Published in "Proceedings of CLEF 2018 Working Notes", 2018, 10-14 September, Avignon, France, which should be cited to refer to this work.

Overview of ImageCLEF 2018 Medical Domain Visual Question Answering Task

Sadid A. Hasan[1], Yuan Ling[1], Oladimeji Farri[1], Joey Liu[1], Henning Müller[2], and Matthew Lungren[3]

[1] Artificial Intelligence Lab, Philips Research North America, Cambridge, MA, USA
{firstname.lastname,dimeji.farri}@philips.com
[2] University of Applied Sciences Western Switzerland (HES-SO), Sierre, Switzerland
henning.mueller@hevs.ch
[3] Department of Radiology, Stanford University, Stanford, CA, USA
mlungren@stanford.edu

Abstract. This paper presents an overview of the inaugural edition of the ImageCLEF 2018 Medical Domain Visual Question Answering (VQA-Med) task. Inspired by the recent success of visual question answering in the general domain, a pilot task was proposed this year to focus on visual question answering in the medical domain. Given medical images accompanied with clinically relevant questions, participating systems were tasked with answering the questions based on the visual image content. A dataset of 6,413 question-answer pairs accompanied with 2,866 medical images extracted from PubMed Central articles was provided; from which, 5,413 question-answer pairs with 2,278 medical images were used for training, 500 question-answer pairs with 324 medical images were used for validation, and 500 questions with 264 medical images were used for testing. Among 28 registered participants, 5 groups submitted a total of 17 runs, indicating a considerable interest in the VQA-Med task.

Keywords: ImageCLEF 2018, Visual Question Answering, Medical Image Interpretation, Question Generation

1 Introduction

With the increasing interest in artificial intelligence (AI) to support clinical decision making and improve patient engagement, opportunities to generate and leverage algorithms for automated medical image interpretation are currently being explored [6, 5]. Since patients may now access structured and unstructured data related to their health via patient portals, such access also motivates the need to help them better understand their conditions regarding their available data, including medical images.

The clinicians' confidence in interpreting complex medical images can be significantly enhanced by a "second opinion" provided by an automated system. In addition, patients may be interested in the morphology/physiology and disease-status of anatomical structures around a lesion that has been well characterized

by their healthcare providers and they may not necessarily be willing to pay significant amounts for a separate office- or hospital visit just to address such questions. Although patients often turn to web search engines to disambiguate complex terms or obtain answers to confusing aspects of a medical image, results from search engines may be nonspecific, erroneous and misleading, or overwhelming in terms of the volume of information.

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[4] [7], we propose a pilot task as part of the ImageCLEF 2018 evaluation campaign[5] [8] to focus on visual question answering in the medical domain (VQA-Med). Given a medical image accompanied with a clinically relevant question, participating systems are tasked with answering the question based on the visual image content.

This paper presents an overview of the VQA-Med task at ImageCLEF 2018. Section 2 introduces the task and Section 3 presents details of the provided corpus. A description of the evaluation methodology is provided in Section 4. We discuss the participant submissions with results in Section 5. Finally, we conclude the paper in Section 6.

2 Task

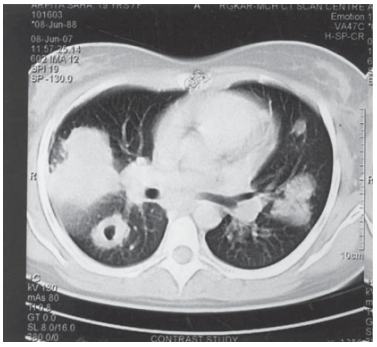
In the inaugural edition we propose a pilot task of visual question answering in the medical domain (VQA-Med) as part of the ImageCLEF 2018 evaluation campaign [8]. Given medical images accompanied with clinically relevant questions, participating systems are tasked with answering the questions based on the visual image content. Figure 1 shows a few example images with associated questions and ground truth answers.

3 Corpus

To create the datasets for the proposed VQA-Med task, we consider medical images along with their captions extracted from PubMed Central articles[6] (essentially a subset of the ImageCLEF 2017 caption prediction task [6]).

We use a semi-automatic approach to generate question-answer pairs from captions of the medical images. First, we automatically generate all possible question-answer pairs from captions using a rule-based question generation (QG) system[7] [4]. The system consists of four modules to automate question generation: 1) sentence simplification, which utilizes clauses, subject, predicate, and verbs to split a long, complex sentence (i.e. the captions associated with the medical images) into multiple simple sentences via lexical alternation and appositive

[4] http://www.visualqa.org/
[5] http://www.imageclef.org/2018
[6] https://www.ncbi.nlm.nih.gov/pmc/
[7] http://www.cs.cmu.edu/~ark/mheilman/questions/



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. 1. Example images with associated question-answer pairs.

identification, 2) answer phrase identification, which identifies relevant phrases from the simple sentences such that corresponding questions can be generated, 3) question generation, where the answer phrases are used to generate possible question phrases through decomposition of the main verb, inversion of the subject and auxiliary verb, and inserting one of the possible question phrases in place of the answer phrases, and 4) candidate questions ranking, where a ranking model is trained to rank the generated candidate questions.

The candidate questions generated via the automatic approach may be noisy as the defined rules may not adequately capture the complex characteristics of medical domain terminologies (clinical concepts) and in particular, the unique

writing style of the medical image captions in biomedical articles. Therefore, two expert human annotators manually check all generated question-answer pairs associated with the medical images in two passes. In the first pass, one annotator proofreads all question-answer pairs and resolves related noises accrued by the aforementioned four modules of the automatic QG system to ensure syntactic and semantic correctness. In the second pass, the other annotator, an expert in clinical medicine, verified all question-answer pairs to form well-curated validation and test sets by ensuring their clinical relevance with respect to associated medical images.

The final curated corpus is comprised of 6,413 question-answer pairs associated with 2,866 medical images. The overall set is 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.

4 Evaluation Methodology

The evaluation of the participant systems of the VQA-Med task is conducted based on three metrics: BLEU, WBSS (Word-based Semantic Similarity), and CBSS (Concept-based Semantic Similarity).

BLEU [1] is used to capture the similarity between a system-generated answer and the ground truth answer. Each answer is converted to lower-case, all punctuations are removed, and the answer is tokenized[8] to individual words. Stopwords are removed using NLTK's[9] English stopword list. Snowball stemming[10] is applied to increase the coverage of overlaps. The overall methodology and resources for the BLEU metric are essentially similar to the ImageCLEF 2017 caption prediction task[11].

Following a recent algorithm to calculate semantic similarity in the biomedical domain [2], we create the WBSS metric based on Wu-Palmer Similarity (WUPS[12]) [3] with WordNet ontology in the backend. 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.

[8] http://www.nltk.org/_modules/nltk/tokenize/punkt.html#PunktLanguageVars.word_tokenize
[9] http://nltk.org/
[10] http://snowball.tartarus.org/texts/introduction.html
[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 Results and Discussion

We received a total of 17 result submissions by 5 different teams from across the world. Table 1 gives an overview of all participants and the number of submitted runs. Note that, there was a limit of maximum 5 run submissions per team. All submitted runs were automatic runs denoting the fact that all participating systems automatically generated answers to the provided questions in the test set.

Overall, most participants used deep learning techniques to build their VQA-Med systems. In particular, participant systems [14–16, 18] leveraged sequence to sequence learning and encoder-decoder-based frameworks [9–11] utilizing deep convolutional neural networks (CNN) to encode medical images (with or without using pre-trained models such as VGG [12], ResNet [13] etc.) and recurrent neural networks (RNN) to generate question encodings (with or without using pre-trained word embeddings). Some participants formulated the VQA-Med task as a multi-label multi-class classification problem [14, 17] while others considered it as a generation task [18]. Participants also used attention-based mechanisms [15–17] 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) [15], the use of advanced techniques such as multimodal compact bilinear (MCB) pooling [15] or multimodal factorized bilinear (MFB) pooling [17] to combine multimodal features, the use of embedding based topic modeling (ETM) [17], and 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 overall results of the participating systems are presented in Table 2 to Table 4 for the three different metrics in a descending order of the scores (the higher the better). The relatively low BLEU scores and WBSS scores of the runs 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.

Table 1. Participating groups.

Team	**Institution**	**#Runs**
FSTT [14]	Abdelmalek Essaadi University, Faculty of Sciences and Techniques, Tangier, Morocco	2
JUST [18]	Jordan University of Science and Technology, Jordan	3
NLM [15]	Lister Hill National Center for Biomedical Communications, National Library of Medicine, Bethesda, MD, USA	5
TU [16]	Tokushima University, Japan	3
UMMS [17]	University of Massachusetts Medical School, Worcester, MA, USA	4

Table 2. BLEU scores of all submitted runs.

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

Table 3. WBSS scores of all submitted runs.

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

Table 4. CBSS scores of all submitted runs.

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

6 Conclusion

This paper presented an overview of the inaugural Medical Domain Visual Question Answering (VQA-Med) challenge conducted as a part of the ImageCLEF 2018 evaluation campaign. We discussed participant submissions and results, which demonstrated the challenges and complexities of the VQA-Med task. In the future, we would consider the interesting data analyses and improvement suggestions presented in [15–17] and plan to increase the dataset size to leverage the power of advanced deep learning algorithms towards improving the state-of-the-art in visual question answering in the medical domain.

References

1. Papineni, K.; Roukos, S.; Ward, T.; Zhu, W. J. (2002). BLEU: a method for automatic evaluation of machine translation (PDF). ACL-2002: 40th Annual meeting of the Association for Computational Linguistics. pp. 311318.
2. Soancolu, G., ztrk, H., & zgr, A. (2017). BIOSSES: a semantic sentence similarity estimation system for the biomedical domain.Bioinformatics,33(14), i49-i58.
3. Wu, Z., & Palmer, M. (1994, June). Verbs semantics and lexical selection. InProceedings of the 32nd annual meeting on Association for Computational Linguistics(pp. 133-138). Association for Computational Linguistics.
4. Heilman, M., & Smith, N. A. (2009). Question Generation via Overgenerating Transformations and Ranking. Language Technologies Institute, Carnegie Mellon University Technical Report CMU-LTI-09-013.
5. Ionescu, B., Mller, H., Villegas, M., Arenas, H., Boato, G., Dang-Nguyen, D., Dicente Cid, Y., Eickhoff, C., Garcia Seco de Herrera, A., Gurrin, C., Islam, B., Kovalev, V., Liauchuk, V., Mothe, J., Piras, L., Riegler, M., and Schwall, I. (2017).

Overview of ImageCLEF 2017: Information extraction from images. Experimental IR Meets Multilinguality, Multimodality, and Interaction. 8th International Conference of the CLEF Association, CLEF, Springer LNCS 10456.
6. Eickhoff, C., Schwall, I., Garca Seco de Herrera, A., and Mller, H. (2017). Overview of ImageCLEFcaption 2017 - Image Caption Prediction and Concept Detection for Biomedical Images. CLEF 2017 Labs Working Notes, CEUR Workshop Proceedings.
7. Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Lawrence Zitnick, C., and Parikh, D. (2015). VQA: Visual Question Answering. ICCV.
8. Ionescu, B., Mller, H., Villegas, M., Garca Seco de Herrera, A., Eickhoff, C., Andrearczyk, V., Dicente Cid, Y., Liauchuk, V., Kovalev, V., Hasan, S. A., Ling, Y., Farri, O., Liu, J., Lungren, M., Dang-Nguyen, D., Piras, L., Riegler, M., Zhou, L., Lux, M., and Gurrin, C. (2018). Overview of ImageCLEF 2018: Challenges, Datasets and Evaluation. Proceedings of the 9th International Conference of the CLEF Association, CLEF.
9. Bahdanau, D., Cho, K., and Bengio, Y. (2015). Neural machine translation by jointly learning to align and translate. ICLR.
10. Sutskever, I., Vinyals, O., and Le, Q. V. (2014). Sequence to Sequence Learning with Neural Networks. NIPS: 3104-3112.
11. Cho, K., van Merrienboer, B., Glehre, C., Bahdanau, D., Bougares, F., Schwenk, H., and Bengio, Y. (2014). Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. EMNLP: 1724-1734.
12. Simonyan, K., and Zisserman, A. (2014). Very Deep Convolutional Networks for Large-Scale Image Recognition. arXiv:1409.1556.
13. He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep Residual Learning for Image Recognition. CVPR: 770-778.
14. Allaouzi, I., Benamrou, B., Ben Ahmed, M. (2018). Deep Neural Networks and Decision Tree classifier for Visual Question Answering in the medical domain. CLEF 2018 Labs Working Notes, CEUR Workshop Proceedings.
15. Ben Abacha, A., Gayen, S., J Lau, J., Rajaraman, S., and Demner-Fushman, D. (2018). NLM at ImageCLEF 2018 Visual Question Answering in the Medical Domain. CLEF 2018 Labs Working Notes, CEUR Workshop Proceedings.
16. Zhou, Y., Kang, X., and Ren, F. (2018). Employing Inception-Resnet-v2 and Bi-LSTM for Medical Domain Visual Question Answering. CLEF 2018 Labs Working Notes, CEUR Workshop Proceedings.
17. Peng, Y., Liu, F., and Rosen, M. (2018). UMass at ImageCLEF Medical Visual Question Answering (Med-VQA) 2018 Task. CLEF 2018 Labs Working Notes, CEUR Workshop Proceedings.
18. Talafha, B., and Al-Ayyoub, M. (2018). JUST at VQA-Med: A VGG-Seq2Seq Model. CLEF 2018 Labs Working Notes, CEUR Workshop Proceedings.