Yunnan University at VQA-Med 2021: Pretrained BioBERT for Medical Domain Visual Question Answering

Qian Xiao,Xiaobing Zhou* ,Ya Xiao and Kun Zhao

Yunnan University, Kunming, China

Corresponding author: zhouxb@ynu.edu.cn

Abstract. This paper describes the submission of the Yunnan University team in the Visual Question Answering task of the ImageCLEF 2021 VQA medical image challenge. According to the analysis of the dataset, we regard this task as a classification task. Firstly, we use the pre-trained VGG16 model, Global Average Pooling (GAP), and image enhancement technology to process and extract the image features. Secondly, we use BioBERT, which is pre-trained with biomedical text, to extract all the semantic features. BioBERT and BERT have the same model structure, but BioBERT have better performance in extracting medical text features. Thirdly, the semantic features and image features are fused by Multi-modal Factorized High-order (MFH) Pooling. Finally, the fused features are input into a fully connected layer for classification. Our method achieved an accuracy score of 0.362 and a BLEU score of 0.402 and ranked 2nd among all the participating teams in the VQA-Med task at ImageCLEF 2021. Our code is publicly available[1].

Keywords: BioBERT·VGG Network·Global Average Pooling·VQA-Med

1 Introduction

The visual question answering (VQA) task aims to answer questions according to the content of the corresponding image. It involves data processing technology in the field of computer vision(CV) and natural language processing(NLP). The dataset of the VQA task is composed of medical images and related question-answer pairs. The main task of the VQA system is to input images and questions into the system and predict an answer according to the questions.

For the general domain VQA, there are a large number of datasets, many advanced models, and technologies to solve this task. With the increasing interest in the application of artificial intelligence technology in the medical field, VQA has attracted people's attention in medical field, because it can support doctors' clinical decisions and enhance patients' understanding of their symptoms from medical images, especially in patient-centered medical care.

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

[1]code: https://github.com/huanhuan414/YNU-at-ImageCLEF-VQA-Med-2021

Compared with VQA for the general domain, VQA for the medical domain is a more challenging task. Firstly, due to the high cost of collecting valid data, the available medical data for training is limited. For the general domain VQA, we can easily obtain thousands of images with guaranteed quality. Secondly, the words used in question and answer matching or medical report are quite different from the language used in daily life, and they are more professional.

In the following, we first describe the work related to the VQA Med task in Section 2. Then the dataset provided by ImageCLEF 2021 is described in Section 3. In Section 4, we describe the details of our proposed method, and then we describe the experiments in Section 5. We finally conclude this paper in Section 6.

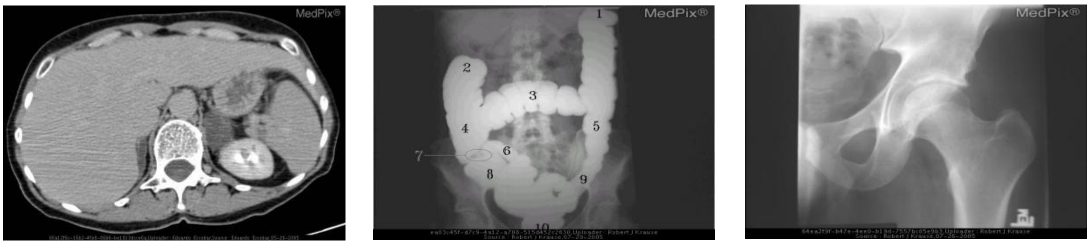
2 Related Work

For the medical field VQA, this task is more challenging because it requires specialized medical datasets and expert doctors to understand the data. The VQA-Med competition started in 2018, and since then, it has provided a medical dataset for VQA tasks every year. In 2019, the Zhejiang University team [3] proposed a convolutional neural network based on VGG16 [7] network and global average pool strategy [9] to extract visual features. The proposed method can effectively capture medical image features in a small training set. The semantic features of the proposed problem are encoded by the BERT [11] model. Then, the common attention mechanism is used to fuse the two enhanced features. In the end, their proposed model ranked 1st among all participating groups of ImageCLEF2019 with an accuracy of 0.624 and a BLEU score of 0.644. In the same year, the cooperative team [15] of Umea University, Sweden, and the University of Bern, Switzerland proposed to use a bilinear model to aggregate and synthesize the extracted image and question features. At the same time, they used an attention scheme to focus on the relevant input context, and further enhanced it by using a set of trained models. Their proposed method ranked 3rd among all the participating groups.

In the third edition of the VQA-Med challenge in 2020, the AIML team [2] used a knowledge reasoning method called skeleton-based sentence mapping (SSM). Using all the questions and answers, they derived a set of classifiable tasks and infered the corresponding tags. At the same time, a classification and task standardization method is proposed to optimize multiple tasks in a single network, which makes it possible to apply multi-scale and multi-architecture integration strategies for robust prediction. In the end, they ranked 1rd among all the participating teams in ImageCLEF2020. The main method of the Inception team [5] is to use the pre-trained VGG16 model, remove the last layer (softmax layer), freeze all layers (except the last four layers) and one of the data enhancement technologies such as geometric transformation, flipping, filling or random erasure of the image. In the end, the Inception team ranked 2nd among all the participating teams in ImageCLEF2020.

3 Data Description

The VQA-Med dataset [16] provided by ImageCLEF 2021 [17] consists of 4000 radiologic images and training sets of question-answer pairs, 3500 train sets, 500 verification sets. Figure. 1 shows three sample examples from VQA-Med 2021 dataset.



Q:what is the primary abnormality in this image?
A:adrenal adenoma

Q:is this a normal gastrointestinal image?
A:yes

Q:is there an abnormality in the x-ray?
A:no

Fig.1. Three examples of ImageCLEF 2021 VQA-Med dataset

ImageCLEF 2019 dataset [4] can be used as additional training data, which contains 3200 medical images and 12792 question-answer pairs associated with images. However, unlike the VQA-Med2021 dataset, it focuses on four main categories of problems: modal, plan, organ system, abnormaly. In this paper, we extend the VQA-Med2021 training set with 473 images and question-answer pairs. Secondly, to further expand the VQA-Med2021 training set, we also add 500 verification sets in VQA-Med2021 as additional training data to the training set for model training.

4 Methodology

In this section, we introduce the task model we submitted to the ImageCLEF VQA-Med 2021 competition. Our model consists of four parts: image feature extractor, text feature extractor, Multi-modal Factorized High-order (MFH) Pooling feature fusion with the Co-attention mechanism, and classification model. In this paper, we regard the ImageCLEF VQA-Med 2021 task as a classification task with C categories. C is the result of removing all the repeated answers in the task. Fig.2. shows the structure of our model.

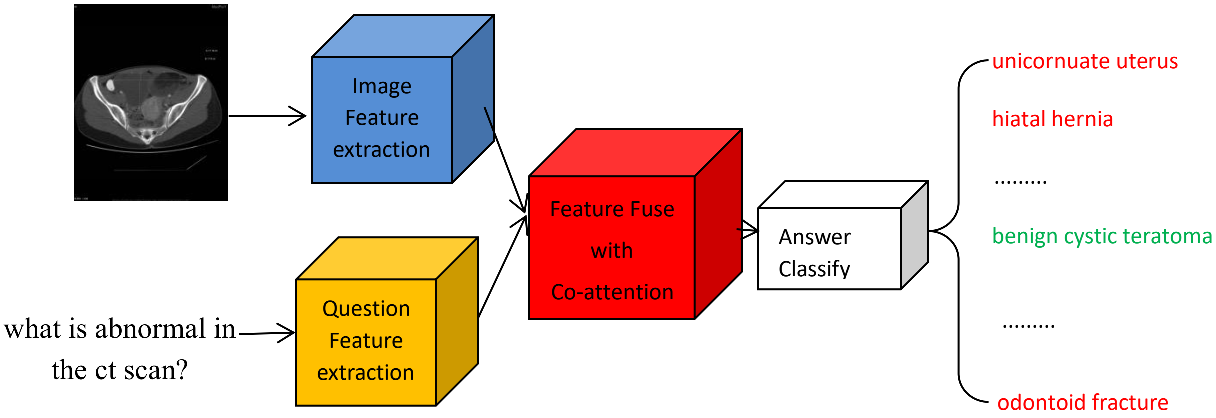


Fig.2. Our model architecture

4.1 Image feature extractor

In our model, we use the pre-trained VGG16 network on the Imagenet dataset to extract features, and the GAP strategy is applied to the VGG16 network to prevent overfitting. GAP averages the output feature of the convolution layer with different channel numbers. The shape of the input model is 224 * 224 * 3, and the shape of the VGG16 network output is 224 * 224 * 64,112 * 112 * 128, 56 * 56 * 258, 28 * 28 * 512, 14 * 14 * 512, 7 * 7 * 512. After averaging the output according to the channel, we get five vectors, 1 * 1 * 64, 1 * 1 * 128, 1 * 1 * 258, 1 * 1 * 512, 1 * 1 * 512, and finally we concatenate these five vectors to get a 1 * 1 * 1472 dimensional vector, and input it to the next network.

4.2 Text feature extractor

BioBERT [10] is used to extract the semantic features of a given question. BioBERT is a pre-trained language representation model for the biomedical field, which has the same network model structure as BERT. Compared with most biomedical text mining models focusing on a single task, BioBERT can achieve the most advanced performance on a variety of biomedical text mining tasks. In order to extract the text features that can represent the question sentence, I encode the question sentence to get the input_ ids, attension_ mask, token_ type_ ids, and then input them into BioBERT to get a 768 dimension vector.

4.3 Feature fusion

Multi-modal feature fusion is one of the most important technologies to improve the performance of the VQA model. For multi-modal feature fusion, most existing methods use a simple linear model to combine image visual features with text semantic features. This paper uses the multi-modal factorized high order pooling (MFH) [8] method, which can fuse multi-modal features with less computational cost. At the same time, the Co-attention mechanism can help the model learn the important parts of visual features and text features, and can better notice the important parts of features while ignoring irrelevant information. In this part, the received 1 * 1 * 1472 dimension image features and 768 dimension text features are sent to the MFH module based on the Co-attention mechanism. Finally, a 2000 dimension fusion feature is obtained and input to the next network.

4.4 Answer prediction

According to the analysis of the dataset of this competition, we regard this task as a classification task. In this part, the received fusion features of 2000 dimensions are first input into a dropout layer with P = 0.3, and then connected to a fully connected layer for final classification prediction.

5 Experiments

Our model trained 350 epochs on GTX2080Ti for about 7 hours. In this part, we will introduce the training process and parameters in detail.

5.1 Train data extension

In addition to the dataset provided by ImageCLEF 2021 VQA-Med task, we also add the abnormal subset of the ImageCLEF 2019 VQA-Med task training set, which contains 473 images and question-answer pairs as the training set of this task. In the ImageCLEF 2019 VQA-Med training set data, only the problem that exists in the ImageCLEF 2021 VQA-Med task test set will be added to this training task as training data.

5.2 Hyperparameter

In order to achieve the best performance of the model in the validation set, the parameters are adjusted several times. Finally, we use binary cross-entropy loss function, Adamax optimizer, dropout with P = 0.1, and initial learning rate of 1e-3. Secondly, the default super parameter setting of MFH (with Co-attention) is used in the multi-modal feature fusion part.

5.3 Evaluation

In VQA-Med2021, accuracy and BLEU are used as the evaluation criteria. The accuracy represents the correct sample in all samples, and the BLEU score measures the similarity between the real answer and the predicted answer. The maximum accuracy we achieved in the validation set is 66.8%. Fig.3 shows the change curve of train accuracy and valid accuracy in the training process. In a total of ten valid submissions, the VGG16 (with GAP) + BioBERT + MFH pooling (with Co-attention) + linear classification layer model proposed in this paper finally achieves an accuracy score of 0.362 and a BLEU score of 0.402. The entries of this paper won second place in this competition. The results of the competition have been shown in Table 1. Our team ID is Zhao_Ling_Ling.

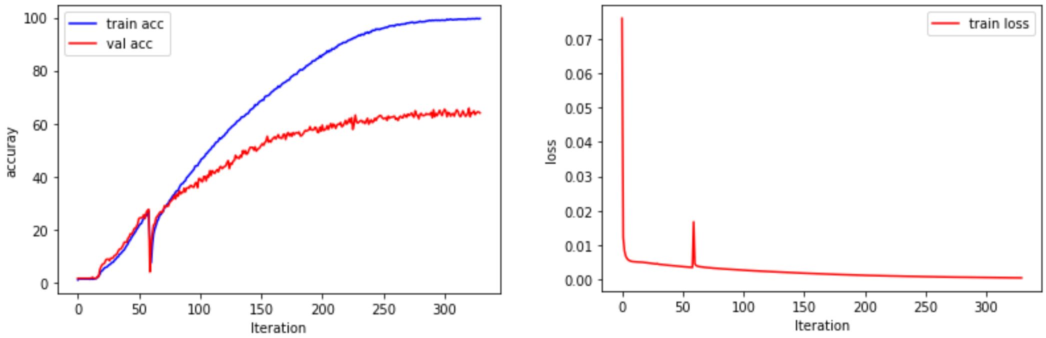


Fig.3. Accuracy、Loss transition by training epoch

Table1. Official results of ImageCLEF VQA-Med 2021

Participants	Accuracy	BLEU
duadua	0.382	0.416
Zhao_Ling_Ling	0.362	0.402

TeamS	0.348	0.391
jeanbenoit_delbrouck	0.348	0.384

6 Conclusion

In this paper, we describe the model we submitted to the ImageCLEF 2021 VQA-Med challenge. We use BioBERT to extract text features. BioBERT has a better performance than BERT in biomedical text extraction. In addition, the application of Multi-modal Factorized High-order (MFH) Pooling with the Co-attention mechanism also makes the model get better performance in this task. For future work, we will continue to improve the current network, introduce some more advanced methods and apply them to other data sets and tasks.

References

[1] Zhan, Li-Ming, Bo Liu, Lu Fan, Jiaxin Chen, and Xiao-Ming Wu. "Medical Visual Question Answering via Conditional Reasoning." In Proceedings of the 28th ACM International Conference on Multimedia, pp. 2345-2354. 2020.

[2] Liao, Zhibin, Qi Wu, Chunhua Shen, Anton van den Hengel, and Johan Verjans. "Aiml at vqa-med 2020: Knowledge inference via a skeleton-based sentence mapping approach for medical domain visual question answering." CLEF, 2020: 78.

[3] Yan, Xin, Lin Li, Chulin Xie, Jun Xiao, and Lin Gu. "Zhejiang University at ImageCLEF 2019 Visual Question Answering in the Medical Domain." In CLEF (Working Notes). 2019: 85.

[4] Abacha, Asma Ben, Sadid A. Hasan, Vivek V. Datla, Joey Liu, Dina Demner-Fushman, and Henning Müller. "VQA-Med: Overview of the Medical Visual Question Answering Task at ImageCLEF 2019." In CLEF (Working Notes). 2019: 272.

[5] Al-Sadi, Aisha, Hana Al-Theiabat, and Mahmoud Al-Ayyoub. "The inception team at vqa-med 2020: Pretrained vgg with data augmentation for medical vqa and vqg." CLEF, 2020: 69.

[6] Jung, Bumjun, Lin Gu, and Tatsuya Harada. "bumjun jung at vqa-med 2020: Vqa model based on feature extraction and multi-modal feature fusion." CLEF, 2020 : 87.

[7] Simonyan, Karen, and Andrew Zisserman. "Very deep convolutional networks for large-scale image recognition." *arXiv preprint arXiv:1409.1556* (2014).

[8] Yu, Zhou, Jun Yu, Chenchao Xiang, Jianping Fan, and Dacheng Tao. "Beyond bilinear: Generalized multimodal factorized high-order pooling for visual question answering." IEEE transactions on neural networks and learning systems 29, no. 12 (2018): 5947-5959.

[9] Lin, Min, Qiang Chen, and Shuicheng Yan. "Network in network." *arXiv preprint arXiv:1312.4400* (2013).

[10] Alsentzer, Emily, John R. Murphy, Willie Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew McDermott. "Publicly available clinical BERT embeddings." arXiv preprint arXiv:1904.03323 (2019).

[11] Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. "Bert:

Pre-training of deep bidirectional transformers for language understanding." arXiv preprint arXiv:1810.04805 (2018).

[12] He, Kaiming, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. "Deep residual learning for image recognition." In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770-778. 2016.

[13] Staudemeyer, Ralf C., and Eric Rothstein Morris. "Understanding LSTM--a tutorial into Long Short-Term Memory Recurrent Neural Networks." arXiv preprint arXiv:1909.09586 (2019).

[14] Cho, Kyunghyun, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. "Learning phrase representations using RNN encoder-decoder for statistical machine translation." arXiv preprint arXiv:1406.1078 (2014).

[15] Vu, Minh, Raphael Sznitman, Tufve Nyholm, and Tommy Löfstedt. "Ensemble of streamlined bilinear visual question answering models for the imageclef 2019 challenge in the medical domain." In CLEF 2019-Conference and Labs of the Evaluation Forum, Lugano, Switzerland, Sept 9-12, 2019, vol. 2380. 2019.

[16] A. Ben Abacha, M. Sarrouti, D. Demner-Fushman, S. A. Hasan, H. Müller, Overview of the vqa-med task at imageclef 2021: Visual question answering and generation in the medical domain, in: CLEF 2021 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org,Bucharest, Romania, 2021.

[17] B. Ionescu, H. Müller, R. Peteri, A. Ben Abacha, M. Sarrouti, 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.Ştefan, 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 MeetsMultilinguality, Multimodality, and Interaction, Proceedings of the 12th International Conference of the CLEF Association (CLEF 2021), LNCS Lecture Notes in ComputerScience, Springer, Bucharest, Romania, 2021.