Overview of the ImageCLEF 2018 Caption Prediction Tasks

Alba G. Seco de Herrera[1], Carsten Eickhoff[2],
Vincent Andrearczyk[3] and Henning Müller[3,4]

[1] University of Essex, UK;
[2] Brown University, Providence RI, United States;
[3] University of Applied Sciences Western Switzerland (HES–SO), Sierre, Switzerland;
[4] University of Geneva, Switzerland.
alba.garcia@essex.ac.uk

Abstract. The caption prediction task is in 2018 in its second edition after the task was first run in the same format in 2017. For 2018 the database was more focused on clinical images to limit diversity. As automatic methods with limited manual control were used to select images, there is still an important diversity remaining in the image data set. Participation was relatively stable compared to 2017. Usage of external data was restricted in 2018 to limit critical remarks regarding the use of external resources by some groups in 2017. Results show that this is a difficult task but that large amounts of training data can make it possible to detect the general topics of an image from the biomedical literature. For an even better comparison it seems important to filter the concepts for the images that are made available. Very general concepts (such as "medical image") need to be removed, as they are not specific for the images shown, and also extremely rare concepts with only one or two examples can not really be learned. Providing more coherent training data or larger quantities can also help to learn such complex models.

Keywords: Caption prediction, Image understanding, radiology

1 Introduction

The caption task described in this paper is part of the ImageCLEF[5] benchmarking campaign [1–4], a framework where researchers can share their expertise and compare their methods based on the exact same data and evaluation methodology in an annual rhythm. ImageCLEF is part of CLEF[6] (Cross Language Evaluation Forum). More on the 2018 campaign in general is described in Ionescu et al. [5] and on the related medical tasks in [6, 7]. In general, ImageCLEF aims at building tasks that are related to clear information needs in medical or non-medical environments [8, 9]. Relationships also exist with the LifeCLEF and CLEFeHealth labs [10, 11].

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

[6] http://www.clef-campaign.org/

The caption task started in 2016 as a pilot task. In 2016, the task was part of the medical image classification task [12, 13], although it unfortunately did not have any participants, also because the questions of the task were not strongly developed at the time. Since 2017, the caption task has been running in the current format. The motivation of this task is the strong increase in available images from the biomedical literature that is growing at an exponential rate and is made available via the PubMed Central® (PMC)[7] repository. As the data set is dominated by compound figures and many general graphs, ImageCLEF has addressed the analysis of compound figures in the past [13]. To extract the image types a hierarchy was created [14], and as training data for these image types are available the global data set of over 5 million images can be filtered to a more homogeneous set containing mainly radiology images as is described in the data preparation section (Section 3). The ImageCLEF caption task aims at better understanding the images in the biomedical literature and extract concepts and captions based only on the visual information of the images (see Figure 1). A further description of the task can be found in Section 2.

This paper presents an overview of the ImageCLEF caption task 2018 including the task and participation in Section 2, the dataset in Section 3 and an explanation of the evaluation framework in Section 4. The participant approaches are described in Section 5, followed by a discussion and the conclusions in Sections 6.

2 Tasks and Participation

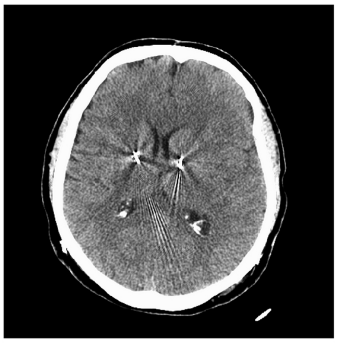
Following 2017 format, the 2018 caption task contains two subtasks, a concept detection subtask that aims at extracting UMLS (Unified Medical Language System®) Concept Unique Identifiers (CUIs) from the images automatically based on the training data made available and a caption prediction subtask that requires to predict a precise text caption for the images in the test data set. Table 1 shows the 8 participants of the task who submitted 44 runs, 28 to the concept detection subtask and 16 to the caption prediction subtask. Three of the groups already participated in 2017, showing that the majority of the participant were new to the task.

It is interesting that despite the fact that the output of the concept detection task can be used for the caption prediction task, none of the participant used such an approach and only two groups participated in both tasks.

Concept Detection As a first step towards automatic image caption and scene understanding, this subtask aims at automatically extracting high-level biomedical concepts (CUIs) from medical images using only the visual content. This approach provides the participating systems with a solid initial building block for image understanding by detecting relevant individual components from which

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

Image:



Concept detection:

- C0589121: Follow-up visit
- C0018946: Hematoma, Subdural
- C1514893: physiologic resolution
- C0546674: Sorbitol dehydrogenase measurement
- C4038402: Low resolution
- C0374531: Postoperative follow-up visit, normally included in the surgical package, to indicate that an evaluation and management service was performed during a postoperative period for a reason(s) related to the original procedure
- C0202691: CAT scan of head
- C1320307: Urgent follow-up
- C3694716: follow-up visit date

Caption prediction: *CT head* at *follow-up visit* demonstrates *resolution* of *SDH.*

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 (Unified Medical Language System$^{®}$) Concept Unique Identifiers (CUIs).

full captions can be composed. The detected concepts are evaluated with a metric based on precision and recall using the concepts extracted from the ground truth captions (see Section 4).

Caption Prediction In this subtask the participants need to predict a coherent caption for the entire medical image. The prediction can be based on the concepts detected in the first subtask as well as the visual analysis of their interaction in the image. Rather than the mere detection of visual concepts, this subtask requires to analyze the interplay of visible elements.

The evaluation of this second subtask is based on metrics such as BLEU scores independent from the first subtask and designed to be robust to variability in style and wording (see Section 4).

Table 1. Participating groups in the 2018 ImageCLEF caption task. Participants marked with a star participated also in 2017.

Team	Institution	# Concept detection	# Caption prediction
UA.PT_Bioinformatics [15] *	DETI - Institute of Electronics and Informatics Engineering, University of Aveiro, Portugal	9	
ImageSem [16]	Institute of Medical Information, Chinese Academy of Medical Sciences/Peking Union Medical College, Beijing, China	7	8
IPL [17] *	Information Processing Laboratory, Athens University of Economics and Business, Athens, Greece	8	
CS MS [18] *	Computer Science Department, Morgan State University, Baltimore, MD, USA	1	1
AILAB	University of the Aegean, Greece	3	
UMass [19]	Umass Medical School, Worcester, MA, USA		5
KU Leuven [20]	Department of Computer Science, KU Leuven, Leuven, Belgium		1
WHU	Wuhan University, Wuhan, Hubei, China		1

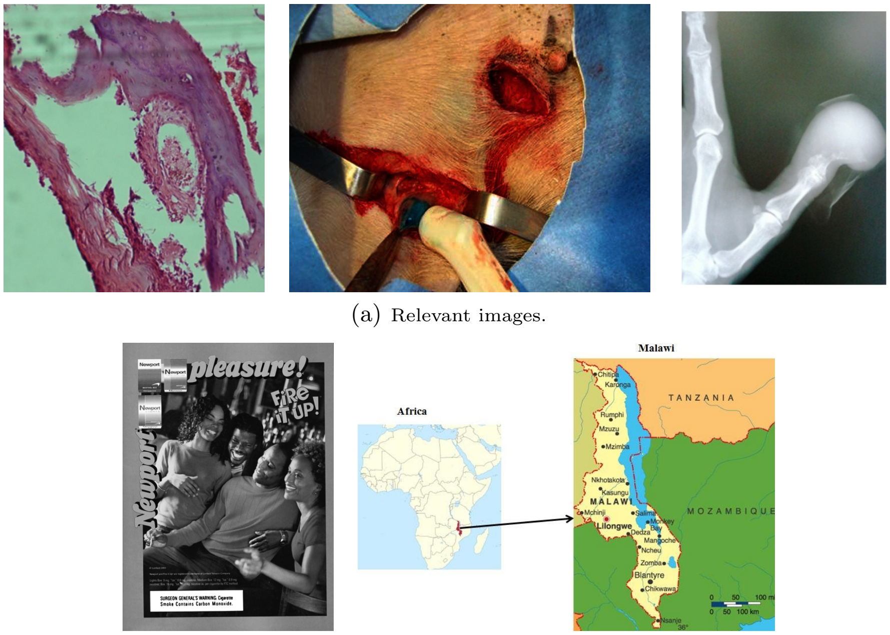
3 Collection

Similarly to previous years, the experimental corpus is derived from scholarly biomedical articles on PMC from which we extract figures and their corresponding captions. As PMC contains many compound and non-clinical figures, we extract a subset of mainly clinical figures to remove noise from the data and focus the challenge on useful radiology/clinical images. The subset was created using a fully automated method based on deep multimodal fusion of Convolutional Neural Networks (CNNs) to classify all 5.8 million images of PMC from 2017 into image types, as described in [21]. This lead to a more homogeneous set of figures than in the 2017 ImageCLEF caption task but diversity still remained high. Besides the removal of many general graphs, also the number of compound figures (i.e. images containing more than one sub figure) was much lower than in 2017. Figure 2 shows some examples of the images contained in the collection including some of the noise that still remained in the data.

In total, the collection is comprised of 232,305 image-caption pairs[8]. This overall set is further split into disjunct training (222,305 pairs) and test (10,000 pairs) sets. For the concept detection subtask, the QuickUMLS library [22] was used to identify UMLS concepts mentioned in the caption text. As a result 111,155 unique UMLS concepts were extracted from the training set.

Table 2 shows examples of the concepts. The average number of concepts per image in the training set is 30 varying between 1 and 1,276. In the training set 2,577 images are labeled with only 1 concept and 3,162 with only 2. Despite the collection being carefully created, there are still non-clinical images (see Figure 2), as all processing was automatic with only limited human checked. There are also non-relevant concepts for the task extracted, again linked to the fact that the data analysis was fully automatic with limited manual quality control. The concepts such as "and","medical image" or "image" are not relevant for

[8] Nine pairs were removed after the challenge started due to incorrect duplicates in the PMC figures.



(a) Relevant images.

(b) Irrelevant images.

Fig. 2. Example of (a) relevant images and (b) problematic images in the 2018 ImageCLEF caption task collection.

the task and not useful to predict from the visual information. Some of the concepts are redundant such as “Marrow - Specimen Source Codes” and “Marrow”. Regardless of the limitations of the annotation the majority of concepts was of good quality and helps to understand the content of the images, as for example the concepts “Marrow” or “X-ray”.

4 Evaluation Methodology

The performance evaluation follows the approach in the previous edition [23] in evaluating both subtasks separately. For the concept detection subtask, the balanced precision and recall trade-off were measured in terms of F_1 scores. Python’s scikit-learn (v0.17.1-2) library was used. Micro F_1 is calculated per image and then the average across all test images is taken as the final measure.

Caption prediction performance is assessed on the basis of BLEU scores [24] 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, Snowball stemming was applied. 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 test images.

Table 2. UMLS (Unified Medical Language System$^{\circledR}$) Concept Unique Identifiers (CUIs) in the 2018 ImageCLEF caption and the number of images labeled with each of those concepts in the training set.

CUI code	concept	number of images
C1550557	RelationshipConjuntion - and	77,003
C1706368	And - dosing instruction fragment	77,003
C1704254	Medical Image	20,165
C1696103	image-dosage form	20,164
C1704922	image	20,164
C3542466	image (foundation metadata concept)	20,164
C1837463	Narrow face	19,491
C1546708	Marrow - Specimen Source Codes	19,253
C0376152	Marrow	19,253
C0771936	Yarrow flower extract	19,079
...	...	...
C0040405	X-Ray Computed Tomography	15,530
...	...	...
C1261259	Wright stain	12,217
C1510508	wright stain	12,137
...	...	...
C1306645	Plain x-ray	10,390
...	...	...
C0412620	CT of abdomen	8,037
C0202823	Chest CT	7,917
...	...	...
C0400569	Simple suture of pancreas	1
C0209088	4-methylcyclohexylamine	1
C0400569	Closed fracture of neck of femur	1

The source code of both evaluation scripts is available on the task Web page at http://imageclef.org/2018/caption.

5 Results

This section shows the results achieved by the participants in both subtasks. Table 3 contains the results of the concept detection subtask and Table 4 contains the results of the caption prediction subtask. None of the participants used external data this year and despite less noise in the 2018 data, no better results were achieved in 2018 compared to 2017, maybe also due to the fact that no external data were used.

5.1 Results for the Concept Detection subTask

28 runs were submitted by 5 groups (see Section 2) to the Concept detection subtasks. Table 2 shows the details of the results. Several approaches were used for the concept detection task, ranging from retrieval systems [16] to deep neural networks. Most research groups implemented at least one approach based on

Table 3. Concept detection performance in terms of F_1 scores.

Team	Run	$MeanF_1$
UA.PT_Bioinformatics	aae-500-o0-2018-04-30_1217	0.1102
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
CS MS	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

deep learning [15, 16, 18], including recurrent networks, various deep CNNs and generative adversarial networks.

Best results were achieved by UA.PT Bioinformatics [15] by applying an adversarial auto-encoder for unsupervised feature learning. They also experimented with a traditional bag of words algorithm, using Oriented FAST and rotated BRIEF (ORB) key point descriptors. UA.PT employed two classification algorithms for concept detection over the learned feature spaces, namely a logistic regression and a variant of k-nearest neighbor (k-NN). Test results showed a best mean F1 score of 0.1102 for linear classifiers, by using the features of the adversarial auto-encoder.

ImageSem [16]was the group following UA.PT Bioinformatics in the ranking. ImageSem was the only group using a retrieval approach, which was more popular in 2017. This approach is based on the open-source Lucene Image Retrieval (LIRE) system used in combination with Latent Dirichlet Allocation (LDA) for clustering concepts of the similar images. ImageSem also experimented with a pre-trained CNN fine-tuned to predict a selected subset of concepts.

IPL [17] proposed a k-NN classifier using two image representation models. One of the methods used is a bag of visual words with dense Scale Invariant Feature Transform (SIFT) descriptors using 4,096 clusters. A second method uses a generalized bag of colors, dividing the image into a codebook of regions of homogeneous colors with 100 or 200 clusters.

The CS MS group [18] used an encoder-decoder model based on a multimodal Recurrent Neural Networks (RNNs). The encoded captions were the input to the RNN via word embedding, while deep image features were encoded via a pre-trained CNN. The combination of the two encoded inputs was used to generate the concepts.

The AILAB used a multimodal deep learning approach based. Instead of using the 220K images, AILAB only used a subset of 4,000 images with feature generation. The visual features are extracted by a pre-trained CNN, while the text features are obtained by word embedding, followed by a Long Short-Term Memory (LSTM) network. The two modalities are then merged and processed by a dense layer to make a final concept prediction.

5.2 Results for the Caption Prediction Task

16 runs were submitted by 5 groups (see Section 2) to the caption prediction subtask. Table 4 shows the details of the results.

ImageSem [16] achieved best results (0.2501 mean BLEU score) using the image retrieval method described in the previous section to combine captions of similar images. Preferred concepts, detected in the concept subtask by high CNN or LDA scores, were also used in other runs to improve the caption generation.

UMass [19] explored and implemented an encoder-decoder framework to generate captions. For the encoder, deep CNN features are used while an LSTM network is used for the decoder. The attention mechanism was also experimented on a smaller sample to evaluate its impact on the model fitting and prediction performance.

Table 4. 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
UMass	test_captions_output4_13_epoch	0.1799
UMass	test_captions_output2_12_epoch	0.1763
CS MS	result_captio	0.1725
UMass	test_captions_output1	0.1696
UMass	test_captions_output5_13_epoch	0.1597
UMass	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

As mentioned in the previous section for concept detection, CS MS [18] also used a similar multimodal deep learning method for caption prediction. A CNN feature extraction of the images was combined with an LSTM on top of word embeddings of the captions. A decoder made of two fully-connected layers produces the captions.

Instead of generating textual sequences directly from images, KU Leuven [20] first learn a continuous representation space for the captions. The representation space is learned by an adverserially regularized autoencoder (ARAE) [25], combining a GAN and the auto-encoder. Subsequently, the task is reduced to learning the mapping from the images to the continuous representation, which is performed by a CNN. The decoder learned in the first step decodes the mapping to a caption for each image.

WHU also used a simple LSTM network that produces a caption by generating one word at every time step conditioned on a context vector, the previous hidden state and the previously generated words.

6 Discussion and Conclusions

The 2018 caption prediction task of ImageCLEF attracted a similar number of participants compared to 2017. No external resources were used, making the task hard, which is also show in the results that overall were lower compared to 2017 despite the training and test data being more homogeneous, which should make the task slightly easier.

Most of the participants used deep learning approaches, but the used networks and architectures varied very strongly. This shows that there is still much research required and that the potential is high to improve results. Also more conventional features extraction and approaches based on retrieval delivered good results, showing also that there are many different ways for creating good models.

The limited participation was partly also linked to the large amount of data made available that caused problems for some research groups. The data set also remains noisy. Only more manual control can likely help creating cleaner data and thus maybe make results of automatic approaches more coherent. Even larger data sets could also help in this direction and really allow to create models for at least more frequently extracted concepts.

References

1. 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)
2. 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
3. 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. Volume 3491 of Lecture Notes in Computer Science (LNCS)., Bath, UK, Springer (2005) 597–613
4. Caputo, B., Muller, H., Thomee, B., Villegas, M., Paredes, R., Zellhofer, D., Goeau, H., Joly, A., Bonnet, P., Gomez, J.M., et al.: Imageclef 2013: the vision, the data and the open challenges. In: International Conference of the Cross-Language Evaluation Forum for European Languages, Springer (2013) 250–268
5. Ionescu, B., Müller, H., Villegas, M., de Herrera, A.G.S., Eickhoff, C., Andrearczyk, V., Cid, Y.D., Liauchuk, V., Kovalev, V., Hasan, S.A., Ling, Y., Farri, O., Liu, J., Lungren, M., Dang-Nguyen, D.T., Piras, L., Riegler, M., Zhou, L., Lux, M., Gurrin, C.: Overview of ImageCLEF 2018: Challenges, datasets and evaluation. In: Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Ninth International Conference of the CLEF Association (CLEF 2018), Avignon, France, LNCS Lecture Notes in Computer Science, Springer (September 10-14 2018)
6. Dicente Cid, Y., Liauchuk, V., Kovalev, V., , Müller, H.: Overview of ImageCLEFtuberculosis 2018 - detecting multi-drug resistance, classifying tuberculosis type, and assessing severity score. In: CLEF2018 Working Notes. CEUR Workshop Proceedings, Avignon, France, CEUR-WS.org <http://ceur-ws.org> (September 10-14 2018)
7. 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, Avignon, France, CEUR-WS.org <http://ceur-ws.org> (September 10-14 2018)
8. Markonis, D., Holzer, M., Dungs, S., Vargas, A., Langs, G., Kriewel, S., Müller, H.: A survey on visual information search behavior and requirements of radiologists. Methods of Information in Medicine **51**(6) (2012) 539–548

9. Müller, H., Despont-Gros, C., Hersh, W., Jensen, J., Lovis, C., Geissbuhler, A.: Health care professionals' image use and search behaviour. In: Proceedings of the Medical Informatics Europe Conference (MIE 2006). IOS Press, Studies in Health Technology and Informatics, Maastricht, The Netherlands (aug 2006) 24–32
10. 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)
11. Goeuriot, L., Kelly, L., Li, W., Palotti, J., Pecina, P., Zuccon, G., Hanbury, A., Jones, G., Müller, H.: Share/clef ehealth evaluation lab 2014, task 3: User-centred health information retrieval clef ehealth overview. In: CLEF Proceedings, Springer LNCS (2014)
12. Villegas, M., Müller, H., de Herrera, A.G.S., Schaer, R., Bromuri, S., Gilbert, A., Piras, L., Wang, J., Yan, F., Ramisa, A., et al.: General overview of imageclef at the clef 2016 labs. In: International conference of the cross-language evaluation forum for European languages, Springer (2016) 267–285
13. García Seco de Herrera, A., Schaer, R., Bromuri, S., Müller, H.: Overview of the ImageCLEF 2016 medical task. In: Working Notes of CLEF 2016 (Cross Language Evaluation Forum). (September 2016)
14. 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)
15. 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, Avignon, France, CEUR-WS.org <http://ceur-ws.org> (September 10-14 2018)
16. 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, Avignon, France, CEUR-WS.org <http://ceur-ws.org> (September 10-14 2018)
17. Valavanis, L., Kalamboukis, T.: IPL at ImageCLEF 2018: A kNN-based concept detection approach. In: CLEF2018 Working Notes. CEUR Workshop Proceedings, Avignon, France, CEUR-WS.org <http://ceur-ws.org> (September 10-14 2018)
18. 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, Avignon, France, CEUR-WS.org <http://ceur-ws.org> (September 10-14 2018)
19. Su, Y., Liu, F.: UMass at ImageCLEF caption prediction 2018 task. In: CLEF2018 Working Notes. CEUR Workshop Proceedings, Avignon, France, CEUR-WS.org <http://ceur-ws.org> (September 10-14 2018)
20. Spinks, G., Moens, M.F.: Generating text from images in a smooth representation space. In: CLEF2018 Working Notes. CEUR Workshop Proceedings, Avignon, France, CEUR-WS.org <http://ceur-ws.org> (September 10-14 2018)
21. 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)
22. Soldaini, L., Goharian, N.: Quickumls: a fast, unsupervised approach for medical concept extraction. In: MedIR workshop, sigir. (2016)
23. Eickhoff, C., Schwall, I., Garcıa Seco de Herrera, A., Müller, H.: Overview of imageclefcaption 2017-image caption prediction and concept detection for biomedical images. CLEF working notes, CEUR (2017)

24. 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
25. Zhao, J.J., Kim, Y., Zhang, K., Rush, A.M., LeCun, Y.: Adversarially regularized autoencoders for generating discrete structures. CoRR, abs/1706.04223 (2017)