NLM at ImageCLEF 2018 Visual Question Answering in the Medical Domain

Asma Ben Abacha, Soumya Gayen, Jason J Lau, Sivaramakrishnan Rajaraman, and Dina Demner-Fushman

Lister Hill National Center for Biomedical Communications,
National Library of Medicine, Bethesda, MD, USA.
asma.benabacha@nih.gov, soumya.gayen@nih.gov, ddemner@mail.nih.gov

Abstract. This paper describes the participation of the U.S. National Library of Medicine (NLM) in the Visual Question Answering task (VQA-Med) of ImageCLEF 2018. We studied deep learning networks with state-of-the-art performance in open-domain VQA. We selected Stacked Attention Network (SAN) and Multimodal Compact Bilinear pooling (MCB) for our official runs. SAN performed better on VQA-Med test data, achieving the second best WBSS score of 0.174 and the third best BLEU score of 0.121. We discuss the current limitations and future improvements to VQA in the medical domain. We analyze the use of automatically generated questions and images selected from the literature based on ImageCLEF data. We describe four areas of improvements dedicated to medical VQA: (i) designing goal-oriented VQA systems and datasets (e.g. clinical decision support, education), (ii) generating and categorizing medical/clinical questions, (iii) selecting (clinically) relevant images, and (iv) capturing the context and the medical knowledge.

Keywords: Visual Question Answering, Deep Learning, Medical Images, Medical Questions and Answers

1 Introduction

This paper describes the participation of the U.S. National Library of Medicine[1] (NLM) in the Visual Question Answering task[2] (VQA-Med) of ImageCLEF 2018 [1]. ImageCLEF is an evaluation campaign that is being organized as part of the CLEF[3] initiative labs since 2003.

This year, the VQA-Med task [2] was introduced for the first time, inspired by the open-domain VQA challenges[4] that started in 2015. Given a medical image and a natural language question about the image, participating systems are tasked with answering the question based on the visual image content. Three datasets were provided for training, validation and testing.

[1] http://www.nlm.nih.gov
[2] http://www.imageclef.org/2018/VQA-Med
[3] http://www.clef-initiative.eu
[4] http://visualqa.org

In our experiments, we used two deep learning VQA models: Stacked Attention Network (SAN) and Multimodal Compact Bilinear pooling (MCB). For image processing, both VQA models use Convolutional Neural Networks (CNNs). SAN uses VGG-16 and MCB uses ResNet-152 and ResNet-50, pre-trained on the ImageNet database[5]. Image features are extracted from the last pooling layer of the CNNs. For question processing, both VQA models use LSTMs without pre-trained embeddings. Question vectors are extracted from the final hidden layer of the LSTMs. We submitted five runs using these two VQA models[6].

The rest of the paper is organized as follows: Section 2 describes the datasets provided in the scope of the VQA-Med challenge. Section 3 presents the deep learning networks that we selected and used for VQA in the medical domain. Section 4 describes our method to fine-tune pre-trained CNNs on modality classification. Section 5 provides a description of the submitted runs. Section 6 presents the official results. Finally, we discuss the VQA task, the data, and future improvements in Section 7.

2 Data Description

Given an image and a natural language question, the VQA task consists in providing an accurate natural language answer based on the content of the image. Figure 1 shows an example from VQA-Med data.

In the scope of the VQA-Med challenge, three datasets were provided:

- The training set contains 5,413 question-answer pairs associated with 2,278 training images.
- The validation set contains 500 question-answer pairs associated with 324 validation images.
- The test set contains 500 questions associated with 264 test images.

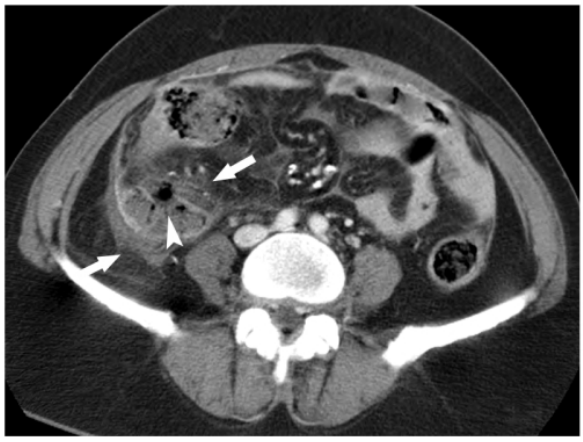
By analyzing the questions manually, four main types of questions could be identified:

1. **Location**, e.g. where is the lesion located? where is the abnormality found? where is the lesion seen?
2. **Finding**, e.g. what does the mri show? what does the ct scan confirm? what does the image demonstrate? what does the ct chest reveal? what is the lesion suggestive of?
3. **Yes/No questions**, e.g. is the brain magnetic resonance imaging normal? are the opacities in both lungs? does the image demonstrate cerebral infarction? Could the recess be mistaken for subcarinal lymphadenopathy?
4. **Other questions**, e.g. what kind of image is this? what is the mass invading? what other abnormality can be seen in the image?

[5] http://www.image-net.org

[6] For the SAN model, we utilized Torch and the computational resources of the NIH HPC Biowulf cluster (http://hpc.nih.gov). We thank Wolfgang Resch for his support.

Image:



Question: what does transverse ct image demonstrate?
Answer: focal defect in inflamed appendiceal wall and periappendiceal inflammatory stranding.

Fig. 1: Example of a medical image and the associated question and answer from the training set of ImageCLEF 2018 VQA-Med.

3 Visual Question Answering Networks

Recently, VQA has been widely addressed with the introduction of new datasets such as the open-domain VQA 1.0 dataset of 250K images, over 760K questions, and around 10M answers [3], and new applications such as answering visual questions for blind people [4].

Different deep networks and attention mechanisms have been applied to open-domain VQA [5, 6]. We studied several VQA networks and selected the following models for our participation in VQA-Med 2018.

3.1 Stacked Attention Network

The Stacked Attention Network (SAN) [7] was proposed to allow multi-step reasoning for answer prediction. SAN includes three components: (i) the image model based on a CNN to extract high level image representations, (ii) the question model using an LSTM to extract a semantic vector of the question and (iii) the stacked attention model which locates the image regions that are relevant to answer the question. The SAN model achieves an Accuracy of 57.3% on VQA 1.0 test data.

For the image model, we used the last pooling layer of VGG-16 pre-trained on imageNet as image features. For the question model, we used the last LSTM

layer as question features. The image features and the question vector were used to generate the attention distribution over the regions of the image.

The first attention layer of the SAN is then computed to capture the correlations between the tokens of the question and the regions in the image. Multimodal pooling is performed to generate a combined question and image vector that is then used as the query for the image in the next layer. We used two attention layers, as it showed better results in open-domain VQA. The last step is answer prediction. For a set of N answer candidates, the answer prediction task is modeled as N-class classification problem and performed using a one-layer neural network. Answers are predicted using Softmax probabilities.

3.2 Multimodal Compact Bilinear pooling

Multimodal Compact Bilinear pooling (MCB) [8] is the winner of the CVPR-2016 VQA Workshop challenge using an attention mechanism that implicitly computes the outer product of visual and textual vectors. MCB architecture contains: (i) a CNN image model, (ii) an LSTM question model, and (iii) MCB pooling that first predicts the spatial attention and then combines the attention representation with the textual representation to predict the answers.

For the image model, we used ResNet-152 and ResNet-50 pre-trained on imageNet. For the question model, a 2-layer (1024 units in each layer) LSTM model is used. Concatenated output from both layers (2048 units) forms the input to the next pooling layer. MCB pooling is then used to combine both image and textual vectors to produce a multimodal representation. To incorporate attention, MCB pooling is used again to merge the multimodal representation with the textual representation for each spatial grid location. We also fine-tuned ResNet-50 on modality classification. Our method is described in the following section.

4 Fine-tuning pre-trained CNNs on modality classification

CNNs are shown to deliver promising results with an increase in the availability of annotated data and computational resources. However, with the scarcity of annotated data, especially in the case of medical imagery, transfer learning is preferred where the CNNs are pre-trained on a huge selection of stock photographic images like ImageNet that contains more than 15 million annotated images belonging to 20K categories. These pre-trained models learn generic features from these large-scale collections, the knowledge can be transferred to the current task. This transfer of knowledge is generic rather than unique to the task under study. Pre-trained CNNs are fine-tuned [9] and/or used as feature extractors [10] for a variety of visual recognition tasks to improve performance.

For ImageClef, we selected five categories of modalities that are the most relevant to VQA-Med data. We used a set of 54,200 medical images associated with the categories: CT (17k images), MRI (12.7k images), XRAY (20k

images), COMPOUND (1,1k images), and Other (3.4k images). We evaluated the performance of a pre-trained ResNet-50 (winner of ILSVRC 2015) toward classifying these modalities. The hyper-parameters were optimized by a method based on randomized grid search. The search ranges were initialized to [1e-7 1e-2], [0.8 0.95] and [1e-10 1e-1] for the learning rate, momentum and L2- weight decay parameters, respectively. The convolutional part of the pre-trained CNNs was instantiated, the pre-trained weights were loaded, a fully connected model was added, the pre-trained layers were frozen up to the deepest convolutional layer and the fully connected model was trained on the extracted features. We then fine-tuned the model from the layer (res5c-branch2c) rather than the entire model to prevent overfitting since the entire CNN has an extremely high entropic capacity and the tendency to overfit. We empirically found that fine-tuning the model from the aforementioned layer improved the classification performance. The model achieved an accuracy of 99.19% in classifying the different imaging modalities.

5 Submitted Runs

We submitted five automatic runs to ImageCLEF 2018 VQA-Med:

- Run1-SAN: This run used the stacked attention network (SAN) with 2 attention layers. The VQA model was trained on imageClef VQA-Med training and validation datasets (10k iterations). VGG-16 pre-trained on ImageNet was used for image features. No additional resources were used.
- Run2-SAN: Same as run number 1, with 12k iterations.
- Run3-SAN: Same as run number 1, with 4k iterations.
- Run4-MCB: This run used the MCB model trained on the training and validation datasets (50k iterations). ResNet-152 was used for image features. No additional resources were used.
- Run5-MCB: Similar to run number 4 (20k iterations). For image features, ResNet-50 was fine-tuned on modality classification.

6 Official Results

In VQA-Med 2018, two metrics were used to evaluate the submitted runs: (i) a new metric called Word-based Semantic Similarity (WBSS), inspired by previous efforts [11, 12] and (ii) the BLEU score [13].

Table 1 shows our official results in the VQA-Med task. The SAN model provided the best results with 0.174 WBSS score and 0.121 BLEU score. Changing the number of iterations had a small impact on the results (Run1, Run2 and Run3). The fourth run used the MCB model and ResNet-152 pre-trained on imageNet for image features. Fine-tuning ResNet-50 on modality classification (Run5) had slightly improved the results of the MCB model.

Figure 2 presents the results of the five participating teams. Our best overall result was obtained by Run1-SAN, achieving the second best WBSS score of 0.1744 and the third best BLEU score of 0.121 in the VQA-Med challenge.

NLM Runs	WBSS Score	BLEU Score
Run1-SAN	0.174	0.121
Run2-SAN	0.168	0.108
Run3-SAN	0.157	0.106
Run4-MCB	0.130	0.083
Run5-MCB	0.144	0.085

Table 1: Official Results of ImageClef 2018 VQA-Med: NLM Runs.

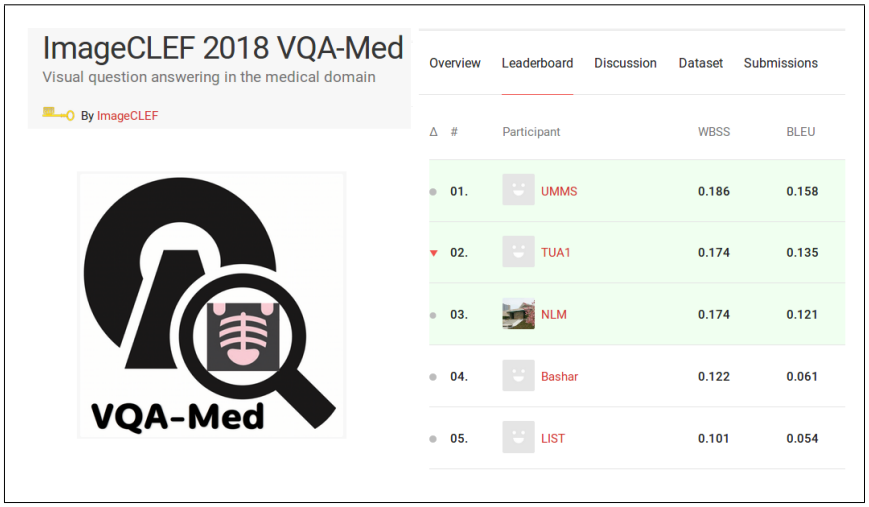


Fig. 2: Official Results of ImageClef 2018 VQA-Med: Participating Teams.

7 Discussion

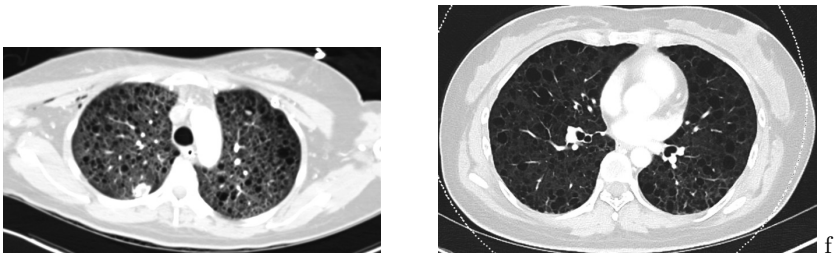
Systems capable of understanding clinical images to the extent of answering questions about the images could support clinical education, clinical decisions and patient education. ImageCLEF VQA-Med is an essential step towards achieving these goals. Our goal for this first challenge was to evaluate the currently existing VQA systems.

7.1 Data-Driven Analysis

While ImageCLEF VQA-Med represents one of the first attempts for VQA to enter the medical domain, and is an excellent starting point, the dataset has several limitations, illustrated by the following example from VQA-Med training set:

– Question: who does ct chest demonstrate interval resolution of without cardiophrenic sparing?

– Answer: pneumothoraces and persistent
– Image: Figure 3b.



(a) First image of the article.

(b) Second image of the article, used in ImageClef 2018 VQA-Med

Fig. 3: Images extracted from a PMC article[7]

This example illustrates some of the issues facing medical VQA:

(1) *Tradeoffs between time-consuming manual construction of datasets and automatically constructing datasets using readily-available resources.* Artificial questions do not always make sense, to the point where we cannot reason what the question was trying to ask. Searching PubMed, we found that the original caption for this image is:

– *"CT chest (axial slices in lung window) demonstrating interval resolution of pneumothoraces and persistent, diffuse numerous thin-walled pulmonary cysts without cardiophrenic sparing."*

Maybe the question was supposed to be *'what does the ct chest demonstrate interval resolution of?'* A more natural way to ask the question might be *'the ct chest demonstrates resolution of what?'*.

(2) *The context and domain knowledge needed to answer the question.* A radiologist may not have been able to answer the above question given only Figure 3b because so much context was lost. The article is a case report of a patient experiencing acute respiratory distress syndrome and Figure 3b is a follow-up image taken 8 months after the patient was discharged. Figure 3a is needed to correctly answer the question with pneumothoraces.

(3) *The clinical task that VQA aims to support.* The purpose of radiology images and captions in PubMed are to summarize and present an interesting case to the scientific community. This differs from clinical radiology reports in hospitals where they assist clinical decision making and direct patient care. If VQA tools are to assist clinical processes, the design of the dataset needs to have aligned goals.

[7] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4638149

7.2 Challenges and Future Improvements

The VQA community has greatly benefited from large datasets such as ImageNet [14] and MS-COCO [15] used to construct public domain VQA datasets. Similarly, specialized medical image resources can be leveraged for VQA, however, there are some fundamental challenges that need to be addressed to apply VQA to a more specialized domain such as Medicine.

If we are to develop tools for VQA to assist clinicians, we will need to better understand what clinincians might ask and how **clinical questions** are naturally phrased. We also need insight to what questions are relevant in different contexts. Just like a question about weather would not make sense for an image of an indoor space, a question about the brain would not make sense for a CT of the chest.

Categorization of question and answer types will help both to direct clinical usefulness and support evaluation of different algorithm designs. Question categorization can help give insight if certain algorithms are good at characterizing the intensity of a tumor while others may be better at counting the number of ribs.

Clinically relevant images are also needed. If the goal of medical VQA is to assist clinical processes, then we need to have images used in clinical settings. Many of the PubMed images are designed for the scientific community. Composite images of multiple images or 3D reconstructions may rarely be relevant for clinical decision-making and supporting direct patient care.

Finally, **context** and background knowledge might play a more important role in medicine than in the open domain. For example, radiologists will know to look for internal bleeding in a head CT scan with contrast even if they are not given the patient's history or symptoms. This context is implicit yet will influence the types of clinically relevant questions and answers. Understanding how we can capture these types of context and the necessary minimum amount are important future goals for benefiting clinical processes.

8 Conclusions

This paper describes our participation in the ImageCLEF 2018 VQA-Med task. We tested two publicly available VQA networks. We achieved the second best WBSS score in the challenge using the SAN model. Our results also showed the positive impact of (i) performing visual attention to learn image regions relevant to answer the clinical questions, and (ii) using a multiple-layer SAN to query an image multiple times and infer the answer progressively. This first VQA challenge in the medical domain allowed testing the combination of language models and vision models in a restricted domain.

Several tracks can be investigated for future improvement including adding questions asked by health care professionals, and using medical terminologies in the linguistic analysis of the questions in deep networks. Based on a new clinical VQA dataset that was recently created [16], we plan to study further the automatic construction of clinical images and related question-answer pairs.

Acknowledgments

This research was supported in part by the Intramural Research Program of the National Institutes of Health (NIH), National Library of Medicine (NLM), and Lister Hill National Center for Biomedical Communications (LHNCBC).

This research was partially made possible through the NIH Medical Research Scholars Program, a public-private partnership supported jointly by the NIH and generous contributions to the Foundation for the NIH from the Doris Duke Charitable Foundation, Genentech, the American Association for Dental Research, the Colgate-Palmolive Company, Elsevier, alumni of student research programs, and other individual supporters via contributions to the Foundation for the NIH. For a complete list, please visit the Foundation website[8].

References

1. 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)
2. 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)
3. Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C.L., Parikh, D.: VQA: Visual Question Answering. In: International Conference on Computer Vision (ICCV). (2015)
4. Gurari, D., Li, Q., Stangl, A.J., Guo, A., Lin, C., Grauman, K., Luo, J., Bigham, J.P.: Vizwiz grand challenge: Answering visual questions from blind people. CoRR **abs/1802.08218** (2018)
5. Kafle, K., Kanan, C.: Visual question answering: Datasets, algorithms, and future challenges. CoRR **abs/1610.01465** (2016)
6. Lu, J., Yang, J., Batra, D., Parikh, D.: Hierarchical question-image co-attention for visual question answering. In: Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain. (2016) 289–297
7. Yang, Z., He, X., Gao, J., Deng, L., Smola, A.J.: Stacked attention networks for image question answering. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016. (2016) 21–29
8. Fukui, A., Park, D.H., Yang, D., Rohrbach, A., Darrell, T., Rohrbach, M.: Multimodal compact bilinear pooling for visual question answering and visual grounding.

[8] http://fnih.org/what-we-do/current-education-and-training-programs/mrsp

In: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016. (2016) 457–468
9. Razavian, A.S., Azizpour, H., Sullivan, J., Carlsson, S.: CNN features off-the-shelf: An astounding baseline for recognition. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR Workshops 2014, Columbus, OH, USA, June 23-28, 2014. (2014) 512–519
10. Rajaraman, S., Antani, S., Poostchi, M., Silamut, K., Hossain, M., Maude, R., Jaeger, S., Thoma, G.: Pre-trained convolutional neural networks as feature extractors toward improved malaria parasite detection in thin blood smear images. PeerJ **6:e4568** (2018)
11. Soganciogl u, G., Öztürk, H., Özgür, A.: BIOSSES: a semantic sentence similarity estimation system for the biomedical domain. Bioinformatics **33**(14) (2017) i49–i58
12. Wu, Z., Palmer, M.S.: Verb semantics and lexical selection. In: 32nd Annual Meeting of the Association for Computational Linguistics, 27-30 June 1994, New Mexico State University, Las Cruces, New Mexico, USA, Proceedings. (1994) 133–138
13. Papineni, K., Roukos, S., Ward, T., Zhu, W.: Bleu: a method for automatic evaluation of machine translation. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, July 6-12, 2002, Philadelphia, PA, USA. (2002) 311–318
14. Deng, J., Dong, W., Socher, R., Li, L., Li, K., Li, F.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009), 20-25 June 2009, Miami, Florida, USA. (2009) 248–255
15. Lin, T., Maire, M., Belongie, S.J., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft COCO: common objects in context. In: Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V. (2014) 740–755
16. Lau, J.J., Gayen, S., Ben Abacha, A., Demner-Fushman, D.: A dataset of clinically generated visual questions and answers about radiology images. Submitted to Scientific Data (2018)