Overview of the ImageCLEFmed 2020 Concept Prediction Task: Medical Image Understanding

Obioma Pelka[1,2][0000−0001−5156−4429], Christoph M. Friedrich[1,3][0000−0001−7906−0038], Alba G. Seco de Herrera[4][0000−0002−6509−5325], and Henning Müller[5,6][0000−0001−6800−9878]

[1] Department of Computer Science, University of Applied Sciences and Arts Dortmund, Germany
{obioma.pelka,christoph.friedrich}@fh-dortmund.de
[2] Department of Diagnostic and Interventional Radiology and Neuroradiology, University Hospital Essen, Germany
[3] Institute for Medical Informatics, Biometry and Epidemiology (IMIBE), University Hospital Essen, Germany
[4] University of Essex, UK
alba.garcia@essex.ac.uk
[5] University of Applied Sciences Western Switzerland (HES-SO), Switzerland
henning.mueller@hevs.ch
[6] University of Geneva, Switzerland

Abstract. This paper describes the ImageCLEFmed 2020 Concept Detection Task. After first being proposed at ImageCLEF 2017, the medical task is in its 4th edition this year, as the automatic detection from medical images still remains a challenging task. In 2020, the format remained the same as in 2019, with a single sub-task. The concept detection task is part of the medical tasks, alongside the tuberculosis and visual question and answering tasks. Similar to the 2019 edition, the data set focuses on radiology images rather than biomedical images, however with an increased number of images. The distributed images were extracted from the biomedical open access literature (PubMed Central). The development data consists of 65,753 training and 15,970 validation images. Each image has corresponding Unified Medical Language System (UMLS®) concepts, that were extracted from the original article image captions. In this edition, additional imaging acquisition technique labels were included in the distributed data, which were adopted for pre-filtering steps, concept selection and ensemble algorithms. Most applied approaches for the automatic detection of concepts were deep learning based architectures. Long short-term memory (LSTM) recurrent neural networks (RNN), adversarial auto-encoder, convolutional neural networks (CNN) image encoders and transfer learning-based multi-label classification models were adopted. The performances of the submitted models (best score 0.3940) were evaluated using F1-scores computed per image and averaged across all 3,534 test images.

Keywords: Concept Detection · Computer Vision · ImageCLEF 2020 · Image Understanding · Image Modality · Radiology

1 Introduction

In this paper, the approaches for the detection of Unified Medical Language System (UMLS®) concepts present in radiology images are presented. The task is part of the ImageCLEF[1] bench-marking campaign, that is part of the Cross Language Evaluation Forum[2] (CLEF). Since 2003, the ImageCLEF bench-marking campaign has been proposing several image understanding tasks from different domains every year [4, 15, 11]. Detailed information on other proposed tasks at the ImageCLEF 2020 can be found in Ionescu et al. [9].

The concept detection task in this year is the fourth edition. At ImageCLEFmed Caption 2017 [3] and ImageCLEFmed Caption 2018 [7], the task was comprised of two (2) sub-tasks: concept detection and caption prediction. The format changed in ImageCLEFmed Caption 2019 [16] with the single task of concept detection and remained that way this year at ImageCLEFmed Caption 2020. New in this edition is that the imaging modality is given for each image both in the development and evaluation sets.

As there is an increasing number of medical images available without metadata, for example in the scientific literature, there is an essential need to create systems that can automatically generate such information, hence making the content of these data sets more useful. The purpose of the ImageCLEFmed 2020 concept detection task was to create a platform for the evaluation of systems capable of automatically creating UMLS®concepts of a given radiology image. These predicted information is applicable for data sets that either not labeled or structured, but also for medical data sets lacking textual metadata, as multi-modal approaches prove to obtain better results regarding several image classification tasks [18, 19].

The manual interpretation and generation of knowledge from medical images is not only time-consuming and prone to error, but also impractical. Therefore, the modeling systems that can automatically map visual content present in the images to concise textual representations is a necessity, in regards to efficient information retrieval and image classification.

For development data, both the development and test sets from the ImageCLEFmed Caption 2019 [16] was distributed. This data set is a subset of the Radiology Object in COntext data set (ROCO) [17] and contains solely radiology images that originate from the PubMed Central (PMC) Open Access Subset[3] [20]. Several UMLS®Concept Unique Identifiers (CUIs) are included to each image. The test set used for official evaluation was created in the same manner as proposed in Pelka et al. [17], for generalization purposes.

[1] http://imageclef.org/ [last accessed: 28.07.2020]
[2] http://www.clef-initiative.eu/ [last accessed: 28.07.2020]
[3] https://www.ncbi.nlm.nih.gov/pmc/tools/openftlist/[last accessed: 28.07.2020]

This paper presents an overview of the ImageCLEFmed 2020 Concept Detection Task. Section 2 contains the task description and lists the participating teams. An explorative analysis computed on the distributed development and test data sets is described in Section 3. The framework used to evaluate the submission runs is explained in Section 4. Section 5 displays the modeling approaches applied by the participating teams and the obtained scores, and is followed by discussion and conclusions in Section 6.

2 Task and Participation

Similar to the ImageCLEF caption task in 2019 [16], in ImageCLEF Caption 2020 the focus is on the automatic detection of concepts in a large corpus of radiology images. The proposed task aims to interpret and summarise insights gained from medical images and therefore provide tools for radiology image understanding. The distributed images in both development and evaluation data sets originate from biomedical articles extracted from the PubMed Central (PMC) Open Access Subset[20]. To each radiology image in the distributed data sets, UMLS®CUIs are included. These concepts are generated from the the original image captions found in the articles. Figure 1 displays an example of an image in the distributed data sets. In comparison to the previous tasks, the following improvements were made:

- The imaging modality was included.
- The focus remained on radiology images as in ImageCLEF 2019 .
- The number of concepts was decreased by preprocessing the captions prior to concept extraction.

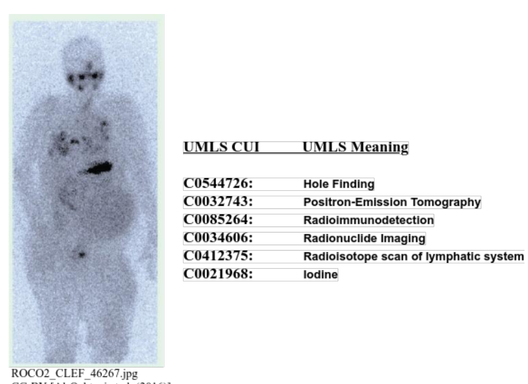


Fig. 1. Example of a radiology image with the corresponding extracted UMLS®CUIs.

The automatic detection of concepts present in images is a fundamental step towards scene understanding and hence image captioning, as the presence of applicable biomedical concepts can be detected and located. As the usage of multi-modal representations (visual and textual) for image classification tasks helps to achieve good performance [19], the automatically generated concepts can be adopted for this purpose. In addition, the concepts can also be used for context-based image analysis, as well as for information retrieval. The detected concepts are evaluated image-wise with precision and recall scores from the ground truth, which is described in Section 4.

Table 1. Participating groups of the ImageCLEF 2020 Concept Detection Task. Teams with previous participation in 2019 are marked with an asterisk.

Team	Institution	Runs
AUEB NLP Group* [12]	Department of Informatics, Athens University of Economics and Business, Athens, Greece	3
PwC_Healthcare [24]	PricewaterhouseCoopers US Advisory, Mumbai, India	9
Essex [6]	School of computer Science and Electronic Engineering, University of Essex, Essex, United Kingdom	9
IML_DFKI [10]	Interactive Machine Learning Group, German Research Center for Artificial Intelligence (DFKI), Saarbrücken, Germany	5
TUC_MC [25]	Technische Unversität Chemnitz, Chemnitz, Germany	10
Morgan_CS [14]	Computer Science Department, Morgan State University, Baltimore, Maryland, United States of America	10
CSE_SSN [2]	Department of Computer Science and Engineering, SSN College of Engineering, Chennai, India	1

In the ImageCLEF 2020 concept detection task a total of 23 unique teams registered in AICrowd and downloaded the End-User-Agreement. This license is needed to obtain access to both development and evaluation data. 57 graded runs were submitted for evaluation by 7 teams from the following countries: Germany, United Kingdom, India, Greece and United States of America, which is listed in Table 2. Each of the groups was allowed 10 graded runs and 5 faulty runs altogether. 10 of the submitted runs were faulty and were not used for the official evaluation.

3 Data Set

As in previous editions, the data set distributed for the task originates from biomedical articles of the PMC Open Access subset [20]. The development data set contains training and validation sets with 65,753 and 15,970 images, respectively. These images are subsets of the multi-modal image data set Radiology

Objects in COntext (ROCO), which is presented in Pelka et al. [17]. ROCO has two classes: Radiology and Out-Of-Class. The first contains 81,825 radiology images and was adopted for the proposed task. It includes several medical imaging modalities such as, Computed Tomography (CT), Ultrasound, X-Ray, Fluoroscopy, Positron Emission Tomography (PET), Mammography, Magnetic Resonance Imaging (MRI), Angiography and PET-CT.

The development data of the 2020 task includes the ImageCLEF caption 2019 development data set (archiving date: until 31.01.2018) and the official evaluation set (archiving date: 01.02.2018 - 01.02.2019). To avoid an overlap with images distributed in previous ImageCLEF medical tasks, the test set for ImageCLEF 2020 was created with a subset of PMC Open Access (archiving date: 01.02.2019 - 01.02.2020). The same procedures applied for the creation of the ROCO data set were applied for the test set as well. An analysis of the distributed data can be seen in Table 2.

Table 2. Analysis on data distribution for ImageCLEFmed 2020 Concept Detection Task.

Imaging Technique	Train	Validation	Test	Sum
DRAN: Angiography	4,713	1,132	325	6,170
DRCO: Combined modalities in one image	487	73	49	609
DRCT: Computerized Tomography	20,031	4,992	1,140	26,163
DRMR: Magnetic Resonance	11,447	2,848	562	14,857
DRPE: Positron emission tomography	502	74	38	614
DRUS: Ultrasound	8,629	2,134	502	11,265
DRXR: X-Ray, 2D radiography	18,944	4,717	918	24,579
Sum	65,753	15,970	3534	84,257

From the PMC Open Access subset [20], a total of 6,031,814 image - caption pairs were extracted in January 2018. Compound figures, which are images with more than one subfigure, were removed using deep learning as proposed in Koitka et al. [13]. The non-compound images were further split into radiology and non-radiology, as the focus was on radiology. Semantic knowledge of object interplay present in the images were extracted in the form of UMLS®Concepts using the QuickUMLS library [23]. The image captions from the biomedical articles served as basis for the extraction of the concepts. The text pre-processing steps applied are described in Pelka et al. [17]. Using deep learning systems as proposed in Koitka et al. [13], the radiology images were further split into seven (7) imaging modality classes. This information can be used for filtering steps prior to model training, as well as for model fine-tuning.

An additional UMLS®CUI denoting the imaging technique modality was added to each image. Figure 2 shows example images from the development data set, according to image modality and additional UMLS®CUI. Similarly to the caption task in 2019 [16], concepts with very high frequency (>13,000), as well as redundant synonyms were removed. This lead to a reduction of concepts per image in comparison to the previous years, from 5,528 in 2019 [16] to 3,047 in 2020. Not all concepts in the ground truth can be visually seen, for example the concept 'Hole Finding' in Fig. 2 can not be detected from the image. Images in the training, validation and test sets have [1-140], [1-142] and [1-95] concepts, respectively. All concepts in the validation and test sets also exist in the training set.

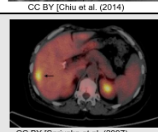
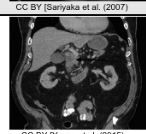
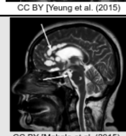
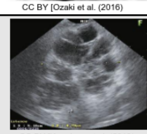
Modality	Example	UMLS CUI	UMLS Meaning
DRAN	 CC BY [Chiu et al. (2014)	C002978	Angiogram
DRCO	 CC BY [Sariyaka et al. (2007)	N/A	N/A
DRCT	 CC BY [Yeung et al. (2015)	C0040398 C0040405	Tomography X-Ray Computed Tomography
DRMR	 CC BY [Mahale et al. (2015)	C0024485	Magnetic Resonance Imaging
DRPE	 CC BY [Ozaki et al. (2016)	C0032743	Positron-Emission Tomography
DRUS	 CC BY [Kongkam et al. (2013)	C0041618	Ultrasonography
DRXR	 CC BY [Tulay et al. (2018)	C0043299	Diagnostic radiologic examination

Fig. 2. Examples of radiology images distributed at the ImageCLEF 2020 concept detection task, showing the seven imaging modalities. All images were randomly selected from the development data set.

Table 3. UMLS®(An excerpt of Unified Medical Language System ®) Concept Unique Identifiers (CUIs) distributed for the task with their respective occurrences. The concepts were randomly chosen in a descending order. All listed concepts were distributed in the training set.

CUI	Concept	Occurrence
C0040398	Tomography	20,031
C0040405	X-Ray Computed Tomography	20,031
C0043299	Diagnostic radiologic examination	18,944
C0024485	Magnetic Resonance Imaging	11,447
C0041618	Ultrasound	8,629
C0441633	Scanning	6733
C0043299	Diagnostic radiologic examination	6321
C1962945	Radiographic imaging procedure	6318
C0040395	Tomography	6235
C0034579	Panoramic Radiography	6127
C0817096	Chest	5981
C0040405	X-Ray Computed Tomography	5801
C1548003	Diagnostic Service Section ID - Radiograph	5159
...	...	...
C0000726	Abdomen	2297
...	...	...
C2985765	Enhancement Description	1084
...	...	...
C0228391	Structure of habenulopeduncular tract	672
C0729233	Dissecting aneurysm of the thoracic aorta	652
...	...	...
C0771711	Pancreas extract	456
...	...	...
C1704302	Permanent premolar tooth	177
...	...	...
C0042070	Urography	67
C0085632	Apathy	67
C0267716	Incisional hernia	67
...	...	...
C0081923	Cardiocrome	1
C0193959	Tonsillectomy and adenoidectomy	1

4 Evaluation Methodology

For all 3,534 radiology images distributed in the test set, UMLS®CUIs have to be predicted by the participating teams automatically. As in the previous years [3, 7, 16], the model performance was measured using the balanced precision and recall trade-off in terms of F1-score. The default implementation of the Python scikit-learn (v0.17.1-2) library was applied to compute the F-scores per image and average them across all test images.

The maximum number of concepts allowed per image was set to 150. This limitation was chosen as the training, validation and test set contain a maximum of 140, 142 and 95 concepts per image. Each group could have a maximum of 15 submission, with 10 valid and 5 faulty. Faulty submissions may include:

- Same image id more than once
- Wrong image id
- Too many concepts
- Same concept more than once
- Not all test images included

All submission runs were uploaded by the participating teams and evaluated with AICrowd[4]. The source code of the evaluation tool is available on the ImageCLEF web page[5].

5 Results

The overall performance achieved by the concepts detection models submitted by the 7 participating teams are listed and discussed in this section. In Table 4, the submission run with best performance per team is shown. An additional evaluation regarding the imaging modality was done internally, after the official concept detection evaluation process. The accuracy (%) across all images in the test set was computed and is listed in Table 6. Compared to the previous editions, there is an improvement regarding the F1-Score of the submitted concept detection models, from 0.1583 in ImageCLEF 2017 [3], 0.1108 in ImageCLEF 2018 [7] and 0.2823 in ImageCLEF 2019 [16] to 0.3940 in 2020.

The AUEB NLP Group [12] from the Athens University of Economics achieved the overall highest F1-Score of 0.3940 for the detection of concepts for the images in the official evaluation test set. Their three (3) submission runs ranked 1st, 2nd and 6th of all 47 submitted runs. The submitted systems are a variation of CheXNet [26] with DenseNet-121 [8] and followed by a feed-forward Neural Network (FFNN), which acts as the classifier layer on the top [12]. The system was first pre-trained on the ImageNet data set [21] and then fine-tuned using the ImageCLEF 2020 concept detection development data set. Several ensemble methods such as the intersection and union of predicted concepts were experimented. The system with the intersection of concepts achieved the overall highest F1-Score.

The overall 2nd ranked participating team is PwC_Healthcare group from PricewaterhouseCoopers with a total number of nine (9) submitted runs. The adopted approaches range from Convolutional Neural Network (CNN) architectures, to Natural Language Processing techniques, as well as clustering algorithms [24]. The group's three (3) best systems ranked 3rd, 4th and 5th. Several pre-processing approaches such as range and intensity normalization and

[4] https://www.aicrowd.com/challenges/imageclef-2020-caption-concept-detection [last accessed: 26.07.2020]

[5] https://www.imageclef.org/system/files/ImageCLEF-ConceptDetection-Evaluation.zip [last accessed: 26.07.2020]

data augmentation were adopted prior to training the models [24]. Multi-modal approaches were experimented to incorporate the concept imbalanced distribution and a novel approach of band classification was applied. This classification method first clusters the vocabulary of concepts into bands and then creates for each band a classification architecture [24].

Table 4. Performance of the participating teams in the ImageCLEF 2020 concept detection task in regards to correctly predicting concepts of the images in the test set. The best run per team is selected. Teams with previous participation in 2019 are marked with an asterisk.

Team	Institution	F1-Score
AUEB NLP Group* [12]	Department of Informatics, Athens University of Economics and Business, Athens, Greece	0.3940
PwC_Healthcare [24]	PricewaterhouseCoopers US Advisory, Mumbai, India	0.3924
Essex [6]	School of computer Science and Electronic Engineering, University of Essex, Essex, United Kingdom	0.3808
IML_DFKI [10]	Interactive Machine Learning Group, German Research Center for Artificial Intelligence (DFKI), Saarbrücken, Germany	0.3745
TUC_MC [25]	Technische Unversität Chemnitz, Chemnitz, Germany	0.3512
Morgan_CS [14]	Computer Science Department, Morgan State University, Baltimore, Maryland, United States of America	0.1673
CSE_SSN [2]	Department of Computer Science and Engineering, SSN College of Engineering, Chennai, India	0.1347

The third best participating team was from the University of Essex, with an overall F1-Score of 0.381. The proposed approach adopts pre-trained DenseNet models [8] for the extraction of relevant features. The additional information on the imaging modality was used for fine-tuning by adding a fully connected layer to the DenseNet-121 model and thereby transforming it into a multi-label classification model [6]. Several concept selection strategies, such as distance and ranked based methods, were applied to a given query image from the test set. The group's five best runs of the nine submitted runs ranked 6th to 10th among all submissions.

Five runs were submitted by the IML group from the German Research Center for Artificial Intelligence, with the best F1-Score of 0.3745, and the 4th best team. Multiple deep learning systems such as VGG16 [22], ResNet50 [5] and DenseNet169 [8], which were pre-trained on the ImageNet data set, were applied for modeling the concept detection systems. The task was addressed as a multi one-hot encoding with a final prediction layer of 3,047 sigmoidal activation units and several fine-tuning steps, such as data augmentation, hyper-parameter settings, were undertaken [10].

Table 5: Concept detection performance in terms of all submitted runs for the ImageCLEF 2020 Concept Detection Task

Group Name	Submission Run	F1-Score
AUEB NLP Group	InterceptCheXNetCheckpoints.csv	0.3940
AUEB NLP Group	BestOf.csv	0.3933
PwC_Healtcare	folderwise_KNN_resnet101_test_pred.csv	0.3924
PwC_Healtcare	combined_test_pred_v1.csv	0.3889
PwC_Healtcare	folder_wise_test_pred_v1.csv	0.3889
AUEB_NLP_Group	UnionCheXNetCheckpoints.csv	0.3870
Essex	submit_run3.csv	0.3808
Essex	submit_run5.csv	0.3805
Essex	submit_run1.csv	0.3797
Essex	cp99_all_modified.txt	0.3785
Essex	c99_all_man.txt	0.3777
IML_DFKI	imageclefmed2020-test-vgg16-f1-bce-nomissing-iml.txt	0.3745
IML_DFKI	imageclefmed2020-test-vgg16-f1-bce-iml.txt	0.3744
PwC_Healtcare	combined_test_pred_new.csv	0.3681
PwC_Healtcare	NLP_clusters_test_pred.csv	0.3668
PwC_Healtcare	knn_t117_test_pred.csv	0.3666
IML_DFKI	imageclefmed2020-test-resnet50-iml.txt	0.3652
IML_DFKI	imageclefmed2020-test-vgg16-iml.txt	0.3631
IML_DFKI	imageclefmed2020-test-densenet169-iml.txt	0.3602
TUC_MC	model_thr0_18.csv	0.3512
TUC_MC	streamlined1_thr0_25.csv	0.3486
TUC_MC	streamlined1_thr0_20.csv	0.3486
TUC_MC	2streamlined1.csv	0.3486
TUC_MC	basemodel_thr0_20.csv	0.3474
TUC_MC	model_low_lr_thr0_20.csv	0.3455
Essex	submit_run2.csv	0.3449
TUC_MC	streamlined1_nomax.csv	0.3448
TUC_MC	basemodel.csv	0.3435
TUC_MC	streamlined1_thr0_12.csv	0.3423
PwC_Healtcare	f1_band_test_t025_pred.csv	0.3379
Essex	cp98_all.txt	0.3370
TUC_MC	model_weighting.csv	0.3325
PwC_Healtcare	NLP_test_pred_fixed.csv	0.3163
Essex	canberra_all_modified.txt	0.2804
PwC_Healtcare	combined_wo_folder_test.csv	0.2655
Essex	cp95_all.txt	0.2459

Morgan_CS	MSU_dense_fcn.txt	0.1673
Morgan_CS	MSU_dense_fcn_4.txt	0.1591
Morgan_CS	MSU_dense_resnet_fcn_1.txt	0.1534
Morgan_CS	MSU_dense_resnet_fcn_1.txt	0.1447
Morgan_CS	MSU_dense_feat.txt	0.1395
CSE_SSN	captions_output.txt	0.1347
Morgan_CS	_MSU_dense_feat.txt	0.1284
Morgan_CS	MSU_dense_fcn_2.txt	0.0943
Morgan_CS	MSU_dense_fcn_3.txt	0.0894
Morgan_CS	MSU_autoenc_fcn.txt	0.0634
Morgan_CS	MSU_lstm_dense_fcn.txt	0.0625

TUC_MC, a media computing group from the Chemnitz University of Technology ranked 5th best participating team. The highest F1-Score from the ten submitted runs was 0.3745. The adopted deep learning model was based on the Xception architecture [1] with weights pre-trained on ImageNet. The submitted runs use the same model base structure, however the hyper-parameters are varied in regards to last layer threshold and max-pooling in the highest layers [25].

Ten runs were submitted by Morgan_CS, a group from the computer science department at the Morgan State University. The best achieved F1-Score was 0.1673, by approaching the concept detection task as a multi-label classification problem [14]. Classifiers were trained with deep features extracted with the deep learning system DenseNet169 and ResNet50 and pre-trained on ImageNet. Other methods experimented include a recurrent concept sequence generator that was modelled using a multimodal technique of fusing text and image features for recurrent sequence prediction.

CSE_SSN from the department of computer science of the SSN College of Engineering Chennai submitted one (1) run for official evaluation and achieved the average F1-Score of 0.1347 on all images in the test set. Similar to several participating teams, the concept detection task was addressed as a convolution neural network multi-label classification problem [2]. The imaging modality distributed was applied for pre-processing and model fine-tuning steps.

An ex-post evaluation was computed on all submitted runs. The aim was to compute the performance on correctly predicting the imaging modality. All images in the development and test set were assigned concepts that denote the acquisition technique, as shown in Figure 2. The images belonging to the imaging modality 'DRCO: Combined modalities in one image' were not considered for evaluation. For all images in the test set, we computed the presence of these concepts in the submission runs using this additional information. The best performance grouped per team is listed in Table 6 and the complete evaluation in Table 7.

Table 6. Performance of the participating teams in the ImageCLEF 2020 concept detection task on correctly predicting the imaging modality of the images in the test set. The best run per team is selected. Teams with a previous participation in 2019 are marked with an asterisk.

Team	Institution	Accuracy (%)
PwC_Healthcare [24]	PricewaterhouseCoopers US Advisory, Mumbai, India	62.08
AUEB NLP Group* [12]	Department of Informatics, Athens University of Economics and Business, Athens, Greece	59.73
Essex [6]	School of computer Science and Electronic Engineering, University of Essex, Essex, United Kingdom	56.34
TUC_MC [25]	Technische Unversität Chemnitz, Chemnitz, Germany	50.08
IML_DFKI [10]	Interactive Machine Learning Group, German Research Center for Artificial Intelligence (DFKI), Saarbrücken, Germany	47.06
Morgan_CS [14]	Computer Science Department, Morgan State University, Baltimore, Maryland, United States of America	02.06
CSE_SSN [2]	Department of Computer Science and Engineering, SSN College of Engineering, Chennai, India	01.39

Table 7: Modality classification performance in terms of all submitted runs for the ImageCLEF 2020 Concept Detection Task

Group Name	Submission Run	Acc(%)
PwC_Healtcare	NLP_clusters_test_pred.csv	62.08
AUEB_NLP_Group	InterceptCheXNetCheckpoints.csv	59.73
AUEB_NLP_Group	BestOf.csv	59.48
essexgp2020	cp99_all_modified.txt	56.34
essexgp2020	c99_all_man.txt	55.69
AUEB_NLP_Group	UnionCheXNetCheckpoints.csv	55.23
PwC_Healtcare	folderwise_KNN_resnet101_test_pred.csv	54.70
PwC_Healtcare	folder_wise_test_pred_v1.csv	52.43
PwC_Healtcare	combined_test_pred_v1.csv	52.43
essexgp2020	submit_run3.csv	50.93
TUC_MC	streamlined1_thr0_25.csv	50.08
essexgp2020	submit_run1.csv	49.29
essexgp2020	submit_run5.csv	48.84
TUC_MC	model_low_lr_thr0_20.csv	48.22
iml	imageclefmed2020-test-densenet169-iml.txt	47.06
iml	imageclefmed2020-test-vgg16-f1-bce-iml.txt	46.94

iml	imageclefmed2020-test-vgg16-f1-bce-nomissing-iml.txt	46.94
iml	imageclefmed2020-test-resnet50-iml.txt	46.83
iml	imageclefmed2020-test-vgg16-iml.txt	45.47
TUC_MC	model_thr0_18.csv	44.88
TUC_MC	basemodel_thr0_20.csv	44.74
PwC_Healtcare	combined_test_pred_new.csv	42.05
PwC_Healtcare	knn_t117_test_pred.csv	41.34
TUC_MC	streamlined1.csv	41.23
TUC_MC	streamlined1_thr0_20.csv	41.23
TUC_MC	basemodel.csv	39.30
essexgp2020	submit_run2.csv	38.88
TUC_MC	model_weighting.csv	38.88
TUC_MC	streamlined1_nomax.csv	37.35
TUC_MC	streamlined1_thr0_12.csv	35.94
PwC_Healtcare	f1_band_test_t025_pred.csv	34.27
essexgp2020	cp98_all.txt	19.78
PwC_Healtcare	combined_wo_folder_test.csv	14.60
essexgp2020	canberra_all_modified.txt	11.83
PwC_Healtcare	NLP_test_pred_fixed.csv	10.67
essexgp2020	cp95_all.txt	02.86
Morgan_CS	MSU_dense_fcn.txt	02.07
Morgan_CS	MSU_dense_fcn_4.txt	01.75
Morgan_CS	MSU_dense_resnet_fcn_1.txt	01.75
Morgan_CS	MSU_dense_feat.txt	01.75
Morgan_CS	MSU_autoenc_fcn.txt	01.58
Morgan_CS	MSU_dense_resnet_fcn_1.txt	01.50
Morgan_CS	MSU_lstm_dense_fcn.txt	01.44
Morgan_CS	MSU_dense_fcn_2.txt	01.41
saradadevi	captions_output.txt	01.39
Morgan_CS	MSU_dense_feat.txt	01.39
Morgan_CS	MSU_dense_fcn_3.txt	01.39

6 Conclusion

This paper presents an overview of applied approaches and their performance, as well as the task description, participation and distributed data set for the ImageCLEF 2020 concept detection task. Similar to the 2019 edition, the results this year show that there is an improvement in the achieved F1-scores (best score 0.3940). In this edition, not only does the dataset contain an increased number of images, the number of concepts were reduced to be more precise and additional modality information was distributed. In the previous editions, the overall best F1-Scores were 0.2823 in Image-med Caption 2019, 0.1108 in ImageCLEFmed Caption 2018 and 0.1583 in ImageCLEFmed Caption 2017. Almost all participating groups were new to the task, with only one team that participated in ImageCLEF caption 2019. The seven participating teams are affiliated to institutions from 5 countries, which shows the continuing research interest to this challenging task.

Most of the submitted runs are based on deep learning architectures. The pre-trained models DenseNet-121, ResNet50 and VGG16 on the ImageNet and CheXNet were used to extract relevant visual representation for the images. Multiple pre-processing steps such as concept filtering, data augmentation and image enhancement were applied to optimize the input for the predicting systems. Long short-term memory (LSTM) recurrent neural networks (RNN), adversarial auto-encoders, CNN image encoders and transfer learning-based multi-label classification models were the frequently used approaches.

As the focus in the caption task 2019 was reduced from biomedical images to solely radiology images, a reduction of the extracted concepts from 111,155 to 5,528 was observed. We added this year an additional label denoting the imaging modality of the images. This extra information was used by several teams for pre-filtering steps prior to training the models, concept selection and for ensemble algorithms. The class imbalance in the distributed data set proved to be challenging for several teams. However, medical data and diseases are also usually unbalanced with a few conditions happening very frequently and most being very rare.

In future work, an extensive review of the clinical relevance for the concepts in the development data should be explored. As the concepts originate from the natural language captions, not all concepts have high clinical utility. Medical journals also have very different policies in terms of checking figure captions. We believe this will assist in creating more efficient systems for automated medical data analysis.

References

1. Chollet, F.: Xception: Deep Learning with Depthwise Separable Convolutions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR, Honolulu, USA, July 22-25, 2017. pp. 1800–1807 (07 2017). https://doi.org/10.1109/CVPR.2017.195
2. Devi, S., S, K.: ImageCLEF 2020: Image Caption Prediction using Multilabel Convolutional Neural Network. In: CLEF2020 Working Notes. CEUR Workshop Proceedings, CEUR-WS.org, Thessaloniki, Greece (September 22-25 2020)
3. Eickhoff, C., Schwall, I., de Herrera, A.G.S., Müller, H.: Overview of ImageCLEFcaption 2017 - Image Caption Prediction and Concept Detection for Biomedical Images. In: Working Notes of CLEF 2017 - Conference and Labs of the Evaluation Forum, Dublin, Ireland, September 11-14, 2017. (2017), http://ceur-ws.org/Vol-1866/invited_paper_7.pdf
4. Ferro, N., Peters, C.: Information Retrieval Evaluation in a Changing World Lessons Learned from 20 Years of CLEF: Lessons Learned from 20 Years of CLEF (01 2019)
5. He, K., Zhang, X., Ren, S., Sun, J.: Deep Residual Learning for Image Recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recog-

nition, CVPR, Las Vegas, USA, June 26 - July 1, 2016. pp. 770–778 (06 2016). https://doi.org/10.1109/CVPR.2016.90
6. de Herrera, A.G.S., Andrade, F.P., Bentley, L., Compean, A.A.: Essex at ImageCLEFcaption 2020 task. In: CLEF2020 Working Notes. CEUR Workshop Proceedings, CEUR-WS.org, Thessaloniki, Greece (September 22-25 2020)
7. de Herrera, A.G.S., Eickhoff, C., Andrearczyk, V., Müller, H.: Overview of the ImageCLEF 2018 Caption Prediction Tasks. In: Working Notes of CLEF 2018 - Conference and Labs of the Evaluation Forum, Avignon, France, September 10-14, 2018. (2018), http://ceur-ws.org/Vol-2125/invited_paper_4.pdf
8. Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q.: Densely Connected Convolutional Networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR, Honolulu, USA, July 22-25, 2017. pp. 2261–2269 (July 2017). https://doi.org/10.1109/CVPR.2017.243
9. Ionescu, B., Müller, H., Péteri, R., Abacha, A.B., Datla, V., Hasan, S.A., Demner-Fushman, D., Kozlovski, S., Liauchuk, V., Cid, Y.D., Kovalev, V., Pelka, O., Friedrich, C.M., de Herrera, A.G.S., Ninh, V.T., Le, T.K., Zhou, L., Piras, L., Riegler, M., l Halvorsen, P., Tran, M.T., Lux, M., Gurrin, C., Dang-Nguyen, D.T., Chamberlain, J., Clark, A., Campello, A., Fichou, D., Berari, R., Brie, P., Dogariu, M., Ştefan, L.D., Constantin, M.G.: Overview of the ImageCLEF 2020: Multimedia Retrieval in Medical, Lifelogging, Nature, and Internet Applications. In: Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the 11th International Conference of the CLEF Association (CLEF 2020), vol. 12260. LNCS Lecture Notes in Computer Science, Springer, Thessaloniki, Greece (September 22-25 2020)
10. Kalimuthu, M., Nunnari, F., Sonntag, D.: A Competitive Deep Neural Network Approach for the ImageCLEFmed Caption 2020 Task. In: CLEF2020 Working Notes. CEUR Workshop Proceedings, CEUR-WS.org, Thessaloniki, Greece (September 22-25 2020)
11. Kalpathy-Cramer, J., de Herrera, A.G.S., Demner-Fushman, D., Antani, S.K., Bedrick, S., Müller, H.: Evaluating performance of biomedical image retrieval systems - An overview of the medical image retrieval task at ImageCLEF 2004-2013. Comp. Med. Imag. and Graph. **39**, 55–61 (2015). https://doi.org/10.1016/j.compmedimag.2014.03.004, https://doi.org/10.1016/j.compmedimag.2014.03.004
12. Karatzas, B., Pavlopoulos, J., Kougia, V., Androutsopoulo, I.: AUEB NLP Group at ImageCLEFmed Caption 2020. In: CLEF2020 Working Notes. CEUR Workshop Proceedings, CEUR-WS.org, Thessaloniki, Greece (September 22-25 2020)
13. Koitka, S., Friedrich, C.M.: Optimized Convolutional Neural Network Ensembles for Medical Subfigure Classification. 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 at the 8th International Conference of the CLEF Association, Dublin, Ireland, September 11-14, 2017, Lecture Notes in Computer Science (LNCS) 10456. pp. 57–68. Springer International Publishing, Cham (2017). https://doi.org/10.1007/978-3-319-65813-1_5
14. Lyode, O., Rahman, M.: Concept Detection in Biomedical Images with Deep Learning Based Multilabel Classification. In: CLEF2020 Working Notes. CEUR Workshop Proceedings, CEUR-WS.org, Thessaloniki, Greece (September 22-25 2020)
15. Müller, H., Clough, P.D., Deselaers, T., Caputo, B. (eds.): ImageCLEF, Experimental Evaluation in Visual Information Retrieval. Springer (2010). https://doi.org/10.1007/978-3-642-15181-1

16. Pelka, O., Friedrich, C.M., de Herrera, A.G.S., Müller, H.: Overview of the ImageCLEFmed 2019 Concept Detection Task. In: Cappellato, L., Ferro, N., Losada, D.E., Müller, H. (eds.) Working Notes of CLEF 2019 - Conference and Labs of the Evaluation Forum, Lugano, Switzerland, September 9-12, 2019. CEUR Workshop Proceedings, vol. 2380. CEUR-WS.org (2019), http://ceur-ws.org/Vol-2380/paper_245.pdf
17. Pelka, O., Koitka, S., Rückert, J., Nensa, F., Friedrich, C.M.: Radiology Objects in COntext (ROCO): A Multimodal Image Dataset. In: Intravascular Imaging and Computer Assisted Stenting - and - Large-Scale Annotation of Biomedical Data and Expert Label Synthesis - 7th Joint International Workshop, CVII-STENT 2018 and Third International Workshop, LABELS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16, 2018, Proceedings. pp. 180–189 (2018). https://doi.org/10.1007/978-3-030-01364-6_20, https://doi.org/10.1007/978-3-030-01364-6_20
18. 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
19. Pelka, O., Nensa, F., Friedrich, C.M.: Variations on Branding with Text Occurrence for Optimized Body Parts Classification. In: Proceedings of the 41th Annual International Conference of the IEEE Engineering in Medicine and Biology Society EMBC 2019, Berlin, Germany, July 23-27, 2019. pp. 890–894 (2019). https://doi.org/10.1109/EMBC.2019.8857478
20. 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
21. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A., Fei-Fei, L.: ImageNet Large Scale Visual Recognition Challenge. International Journal of Computer Vision **115** (09 2014). https://doi.org/10.1007/s11263-015-0816-y
22. Simonyan, K., Zisserman, A.: Very Deep Convolutional Networks for Large-Scale Image Recognition. arXiv 1409.1556 (09 2014)
23. Soldaini, L., Goharian, N.: QuickUMLS: a fast, unsupervised approach for medical concept extraction. In: MedIR Workshop, SIGIR (2016)
24. Sonker, R., Mishra, A., Bansal, P., Pattnaik, A.: Techniques for Medical Concept Detection from Multi-Modal Images. In: CLEF2020 Working Notes. CEUR Workshop Proceedings, CEUR-WS.org, Thessaloniki, Greece (September 22-25 2020)
25. Udas, N., Beuth, F., Kowerko, D.: TUC MC group at ImageCLEFmed 2020 concept detection task using Xception models. In: CLEF2020 Working Notes. CEUR Workshop Proceedings, CEUR-WS.org, Thessaloniki, Greece (September 22-25 2020)
26. Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M., Summers, R.M.: ChestX-Ray8: Hospital-Scale Chest X-Ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR, Honolulu, USA, July 22-25, 2017. pp. 3462–3471 (2017)