Overview of the VQA-Med Task at ImageCLEF 2020: Visual Question Answering and Generation in the Medical Domain

Asma Ben Abacha[1], Vivek V. Datla[2], Sadid A. Hasan[3], Dina Demner-Fushman[1], and Henning Müller[4]

[1] Lister Hill Center, National Library of Medicine, USA
[2] Philips Research Cambridge, USA
[3] CVS Health, USA
[4] University of Applied Sciences Western Switzerland, Sierre, Switzerland
asma.benabacha@nih.gov, vivek.datla@philips.com, sadidhasan@gmail.com, ddemner@mail.nih.gov, henning.mueller@hevs.ch

Abstract. This paper presents an overview of the Medical Visual Question Answering (VQA-Med) task at ImageCLEF 2020. This third edition of VQA-Med included two tasks: (i) Visual Question Answering (VQA), where participants were tasked with answering abnormality questions from the visual content of radiology images and (ii) Visual Question Generation (VQG), consisting of generating relevant questions about radiology images based on their visual content. In VQA-Med 2020, 11 teams participated in at least one of the two tasks and submitted a total of 62 runs. The best team achieved a BLEU score of 0.542 in the VQA task and 0.348 in the VQG task.

Keywords: Visual Question Answering, Visual Question Generation, Data Creation, Radiology Images, Medical Questions and Answers

1 Introduction

With the increasing interest in artificial intelligence technologies to support clinical decision making and improve patient engagement, opportunities to generate and leverage algorithms for automated medical image interpretation are being explored at a faster pace. The clinicians' confidence in interpreting complex medical images can be enhanced by a "second opinion" provided by an automated system. Also, since patients may now access structured and unstructured data related to their health via patient portals, such access motivates the need to help them better understand their conditions regarding their available data, including medical images.

To offer more training data and evaluation benchmarks, we organized the first visual question answering (VQA) task in the medical domain in 2018 [4], and continued the task in 2019 [2] as part of the ImageCLEF initiatives [6]. Following the strong engagement from the research community in both editions of VQA in the medical domain (VQA-Med) and the ongoing interests from both the computer vision and the medical informatics communities, we continued the task this year (VQA-Med 2020) within the scope of ImageCLEF-2020 initiatives [5] by putting an enhanced focus on answering questions about abnormalities from the visual content of associated radiology images. Furthermore, we introduced an additional task this year, visual question generation (VQG), consisting of generating relevant questions about radiology images.

2 Task Description

For the visual question answering task, similar to 2019, given a radiology medical image accompanied by a clinically relevant question, participating systems were tasked with answering the question based on the visual image content. In VQA-Med 2020, we specifically focused on questions about abnormality (e.g., "what is most alarming about this ultrasound image?"), which can be answered from the image content without requiring additional medical knowledge or domain-specific inference. Additionally, the visual question generation (VQG) task was introduced for the first time in this third edition of the VQA-Med challenge. This task required participants to generate relevant natural language questions about radiology images using their visual content.

3 Data Creation

3.1 VQA Data

For the visual question answering task, we automatically constructed the training, validation, and test sets by: (i) applying several filters to select relevant images and associated annotations, and, (ii) creating patterns to generate the questions and their answers. We selected relevant medical images from the MedPix[5] database with filters based on their captions, localities, and diagnosis methods. We selected only the cases where the diagnosis was made based on the image. Examples of the selected diagnosis methods include: CT/MRI imaging, angiography, characteristic imaging appearance, radiographs, imaging features, ultrasound, and diagnostic radiology.

Finally, we selected the list of abnormalities to be used to create the question-answer pairs. The final list covers 330 medical problems; each problem occurs at least 10 times in the created VQA data.

Examples of medical problems (and their frequency) in the VQA data:

- pulmonary embolism (114),

[5] `https://medpix.nlm.nih.gov/`

– acute appendicitis (109),
– angiomyolipoma (68),
– osteochondroma (63),
– adenocarcinoma of the lung (60),
– sarcoidosis (58).

The VQA training set includes 4,000 radiology images with 4,000 Question-Answer (QA) pairs. The validation set consists of 500 radiology images with 500 QA pairs. The test set includes 500 radiology images and 500 questions. To further ensure the quality of the data, the test set was manually validated by a medical doctor. Figure 1 presents examples from the VQA-Med-2020 test set. The participants were also encouraged to utilize the VQA-Med-2019 dataset as additional training data.

3.2 VQG Data

For the visual question generation task, we automatically constructed the training, validation, and test sets in a similar fashion by using a separate collection of radiology images and their associated captions. We semi-automatically generated questions from the image captions first by using a rule-based sentence-to-question generation approach[6], and then, three annotators manually curated the list of question-answer pairs by removing or editing the noises related to grammatical inconsistencies. The final curated corpus for the VQG task was comprised of 780 radiology images with 2,156 associated questions (and answers) for training, 141 radiology images with 164 questions for validation, and 80 radiology images for testing.

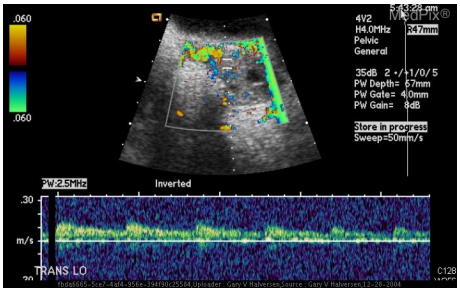
4 Submitted Runs

Out of 47 online registrations, 30 participants submitted signed end user agreement forms. Finally, 11 groups submitted a total of 49 successful runs for the VQA task[7] (cf. Figure 2), while 3 groups submitted a total of 13 successful runs for the VQG task[8], indicating a notable interest in the VQA-Med 2020 challenge. Table 1 and Table 2 give an overview of all participants and the number of submitted runs (please note that were allowed only 5 runs per team).

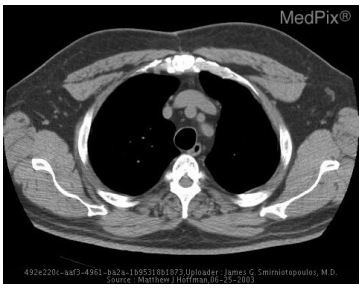
5 Results

Similar to the evaluation setup of the VQA-Med 2019 challenge [2], the evaluation of the participant systems for the VQA task in the VQA-Med 2020 challenge is also conducted based on two primary metrics: accuracy and BLEU. We used

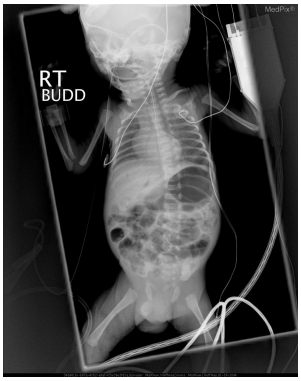
[6] http://www.cs.cmu.edu/~ark/mheilman/questions/
[7] https://www.aicrowd.com/challenges/imageclef-2020-vqa-med
[8] https://www.aicrowd.com/challenges/imageclef-2020-vqa-med-vqg



(a) **Q**: what abnormality is seen in the image? **A**: ovarian torsion



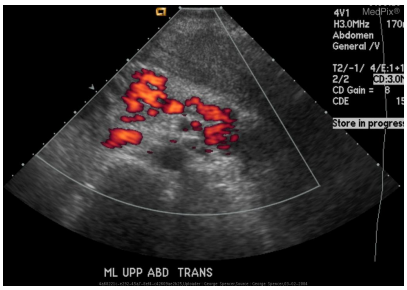
(b) **Q**: what is abnormal in the ct scan? **A**: partial anomalous pulmonary venous return



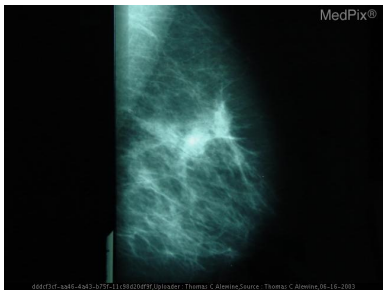
(c) **Q**: what is the primary abnormality in this image? **A**: necrotizing enterocolitis

(d) **Q**: is the x-ray normal? **A**: no

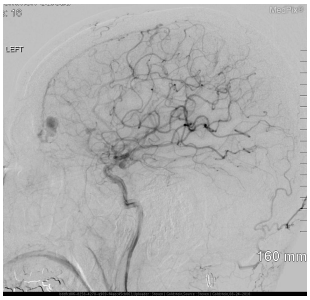
(e) **Q**: what abnormality is seen in the image? **A**: ollier's disease, enchondromatosis



(f) **Q**: what is abnormal in the ultrasound? **A**: cirrhosis of the liver



(g) **Q**: what is abnormal in the mammograph? **A**: infiltrating ductal carcinoma



(h) **Q**: what is the primary abnormality in this image? **A**: dural fistula, avf

Fig. 1: Examples from the Test Set of the VQA Task

Table 1: Participating groups in the VQA-Med 2020 VQA task.

Team	*Institution (s)*	*# Runs*
bumjun_jung [7]	University of Tokyo and RIKEN AIP (Japan)	5
dhruv_sharma	Virginia Tech (USA)	1
HCP-MIC [3]	School of Data and Computer Science, Sun Yat-Sen University (China)	5
HARENDRAKV [14]	Vadict Innovations and Quest Global (India)	5
kdevqa [13]	Toyohashi University of Technology (Japan)	4
NLM [11]	U.S. National Library of Medicine (USA)	5
sheerin	Individual participation (India)	5
Shengyan [9]	School of Information Science and Engineering, Yunnan University (China)	5
TheInceptionTeam [1]	Jordan University of Science and Technology (Jordan)	5
umassmednlp	University of Massachusetts Medical School (USA)	4
AIML [8]	The Australian Institute for Machine Learning, University of Adelaide and South Australian Health and Medical Research Institute (Australia)	5

Table 2: Participating groups in the VQA-Med 2020 VQG task.

Team	*Institution (s)*	*# Runs*
NLM [11]	U.S. National Library of Medicine (USA)	3
TheInceptionTeam [1]	Jordan University of Science and Technology (Jordan)	5
AIML [8]	The Australian Institute for Machine Learning, University of Adelaide and South Australian Health and Medical Research Institute (Australia)	5

an adapted version of accuracy from the general domain VQA[9] task that strictly considers exact matching of a participant provided answer and the ground truth answer. To compensate for the strictness of the accuracy metric, BLEU [10] is used to capture the word overlap-based similarity between a system-generated answer and the ground truth answer. The overall methodology and resources for the BLEU metric are essentially similar to last year's VQA task [2]. The BLEU metric is also used to evaluate the submissions for the VQG task, where we essentially compute the word overlap-based average similarity score between the system-generated questions and the ground truth question for each given test image. The overall results of the participating systems are presented in Table 3 and Table 4 in a descending order of the accuracy and average BLEU scores respectively (the higher the better).

6 Discussion

Similar to the last two years, participants continued to use state-of-the-art deep learning techniques to build their VQA-Med systems for both VQA and VQG

[9] `https://visualqa.org/evaluation.html`

Table 3: Maximum Accuracy and Maximum BLEU Scores for VQA Task (out of each team's submitted runs).

Team	*Accuracy*	*BLEU*
AIML	0.496	0.542
TheInceptionTeam	0.480	0.511
bumjun_jung	0.466	0.502
HCP-MIC	0.426	0.462
NLM	0.400	0.441
HARENDRAKV	0.378	0.439
Shengyan	0.376	0.412
kdevqa	0.314	0.350
sheerin	0.282	0.330
umassmednlp	0.220	0.340
dhruv_sharma	0.142	0.177

Table 4: Maximum Average BLEU Scores for VQG Task (out of each team's submitted runs).

Team	*Average BLEU*
AIML	0.348
TheInceptionTeam	0.339
NLM	0.116

tasks [4, 2]. In particular, most systems leveraged encoder-decoder architectures with, e.g., deep convolutional neural networks (CNNs) like VGGNet or ResNet. A variety of pooling strategies were explored, e.g., global average pooling to encode image features and transformer-based architectures like BERT or recurrent neural networks (RNN) to extract question features (for the VQA task). Various types of attention mechanisms are also used coupled with different pooling strategies such as multimodal factorized bilinear (MFB) pooling or multi-modal factorized high-order pooling (MFH) in order to combine multimodal features followed by bilinear transformations to finally predict the possible answers in the VQA task and generate possible question words in the VQG task. Additionally, the top performing systems first classified the questions into two types: yes/no, and abnormality, then added another multi-class classification framework for abnormality-related question answering, while using the same backbone architecture along with utilizing additional training data, leading to better results.

Analyses of the results in Table 3 suggest that in general, participating systems performed well for the VQA task and achieved better accuracy relatively compared to last year's results for answering abnormality-related questions [2]. They obtained slightly lower BLEU scores as we focused on only abnormality questions this year that are generally complex than modality, plane, or organ category questions given in the last year. Overall, the VQA task results obtained this year entail the robustness of the provided dataset compared to last year's task due to the enhanced focus on the abnormality-related questions for corpus

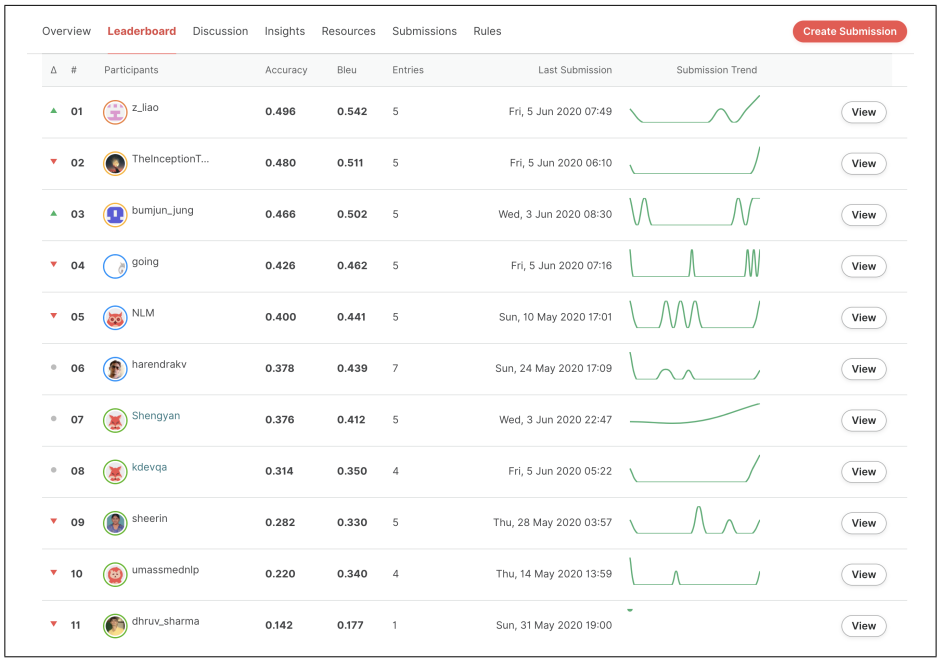


Fig. 2: Results of the VQA Task on the AICrowd platform

creation. For the VQG task, results in Table 4 suggest that the task was comparatively more challenging than the VQA task as the systems achieved lower BLEU scores. As BLEU is not the ideal metric to semantically compare the generated questions with the ground-truth questions, this could also urge the necessity of an embedding-based similarity metric to be explored in the future edition of this task.

7 Conclusion

In this paper, we presented the VQA-Med 2020 tasks, datasets, and official results. We created new datasets for the visual question generation and visual question answering tasks with a focus on questions about abnormality. In the VQA task, the best team achieved 0.542 BLEU score and 0.496 accuracy. The VQG task was more challenging, with a best BLEU score of 0.348. In the future editions of VQA-Med, we will focus on expanding the VQG dataset with more images and questions [12] to enable effective development of deep learning models and on designing new evaluation metrics for both tasks.

References

1. Al-Sadi, A., Al-Theiabat, H., Al-Ayyoub, M.: The inception team at vqa-med 2020: Pretrained vgg with data augmentation for medical vqa and vqg. In: CLEF 2020 Working Notes. CEUR Workshop Proceedings, CEUR-WS.org, Thessaloniki, Greece (September 22-25 2020)
2. Ben Abacha, A., Hasan, S.A., Datla, V.V., Liu, J., Demner-Fushman, D., Müller, H.: Vqa-med: Overview of the medical visual question answering task at imageclef 2019. In: 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)
3. Chen, G., Gong, H., Li, G.: Hcp-mic at vqa-med 2020: Effective visual representation for medical visual question answering. In: CLEF 2020 Working Notes. CEUR Workshop Proceedings, CEUR-WS.org, Thessaloniki, Greece (September 22-25 2020)
4. Hasan, S.A., Ling, Y., Farri, O., Liu, J., Müller, H., Lungren, M.: Overview of imageclef 2018 medical domain visual question answering task. In: Working Notes of CLEF 2018 - Conference and Labs of the Evaluation Forum, Avignon, France, September 10-14, 2018. (2018)
5. Ionescu, B., Müller, H., Péteri, R., Ben Abacha, A., 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 lifelogging, medical, 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)
6. Ionescu, B., Müller, H., Péteri, R., Cid, Y.D., Liauchuk, V., Kovalev, V., Klimuk, D., Tarasau, A., Ben Abacha, A., Hasan, S.A., Datla, V., Liu, J., Demner-Fushman, D., Dang-Nguyen, D.T., Piras, L., Riegler, M., Tran, M.T., Lux, M., Gurrin, C., Pelka, O., Friedrich, C.M., de Herrera, A.G.S., Garcia, N., Kavallieratou, E., del Blanco, C.R., Rodríguez, C.C., Vasillopoulos, N., Karampidis, K., Chamberlain, J., Clark, A., Campello, A.: ImageCLEF 2019: Multimedia retrieval in medicine, lifelogging, security and nature. In: Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the 10th International Conference of the CLEF Association (CLEF 2019), LNCS Lecture Notes in Computer Science, Springer, Lugano, Switzerland (September 9-12 2019)
7. Jung, B., Gu, L., Harada, T.: bumjun_jung at vqa-med 2020: Vqa model based on feature extraction and multi-modal feature fusion. In: CLEF 2020 Working Notes. CEUR Workshop Proceedings, CEUR-WS.org, Thessaloniki, Greece (September 22-25 2020)
8. Liao, Z., Wu, Q., Shen, C., van den Hengel, A., Verjans, J.: Aiml at vqa-med 2020: Knowledge inference via a skeleton-based sentence mapping approach for medical domain visual question answering. In: CLEF 2020 Working Notes. CEUR Workshop Proceedings, CEUR-WS.org, Thessaloniki, Greece (September 22-25 2020)
9. Liu, S., Ding, H., Zhou, X.: Shengyan at vqa-med 2020: An encoder-decoder model for medical domain visual question answering task. In: CLEF 2020 Working Notes.

CEUR Workshop Proceedings, CEUR-WS.org, Thessaloniki, Greece (September 22-25 2020)
10. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: BLEU: a method for automatic evaluation of machine translation. In: Proceedings of the 40th annual meeting on association for computational linguistics. pp. 311–318. Association for Computational Linguistics (2002)
11. Sarrouti, M.: Nlm at vqa-med 2020: Visual question answering and generation in the medical domain. In: CLEF 2020 Working Notes. CEUR Workshop Proceedings, CEUR-WS.org, Thessaloniki, Greece (September 22-25 2020)
12. Sarrouti, M., Ben Abacha, A., Demner-Fushman, D.: Visual question generation from radiology images. In: Proceedings of the first workshop on Advances in Language and Vision Research (ALVR). Association for Computational Linguistics, Seattle, Washington (July 2020), https://alvr-workshop.github.io/proceedings/ALVR_2020_15_Paper.pdf
13. Umada, H., Aono, M.: kdevqa at vqa-med 2020: focusing on glu-based classification. In: CLEF 2020 Working Notes. CEUR Workshop Proceedings, CEUR-WS.org, Thessaloniki, Greece (September 22-25 2020)
14. Verma, H.K., S., S.R.: Harendrakv at vqa-med 2020: Sequential vqa with attention for medical visual question answering. In: CLEF 2020 Working Notes. CEUR Workshop Proceedings, CEUR-WS.org, Thessaloniki, Greece (September 22-25 2020)