HCP-MIC at VQA-Med 2020: Effective Visual Representation for Medical Visual Question Answering

Guanqi Chen, Haifan Gong, and Guanbin Li*

School of Data and Computer Science, Sun Yat-sen University, Guangzhou, China
chengq26@mail2.sysu.edu.cn, haifangong@outlook.com,
liguanbin@mail.sysu.edu.cn

Abstract. This paper describes our submission for the Medical Domain Visual Question Answering Task of ImageCLEF 2020. We desert complex cross-modal fusion strategies and concentrate on how to capture the effective visual representation, due to the information inequality between images and questions in this task. Based on the observation of long-tailed distribution in the training set, we utilize the bilateral-branch network with a cumulative learning strategy to tackle this issue. Besides, to alleviate the issue of limited training data, we design an approach to extend the training set by Kullback-Leibler divergence. Our proposed method achieved the score with 0.426 in accuracy and 0.462 in BLEU, which ranked 4th in the competition. Our code is publicly available[1].

1 Introduction

Visual Question Answering (VQA) aims at answering questions according to the content of corresponding images. In recent years, researchers have made great progress in the VQA task with many effective methods and large-scale datasets. With the purpose of supporting clinical decision making and improving patient engagement, the VQA task is introduced into the medical field. To promote the development of medical VQA, ImageCLEF [10] organizes 3rd edition of the Medical Domain Visual Question Answering Task [3] (see examples in Figure 1). Compared to general VQA, the valid medical data for training is limited in ImageCLEF 2020 VQA-Med task. Besides, it focuses particularly on questions about abnormalities, which is different from previous editions of the VQA-Med task. We argue that the semantic information from questions is finite due to the single theme of the ImageCLEF 2020 VQA-Med task. However, there are many kinds of abnormal medical images, which need effective visual representation to distinguish them.

* Corresponding author is Guanbin Li.

[1] Code: https://github.com/haifangong/HCP-MIC-at-ImageCLEF-VQA-Med-2020.

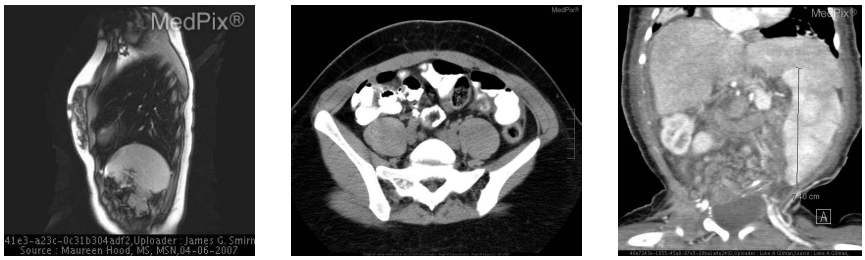

Fig. 1. Three examples of image and corresponding question-answer pair in the ImageCLEF 2020 VQA-Med training set.

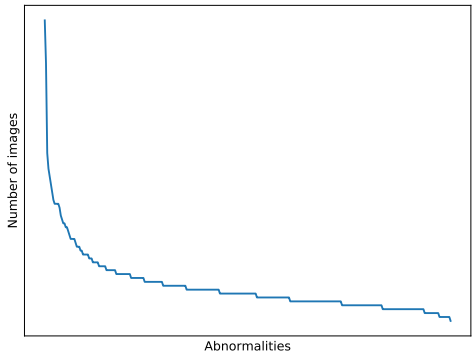

Fig. 2. Long-tailed distribution in the ImageCLEF 2020 VQA-Med training set.

In this paper, we describe the method we developed to deal with the above concerns. Based on the observation of the questions, we divide them into three groups, and utilize a pre-trained BioBERT [12] to classify them. As for visual representation, we map abnormalities to medical images, and discover the phenomenon of the long-tailed distribution in the training set (as shown in Figure 2). Thus, we apply the bilateral-branch network with a cumulative learning strategy [19] to obtain effective visual representation. In addition, we propose a retrieval-based candidate answer selection algorithm to further improve the performance. Last but not least, to alleviate the issue of limited training data, we design an approach to expand the training set by Kullback-Leibler (KL) divergence.

2 Related Work

The common framework for VQA systems is composed of four parts: an image encoder, a question encoder, a cross-modal fusion strategy, and an answer predictor. Many researchers highlight and explore the cross-modal fusion strategy for a better combination of visual and linguistic information. Some works [6,11] utilize compact bilinear pooling methods to capture the joint representation between images and questions. Yang et al. [16], Cao et al. [4], and Anderson et al. [2] exploited the question information to attend the corresponding sub-region of the image. [17] proposed a co-attention mechanism between images and questions to obtain better multi-modal alignment and representation. However, based on the observation that the ImageCLEF 2020 VQA-Med task focuses particularly on questions about abnormalities, we argue that the abnormalities rely on the information from images rather than questions. Thus, we desert the complex cross-modal fusion strategy due to the information inequality between images and questions. And we concentrate on how to obtain an effective visual representation.

As for visual representation in VQA systems, the bottom-up feature representation [2] based on deep CNNs is adopted by many works. [2] utilized Faster R-CNN [15] to capture region-specific features in a bottom-up attention way, which boosted the performance of VQA and image captioning tasks. However, since not all radiology images contain object-level annotations, medical VQA systems usually apply a CNN to extract grid-like feature maps as visual representations. In the ImageCLEF 2020 VQA-Med task, we discover that there exists a long-tailed distribution phenomenon in the training set. Therefore, we adopt the bilateral-branch network with a cumulative learning strategy to obtain effective visual representation.

3 Datasets

In ImageCLEF 2020 VQA-Med task, the dataset includes a training set of 4000 radiology images with 4000 question-answer (QA) pairs, a validation set of 500 radiology images with 500 QA pairs, and a test set of 500 radiology images with 500 QA pairs. The questions mainly focus on the abnormalities of medical images, and they can be divided into two forms. One is making inquiries about the existence of abnormalities in the picture, and another is making inquiries about the abnormal type. Figure 1 shows three examples in the VQA-Med dataset.

The VQA-Med-2019 dataset [1] can be used as additional training data, whose training set contains 3200 medical images associated with 12792 QA pairs. However, different from the VQA-Med-2020 dataset, it focuses on four main categories of questions: Modality, Plane, Organ system, and Abnormality. In this paper, we only leverage its Abnormality subset to extend the VQA-Med-2020 training set .

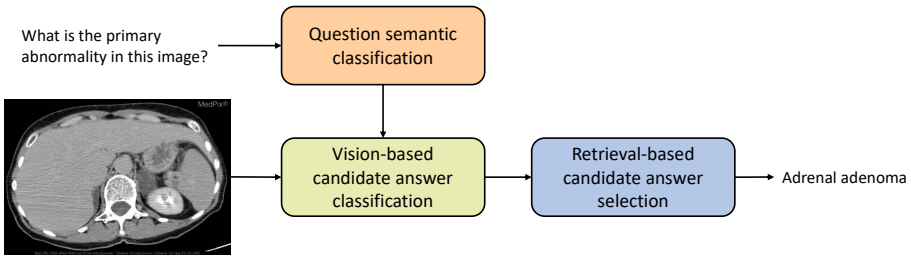

Fig. 3. Overview of the proposed medical VQA framework.

4 Methodology

As shown in Figure 3, our medical VQA framework consists of three parts: question semantic classification, vision-based candidate answer classification, and retrieval-based candidate answer selection. We train the first two parts separately and then connect all the components to predict the final answer in the inference phase. Besides, we design a distribution-based algorithm to expand the training set for further improving the performance of the model.

4.1 Question semantic classification

According to different answer forms, we divide all the questions into two categories: open-ended questions (e.g., Figure 1(c)), and closed-ended questions (e.g., Figure 1(a)(b)). Based on different semantic information, the closed-ended questions can be further separated into two classes: closed-ended abnormal questions (e.g., Figure 1(b)) representing whether the image is abnormal, and closed-ended normal questions (e.g., Figure 1(a)) denoting whether the image looks normal. In all, we need to classify the question sentence into three categories: open-ended questions, closed-ended abnormal questions, and closed-ended normal questions.

For question semantic classification, a pre-trained BioBERT is adopted to classify the questions. Unlike the conventional BERT [5], the BioBERT is a domain-specific language representation model pre-trained on large-scale biomedical corpora. Based on BioBert, we send the 768-dimensional vector, the output of BioBERT, into a 2-layer MLP to obtain the classification score of the input question.

4.2 Vision-based candidate answer classification

In this part, we need to classify the answer according to the radiology image and the category of question. For the closed-ended questions, we apply the ResNet-34 [7] to distinguish normal and abnormal medical images. Then, combining with the fine-grained category information of closed-ended questions, we can simply choose the answer from “yes” or “no”.

As for the open-ended questions, we need to predict the specific abnormalities of the input images in a classification way. Due to the long-tailed distribution among the candidate answers in the training set, we apply the bilateral-branch network (BBN) with a cumulative learning strategy to deal with this problem. In BBN, there are two branches, one is called "conventional learning branch", and another is called "re-balancing branch". The conventional learning branch is for representation learning while the re-balancing is for classifier learning. In the meanwhile, a novel cumulative learning strategy is proposed for adjusting bilateral learning. It is worth noting that, inspired by the attention mechanism, many advanced residual networks have been proposed, such as SE-Net [9], SK-Net [13], NLCE-Net [8]. In this work, we replace the original ResNet in BBN with ResNeSt [18].

4.3 Retrieval-based candidate answer selection

As for the open-ended questions, we discover that the top-5 score is about 10% higher than the top-1 score for the open-ended questions the training procedure. To alleviate this issue, we apply the retrieval-based top-5 answer selection to further improve the performance. The schedule is designed into three steps. The first step is to create a feature dictionary of each class based on the training set. It is worth noting that those features are extracted from the BBN. The second one is to calculate the feature-level cosine similarity between the input sample and all the training samples belong to the top-5 categories. Then, we treat the answer of the most similar training sample as the final prediction.

4.4 Expanding the training set by Kullback-Leibler divergence

Since the valid medical data for training is limited in ImageCLEF 2020 VQA-Med task and external datasets are allowed to use, we expand the training set with the data from the VQA-Med-2019 dataset. Before extending the training set, we define the distribution of the VQA-Med-2020 training set as P_{tr}, which is obtained by:

$$P_{tr} = \frac{n_k}{\sum_{j=1}^{C} n_j} \tag{1}$$

where k and j are the indexes of category, C denotes the number of catergories, and n represents the number of samples with same category. And we exploit the same way to calculate the distribution of the validation set P_v. The KL divergence between P_{tr} and P_v is defined as:

$$\mathcal{D}_{\mathrm{KL}}(P_v||P_{tr}) = \sum_k P_v(k) \log \frac{P_v(k)}{P_{tr}(k)} \tag{2}$$

Then we expand the training set by the following steps. For each sample in the the Abnormality training subset of the VQA-Med-2019 dataset, we assume that it is added to the VQA-Med-2020 training set. Then, we calculate the distribution of new training set $\hat{P}_{tr}$ and the KL divergence $\mathcal{D}_{\mathrm{KL}}(P_v||\hat{P}_{tr})$. Lastly, we extend the training set with the sample if $\mathcal{D}_{\mathrm{KL}}(P_v||\hat{P}_{tr})$ is lower than $\mathcal{D}_{\mathrm{KL}}(P_v||P_{tr})$.

Table 1. Official results of the ImageCLEF 2020 VQA-Med task.

Teams	Accuracy	BLEU
z_liao	0.496	0.542
TheInceptionTeam	0.480	0.511
bumjun_jung	0.466	0.502
Ours	0.426	0.462
NLM	0.400	0.441

Table 2. Ablation study on the VQA-Med-2020 validation set.

Methods	Accuracy	Boost
Baseline	36.6%	-
+BBN-ResNet-34	51.0%	+14.4%
+Training Set Expansion by KL Divergence	54.0%	+3.0%
+BBN-ResNeSt-50	55.0%	+1.0%
+Image Center Cropping	56.6%	+1.6%
+Retrieval-based Candidate Answer Selection	57.2%	+0.6%

5 Experiments

5.1 Implementation details

As for training data, we leverage the whole VQA-Med-2020 training set with 4000 questions to train the BioBERT for question semantic classification. We leverage the extended dataset to train the vision-based model. Among them, 303 images are used to train a ResNet-34 to determine whether the images are abnormal or not, and 4039 images are used to train a BBN to recognize the abnormalities. Besides, a center cropping operation is applied to the input image.

We train those models which are mentioned above separately with corresponding cross-entropy losses. And the optimizer we used is SGD with momentum which is set to 0.9. The initial learning rate is set to 0.08, and the weight decay is 4e-4. We select the best model based on the performance on the validation set.

5.2 Evaluation

The VQA-Med competition uses accuracy and BLEU [14] as the evaluation metrics. Accuracy is calculated as the number of correct predicted answers over total answers. BLEU measures the similarity between the predicted answers and ground truth answers. As shown in Table 1, we achieved an accuracy of 0.426 and a BLEU score of 0.462 in the VQA-Med-2020 test set, which won the 4th place in the competition.

5.3 Abaltion study

In this section, we study some contributions of our proposed method on the VQA-Med-2020 validation set, which is shown in Table 2. The baseline represents the

method that contains a BioBERT for question semantic classification and two ResNet-34 models for vision-based candidate answer classification. And we train the baseline with the original VQA-Med-2020 training set.

Firstly, we replace a ResNet-34 with a BBN-ResNet-34 to better recognize the abnormalities, which surpasses the baseline by 14.4%. We expand the training set by KL divergence, which brings an improvement of 3.0%. The performance is further boosted by 1.0%, using a powerful ResNeSt-50 backbone. Then, we apply a center cropping operation to the input image for reducing noise, which leads to 1.6% improvement. The strategy of retrieval-based candidate answer selection brings a performance gain of 0.6%. Finally, we achieve 57.2% accuracy on the VQA-Med-2020 validation set.

6 Conclusion

In this paper, we describe the method we submitted in ImageCLEF 2020 VQA-Med task. Considering the information inequality between images and questions in this task, we desert complex cross-modal fusion strategies. We adopt the bilateral-branch network with a cumulative learning strategy to handle the long-tailed problem for effective visual representation. Besides, to alleviate the issue of limited training data, we design an approach to extend the training set by Kullback-Leibler divergence. In addition, we propose a retrieval-based candidate answer selection module to further boost the performance. Our proposed method achieves great results with an accuracy of 0.426 and a BLEU score of 0.462.

References

1. Abacha, A.B., Hasan, S.A., Datla, V., Liu, J., Demner-Fushman, D., Müller, H.: Vqa-med: Overview of the medical visual question answering task at imageclef 2019. In: CLEF (2019) 3
2. Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., Zhang, L.: Bottom-up and top-down attention for image captioning and visual question answering. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 6077–6086 (2018) 2
3. 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) 1
4. Cao, Q., Liang, X., Li, B., Li, G., Lin, L.: Visual question reasoning on general dependency tree. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2018) 2
5. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. In: NAACL-HLT (2019) 4.1
6. Fukui, A., Park, D.H., Yang, D., Rohrbach, A., Darrell, T., Rohrbach, M.: Multimodal compact bilinear pooling for visual question answering and visual grounding. arXiv preprint arXiv:1606.01847 (2016) 2

7. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016) 4.2
8. He, X., Yang, S., Li, G., Li, H., Chang, H., Yu, Y.: Non-local context encoder: Robust biomedical image segmentation against adversarial attacks. In: AAAI (2019) 4.2
9. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 7132–7141 (2018) 4.2
10. 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) 1
11. Kim, J.H., Jun, J., Zhang, B.T.: Bilinear attention networks. In: Advances in Neural Information Processing Systems. pp. 1564–1574 (2018) 2
12. Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C.H., Kang, J.: BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics (09 2019). https://doi.org/10.1093/bioinformatics/btz682 1
13. Li, X., Wang, W., Hu, X., Yang, J.: Selective kernel networks. In: IEEE Conference on Computer Vision and Pattern Recognition (2019) 4.2
14. 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. pp. 311–318. ACL (2002) 5.2
15. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. In: Advances in neural information processing systems. pp. 91–99 (2015) 2
16. Yang, Z., He, X., Gao, J., Deng, L., Smola, A.: Stacked attention networks for image question answering. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 21–29 (2016) 2
17. Yu, Z., Yu, J., Cui, Y., Tao, D., Tian, Q.: Deep modular co-attention networks for visual question answering. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2019) 2
18. Zhang, H., Wu, C., Zhang, Z., Zhu, Y., Zhang, Z., Lin, H., Sun, Y., He, T., Mueller, J., Manmatha, R., Li, M., Smola, A.J.: Resnest: Split-attention networks. CoRR **abs/2004.08955** (2020) 4.2
19. Zhou, B., Cui, Q., Wei, X.S., Chen, Z.M.: Bbn: Bilateral-branch network with cumulative learning for long-tailed visual recognition. In: The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2020) 1