.csv	0.5049	26	0.5000	29
MedGIFT	MDR_AllFeats_std_euclidean_TST.csv	0.5039	27	0.5424	19
LIST	MDRLIST.txt	0.5029	28	0.4576	37
UniversityAlicante	testOFFullVersion2.csv	0.4971	29	0.4958	31
MedGIFT	MDR_HOG_mean_correlation_TST.csv	0.4941	30	0.5551	14
MedGIFT	MDR_Riesz_AllCols_correlation_TST.csv	0.4855	31	0.5212	22
UniversityAlicante	testOpticalFlowFull.csv	0.4845	32	0.5169	23
MedGIFT	MDR_Riesz_mean_euclidean_TST.csv	0.4824	33	0.5297	21
UniversityAlicante	testFrequency.csv	0.4781	34	0.4788	34
UniversityAlicante	testflowI.csv	0.4740	35	0.4492	39
MedGIFT	MDR_HOG_AllCols_euclidean_TST.csv	0.4693	36	0.5720	6
VISTA@UEvora	MDR-Run-06-Sk-SL.txt	0.4661	37	0.4619	35
MedGIFT	MDR_AllFeats_AllCols_correlation_TST.csv	0.4568	38	0.5085	27
VISTA@UEvora	MDR-Run-04-Mix-Vote-L-RT-RF.txt	0.4494	39	0.4576	38

5 The VQA-Med Task

5.1 Task Description

Visual Question Answering is a new and exciting problem that combines natural language processing and computer vision techniques. Inspired by the recent success of visual question answering in the general domain[7] [3], we propose a pilot task to focus on visual question answering in the medical domain (VQA-Med). Given medical images accompanied with clinically relevant questions, participating systems were tasked with answering questions based on the visual image content. Figure 2 shows a few example images with associated questions and ground truth answers.

[7] http://www.visualqa.org/

Table 8: Results for the TBT subtask.

Group Name	Run	Kappa	Rank Kappa	Acc	Rank Acc
UIIP_BioMed	TBT_run_TBdescs2_zparts3_thrprob50_rf150.csv	**0.2312**	1	**0.4227**	1
fau_ml4cv	TBT_m4_weighted.txt	0.1736	2	0.3533	10
MedGIFT	TBT_AllFeats_std_euclidean_TST.csv	0.1706	3	0.3849	2
MedGIFT	TBT_Riesz_AllCols_euclidean_TST.csv	0.1674	4	0.3849	3
VISTA@UEvora	TBT-Run-02-Mohan-RF-F20I1500S20-317.txt	0.1664	5	0.3785	4
fau_ml4cv	TBT_m3_weighted.txt	0.1655	6	0.3438	12
VISTA@UEvora	TBT-Run-05-Mohan-RF-F20I2000S20.txt	0.1621	7	0.3754	5
MedGIFT	TBT_AllFeats_AllCols_correlation_TST.csv	0.1531	8	0.3691	7
MedGIFT	TBT_AllFeats_mean_euclidean_TST.csv	0.1517	9	0.3628	8
MedGIFT	TBT_Riesz_std_euclidean_TST.csv	0.1494	10	0.3722	6
San Diego VA HCS/UCSD	Task2Submission64a.csv	0.1474	11	0.3375	13
San Diego VA HCS/UCSD	TBTTask_2_128.csv	0.1454	12	0.3312	15
MedGIFT	TBT_AllFeats_AllCols_correlation_TST.csv	0.1356	13	0.3628	9
VISTA@UEvora	TBT-Run-03-Mohan-RF-7FF20I1500S20-Age.txt	0.1335	14	0.3502	11
San Diego VA HCS/UCSD	TBTLast.csv	0.1251	15	0.3155	20
fau_ml4cv	TBT_w_combined.txt	0.1112	16	0.3028	22
VISTA@UEvora	TBT-Run-06-Mix-RF-5FF20I2000S20.txt	0.1005	17	0.3312	16
VISTA@UEvora	TBT-Run-04-Mohan-VoteRFLMT-7F.txt	0.0998	18	0.3186	19
MedGIFT	TBT_HOG_AllCols_euclidean_TST.csv	0.0949	19	0.3344	14
fau_ml4cv	TBT_combined.txt	0.0898	20	0.2997	23
MedGIFT	TBT_HOG_std_correlation_TST.csv	0.0855	21	0.3218	18
fau_ml4cv	TBT_m2p01_small.txt	0.0839	22	0.2965	25
MedGIFT	TBT_AllFeats_std_correlation_TST.csv	0.0787	23	0.3281	17
fau_ml4cv	TBT_m2.txt	0.0749	24	0.2997	24
MostaganemFSEI	TBT_mostaganemFSEI_run4.txt	0.0629	25	0.2744	27
MedGIFT	TBT_HOG_std_correlation_TST.csv	0.0589	26	0.3060	21
fau_ml4cv	TBT_modelsimple_lmbdap1_norm.txt	0.0504	27	0.2839	26
MostaganemFSEI	TBT_mostaganemFSEI_run1.txt	0.0412	28	0.2650	29
MostaganemFSEI	TBT_MostaganemFSEI_run2.txt	0.0275	29	0.2555	32
MostaganemFSEI	TBT_MostaganemFSEI_run6.txt	0.0210	30	0.2429	33
UniversityAlicante	3nnconProbabilidad2.txt	0.0204	31	0.2587	30
UniversityAlicante	T23nnFinal.txt	0.0204	32	0.2587	31
fau_ml4cv	TBT_m1.txt	0.0202	33	0.2713	28
LIST	TBTLIST.txt	-0.0024	34	0.2366	34
MostaganemFSEI	TBT_mostaganemFSEI_run3.txt	-0.0260	35	0.1514	37
VISTA@UEvora	TBT-Run-01-sk-LDA-Update-317-New.txt	-0.0398	36	0.2240	35
VISTA@UEvora	TBT-Run-01-sk-LDA-Update-317.txt	-0.0634	37	0.1956	36
UniversityAlicante	T2SVMFinal.txt	-0.0920	38	0.1167	38
UniversityAlicante	SVMirene.txt	-0.0923	39	0.1136	39

5.2 Dataset

We considered medical images along with their captions extracted from PubMed Central articles[8] (essentially a subset of the ImageCLEF 2017 caption prediction task [13]) to create the datasets for the proposed VQA-Med task.

We used a semi-automatic approach to generate question-answer pairs from captions of the medical images. First, we automatically generated all possible question-answer pairs from captions using a rule-based question generation (QG) system[9]. The candidate questions generated via the automatic approach contained noise due to rule mismatch with the clinical domain sentences. Therefore,

[8] https://www.ncbi.nlm.nih.gov/pmc/
[9] http://www.cs.cmu.edu/~ark/mheilman/questions/

Table 9: Results for the SVR subtask.

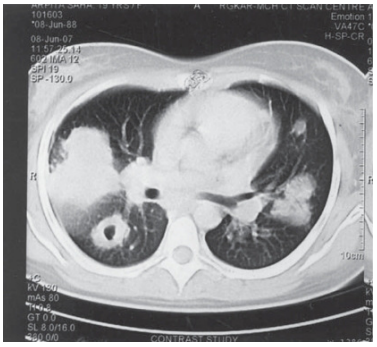
Group Name	Run	RMSE	Rank RMSE	AUC	Rank AUC
UIIP_BioMed	SVR_run_TBdescs2_zparts3_thrprob50_rf100.csv	**0.7840**	1	0.7025	6
MedGIFT	SVR_HOG_std_euclidean_TST.csv	0.8513	2	0.7162	5
VISTA@UEvora	SVR-Run-07-Mohan-MLP-6FTT100.txt	0.8883	3	0.6239	21
MedGIFT	SVR_AllFeats_AllCols_euclidean_TST.csv	0.8883	4	0.6733	10
MedGIFT	SVR_AllFeats_AllCols_correlation_TST.csv	0.8934	5	**0.7708**	1
MedGIFT	SVR_HOG_mean_euclidean_TST.csv	0.8985	6	0.7443	3
MedGIFT	SVR_HOG_mean_correlation_TST.csv	0.9237	7	0.6450	18
MedGIFT	SVR_HOG_AllCols_euclidean_TST.csv	0.9433	8	0.7268	4
MedGIFT	SVR_HOG_AllCols_correlation_TST.csv	0.9433	9	0.7608	2
HHU-DBS	SVR_RanFrst.txt	0.9626	10	0.6484	16
MedGIFT	SVR_Riesz_AllCols_correlation_TST.csv	0.9626	11	0.5535	34
MostaganemFSEI	SVR_mostaganemFSEI_run3.txt	0.9721	12	0.5987	25
HHU-DBS	SVR_RanFRST_depth_2_new_new.txt	0.9768	13	0.6620	13
HHU-DBS	SVR_LinReg_part.txt	0.9768	14	0.6507	15
MedGIFT	SVR_AllFeats_mean_euclidean_TST.csv	0.9954	15	0.6644	12
MostaganemFSEI	SVR_mostaganemFSEI_run6.txt	1.0046	16	0.6119	23
VISTA@UEvora	SVR-Run-03-Mohan-MLP.txt	1.0091	17	0.6371	19
MostaganemFSEI	SVR_mostaganemFSEI_run4.txt	1.0137	18	0.6107	24
MostaganemFSEI	SVR_mostaganemFSEI_run1.txt	1.0227	19	0.5971	26
MedGIFT	SVR_Riesz_std_correlation_TST.csv	1.0492	20	0.5841	29
VISTA@UEvora	SVR-Run-06-Mohan-VoteMLPSL-5F.txt	1.0536	21	0.6356	20
VISTA@UEvora	SVR-Run-02-Mohan-RF.txt	1.0580	22	0.5813	31
MostaganemFSEI	SVR_mostaganemFSEI_run2.txt	1.0837	23	0.6127	22
Middlesex University	SVR-Gao-May4.txt	1.0921	24	0.6534	14
HHU-DBS	SVR_RanFRST_depth_2_Ludmila_new_new.txt	1.1046	25	0.6862	8
VISTA@UEvora	SVR-Run-05-Mohan-RF-3FI300S20.txt	1.1046	26	0.5812	32
VISTA@UEvora	SVR-Run-04-Mohan-RF-F5-I300-S200.txt	1.1088	27	0.5793	33
VISTA@UEvora	SVR-Run-01-sk-LDA.txt	1.1770	28	0.5918	27
HHU-DBS	SVR_RanFRST_depth_2_new.txt	1.2040	29	0.6484	17
San Diego VA HCS/UCSD	SVR9.csv	1.2153	30	0.6658	11
San Diego VA HCS/UCSD	SVRSubmission.txt	1.2153	31	0.6984	7
HHU-DBS	SVR_DTree_Features_Best_Bin.txt	1.3203	32	0.5402	36
HHU-DBS	SVR_DTree_Features_Best.txt	1.3203	33	0.5848	28
HHU-DBS	SVR_DTree_Features_Best_All.txt	1.3714	34	0.6750	9
MostaganemFSEI	SVR_mostaganemFSEI.txt	1.4207	35	0.5836	30
Middlesex University	SVR-Gao-April27.txt	1.5145	36	0.5412	35

two expert human annotators manually checked all generated question-answer pairs associated with the medical images in two passes. In the first pass, syntactic and semantic correctness were ensured while in the second pass, well-curated validation and test sets were generated by verifying the clinical relevance of the questions with respect to associated medical images.

The final curated corpus was comprised of 6,413 question-answer pairs associated with 2,866 medical images. The overall set was split into 5,413 question-answer pairs (associated with 2,278 medical images) for training, 500 question-answer pairs (associated with 324 medical images) for validation, and 500 questions (associated with 264 medical images) for testing.

5.3 Participating Groups and Runs Submitted

Out of 58 online registrations, 28 participants submitted signed end user agreement forms. Finally, 5 groups submitted a total of 17 runs, indicating a consider-



Question: What does the CT scan of thorax show?
Answer: bilateral multiple pulmonary nodules

Question: Is the lesion associated with a mass effect?
Answer: no

Fig. 2: Example images with question-answer pairs in the VQA-Med task.

able interest in the VQA-Med task. Table 10 gives an overview of all participants and the number of submitted runs[10].

5.4 Results

The evaluation of the participant systems of the VQA-Med task was conducted based on three metrics: BLEU, WBSS (Word-based Semantic Similarity), and CBSS (Concept-based Semantic Similarity) [19]. BLEU [26] is used to capture the similarity between a system-generated answer and the ground truth answer.

[10] There was a limit of maximum 5 run submissions per team.

Table 10: Participating groups in the VQA-Med task.

Team	Institution	#Runs
FSTT	Abdelmalek Essaadi University, Faculty of Sciences and Techniques, Tangier, Morocco	2
JUST	Jordan University of Science and Technology, Jordan	3
NLM	Lister Hill National Center for Biomedical Communications, National Library of Medicine, Bethesda, MD, USA	5
TU	Tokushima University, Japan	3
UMMS	University of Massachusetts Medical School, Worcester, MA, USA	4

The overall methodology and resources for the BLEU metric are essentially similar to the ImageCLEF 2017 caption prediction task[11]. The WBSS metric is created based on Wu-Palmer Similarity (WUPS[12]) [43] with WordNet ontology in the backend by following a recent algorithm to calculate semantic similarity in the biomedical domain [31]. WBSS computes a similarity score between a system-generated answer and the ground truth answer based on word-level similarity. CBSS is similar to WBSS, except that instead of tokenizing the system-generated and ground truth answers into words, we use MetaMap[13] via the pymetamap wrapper[14] to extract biomedical concepts from the answers, and build a dictionary using these concepts. Then, we build one-hot vector representations of the answers to calculate their semantic similarity using the cosine similarity measure.

The overall results of the participating systems are presented in Table 11a to Table 11c for the three metrics in a descending order of the scores (the higher the better).

Table 11: Scores of all submitted runs in the VQA-Med task.

(a) BLEU

Team	Run ID	BLEU
UMMS	6113	0.162
UMMS	5980	0.160
UMMS	6069	0.158
UMMS	6091	0.155
TU	5994	0.135
NLM	6084	0.121
NLM	6135	0.108
TU	5521	0.106
NLM	6136	0.106
TU	6033	0.103
NLM	6120	0.085
NLM	6087	0.083
JUST	6086	0.061
FSTT	6183	0.054
JUST	6038	0.048
JUST	6134	0.036
FSTT	6220	0.028

(b) WBSS

Team	Run ID	WBSS
UMMS	6069	0.186
UMMS	6113	0.185
UMMS	5980	0.184
UMMS	6091	0.181
NLM	6084	0.174
TU	5994	0.174
NLM	6135	0.168
TU	5521	0.160
NLM	6136	0.157
TU	6033	0.148
NLM	6120	0.144
NLM	6087	0.130
JUST	6086	0.122
JUST	6038	0.104
FSTT	6183	0.101
JUST	6134	0.094
FSTT	6220	0.080

(c) CBSS

Team	Run ID	CBSS
NLM	6120	0.338
TU	5521	0.334
TU	5994	0.330
NLM	6087	0.327
TU	6033	0.324
FSTT	6183	0.269
FSTT	6220	0.262
NLM	6136	0.035
NLM	6084	0.033
NLM	6135	0.032
JUST	6086	0.029
UMMS	6069	0.023
UMMS	5980	0.021
UMMS	6091	0.017
UMMS	6113	0.016
JUST	6038	0.015
JUST	6134	0.011

[11] http://www.imageclef.org/2017/caption
[12] https://datasets.d2.mpi-inf.mpg.de/mateusz14visualturing/calculate_wups.py
[13] https://metamap.nlm.nih.gov/
[14] https://github.com/AnthonyMRios/pymetamap

5.5 Lessons Learned and Next Steps

In general, participants used deep learning techniques to build their VQA-Med systems [19]. In particular, participant systems leveraged sequence to sequence learning and encoder-decoder-based frameworks utilizing deep convolutional neural networks (CNN) to encode medical images and recurrent neural networks (RNN) to generate question encoding. Some participants used attention-based mechanisms to identify relevant image features to answer the given questions. The submitted runs also varied with the use of various VQA networks such as stacked attention networks (SAN), the use of advanced techniques such as multimodal compact bilinear (MCB) pooling or multimodal factorized bilinear (MFB) pooling to combine multimodal features, the use of different hyperparameters etc. Participants did not use any additional datasets except the official training and validation sets to train their models.

The relatively low BLEU scores and WBSS scores of the runs in the results table denote the difficulty of the VQA-Med task in generating similar answers as the ground truth, while higher CBSS scores suggest that some participants were able to generate relevant clinical concepts in their answers similar to the clinical concepts present in the ground truth answers. To leverage the power of advanced deep learning algorithms towards improving the state-of-the-art in visual question answering in the medical domain, we plan to increase the dataset size in the future editions of this task.

6 The Lifelog Task

6.1 Motivation and Task Setup

An increasingly wide range of personal devices, such as smart phones, video cameras as well as wearable devices that allow capturing pictures, videos, and audio clips pf every moment of life have now become inseparable companions and, considering the huge volume of data created, there is an urgent need for systems that can automatically analyze the data in order to categorize, summarize and also retrieve information that the user may require. This kind of data, commonly referred to as *lifelogs*, gathered increasing attention in recent years within the research community above all because of the precious information that can be extracted from this kind of data and for the remarkable effects in the technological and social field.

Despite the increasing number of successful related workshops and panels (e.g., JCDL 2015[15], iConf 2016[16], ACM MM 2016[17], ACM MM 2017[18]) lifelogging has seldom been the subject of a rigorous comparative benchmarking

[15] http://www.jcdl.org/archived-conf-sites/jcdl2015/www.jcdl2015.org/panels.html
[16] http://irlld2016.computing.dcu.ie/index.html
[17] http://lta2016.computing.dcu.ie
[18] http://lta2017.computing.dcu.ie

exercise as, for example, the lifelog evaluation task at NTCIR-14[19] or last year's edition of the ImageCLEFlifelog task [6]. Also in this second edition of the task we aim to bring the attention of lifelogging to a wider audience and to promote research into some of its key challenges such as on multi-modal analysis of large data collections. The ImageCLEF 2018 LifeLog task [7] aims to be a comparative evaluation of information access and retrieval systems operating over personal lifelog data. The task consists of two sub-tasks and both allow participation independently. These sub-tasks are:

- Lifelog moment retrieval (LMRT);
- Activities of Daily Living understanding (ADLT).

Lifelog moment retrieval task (LMRT)
The participants have to retrieve a number of specific moments in a lifelogger's life. "Moments" were defined as semantic events or activities that happened throughout the day. For example, participants should return the relevant moments for the query "*Find the moment(s) when I was shopping for wine in the supermarket.*" Particular attention should 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.

Activities of daily living understanding task (ADLT)
The participants should analyze the lifelog data from a given period of time (e.g., *"From August 13 to August 16"* or *"Every Saturday"*) and provide a summarization based on the selected concepts provided by the task organizers of Activities of Daily Living (ADL) and the environmental settings / contexts in which these activities take place.

In the following it is possible to see some examples of ADL concepts:

- *"Commuting (to work or another common venue)"*
- *"Traveling (to a destination other than work, home or another common social event)"*
- *"Preparing meals (include making tea or coffee)"*
- *"Eating/drinking"*

Some examples of contexts are:

- *"In an office environment"*
- *"In a home"*
- *"In an open space"*

The summarization is described as the total duration and the number of times the queried concepts happens.

- ADL: "Eating/drinking: 6 times, 90 minutes", "Traveling: 1 time, 60 minutes".
- Context: "In an office environment: 500 minutes", "In a church: 30 minutes".

[19] http://ntcir-lifelog.computing.dcu.ie

Table 12: Statistics of ImageCLEFlifelog2018 Dataset.

Size of the Collection	**18.854** GB
Number of Images	**80,440** images
Number of Known Locations	**135** locations
Concepts	Fully annotated (by Microsoft Computer Vision API)
Biometrics	Fully provided (24×7)
Human Activities	Provided
Number of ADLT Topics	**20** (10 for devset, 10 for testset)
Number of LMRT Topics	**20** (10 for devset, 10 for testset)

6.2 Dataset Employed

This year a completely new multimodal dataset was provided to participants. This consists of 50 days of data from a lifelogger. The data contain a large collection of wearable camera images (1,500-2,500 per day), visual concepts (automatically extracted visual concepts with varying rates of accuracy), semantic content (semantic locations, semantic activities) based on sensor readings (via the Moves App) on mobile devices, biometric information (heart rate, galvanic skin response, calorie burn, steps, etc.), music listening history. The dataset is built based on the data available for the NTCIR-13 - Lifelog 2 task [16]. A summary of the data collection is shown in Table 12.

Evaluation Methodology

For assessing performance in the *Lifelog moment retrieval task* classic metrics were employed. These metrics are:

- Cluster Recall at $X(CR@X)$ — a metric that assesses how many different clusters from the ground truth are represented among the top X results;
- Precision at $X(P@X)$ — measures the number of relevant photos among the top X results;
- F1-measure at $X(F1@X)$ — the harmonic mean of the previous two measures.

Various cut off points were considered, e.g., $X = 5, 10, 20, 30, 40, 50$. Official ranking metric this year was the **F1-measure@10**, which gives equal importance to diversity (via $CR@10$) and relevance (via $P@10$).

Participants were allowed to undertake the sub-tasks in an interactive or automatic manner. For interactive submissions, a maximum of five minutes of search time is allowed per topic. In particular, the organizers would like to emphasize methods that allow interaction with real users (via Relevance Feedback, RF, for example), i.e., beside the best performance, the method of interaction (e.g. the number of iterations using relevance feedback), or innovation level of the method (for example, new way to interact with real users) are encouraged.

In the *Activities of daily living understanding*, the evaluation metric is the percentage of dissimilarity between the ground-truth and the submitted values, measured as average of the time and minute differences, as follows:

Table 13: Submitted runs for ImageCLEFlifelog2018 LMRT task.

Team	Run Name	F1@10
Organizers [45]	Run 1*	0.077
	Run 2*	0.131
	Run 3*,†	0.407
	Run 4*,†	0.378
	Run 5*,†	0.365
AILab-GTI [24]	Subm#1	0.504
	Subm#2	**0.545**
	Subm#3	0.477
	Subm#4	0.536
	Subm#5	0.477
	Subm#6	0.480
	exps5	0.512
	Subm#0†	0.542
Regim Lab [1]	Run 1	0.065
	Run 2	0.364
	Run 3	0.411
	Run 4	0.411
	Run 5	0.424
NLP-Lab [34]	Run 1	0.177
	Run 3	0.223
	Run 4	0.395
	Run 5	0.354
HCMUS [35]	Run 1	0.355
	Run 2	0.479
CAMPUS-UPB [12]	Run 1	0.216
	Run 2†	0.169
	Run 3†	0.168
	Run 4†	0.166
	Run 5†	0.443

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

$$ADL_{score} = \frac{1}{2}\left(max(0, 1 - \frac{|n-n_{gt}|}{n_{gt}}) + max(0, 1 - \frac{|m-m_{gt}|}{m_{gt}})\right)$$

where n, n_{gt} are the submitted and ground-truth values for how many times the events occurred, respectively, and m, m_{gt} are the submitted and ground-truth values for how long (in minutes) the events happened, respectively.

6.3 Participating Groups and Runs Submitted

This year the number of participants was considerably higher with respect to 2017: we received in total 41 runs: 29 (21 official, 8 additional) for LMRT and 12 (8 official, 4 additional) for ADLT, from 7 teams from Brunei, Taiwan, Vietnam, Greece-Spain, Tunisia, Romania, and a multi-nation team from Ireland, Italy, Austria, and Norway. The received approaches range from fully automatic to fully manual, from using a single information source provided by the task to using all information as well as integrating additional resources, from traditional learning methods (e.g. SVMs) to deep learning and ad-hoc rules. Submitted runs and their results are summarized in Tables 13 and 14.

Table 14: Submitted runs for ImageCLEFlifelog2018 ADLT task.

Team	Run Name	Score (% dissimilarity)
Organizers [45]	Run 1*	0.816
	Run 2*,†	0.456
	Run 3*,†	0.344
	Run 4*,†	0.481
	Run 5*,†	0.485
CIE@UTB [8]	Run 1	**0.556**
NLP-Lab [34]	Run 1	0.243
	Run 2	0.285
	Run 3	0.385
	Run 4	0.459
	Run 5	0.479
HCMUS [35]	Run 1	0.059

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 the majority of the approaches this year exploit and combine visual, text, location and other information to solve the task, which is different from last year when often only one type of data was analysed. Furthermore, we learned that lifelogging is following the trend in data analytics, meaning that participants are using deep learning in many cases. However, there still is room for improvement, since the best results are coming from the fine-tuned queries, which means we need more advanced techniques on bridging the gap between the abstract of human needs and the multi-modal data. Regarding the number of the signed-up teams and the submitted runs, we received a significant improvement compared to last year. This shows how interesting and challenging lifelog data is and that it holds much research potential. As next steps we do not plan to enrich the dataset but rather provide richer data and narrow down the application of the challenges (e.g., extend to health-care application).

7 Conclusions

This paper presents a general overview of the activities and outcomes of the ImageCLEF 2018 evaluation campaign. Four tasks were organised covering challenges in: caption prediction, tuberculosis type and drug resistance detection, medical visual question answering and lifelog retrieval.

The participation increased slightly compared to 2017, with over 130 signed user agreements, and in the end 31 groups submitting results. This is remarkable as three of the tasks are only in the second edition and one was in the first edition. 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.

As is now becoming commonplace, many of the participants employ deep neural networks to address all proposed tasks. In the tuberculosis task, the results

in multi-drug resistance are still limited for practical use, though good performance was obtained in the new severity scoring subtask. In the visual question answering task the scores were relatively low, even though some approaches do seem to predict concepts present. In the lifelog task, in contrast to the previous year, several approaches used a combination of visual, text, location and other information.

The use of crowdAI was a change for many of the traditional participants and created many questions and also much work for the task organizers. On the other hand it is a much more modern platform that offers new possibilities, for example continuously running the challenge even beyond the workshop dates. The benefits of this will likely only be seen in the coming years.

ImageCLEF 2018 again brought together an interesting mix of tasks and approaches and we are looking forward to the fruitful discussions at the workshop.

Acknowledgements

Bogdan Ionescu — part of this work was supported by the Ministry of Innovation and Research, UEFISCDI, project SPIA-VA, agreement 2SOL/2017, grant PN-III-P2-2.1-SOL-2016-02-0002.
Duc-Tien Dang-Nguyen, Liting Zhou and Cathal Gurrin — part of this work has emanated from research supported in part by research grants from the Irish Research Council (IRC) under Grant Number GOIPG/2016/741 and Science Foundation Ireland (SFI) under grant number SFI/12/RC/2289.

References

1. Abdallah, F.B., Feki, G., Ezzarka, M., Ammar, A.B., Amar, C.B.: Regim Lab Team at ImageCLEFlifelog LMRT Task 2018 (September 10-14 2018)
2. Andrearczyk, V., Henning, M.: Deep multimodal classification of image types in biomedical journal figures. In: International Conference of the Cross-Language Evaluation Forum (CLEF) (2018)
3. Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C.L., Parikh, D.: VQA: Visual Question Answering. In: International Conference on Computer Vision (ICCV) (2015)
4. Balikas, G., Krithara, A., Partalas, I., Paliouras, G.: BioASQ: A challenge on large–scale biomedical semantic indexing and question answering. In: Multimodal Retrieval in the Medical Domain (MRMD) 2015. Lecture Notes in Computer Science, Springer (2015)
5. 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)
6. Dang-Nguyen, D.T., Piras, L., Riegler, M., Boato, G., Zhou, L., Gurrin, C.: Overview of ImageCLEFlifelog 2017: Lifelog Retrieval and Summarization. In: CLEF 2017 Labs Working Notes. CEUR Workshop Proceedings, CEUR-WS.org <http://ceur-ws.org>, Dublin, Ireland (September 11-14 2017)

7. Dang-Nguyen, D.T., Piras, L., Riegler, M., Zhou, L., Lux, M., Gurrin, C.: Overview of ImageCLEFlifelog 2018: Daily Living Understanding and Lifelog Moment Retrieval. In: CLEF 2018 Working Notes. CEUR Workshop Proceedings, CEUR-WS.org <http://ceur-ws.org>, Avignon, France (September 10-14 2018)
8. Dao, M.S., Kasem, A., Nazmudeen, M.S.H.: Leveraging Content and Context to Foster Understanding of Activities of Daily Living (September 10-14 2018)
9. 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. No. 1390 in CEUR Workshop Proceedings (Apr 2015)
10. Dicente Cid, Y., Kalinovsky, A., Liauchuk, V., Kovalev, V., , Müller, H.: Overview of ImageCLEFtuberculosis 2017 - predicting tuberculosis type and drug resistances. In: CLEF 2017 Labs Working Notes. CEUR Workshop Proceedings, CEUR-WS.org <http://ceur-ws.org>, Dublin, Ireland (September 11-14 2017)
11. 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: CLEF 2018 Working Notes. CEUR Workshop Proceedings, CEUR-WS.org <http://ceur-ws.org>, Avignon, France (September 10-14 2018)
12. Dogariu, M., Ionescu, B.: Multimedia Lab @ CAMPUS at ImageCLEFlifelog 2018 Lifelog Moment Retrieval (September 10-14 2018)
13. Eickhoff, C., Schwall, I., García Seco de Herrera, A., Müller, H.: Overview of ImageCLEFcaption 2017 - image caption prediction and concept detection for biomedical images. In: CLEF 2017 Labs Working Notes. CEUR Workshop Proceedings, CEUR-WS.org <http://ceur-ws.org>, Dublin, Ireland (September 11-14 2017)
14. Goeuriot, L., Kelly, L., Suominen, H., Névéol, A., Robert, A., Kanoulas, E., Spijker, R., Palotti, J., Zuccon, G.: Clef 2017 ehealth evaluation lab overview. In: Jones, G.J., Lawless, S., Gonzalo, J., Kelly, L., Goeuriot, L., Mandl, T., Cappellato, L., Ferro, N. (eds.) Experimental IR Meets Multilinguality, Multimodality, and Interaction. pp. 291–303. Springer International Publishing, Cham (2017)
15. Gravier, G., Bischke, B., Demarty, C.H., Zaharieva, M., Riegler, M., Dellandrea, E., Bogdanov, D., Sutcliffe, R., Jones, G.J., Larson, M.: Working notes proceedings of the mediaeval 2017 workshop. In: MediaEval 2017 Working Notes. CEUR Workshop Proceedings, CEUR-WS.org <http://ceur-ws.org/Vol-1984>, Dublin, Ireland (September 13-15 2017)
16. Gurrin, C., Joho, H., Hopfgartner, F., Zhou, L., Gupta, R., Albatal, R., Dang-Nguyen, D.T.: Overview of NTCIR-13 Lifelog-2 Task. In: Proceedings of the 13th NTCIR Conference on Evaluation of Information Access Technologies (2017)
17. Hanbury, A., Müller, H., Balog, K., Brodt, T., Cormack, G.V., Eggel, I., Gollub, T., Hopfgartner, F., Kalpathy-Cramer, J., Kando, N., Krithara, A., Lin, J., Mercer, S., Potthast, M.: Evaluation–as–a–service: Overview and outlook. ArXiv 1512.07454 (2015)
18. Hanbury, A., Müller, H., Langs, G., Weber, M.A., Menze, B.H., Fernandez, T.S.: Bringing the algorithms to the data: cloud–based benchmarking for medical image analysis. In: CLEF conference. Springer Lecture Notes in Computer Science (2012)
19. 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)

20. García Seco de Herrera, A., Eickhoff, C., Andrearczyk, V., , Müller, H.: Overview of the ImageCLEF 2018 caption prediction tasks. In: CLEF 2018 Working Notes. CEUR Workshop Proceedings, CEUR-WS.org <http://ceur-ws.org>, Avignon, France (September 10-14 2018)
21. Ionescu, B., Müller, H., Villegas, M., Arenas, H., Boato, G., Dang-Nguyen, D.T., Dicente Cid, Y., Eickhoff, C., Garcia Seco de Herrera, A., Gurrin, C., Islam, B., Kovalev, V., Liauchuk, V., Mothe, J., Piras, L., Riegler, M., Schwall, I.: Overview of ImageCLEF 2017: Information extraction from images. In: Experimental IR Meets Multilinguality, Multimodality, and Interaction 8th International Conference of the CLEF Association, CLEF 2017. Lecture Notes in Computer Science, vol. 10456. Springer, Dublin, Ireland (September 11-14 2017)
22. Joly, A., Goëau, H., Glotin, H., Spampinato, C., Bonnet, P., Vellinga, W.P., Lombardo, J.C., Planqué, R., Palazzo, S., Müller, H.: Lifeclef 2017 lab overview: multimedia species identification challenges. In: Proceedings of CLEF 2017 (2017)
23. 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)
24. Kavallieratou, E., del Blanco, C.R., Cuevas, C., García, N.: Retrieving Events in Life Logging (September 10-14 2018)
25. 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)
26. 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)
27. Pinho, E., Costa, C.: Feature learning with adversarial networks for concept detection in medical images: UA.PT Bioinformatics at ImageCLEF 2018. In: CLEF2018 Working Notes. CEUR Workshop Proceedings, CEUR-WS.org <http://ceur-ws.org>, Avignon, France (September 10-14 2018)
28. Rahman, M.M.: A cross modal deep learning based approach for caption prediction and concept detection by cs morgan state. In: CLEF2018 Working Notes. CEUR Workshop Proceedings, CEUR-WS.org <http://ceur-ws.org>, Avignon, France (September 10-14 2018)
29. Sharma, A., Hill, A., Kurbatova, E., van der Walt, M., Kvasnovsky, C., Tupasi, T.E., Caoili, J.C., Gler, M.T., Volchenkov, G.V., Kazennyy, B.Y., Demikhova, O.V., Bayona, J., Contreras, C., Yagui, M., Leimane, V., Cho, S.N., Kim, H.J., Kliiman, K., Akksilp, S., Jou, R., Ershova, J., Dalton, T., Cegielski, P.: Estimating the future burden of multidrug-resistant and extensively drug-resistant tuberculosis in india, the philippines, russia, and south africa: a mathematical modelling study. The Lancet Infectious Diseases 17(7), 707 – 715 (2017), http://www.sciencedirect.com/science/article/pii/S1473309917302475
30. Soldaini, L., Goharian, N.: Quickumls: a fast, unsupervised approach for medical concept extraction. In: MedIR Workshop, SIGIR (2016)
31. Soğancıoğlu, G., Öztürk, H., Özgür, A.: Biosses: a semantic sentence similarity estimation system for the biomedical domain. Bioinformatics 33(14), i49–i58 (2017)
32. Spinks, G., Moens, M.F.: Generating text from images in a smooth representation space. In: CLEF2018 Working Notes. CEUR Workshop Proceedings, CEUR-WS.org <http://ceur-ws.org>, Avignon, France (September 10-14 2018)

33. Su, Y., Liu, F.: UMass at ImageCLEF caption prediction 2018 task. In: CLEF2018 Working Notes. CEUR Workshop Proceedings, CEUR-WS.org <http://ceur-ws.org>, Avignon, France (September 10-14 2018)
34. Tang, T.H., Fu1, M.H., Huang, H.H., Chen, K.T., Chen, H.H.: NTU NLP-Lab at ImageCLEFlifelog 2018: Visual Concept Selection with Textual Knowledge for Understanding Activities of Daily Living and Life Moment Retrieval (September 10-14 2018)
35. Tran, M.T., Truong, T.D., Dinh-Duy, T., Vo-Ho, V.K., Luong, Q.A., Nguyen, V.T.: Lifelog Moment Retrieval with Visual Concept Fusion and Text-based Query Expansion (September 10-14 2018)
36. 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)
37. 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)
38. Valavanis, L., Kalamboukis, T.: IPL at ImageCLEF 2018: A kNN-based concept detection approach. In: CLEF2018 Working Notes. CEUR Workshop Proceedings, CEUR-WS.org <http://ceur-ws.org>, Avignon, France (September 10-14 2018)
39. Villegas, M., Müller, H., Gilbert, A., Piras, L., Wang, J., Mikolajczyk, K., García Seco de Herrera, A., Bromuri, S., Amin, M.A., Kazi Mohammed, M., Acar, B., Uskudarli, S., Marvasti, N.B., Aldana, J.F., Roldán García, M.d.M.: General overview of ImageCLEF at the CLEF 2015 labs. In: Working Notes of CLEF 2015. Lecture Notes in Computer Science, Springer International Publishing (2015)
40. Villegas, M., Müller, H., García Seco de Herrera, A., Schaer, R., Bromuri, S., Gilbert, A., Piras, L., Wang, J., Yan, F., Ramisa, A., Dellandrea, E., Gaizauskas, R., Mikolajczyk, K., Puigcerver, J., Toselli, A.H., Sánchez, J.A., Vidal, E.: General Overview of ImageCLEF at the CLEF 2016 Labs. In: Experimental IR Meets Multilinguality, Multimodality, and Interaction, Lecture Notes in Computer Science, vol. 9822, pp. 267–285. Springer International Publishing (2016)
41. Wang, X., Zhang, Y., Guo, Z., Li, J.: ImageSem at ImageCLEF 2018 caption task: Image retrieval and transfer learning. In: CLEF2018 Working Notes. CEUR Workshop Proceedings, CEUR-WS.org <http://ceur-ws.org>, Avignon, France (September 10-14 2018)
42. Wang, Y.X.J., Chung, M.J., Skrahin, A., Rosenthal, A., Gabrielian, A., Tartakovsky, M.: Radiological signs associated with pulmonary multi-drug resistant tuberculosis: an analysis of published evidences. Quantitative Imaging in Medicine and Surgery 8(2), 161–173 (2018)
43. Wu, Z., Palmer, M.: Verbs semantics and lexical selection. In: Proceedings of the 32nd annual meeting on Association for Computational Linguistics. pp. 133–138. Association for Computational Linguistics (1994)
44. Zhao, J.J., Kim, Y., Zhang, K., Rush, A.M., LeCun, Y.: Adversarially regularized autoencoders for generating discrete structures. CoRR, abs/1706.04223 (2017)
45. Zhou, L., Piras, L., Riegler, M., Lux, M., Dang-Nguyen1, D.T., Gurrin, C.: An interactive lifelog retrieval system for activities of daily living understanding (September 10-14 2018)