NLM at VQA-Med 2020: Visual Question Answering and Generation in the Medical Domain

Mourad Sarrouti

U.S. National Library of Medicine, National Institutes of Health, Bethesda, MD
mourad.sarrouti@nih.gov

Abstract. This paper describes the participation of the U.S. National Library of Medicine (NLM) in Visual Question Answering (VQA) and Visual Question Generation (VQG) tasks of the VQA-Med challenge at ImageCLEF 2020. In the VQA task, I proposed a variational autoencoders model that takes as input a medical question-image pair and generates a natural language answer as output. The encoder consists of a pre-trained CNN model and LSTM to encode the dense vectors of the images and the questions into a latent space, respectively. The decoder network uses LSTM to decode questions from the latent space. I also presented a multi-class image classification-based method for VQA that takes as input an image and returns an answer as output. I used the pre-trained model ResNet-50 with the last layer (the Softmax layer) removed, and added a Softmax layer with different answers as classes. I used the VQA-Med 2019 and VQA-Med 2020 training datasets to train my models. In the VQG task, I presented a variational autoencoders model that takes as input an image and generates a question as output. I also generated new training data from the existing VQA-Med 2020 VQG dataset, based on contextual word embeddings and image augmentation techniques. My best VQA and VQG models achieve 44.1% and 11.6% respectively in terms of BLEU score.

Keywords: Visual Question Answering, Visual Question Generation, Deep Learning, Variational Autoencoders, Natural Language Processing, Computer Vision, ImageCLEF 2020.

1 Introduction

Visual Question Answering (VQA) and Visual Question Generation (VQG) from images is a rising research topic in both fields of natural language processing [12,14,13] and computer vision [11,17,15]. A VQA system takes as input an image and a natural language question about the image, and produces a natural language answer as the output. Whereas a VQG system aims at generating

natural language questions from the images. Both VQA and VQG combine natural language processing that provide an understanding of the question and the ability to produce the answer, and computer vision techniques that provide an understanding of the content of the image.

In contrast to answering and generating visual questions from the content of the image in the open domain, answering and generating visual questions has received little attention in the medical domain. A few recent works have attempted to answer and generate questions about medical images [5,15].

This paper presents the participation of the U.S. National Library of Medicine (NLM) in VQA and VQG tasks of the VQA-Med challenge [4], which is organized by ImageCLEF 2020 [9]. The VQA-Med challenge aims at answering and generating questions about medical images. For the VQA task, I proposed a variational autoencoders model that is tasked with answering a natural language question when shown a medical image. I also presented another VQA model based on multi-class image classification approach. The questions format are repetitive so they might not contribute in answer predictions and only the image can determine the answer. For the VQG task, I introduced and used our recent VQG system based on variational autoencoders that takes as input a medical image and generates a question as output [15]. All my models use the pre-trained convolutional neural networks (CNN), ResNet50 [8], for visual features extractions.

The rest of paper is organized as follows. Section 2 presents the most relevant work. Section 2 describes datasets used in the 2020 VQA-Med challenge. Section 4 presents my proposed models for answering and generating visual questions from medical images. Official results for all models are presented in Section 5. Finally, the paper is concluded in Section 6.

2 Related Work

Open-domain VQA and VQG research areas has received much attention from the research community in recent years. They benefited from the open-domain VQA challenge[1], which takes place regularly every year since 2015. Otherwise, VQA and VQG has been a challenge from the past few years in the medical domain. There has been no significant progress toward this, because of the lack of labelled data and the difficulty of creating such data. Medical images such as radiology images are highly domain-specific, which can only be interpreted by well-educated medical professionals. Since the launch of the VQA-Med challenge at ImageCLEF [7,6], methods in medical VQA continue to evolve to better meet the needs of users visual questions. Many participants systems follow a traditional supervised maximum likelihood estimation (MLE) paradigm that typically relies on a convolutional neural network (CNN) + recurrent neural network (RNN) encoder-decoder formulation (e.g., [3]). They leveraged CNN like VGGNet [16] or ResNet [8] with a variety of pooling strategies to encode

[1] `https://visualqa.org/challenge_2016.html`

image features and RNN to extract question features. Some other participants formulated VQA as multi-class image classification task [2]. In addition, some teams have improved VQA performance using advanced techniques such as the stacked attention networks and multimodal compact bilinear (MCB) pooling [1].

In contrast to VQA, VQG has received little interest so far in the medical domain. More recently, the task of VQG in the medical domain has been studied and explored in [15]. The authors introduced VQGR, a VQG system that is able to generate natural language questions when shown radiology images. They have used variational autoencoders as a neural network and introduced a text data augmentation technique to create more training data.

3 Data Description

3.1 VQA

Given an image and a question expressed in natural language, the VQA task consists in providing an answer based on the image content. The dataset used in VQA-Med 2020 consists of 4,000 radiology images with 4,000 Question-Answer (QA) pairs as training data, 500 radiology images with 500 QA pairs as validation data, and 500 radiology images with 500 questions as test data. Figure 1 shows examples from the VQA-Med 2020 data.

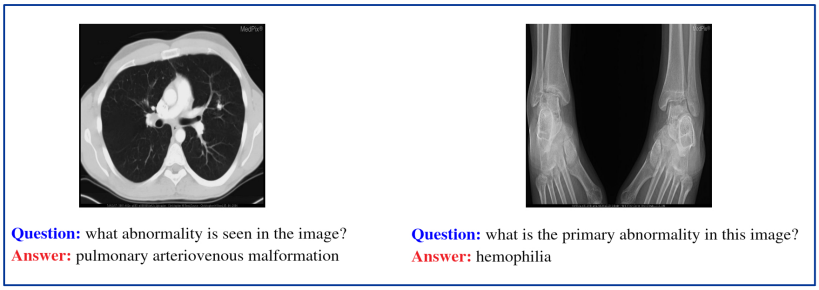
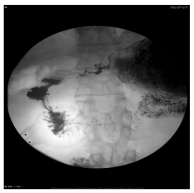


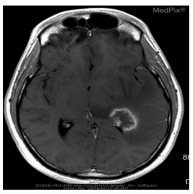
Fig. 1. Example of radiology images and the associated questions and answers from the VQA validation set of ImageCLEF 2020 VQA-Med.

3.2 VQG

Given a radiology image, the VQG task consists in generating a natural language question based on the content of the image. The dataset used in VQA-Med 2020 consists of 780 radiology images with 2,156 associated questions as training data, 141 radiology images with 164 questions as validation data, and 80 radiology images as test data. Figure 2 shows examples from VQA-Med 2020 VQG data.



Question: what does fluoroscopic image from an UGI show?



Question: what is seen in the left temporal lobe adjacent to the posterior horn of the left ventricle?

Fig. 2. Example of radiology images and the associated questions from the VQG training set of ImageCLEF 2020 VQA-Med.

4 Methods

In this section, I present in details the proposed methods for my participation in VQA and VQG task of the 2020 VQA-Med challenge.

4.1 Visual Question Answering

Variational autoencoders-based method for VQA: To address the VQA task of the VQA-Med challenge at ImageCLEF 2020, I proposed a VQA model based on the variational autoencoders approach [10] that takes as input a medical question-image pair and generates a natural language answer as output.

As shown in Figure 3, the proposed model consists of two neural network modules, encoder, and decoder, for learning the probability distributions of data $p(x)$. First, the encoder creates a latent variable z from the image v and the question q, and encodes the dense vectors h_v and h_q into a latent space $z - space$. A CNN is used to obtain the image feature map v, and an LSTM is used to generate the embedded question features q. Then, the model reconstructs the input features L_v, L_q from the $z-$space using a simple Multi Layer Perceptron (MLP) which is a neural network with fully connected layers. I optimize the model by minimizing the following l_2 loss:

$$L_v = ||h_v - \hat{h_v}||_2, L_q = ||h_q - \hat{h_q}||_2 \quad (1)$$

Finally, it uses LSTM decoder to generate the answer $\hat{a}$ from the $z-$space. The decoder takes a sample from the latent dimension z-space, and uses that as an input to output the answer $\hat{a}$. It receives a "start" symbol and proceeds to output an answer word by word until it produces an "end" symbol. Cross Entropy loss function have been used to evaluate the quality of the neural network and to minimize the error L_g between the generated answer $\hat{a}$ and the reference answer a.

The final loss of the proposed VQA model is as follows:

$$Loss = \lambda_1 L_g + \lambda_2 KL + \lambda_3 L_v + \lambda_4 L_q \quad (2)$$

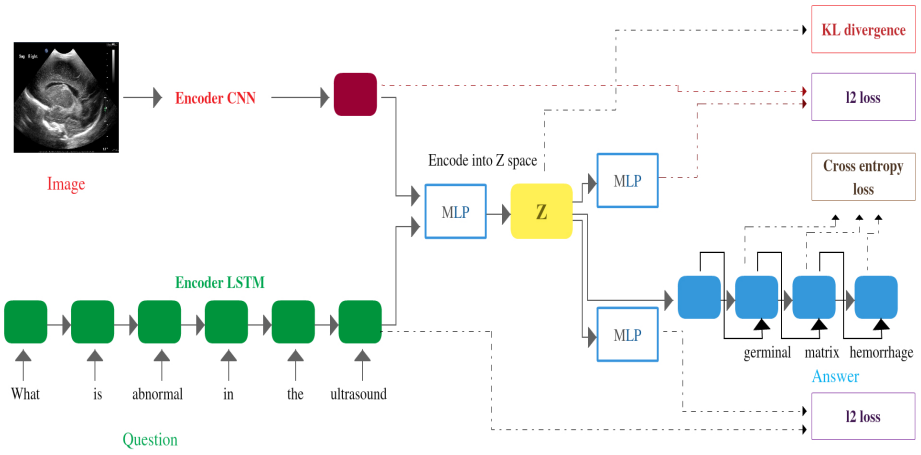


Fig. 3. Overview of the proposed VQA system based on variational autoencoders for radiology images. Input questions are encoded by an LSTM. Images are encoded by a CNN.

where KL is Kullback-Leibler divergence, $\lambda_1, \lambda_2, \lambda_3, \lambda_4$ are hyper-parameters that control the variational loss, the question generation loss, the image reconstruction loss, the question reconstruction loss, respectively.

Multi-class image classification-based method for VQA: I introduced another VQA system based on multi-class image classification approach to solve the VQA task of the VQA-Med challenge at ImageCLEF 2020. The questions are repetitive and have almost the same format and meaning even if they use some different words. So, the questions would not contribute in answer predictions and only the image can determine the answer. Moreover, the majority of questions have a fixed number of candidate answers (332 answers in the whole data) and therefore they can be answered by multi-way classification. Consequently, the VQA task can be equivalently formulated as multi-class classification problems with 332 classes. To do so, I use the pre-trained model ResNet50 with the last layer (the Softmax layer) removed. The output from this part is fed into a Softmax layer with 332 classes (candidate answers).

4.2 Visual Question Generation

To address the VQG task of the VQA-Med challenge at ImageCLEF 2020, I used our recent VQG model based on the variational autoencoders approach [15] that takes as input a medical image and generates a natural language question as output. This model first uses a CNN for obtaining the image feature map v and encoding the dense vectors h_v into a latent (hidden) representation $z-$space. It then reconstructs the inputs from the $z-$space using a simple MLP. Finally, it

uses a decoder LSTM to generate the question $\hat{q}$ from the $z-$space. The decoder takes a sample from the latent dimension z-space, and uses that as an input to output the question $\hat{q}$. I trained this model on the augmented data obtained using contextual word embeddings and image augmentation techniques. I first make use of VQA-Med image-question pairs to generate a heavily augmented dataset for training the question generation model. Each question is tagged with part-of-speech and each candidate word replaced by its most cosine-similar neighbor in a word embedding space based on vocabulary from English Wikipedia, PubMed and PubMedCentral to generate a new augmented question. Each image is also augmented with shifts, flips, rotations, and blurs. More details of this method appairs in [15].

5 Results and Discussion

In this section, I report my official results in the 2020 VQA-Med challenge. The evaluation metrics are accuracy and BLEU for the VQA task, and BLEU for the VQG task. For all models, all images are resized to 224*224, adam optimiser with a learning rate of 0.0001 and a batch size of 32 is used. All models are trained for 20 epochs and the best validation results are used as final results. I implemented these models using PyTorch. The source code are available at `https://github.com/sarrouti/vqa` and `https://github.com/sarrouti/vqa-mcc`.

I submitted five automatic runs to the VQA task at ImageCLEF 2020 VQA-Med:

- Run 1: This run used variational autoencoders. The VQA model was trained on the 2020 VQA-Med data, and without inputs reconstruction. The output length is 3.
- Run 2: This run used variational autoencoders. The VQA model was trained on the 2020 VQA-Med and 2019 VQA-Med training data, and without inputs reconstruction. The output length is 3.
- Run 3: This run used variational autoencoders. The VQA model was trained on the 2020 VQA-Med and 2019 VQA-Med training data, and with inputs reconstruction. The output length is 4.
- Run 4: This run used variational autoencoders. The VQA model was trained on the 2020 VQA-Med and 2019 VQA-Med training data, and with inputs reconstruction. The output length is 10.
- Run 5: This run used multi-class image classification-based method for VQA.

Table 1 shows my official results in the VQA task of the VQA-Med challenge. Run 5 which deals with VQA as a multi-class image classification problem has the best accuracy score and best BLEU score among my submissions. The multi-class image classification approach significantly outperforms the variational autoencoders-based method for VQA. This is likely because the questions were repetitive and all questions have almost the same format and meaning even if they use different words. So, it is expected that the questions would not contribute well in answer generation and only the image can determine the answer. Moreover,

in the VQA-Med 2020 data, the majority of questions have a fixed number of candidate answers and hence can be answered by multi-class classification.

Table 1. Official results of ImageCLEF 2020 VQA-Med: NLM runs for the VQA task.

NLM Runs	Accuracy	BLEU
Run 1	0.232	0.299
Run 2	0.256	0.323
Run 3	0.278	0.321
Run 4	0.286	0.335
Run 5	0.400	0.441

Table 2 presents the results of the best five participating teams in the VQA task of the VQA-Med challenge. My best overall result was obtained by Run 5, achieving the fifth best BLEU score of 0.441 and the fifth best accuracy of 0.40 in the VQA-Med challenge.

Table 2. Official results of the VQA-Med challenge at ImageCLEF 2020: The top five participating teams in the VQA task.

Participants	Accuracy	BLEU
z_liao	0.496	0.542
TheInceptionTeam	0.48	0.511
bumjun_jung	0.466	0.502
going	0.426	0.462
NLM	0.40	0.441

On the other hand, I submitted 3 automatic runs to the VQG task at ImageCLEF 2020 VQA-Med:

- Run 1: This run used the varational autoencoders. The VQG model was trained on the augmented data, and without inputs reconstruction. The output length is 10.
- Run 2: This run used the varational autoencoders. The VQG model was trained on the augmented data, and with inputs reconstruction. The output length is 20.
- Run 3: This run used the varational autoencoders. The VQG model was trained on the augmented data, and with inputs reconstruction. The output length is 30.

Table 3 shows my official results in the VQG task of the VQA-Med challenge at ImageCLEF 2020. Run 2 has the best BLEU score (0.116) among my submissions. Table 4 presents the results of the participating teams. Although many teams have participated in the VQA task of the VQA-Med challenge, only 3 teams

have submitted runs for the VQG task. The results obtained by my VQG system compared with other systems are encouraging and I hope to make improvements in the future. VQG is a challenging task, especially in the medical domain as the available dataset is too small for training efficient VQG models. Small data might require models that have low complexity. Whereas the variational autoencoders model requires a large amount of training data as it tries to learn deeply the underlying data distribution of the input to output new sequences.

Table 3. Official results of ImageCLEF 2020 VQA-Med: NLM runs for the VQG task.

NLM Runs	BLEU
Run 1	0.104
Run 2	0.116
Run 3	0.115

Table 4. Official results of the VQA-Med challenge at ImageCLEF 2020: The participating teams in the VQG task.

Participants	BLEU
z_liao	0.348
TheInceptionTeam	0.339
NLM	0.116

6 Conclusion

In this paper, I described my participation in the VQA and VQG tasks at ImageCLEF VQA-Med 2020. I introduced a variational autoencoders-based method and a multi-class image classification-based method for VQA. My VQA model's best accuracy is 0.40 with 0.441 BLEU score. I also presented a variational autoencoders-based method for VQG. The VQG model achieved 0.116 in terms of BLEU score. In the future, I plan to use the generated questions to advance VQA in the medical domain. I also plan to improve the performance of both VQA and VQG by using the attention mechanism that allows to pay more attention to specific regions that better represent the question instead of the whole image.

Acknowledgements

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

References

1. Abacha, A.B., Gayen, S., Lau, J.J., Rajaraman, S., Demner-Fushman, D.: Nlm at imageclef 2018 visual question answering in the medical domain. In: CLEF (Working Notes) (2018)
2. Al-Sadi, A., Talafha, B., Al-Ayyoub, M., Jararweh, Y., Costen, F.: Just at imageclef 2019 visual question answering in the medical domain. In: CLEF (Working Notes) (2019)
3. Allaouzi, I., Ahmed, M.B., Benamrou, B.: An encoder-decoder model for visual question answering in the medical domain. In: CLEF (Working Notes) (2019)
4. Ben Abacha, A., Datla, V.V., Hasan, S.A., Demner-Fushman, D., Müller, H.: 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 (September 22-25 2020)
5. 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: Cappellato, L., Ferro, N., Losada, D.E., Müller, H. (eds.) 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), http://ceur-ws.org/Vol-2380/paper_272.pdf
6. 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: CLEF 2019 Working Notes. CEUR Workshop Proceedings, CEUR-WS.org <http://ceur-ws.org>, Lugano, Switzerland (September 9-12 2019)
7. Hasan, S.A., Ling, Y., Farri, O., Liu, J., Müller, H., Lungren, M.P.: Overview of imageclef 2018 medical domain visual question answering task. In: CLEF (Working Notes) (2018)
8. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition (2015)
9. 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 medical, lifelogging, nature, and internet applications
10. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114 (2013)
11. Mostafazadeh, N., Misra, I., Devlin, J., Mitchell, M., He, X., Vanderwende, L.: Generating natural questions about an image. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 1802–1813. Association for Computational Linguistics, Berlin, Germany (Aug 2016). https://doi.org/10.18653/v1/P16-1170, https://www.aclweb.org/anthology/P16-1170
12. Sarrouti, M., Alaoui, S.O.E.: A machine learning-based method for question type classification in biomedical question answering. Methods of Information in Medicine **56**(03), 209–216 (2017). https://doi.org/10.3414/me16-01-0116, https://doi.org/10.3414%2Fme16-01-0116
13. Sarrouti, M., Alaoui, S.O.E.: A passage retrieval method based on probabilistic information retrieval model and UMLS concepts in biomedical

question answering. Journal of Biomedical Informatics **68**, 96–103 (apr 2017). https://doi.org/10.1016/j.jbi.2017.03.001, https://doi.org/10.1016%2Fj.jbi.2017.03.001
14. Sarrouti, M., Alaoui, S.O.E.: Sembionlqa: A semantic biomedical question answering system for retrieving exact and ideal answers to natural language questions. Artificial Intelligence in Medicine **102**, 101767 (2020). https://doi.org/https://doi.org/10.1016/j.artmed.2019.101767
15. 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. pp. 12–18. Association for Computational Linguistics, Online (Jul 2020), https://www.aclweb.org/anthology/2020.alvr-1.3
16. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition (2014)
17. Zhang, S., Qu, L., You, S., Yang, Z., Zhang, J.: Automatic generation of grounded visual questions. In: Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence (IJCAI-17). pp. 4235–4243 (2016)