

# Overview of the VQA-Med Task at ImageCLEF 2021: Visual Question Answering and Generation in the Medical Domain

Asma Ben Abacha<sup>1</sup>, Mourad Sarrouiti<sup>1</sup>, Dina Demner-Fushman<sup>1</sup>, Sadid A. Hasan<sup>2</sup> and Henning Müller<sup>3</sup>

<sup>1</sup>National Library of Medicine, USA

<sup>2</sup>CVS Health, USA

<sup>3</sup>University of Applied Sciences Western Switzerland (HES-SO), Sierre, Switzerland

## Abstract

This paper presents an overview of the fourth edition of the Medical Visual Question Answering (VQA-Med) task at ImageCLEF 2021. VQA-Med 2021 includes a task on Visual Question Answering (VQA), where participants are tasked with answering questions from the visual content of radiology images, and a second task on Visual Question Generation (VQG), consisting of generating relevant questions about radiology images. Thirteen teams participated in VQA-Med 2021 and submitted a total of 75 runs. The best teams achieved a BLEU score of 0.416 in the VQA task and 0.383 in the VQG task.

## Keywords

Visual Question Answering, Visual Question Generation, Data Creation, Radiology Images

## 1. Introduction

Visual Question Answering is a challenging and promising problem that combines natural language processing (NLP) and computer vision (CV) techniques. With the increasing interest in artificial intelligence (AI) 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. To offer more training data and evaluation benchmarks, we organized the first visual question answering (VQA) task in the medical domain in 2018 [1], and continued the task in 2019 [2] and 2020 [3].

Following the strong engagement from the research community in the previous editions of VQA in the medical domain (VQA-Med), we continued the task this year within the scope of ImageCLEF 2021 [4], with a focus on answering questions about abnormalities in radiology images. In this edition, we also organized a second task on visual question generation (VQG), consisting of generating relevant natural language questions about radiology images based on their visual content<sup>1</sup>.

---

CLEF 2021 – Conference and Labs of the Evaluation Forum, September 21–24, 2021, Bucharest, Romania

✉ asma.benabacha@nih.gov (A. Ben Abacha); mourad.sarrouiti@nih.gov (M. Sarrouiti); ddemner@mail.nih.gov (D. Demner-Fushman); sadidhasan@gmail.com (S. A. Hasan); henning.mueller@hevs.ch (H. Müller)



© 2021 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR Workshop Proceedings (CEUR-WS.org)

<sup>1</sup><https://www.imageclef.org/2021/medical/vqa>

## 2. Task Description

The two VQA-Med tasks can be described more precisely as follows:

- **Visual question answering (VQA)**<sup>2</sup>: given a radiology image accompanied by a relevant question, participating systems in VQA-Med 2021 were tasked with answering the question based on the visual image content.
- **Visual question generation (VQG)**<sup>3</sup>: given a radiology image, participating systems were tasked with generating relevant natural language questions about the abnormality present in the image.

## 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<sup>4</sup> 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. Finally, we considered the most frequent abnormality question categories to create the data set, which included a training set of 4,500 radiology images with 4,500 question-answer (QA) pairs (the same dataset used in 2020), a new validation set of 500 radiology images with 500 QA pairs, and a new test set of 500 radiology images with 500 questions about Abnormality. To further ensure the quality of the data, the reference answers of the test set were manually validated by a medical doctor. Figure 1 presents examples from the VQA 2021 test set. The participants were also encouraged to utilize the VQA-Med 2019 and 2020 datasets as additional training data.

### 3.2. VQG Data

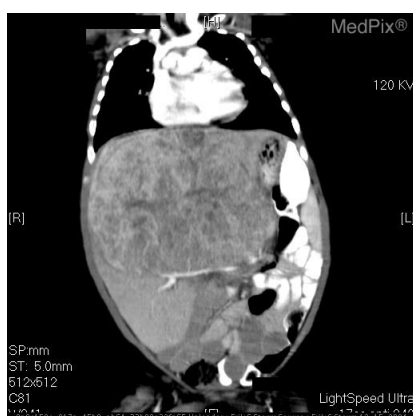
For the visual question generation task, we constructed the validation and test sets semi-automatically. First, we generated questions automatically from the images and their captions using two different approaches. In a first approach, we used only the image and a variational autoencoder model called VQGR [5] trained on the VQA-RAD dataset [6] (A CNN was used to encode the images and an LSTM to decode the questions). The second approach used a T5-based model fine-tuned on the SQuAD and MS MARCO datasets to generate questions from the image captions. Then, a medical doctor curated the list of automatically created questions. The final curated corpus for the VQG task was comprised of 85 radiology images with 200 questions for validation and 100 radiology images with 302 reference questions for the test set. Figure 2 presents examples from the VQG 2021 test set.

---

<sup>2</sup><https://www.aicrowd.com/challenges/imageclef-2021-vqa-med-vqa>

<sup>3</sup><https://www.aicrowd.com/challenges/imageclef-2021-vqa-med-vqg>

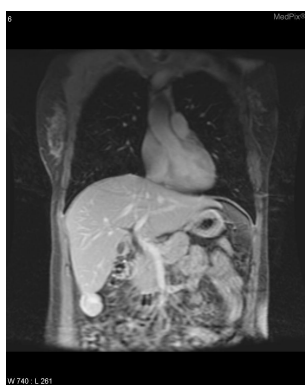
<sup>4</sup><https://medpix.nlm.nih.gov/>



(a) **Q:** What is the primary abnormality in this image? **A:** Large heterogeneously enhancing right hepatic mass with mass effect on left lobe consistent with hepatoblastoma



(b) **Q:** What is abnormal in the x-ray? **A:** enchondroma | Lytic lesion with chondroid matrix of the proximal metadiaphysis of the humerus



(c) **Q:** What is most alarming about this mri? **A:** focal nodular hyperplasia



(d) **Q:** What abnormality is seen in the image? **A:** Enhancing lesion right parietal lobe with surrounding edema

**Figure 1:** Examples from the test set of the VQA 2021 Task

## 4. Submitted Runs

Out of 48 online registrations, 33 participants submitted signed end user agreement forms, and 13 teams submitted a total of 75 successful runs; including 68 runs for the VQA task and 7 runs for the VQG task. Table 1 gives an overview of all participating teams and the number of submitted runs (only 10 runs were allowed per team).

**Table 1**

Participating groups in the VQA-Med 2021 tasks.

| <i>Team</i>          | <i>Institution</i>                                                                   | <i># Valid Runs</i> |
|----------------------|--------------------------------------------------------------------------------------|---------------------|
| Yunnan [7]           | Yunnan University (China)                                                            | 10                  |
| SYSU-HCP [8]         | School of Computer Science and Engineering,<br>Sun Yat-sen University (China)        | 10                  |
| TAM [9]              | South China Normal University (China)                                                | 10                  |
| TeamS [10]           | D4L data4life gGmbH&Hasso Plattner Institute (Germany)                               | 10                  |
| jeanbenoit_delbrouck | Stanford University (USA)                                                            | 10                  |
| sheerin [11]         | Siva Subramaniya Nadar College of Engineering (India)                                | 5                   |
| IALab_PUC [12]       | IALab group of the Pontifical Catholic University (Chile)                            | 5                   |
| Chabbiimen [13]      | REGIM Lab & Higher Institute of Informatics and Communication Technologies (Tunisia) | 5                   |
| SSN_hacML            | SSN College of Engineering, Chennai (India)                                          | 3                   |
| Lijie [14]           | School of Information Science and Engineering, Yunnan University (China)             | 2                   |
| sliencec             | SIE of NCU, Nanchang (China)                                                         | 2                   |
| riven                | SEU, Suzhou (China)                                                                  | 1                   |

**Table 2**

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

| <i>Team</i>          | <i>Accuracy</i> | <i>BLEU</i> |
|----------------------|-----------------|-------------|
| SYSU-HCP             | 0.382           | 0.416       |
| Yunnan University    | 0.362           | 0.402       |
| TeamS                | 0.348           | 0.391       |
| jeanbenoit_delbrouck | 0.348           | 0.384       |
| riven                | 0.332           | 0.361       |
| Lijie                | 0.316           | 0.352       |
| IALab_PUC            | 0.236           | 0.276       |
| TAM                  | 0.222           | 0.255       |
| sliencec             | 0.220           | 0.235       |
| sheerin              | 0.196           | 0.227       |
| SSN_hacML            | 0.000           | 0.002       |
| Baseline 1           | 0.288           | 0.326       |
| Baseline 2           | 0.134           | 0.156       |

**Table 3**

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

| <i>Team</i> | <i>Average BLEU</i> |
|-------------|---------------------|
| Chabbiimen  | 0.383               |
| Baseline    | 0.274               |

## 5. Results

Similar to the evaluation setup of the VQA-Med 2020 challenge [3], the evaluation of the participant systems for the VQA task in VQA-Med 2021 is also conducted based on two primary metrics: accuracy and BLEU. We used an adapted version of accuracy from VQA in the open domain<sup>5</sup> that relies on an exact matching between a participant provided answer and the ground truth answer. To compensate for the strictness of the accuracy metric, BLEU [15] 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. The BLEU metric is also used to evaluate the submissions for the VQG task to compute an overlap-based average similarity score between the system-generated questions and the ground truth question for each given test image<sup>6</sup>.

We prepared three baseline systems for the VQA and VQG tasks. Our VQA baselines are based on a multi-class image classification approach using ResNet50 (baseline 1) and a variational autoencoder model (baseline 2) trained on the VQA-Med data [16]. Our VQG baseline system relies on a variational autoencoder model trained on the VQA-RAD and VQA-Med datasets [5].

The overall results of the participating systems and our baselines are presented in Table 2 and Table 3 in descending accuracy order and average BLEU scores, respectively.

## 6. Discussion

The results in Table 2 show that participating systems performed relatively well for the VQA task, in comparison with the VQG results, presented in Table 3, and suggesting that the VQG task was more challenging. However, the participating systems achieved better BLEU scores compared to last year’s VQG results [3].

The participants’ approaches relied on state-of-the-art deep learning techniques for the VQA and VQG tasks. Most systems used Convolutional Neural Networks (CNNs) for visual feature extraction such as VGGNet, ResNet, and DenseNet. Long-short-term memory (LSTM) networks and Transformer-based models (e.g. BERT, BioBERT) were used to extract question features. Several pooling strategies were explored such as multimodal factorized bilinear "MFB" pooling or multi-modal factorized high-order "MFH" pooling to combine image and question features and generate the answer (e.g. [7, 9]).

Participating teams also applied various attention mechanisms and ensemble methods. For instance, the SYSU-HCP team [8] designed a hierarchical feature extraction structure to capture multi-scale features of radiology images and replaced the fully-connected layers with hierarchical adaptive global average pooling layers. For training, they used three techniques: data augmentation, curriculum learning, and label smoothing. Their final system relied on a multi-architecture ensemble combining the output of eight models and achieving the best accuracy of 0.382 and BLEU score of 0.416.

---

<sup>5</sup><https://visualqa.org/evaluation.html>

<sup>6</sup><https://github.com/abachaa/VQA-Med-2021/tree/main/EvaluationCode>

## 7. Conclusion

In this paper, we presented the ImageCLEF VQA-Med 2021 tasks and official results. We created new datasets for the visual question generation and visual question answering tasks with a more pronounced focus on questions about abnormality. For the VQG task, we explored the use of Deep Learning and Transformer-based models for semi-automatic question generation from the images and their captions. The VQA-Med task attracted high participation in ImageCLEF 2021. The best VQA team achieved 0.416 BLEU score and 0.382 accuracy. For the VQG task, the best BLEU score is 0.383, outperforming the results achieved last year. We hope that these VQA and VQG datasets will encourage further research efforts in multimodal architectures and approaches for radiology image understanding.

## Acknowledgments

This work was partially supported by the intramural research program at the U.S. National Library of Medicine, National Institutes of Health.

## References

- [1] S. A. Hasan, Y. Ling, O. Farri, J. Liu, H. Müller, M. Lungren, 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.
- [2] A. Ben Abacha, S. A. Hasan, V. V. Datla, J. Liu, D. Demner-Fushman, H. Müller, 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, volume 2380 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2019.
- [3] A. Ben Abacha, V. V. Datla, S. A. Hasan, D. Demner-Fushman, H. Müller, Overview of the vqa-med task at imageclef 2020: Visual question answering and generation in the medical domain, in: CLEF 2020 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, Thessaloniki, Greece, 2020.
- [4] B. Ionescu, H. Müller, R. Peteri, A. Ben Abacha, M. Sarrouiti, D. Demner-Fushman, S. A. Hasan, V. Kovalev, S. Kozlovski, V. Liauchuk, Y. Dicente, O. Pelka, A. G. S. de Herrera, J. Jacutprakart, C. M. Friedrich, R. Berari, A. Tauteanu, D. Fichou, P. Brie, M. Dogariu, L. D. Ștefan, M. G. Constantin, J. Chamberlain, A. Campello, A. Clark, T. A. Oliver, H. Moustahfid, A. Popescu, J. Deshayes-Chossart, Overview of the ImageCLEF 2021: Multimedia retrieval in medical, nature, internet and social media applications, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction, Proceedings of the 12th International Conference of the CLEF Association (CLEF 2021), LNCS Lecture Notes in Computer Science, Springer, Bucharest, Romania, 2021.
- [5] M. Sarrouiti, A. Ben Abacha, D. Demner-Fushman, Visual question generation from radiology images, in: Proceedings of the First Workshop on Advances in Language and

Vision Research, Association for Computational Linguistics, Online, 2020, pp. 12–18. URL: <https://www.aclweb.org/anthology/2020.alvr-1.3>.

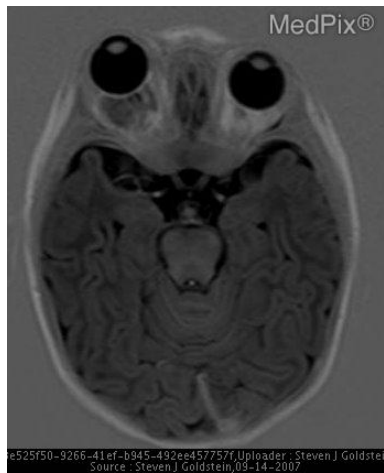
- [6] J. J. Lau, S. Gayen, A. Ben Abacha, D. Demner-Fushman, A dataset of clinically generated visual questions and answers about radiology images, *Scientific Data* 5 (2018). URL: <https://www.nature.com/articles/sdata2018251>.
- [7] Q. Xiao, X. Zhou, Y. Xiao, K. Zhao, Yunnan university at vqa-med 2021: Pretrained biobert for medical domain visual question answering, in: *Working Notes of CLEF 201*, 2021.
- [8] H. Gong, R. Huang, G. Chen, G. Li, Sysu-hcp at vqa-med 2021: A data-centric model with efficient training methodology for medical visual question answering, in: *Working Notes of CLEF 201*, 2021.
- [9] Y. Li, Z. Yang, T. Hao, Tam at vqa-med 2021: A hybrid model with feature extraction and fusion for medical visual question answering, in: *Working Notes of CLEF 201*, 2021.
- [10] S. Eslami, G. de Melo, C. Meinel, Teams at vqa-med 2021: Bbn-orchestra for long-tailed medical visual question answering, in: *Working Notes of CLEF 201*, 2021.
- [11] S. S. N. Mohamed, K. Srinivasan, Imageclef 2021: An approach for vqa to solve abnormality related queries using improved datasets, in: *Working Notes of CLEF 201*, 2021.
- [12] R. Schilling, P. Messina, D. Parra, H. Lobel, Puc chile team at vqa-med 2021: approaching vqa as a classification task via fine-tuning a pretrained cnn, in: *Working Notes of CLEF 201*, 2021.
- [13] I. Chebbi, G. Feki, C. B. Amar, Regim lab at vqa-med 2021: Visual generation of relevant natural language questions from radiology images for anomaly detection, in: *Working Notes of CLEF 201*, 2021.
- [14] J. Li, S. Liu, Lijie at imageclefmed vqa-med 2021: Attention model based on efficient interaction between multimodality, in: *Working Notes of CLEF 201*, 2021.
- [15] K. Papineni, S. Roukos, T. Ward, W.-J. Zhu, 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, pp. 311–318.
- [16] M. Sarrouiti, Nlm at vqa-med 2020: Visual question answering and generation in the medical domain, *CLEF*, 2020.



- (a) **Q1:** What lesion is seen in the mediastinum? **Q2:** Are there any calcifications in the mediastinal mass? **Q3:** Where is the hypodensity consistent with necrosis seen? **Q4:** Are there any enlarged lymph nodes? **Q5:** Where is an enlarged lymph node located?



- (b) **Q1:** Where are the exophytic lesions located? **Q2:** What lesions affect the femur and tibia? **Q3:** Do the lesions involve the knee joint? **Q4:** Do the lesions demonstrate medullary continuity with the bone of origin? **Q5:** Is the fibula deformed? **Q6:** For what disorder are these multiple exostoses diagnostic?



- (c) **Q1:** What causes proptosis of the right eye? **Q2:** What kind of lesion is present in the right orbit? **Q3:** Are the optic nerves and muscles involved? **Q4:** Is the mass homogeneous? **Q5:** What is the lesion suggestive of?



- (d) **Q1:** Where is the thrombus located? **Q2:** Where is collateralization demonstrated? **Q3:** Is the thecal sac effected?

**Figure 2:** Examples from the Test Set of the VQG 2021 Task