Published in Proceedings of CLEF 2017 working Notes, 2017, Dublin, Ireland, 11-14 September, which should be cited to refer to this work.

Overview of ImageCLEFcaption 2017 – Image Caption Prediction and Concept Detection for Biomedical Images

Carsten Eickhoff[1], Immanuel Schwall[1], Alba G. Seco de Herrera[2], and Henning Müller[3]

[1] ETH Zurich, Switzerland;
[2] Lister Hill National Center for Biomedical Communications, National Library of Medicine, Bethesda, USA;
[3] University of Applied Sciences Western Switzerland (HES–SO), Sierre, Switzerland.
c.eickhoff@acm.org

Abstract. This paper presents an overview of the ImageCLEF 2017 caption tasks on the analysis of images from the biomedical literature. Two subtasks were proposed to the participants: a concept detection task and caption prediction task, both using only images as input. The two subtasks tackle the problem of providing image interpretation by extracting concepts and predicting a caption based on the visual information of an image alone. A dataset of 184,000 figure-caption pairs from the biomedical open access literature (PubMed Central) are provided as a testbed with the majority of them as trainign data and then 10,000 as validation and 10,000 as test data. Across two tasks, 11 participating groups submitted 71 runs. While the domain remains challenging and the data highly heterogeneous, we can note some surprisingly good results of the difficult task with a quality that could be beneficial for health applications by better exploiting the visual content of biomedical figures.

Keywords: ImageCLEF 2017, Caption Prediction, Image Understanding, Computer Vision, Radiology

1 Introduction

Interpreting and summarizing the insights gained from medical images such as radiography or biopsy samples is a time-consuming task that involves highly trained experts and often represents a bottleneck in clinical diagnosis pipelines. As a consequence, there is a considerable need for automatic methods that can approximate the mapping from visual information to condensed textual descriptions. ImageCLEF[4] is an evaluation campaign that has being organized as part of the CLEF initiative labs since 2003 [1, 2]. The campaign offers several research tasks that welcome participation from teams around the world and change from year to year [3]. In 2017, the caption task of ImageCLEF 2017 addresses the

[4] http://imageclef.org/

problem of image understanding as a cross-modality matching scenario in which visual content and textual descriptors need to be aligned and concise textual interpretations of medical images are generated. A similar task was proposed in 2016 but without any submission [4], as it is a very challenging task. The task is based on a large-scale collection of figures from open access biomedical journal articles from PubMed Central (PMC)[5]. Each image is accompanied by its original caption and a set of extracted UMLS® (Unified Medical Language System®)[6] Concept Unique Identifiers (CUIs), constituting a natural testbed for this image captioning task. A subset of PMC concentrating on clinical images and limiting the number of compound figures is used.

This paper gives an overview of the caption task at ImageCLEF 2017. Section 2 introduces the two subtasks and Section 3 the data set and ground truth. A description of the evaluation methodology is provided in Section 4. Subsequently, the participant submissions are analysed in Section 5 and Section 6 briefly discusses their respective strengths and weaknesses as well as their implications for academic research and medical practice. Finally, we conclude with an outlook to the possible future of the evaluation campaign in Section 7.

2 Tasks

This first edition of the biomedical image captioning task at ImageCLEF comprises two sub tasks: (1) Concept Detection and (2) Caption Prediction. Figure 1 shows an example image of a tomographic angiography reconstruction along with its relevant concepts as well as the reference caption.

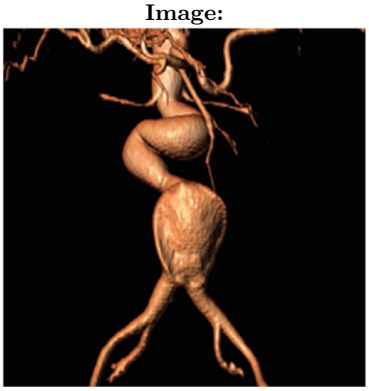
Concept Detection As a first step towards automatic image caption understanding, participating systems are tasked with identifying the presence of relevant biomedical concepts in medical images. Based on the visual image content, this subtask provides the building blocks for the image understanding step by identifying the individual components from which full captions can be composed.

Caption Prediction On the basis of the concept vocabulary detected in the first subtask as well as the visual information of their interaction in the image, participating systems are tasked with composing coherent natural language captions for the entirety of an image. In this step, rather than the mere coverage of visual concepts, detecting the interplay of visible elements is crucial for recreating the original image caption.

[5] PubMed Central (PMC) is a free fulltext archive of biomedical and life sciences journal literature at the U.S. National Institute of Healths National Library of Medicine (NIH/NLM) (see http://www.ncbi.nlm.nih.gov/pmc/).

[6] https://www.nlm.nih.gov/research/umls

Image:



Concept detection:

- C0002940: Aneurysm
- C0002978: angiogram
- C0027530: Neck
- C0087111: Therapeutic procedure
- C0524425: inside the blood vessel
- C0524865: Reconstructive Surgical Procedures
- C3887704: treatment - ActInformationManagementReason

Caption prediction:

Preoperative computed tomographic angiography reconstruction showing hostile neck anatomy amenable to treatment with endovascular aneurysm sealing (EVAS).

Fig. 1. Example of an image and the information provided in the training set.

3 Collection

The experimental corpus is derived from scholarly biomedical articles on PMC from which we extract figures and their corresponding captions. In total, the collection is comprised of 184,614 image-caption pairs. This overall set is further split into disjunct training (164,614 pairs), validation (10,000 pairs) and test (10,000 pairs) sets. For the concept detection sub task, we used the QuickUMLS library [5] to identify all UMLS concepts mentioned in the caption text.

The subset of PMC was created using an automated method to classify all 3 million images of PMC from early 2016 into image types [6] fully automatically. We keep clinical image types and remove compound figures. As PMC contains many compound figures and as the method was fully automatic we have approximately 10-20% of the images that are either compound or non-clinical, which creates noise in the data set and makes the task even more challenging.

4 Evaluation Methodology

The evaluation of both sub tasks is conducted separately. For the concept detection task, we measure the balanced precision and recall trade-off in terms of F_1 scores. To this end, we use Python's scikit-learn (v0.17.1-2) library. We compute micro F_1 per image and average across all test images. A total of 393 reference captions in the test set do not contain any UMLS concepts. The respective images are excluded from the evaluation.

Caption prediction performance is assessed on the basis of BLEU scores [7] using the Python NLTK (v3.2.2) default implementation. Candidate captions are lower cased, stripped of all punctuation and English stop words. Finally, to increase coverage, we apply Snowball stemming. BLEU scores are computed per reference image, treating each entire caption as a sentence, even though it may contain multiple natural sentences. We report average BLEU scores across all 10,000 test images.

The source code of both evaluation scripts is available on the task Web page[7].

5 Results

We received a total of 71 submissions by 11 individual teams. Table 1 gives an overview of all participants and their runs. There was a limit of at most 10 runs per team and sub task and the submissions are roughly evenly split between tasks. The call for contributions did not initially make any assumptions about the kinds of strategies and external data that participants would rely on. As a consequence, in this first edition of the task, we see a broad range of performance scores as well as methods being applied. Evaluation of the results showed that some teams employed methods that were at least partially trained on external resources including PMC articles. Since such approaches cannot be guaranteed to have respected our division into training, validation and test folds and might subsequently leak test examples into the training process, we separately list runs relying exclusively on the official collection as well as those making use of external information.

5.1 Concept Detection

The concept detection task received 37 runs from 9 participating groups. Table 2 lists the performance of all official (no external information used) runs. The global overview of all runs, including those using external information, can be found in Table 3.

The vast majority of runs was purely automatic (A) in nature with only few submissions relying on some form of manual intervention (M). There was no noticeable advantage of relying on manual interventions as all manual runs lie well in the center of the performance score range.

[7] http://imageclef.org/2017/caption

Table 1. Participating Groups.

Team	Institution	# Runs T1	# Runs T2
AAI [8]	AI Lab, University of the Aegean, Mytilene, Greece	1	0
PRNA [9]	Artificial Intelligence Lab, Philips Research North America, Cambridge, MA, USA	3	4
BCSG [10]	Biomedical Computer Science Group, University of Applied Sciences and Arts Dortmund, Germany	0	10
UAPT [11]	Institute of Electronics and Informatics Engineering, University of Aveiro, Portugal	3	0
BMET [12]	School of Information Technologies, University of Sydney, Australia	3	4
IPL [13]	Information Processing Laboratory, Athens University of Economics and Business, Athens, Greece	10	0
ISIA [14]	Key Laboratory of Intelligent Information Processing, Chinese Academy of Sciences, Beijing, China	0	10
MAMI [15]	CNRS, University of Toulouse, France, & University of Antananarivo, Madagascar	2	0
MUPB [16]	University Politehnica of Bucharest, Romania, University of Applied Sciences Western Switzerland, Sierre & University of Geneva, Switzerland	1	0
MSU [17]	Computer Science Department, Morgan State University, Baltimore, MD, USA	4	0
NLM [18]	Lister Hill National Center for Biomedical Communications, National Library of Medicine, Bethesda, MD, USA	10	6

While most teams rely on some form of convolutional neural networks to represent visual information (NLM [18], PRNA [9], BMET [12], AAI [8], MAMI [15], MUPB [16]), some chose more traditional bag-of-visual-words representations (IPL [13], MSU [17]) or even relied on mixtures of both representation types (UAPT [11]). While, on average, CNN-based models seem to deliver more robust results, some of the most competitive submissions are purely based on traditional features.

The use of very deep residual networks (AAI [8], MAMI [15], MUPB [16]), on average, did not introduce significant improvements over shallower CNN versions. On top of the basic image representation approaches, we see a broad range of affiliate techniques used for recognizing bio-medical concepts. PRNA [9] successfully rely on attention models for image understanding which seems to introduce a considerable relative advantage over other model variants. The use of convolutional de-noising auto-encoders (UAPT [11]) for unsupervised representation learning did not seem to lead to considerable improvements.

Several groups included retrieval-based methods that would identify highly visually related images in the official training set (IPL [13], MSU [17]) or an external collection of images (NLM [18]). The captions of such related images are then scanned for bio-medical concepts to be assigned to the candidate image. This approach generally resulted in very good results, among them several of the best-performing submissions for the task.

Table 2. Concept detection performance in terms of F_1 scores without the use of external resources.

Team	Run	Type	F_1
IPL	1494006128917	A	0.1436
IPL	1494006074473	A	0.1418
IPL	1494009510297	A	0.1417
IPL	1494006054264	A	0.1415
IPL	1494009412127	A	0.1414
IPL	1494009455073	A	0.1394
IPL	1494006225031	A	0.1365
IPL	1494006181689	A	0.1364
IPL	1494006414840	A	0.1212
IPL	1494006360623	A	0.1208
BMET	1493791786709	A	0.0958
BMET	1493791318971	A	0.0880
NLM	1494013963830	A	0.0880
NLM	1494014008563	A	0.0868
BMET	1493698613574	A	0.0838
NLM	1494013621939	A	0.0811
NLM	1494013664037	A	0.0695
MSU	1494060724020	M	0.0498
UAPT	1493841144834	M	0.0488
UAPT	1493995613907	M	0.0463
MSU	1494049613114	M	0.0461
MSU	1494048615677	M	0.0434
UAPT	1493976564810	M	0.0414
MSU	1494048330426	A	0.0273
NLM	1494012725738	A	0.0012

Table 3. Concept detection performance of runs using external resources; the exact type of third-party material is indicated.

Team	Run	Type	Resources	F_1
NLM	1494012568180	A	Open-i indexed PubMed	0.1718
NLM	1494012586539	A	Open-i indexed PubMed	0.1648
AAI	1491857120689	A	ImageNet & MS COCO	0.1583
NLM	1494014122269	A	Open-i indexed PubMed	0.1390
NLM	1494012605475	A	Open-i indexed PubMed	0.1228
PRNA	1493823116836	A	ImageNet	0.1208
MAMI	1496127572481	M	ImageNet	0.0462
PRNA	1493823633136	A	ImageNet	0.0234
PRNA	1493823760708	A	ImageNet	0.0215
NLM	1495446212270	A	Open-i indexed PubMed	0.0162
MUPB	1493803509469	A	ImageNet	0.0028
MAMI	1493631868847	M	ImageNet	0.0000

5.2 Caption Prediction

The harder caption prediction task received 34 runs from 5 participating groups. Table 4 lists the performance of all official (no external information used) runs. The global overview of all runs, including those using external information, can be found in Table 5. For this task, no manual runs were submitted.

Most submitted runs are based on the teams' respective contributions to the concept detection task expanded by language modeling capabilities. Often this takes the form of recurrent neural networks (ISIA [14], BCSG [10], PRNA [9], BMET [12]), making the CNN + LSTM combination a frequently-used setup.

As for the first sub task, the use of retrieval-based methods to identify highly visually related images and using their captions as a starting point for candidate caption generation (ISIA [14] MSU [17], NLM [18]) resulted in highly competitive performance.

Table 4. Caption prediction performance in terms of BLEU scores without the use of external resources.

Team	Run	BLEU
ISIA	1493921574200	0.2600
ISIA	1493666388885	0.2507
ISIA	1493922473076	0.2454
ISIA	1494002110282	0.2386
ISIA	1493922527122	0.2315
NLM	1494038340934	0.2247
ISIA	1493831729114	0.2240
ISIA	1493745561070	0.2193
ISIA	1493715950351	0.1953
ISIA	1493528631975	0.1912
ISIA	1493831517474	0.1684
NLM	1494038056289	0.1384
NLM	1494037493960	0.1131
BMET	1493702564824	0.0982
BMET	1493698682901	0.0851
BMET	1494020619666	0.0826
BMET	1493701062845	0.0656

6 Discussion

There are several observations that should be taken into account when analyzing the results presented in the previous section. Most notably, as a consequence of the data source (scholarly biomedical journal articles), the collection contains a considerable amount of noise in the form of compound figures with potentially highly heterogeneous content. In future editions of this task, we will consider

Table 5. Caption prediction performance of runs using external resources; the exact type of third-party material is indicated.

Team	Run	Resources	BLEU
NLM	1494014231230	Open-i indexed PubMed	0.5634
NLM	1494081858362	Open-i indexed PubMed	0.3317
PRNA	1493825734124	ImageNet	0.3211
NLM	1495446212270	Open-i indexed PubMed	0.2646
PRNA	1493824027725	ImageNet	0.2638
PRNA	1493825504037	ImageNet	0.1801
PRNA	1493824818237	ImageNet	0.1107
BCSG	1493885614229	ImageNet	0.0749
BCSG	1493885575289	ImageNet	0.0675
BCSG	1493885210021	ImageNet	0.0624
BCSG	1493885397459	ImageNet	0.0537
BCSG	1493885352146	ImageNet	0.0527
BCSG	1493885286358	ImageNet	0.0411
BCSG	1493885541193	ImageNet	0.0375
BCSG	1493885499624	ImageNet	0.0365
BCSG	1493885708424	ImageNet	0.0326
BCSG	1493885450000	ImageNet	0.0200

using a less diverse source of images such as radiology/pathology in order to reduce the amount of variation in the data.

Secondly, the UMLS concept extraction employed here is a probabilistic process that introduces its own errors. As a consequence, there are several training captions that do not contain any UMLS concepts, making such examples difficult to use for concept detection purposes. In the future, we will rely on more rigorous filtering to ensure good concept coverage across training, validation and test data.

Finally, there should have been a clearer specification of what external material, if any, is permissible for use. The teams employed a wide number of third-party material ranging from general academic collections such as ImageNet, mainly in the form of pre-trained networks, to corpora of scholarly articles. While the former do not represent a major problem, the latter could, conceivably contain the exact image and caption pairs of our test set, the use of which would create a strong advantage and a non-realistic setting for really novel data. The experimental overview shows some evidence of this happening when methods using PubMed Central images in the training step vastly outperform all competitors on both tasks. For this reason, we made the conservative decision to separate between official runs using no external information at all and those that used third-party material. In the future, we will more carefully specify which kind of external material is safe to use. It does make sense to allow for external data to be used but it needs to be made clear that no test data are included.

7 Conclusions

This paper presents an overview of the ImageCLEF 2017 biomedical image captioning task. We consider the sub tasks of concept detection and full caption prediction. The participating groups investigated the use of a wide range of image understanding techniques. Especially neural network methods are highly popular and delivered convincing performance on these hard problems. The individually relatively low scores motivate further homogenization of tasks and collection in future editions of the challenge.

Acknowledgments

This research was supported in part by the Intramural Research Program of the National Institutes of Health (NIH), National Library of Medicine (NLM), and Lister Hill National Center for Biomedical Communications (LHNCBC).

References

1. 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) (2015) 55 – 61
2. Müller, H., Clough, P., Deselaers, T., Caputo, B., eds.: ImageCLEF – Experimental Evaluation in Visual Information Retrieval. Volume 32 of The Springer International Series On Information Retrieval. Springer, Berlin Heidelberg (2010)
3. 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)
4. Villegas, M., Müller, H., Garcia Seco de Herrera, A., Schaer, R., Bromuri, S., Gilbert, A., Piras, L., Wang, J., Yan, F., Ramisa, A., Dellandrea, A., Gaizauskas, R., Mikolajczyk, K., Puigcerver, J., Toselli, A.H., Sanchez, J.A., Vidal, E.: General overview of ImageCLEF at the CLEF 2016 labs. In: CLEF 2016 Proceedings. Volume 10456 of Lecture Notes in Computer Science., Evora. Portugal, Springer (September 2016)
5. Soldaini, L., Goharian, N.: Quickumls: a fast, unsupervised approach for medical concept extraction. In: MedIR Workshop, SIGIR. (2016)
6. Müller, H., Kalpathy-Cramer, J., Demner-Fushman, D., Antani, S.: Creating a classification of image types in the medical literature for visual categorization. In: SPIE Medical Imaging. (2012)
7. 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, Association for Computational Linguistics (2002) 311–318
8. Katsios, D., Kavallieratou, E.: Concept detection on medical images using deep residual learning network. (2017)

9. Hasan, S.A., Ling, Y., Liu, J., Sreenivasan, R., Anand, S., Arora, T., Datla, V.V., Lee, K., Qadir, A., Swisher, C., Farri, O.: PRNA at ImageCLEF 2017 caption prediction and concept detection tasks. (2017)
10. Pelka, O., Friedrich, C.M.: Keyword generation for biomedical image retrieval with recurrent neural networks. (2017)
11. Pinho, E., Figueira Silva, J.a., Ferreira Silva, J.M., Costa, C.: Towards representation learning for biomedical concept detection in medical images: UA.PT bioinformatics in ImageCLEF 2017. (2017)
12. Lyndon, D., Kumar, A., Kim, J.: Neural captioning for the ImageCLEF 2017 medical image challenges. (2017)
13. Valavanis, L., Stathopoulos, S.: IPL at ImageCLEF 2017 concept detection task. CLEF working notes, CEUR (2017)
14. Liang, S., Li, X., Zhu, Y., Li, X., Jiang, S.: ISIA at ImageCLEF 2017 image caption task. (2017)
15. Mothe, J., Ny Hoavy, N., Randrianarivony, M.I.: IRIT & MISA at ImageCLEF 2017 - multi label classification. (2017)
16. Stefan, L.D., Ionescu, B., Müller, H.: Generating captions for medical images with a deep learning multi-hypothesis approach: ImageCLEF 2017 caption task. (2017)
17. Rahman, M., Lagree, T., Taylor, M.: A cross-modal concept detection and caption prediction approach in ImageCLEFcaption track of ImageCLEF 2017. (2017)
18. Ben Abacha, A., García Seco de Herrera, A., Gayen, S., Demner-Fushman, D., Antani, S.: NLM at ImageCLEF 2017 caption task. CLEF working notes, CEUR (2017)