Zhejiang University at ImageCLEF 2019 Visual Question Answering in the Medical Domain

Xin Yan[1], Lin Li[1], Chulin Xie[1], Jun Xiao[1], and Lin Gu [*2]

1.Zhejiang University, Hangzhou, China
{yan_xin,chulinxie}@zju.edu.cn,junx@cs.zju.edu.cn
hanlin_233@163.com
2.National Institute of Informatics (NII), Tokyo, Japan
ling@nii.ac.jp
corresponding author * Lin Gu : ling@nii.ac.jp

Abstract. This paper describes the submission of Zhejiang University for Visual Question Answering task in medical domain (VQA-Med) of ImageCLEF 2019[2]. We propose a novel convolutional neural network (CNN) based on VGG16 network and Global Average Pooling strategy to extract visual features. Our proposed CNN is able to effectively capture the medical image features under small training set. The semantic features of the raised question is encoded by a BERT model. We then leverage a co-attention mechanism to fuse these two features enhanced with jointly learned attention. These vectors then are then fed to a decoder to predict the answer in a manner of classification. Our model achieves the score with 0.624 in accuracy and 0.644 in BLEU, which ranked first among all participating groups in the ImageCLEF 2019 VQA-Med task[2].

Keywords: Visual Question Answering· VGG Network · Global Average Pooling · BERT

1 Introduction

Visual Question Answering (VQA) is a multidisciplinary task involves both Computer Vision (CV) and Natural Language Processing (NLP) techniques. As illsutrated in Fig.1, presented with an image, the VQA system is expected to answer the raised natural language question about it.

In recent years, VQA has been successful in the general domain with a number of effective models and large-scale datasets. With the development of medical digitization, VQA in medical domain is drawing attention because it could not only serve as a supplementary reference for clinical decision, but also help patients better and faster understand their conditions from medical images. To

promote the development, ImageCLEF 2019 organises 2nd edition of the Medical Domain Visual Question Answering Task[2].

However, the VQA is very challenging on the medical task. On one hand, valid medical data for training are limited compared to those in the general domain. On the other hand, the content and focus of medical images are distinct from the general images. The glossary and the presentation of sentences in medical report are also different from language in every-day's discussion.

In this paper, we propose a system to effectively solve the Med-VQA problem in ImageCLEF 2019 challenge[8]. As illustrated in Fig.2, the proposed system comprises the following steps: 1. With a novel Convolutional Neural Network (CNN), visual features are extracted from the input image. 2. BERT[4], a NLP network is applied to capture syntactic patterns of question and encode it into context features. 3. To extenuate the effect of irrelevant or noisy information, attention mechanism is introduced to focus on particular image regions based on language. 4. Feature fusion mechanism is used to integrate the visual and textural feature vectors to generate a jointed representation. 5. These vectors then are then fed to a decoder to predict the answer in a manner of classification.

Our main contribution can be concluded as follows: Firstly, a novel CNN based on VGG16[12] network and Global Average Pooling[10] strategy is proposed to extract visual features under limited training set. Secondly, we use Multi-modal Factorized Bilinear Pooling (MFB)[18] with co-attention to fuse these two features enhanced with jointly learned attention.

2   Related Works

According to recent work on VQA problems in the general domain, the performance of VQA is particularly sensitive to feature fusion strategy of textual and visual information.

Deep Convolution Neural Networks(CNNs), such as VGGNet[12], ResNet[6], Inception, pretrained on large dataset in the general domain has been successfully explored to extract image feature in recent years. When encoding the question, the majority of research use Recurrent Neural Networks(RNNs) and such as long short-term memory (LSTM)[7], grated recurrent units(GRU)[3] to capture syntactic patterns.

For fine-grained image and question representation, attention mechanisms are effective in extracting the localized image or language features, while global features may bring irrelevant or noisy information. Attention mechanisms have been successfully employed in image captioning [16] and machine translation [15], [1]. For VQA task, [17] developed a multiple-layer stacked attention networks (SANs) to query an image multiple times to progressively infer the answer. [14] used image features from bottom-up attention to provide region-specific features. [13] built upon previous VQA models by developing thirteen attention mechanisms and introducing a simplified classifier to the model. [11] put forward a novel "co-attention" mechanism that jointly reasons about visual attention and question attention.

With respect to multi-modal feature fusion, [11] presented a hierarchical co-attention model (Hie+CoAtt) which combines the co-attention multi-modal features using element-wise summations, concatenation, and fully connected layers. [5] proposed to rely on Multimodal Compact Bilinear pooling (MCB) which computes the outer product between two vectors. [9] employed the Multi-modal Low-rank Bilinear (MLB) pooling model to get a joint representation based on the Hadamard product of two feature vectors. [18] developed Multi-modal Factorized Bilinear pooling (MFB) method and combined it with co-attention learning. Our proposed VQA model in medical domain derives inspiration from that architecture (MFB+CoAtt).

3 Data Description

Fig. 1 shows two examples in VQA-Med dataset[2]. In the ImageCLEF 2019 VQA-Med task[2], the dataset are divided into three subsets:

- The training set contains 12792 question-answer pairs associated with 3200 training images.
- The validation set contains 2000 question-answer pairs associated with 500 training images.
- The test set contains 500 question-answer pairs associated with 500 training images.

There are 4 categories of questions: abnormality, modality, organ system and plane.

- Abnormality: the questions are mainly in two forms: 1. inquiry on the existence of abnormalities in the picture, 2. inquiry on the abnormal type.
- Modality, inquiry on the types of medical images such as mri, CT images.
- Organ, inquiry on what organ is shown in the image.
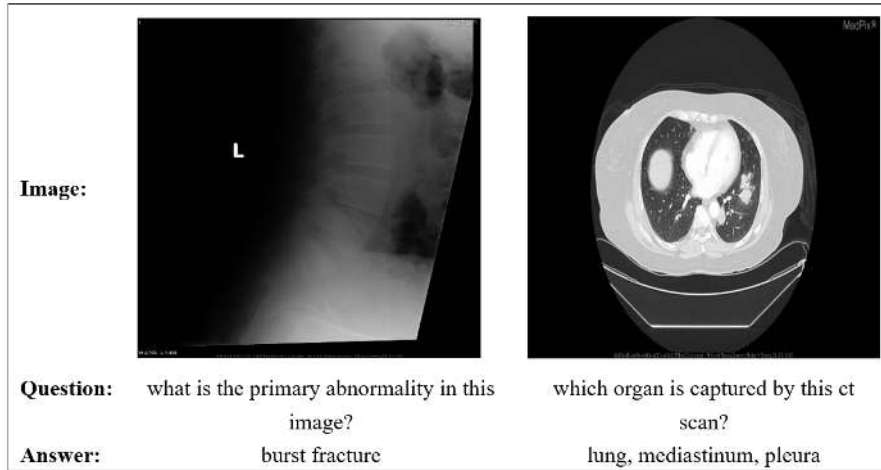- Plane, inquiry on the captured plane such as vertical or horizontal.

Fig. 1. Two examples of medical image and the associated question-answer pair from the ImageCLEF 2019 VQA-Med training set.

4 Methods

In this section, we would introduce our model submitted for the ImageCLEF 2019 VQA-Med task[2]. Our model consists of four modules: image feature extraction, question semantic encoder, feature fuse with co-attention mechanism and answer prediction, which are shown in Figure. 2.
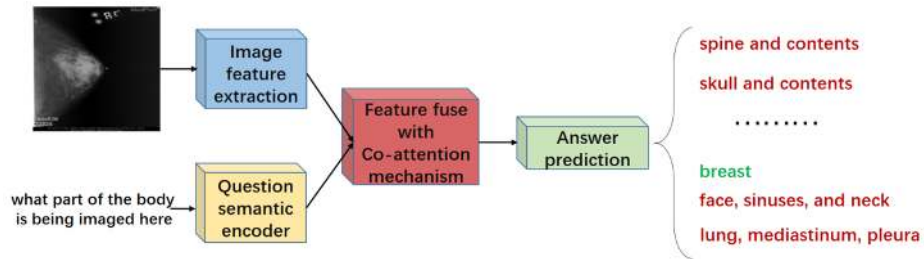

Fig. 2. Our model architecture

4.1 Image feature extraction

In open-domain VQA, the convolutional network like VGGNet[12] or ResNet[6] are usually used to extract image feature map which represents the visual content of the image. In order to extract the feature of medical image, we proposed a new

convolution network that based on VGG16 network(pretrained on ImageNet) and Global Average Pooling[10] strategy. We remove all the fully-connected layers in the VGG16 network and the convolution outputs of different feature scales are concatenated after global average pooling to form a 1984-dimensional vector to represent the image. The architecture is shown in Fig.3. Our experiments show that the new network structure could effectively avoid over-fitting and improve the accuracy of the model.
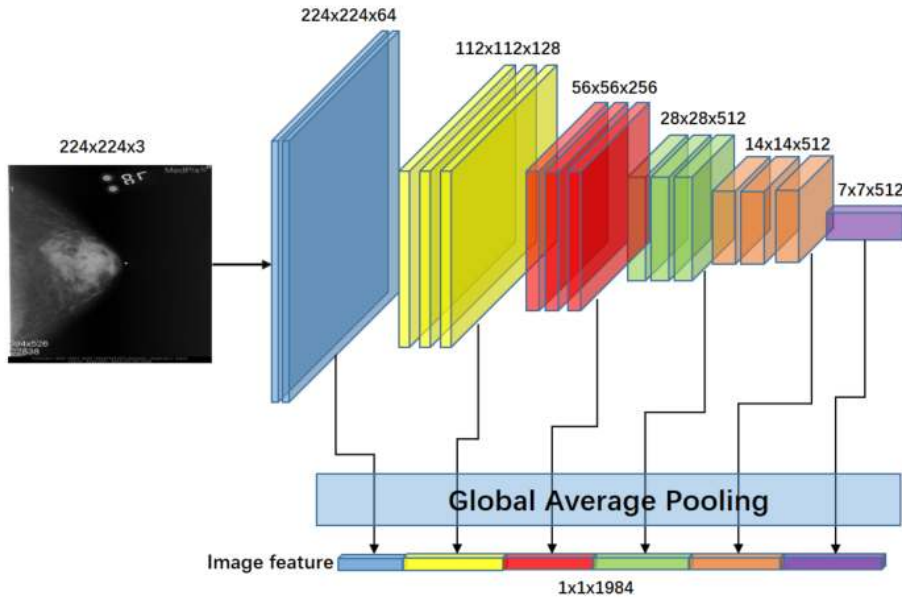

Fig. 3. Image feature extraction

4.2   Question semantic encoder

We propose a question encoder based on the bidirectional encoder representation from transformers(BERT)[4] to get the semantic feature of question. BERT[4] is a pre-trained language representation model proposed by Google. Unlike the context-free model such as Glove which generates a "word embedding" for each word, BERT[4] emphasizes more on the relationships between a word and the other words in a sentence that can effectively avoid polysemy. The model we used is a basic version of BERT[4] which includes 12 layers, 768 hidden variables with a total of 110M parameters. To represent each sentence, we average the last and penultimate layer to obtain a 768-d question feature vector.

4.3 Feature fuse with co-attention mechanism

The strategy to combine visual and semantic feature plays an important role in improving performance of VQA task. Co-attention mechanism assigns weight of importance to features from different regions to avoid the irrelevant information. We therefore use multi-modal factorized bilinear pooling (MFB) [18] with co-attention to fuse the two modalities of features. The network is shown in Fig.4.
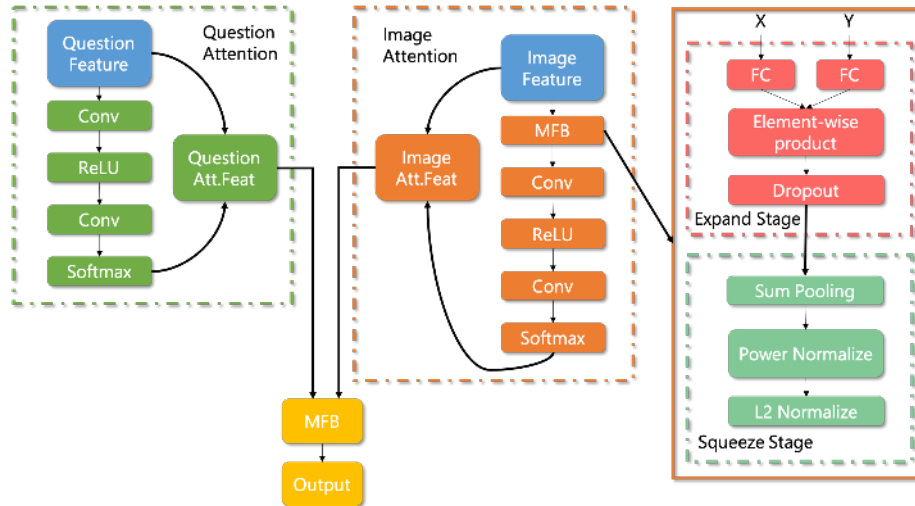

Fig. 4. MFB with co-attention

5 Experiments

5.1 Visual Feature Network

As discussed above, the convolution network with global average pooling[10] could effectively avoid the over-fitting on the small dataset. As shown in Fig.5. Severe over-fitting has occurred in the model without GAP (the left one). As the training progresses, the loss on the validation set decreases and then increases. This did not happen in the model with GAP (the right one) and the model achieves higher accuracy on validation set.

Based on the performance on the validation set, the parameters are set as follows. We use the ADAM optimizer with initial learning rate 1e-4. The regularization coefficient is 1e-3. The dropout coecient in MFB is 0.3 and the batch size is 32. We train the model for 300 epoch on one GTX1080Ti for 1 hours.

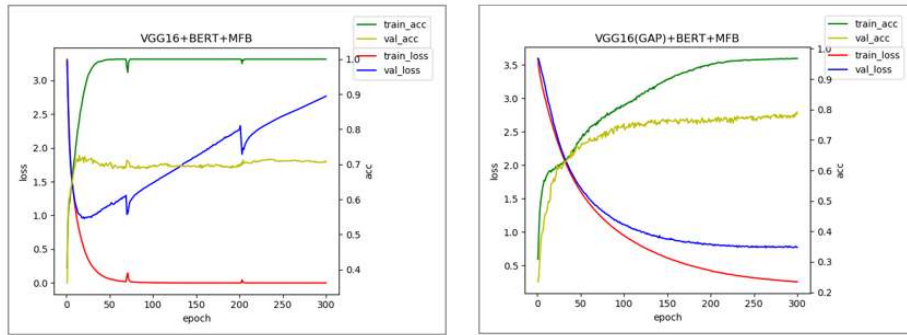**Fig. 5.** Contrast whether GAP is used or not.

5.2 Evaluation

The VQA-Med competition[2] adopted two evaluation indexes, accuracy(strict) and BLEU. Accuracy was used to measure the ratio between the number of correctly classified and the total number of test data set. BLEU measures the similarity between the predicted answer and the actual answer. Based on the above architecture, we submitted six valid runs, among which "VGG16(GAP)+BERT+MFB" achieved the best accuracy score of 0.624 and the best BLEU score of 0.644. The result of the competition is shown in the Figure.6 with the team ID: Hanlin.

Fig. 6. Official Results of ImageCLEF 2019 VQA-Med. The ID of our team is Hanlin.

6 Conclusions

In this paper,we describes the model we submitted in ImageCLEF 2019 VQA-Med task. Our proposed model VGG16(GAP)+BERT+MFB could effectively suppress over-fitting on small data sets. We have achieved the score with 0.624 in accuracy and 0.644 in BLEU on the test set. This performance ranks the first among all participating groups. In futurewe will continue improving the accuracy of our model and evaluating it on more datasets.

References

1. Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *international conference on learning representations*, 2015.
2. Asma Ben Abacha, 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 *CLEF2019 Working Notes*, CEUR Workshop Proceedings, Lugano, Switzerland, September 09-12 2019. CEUR-WS.org <http://ceur-ws.org>.
3. Kyunghyun Cho, 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.
4. Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.
5. Akira Fukui, Dong Huk Park, Daylen Yang, Anna Rohrbach, Trevor Darrell, and Marcus Rohrbach. Multimodal compact bilinear pooling for visual question answering and visual grounding. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 457–468, 2016.
6. Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016.
7. Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.
8. Bogdan Ionescu, Henning Müller, Renaud Péteri, Yashin Dicente Cid, Vitali Liauchuk, Vassili Kovalev, Dzmitri Klimuk, Aleh Tarasau, Asma Ben Abacha, Sadid A. Hasan, Vivek Datla, Joey Liu, Dina Demner-Fushman, Duc-Tien Dang-Nguyen, Luca Piras, Michael Riegler, Minh-Triet Tran, Mathias Lux, Cathal Gurrin, Obioma Pelka, Christoph M. Friedrich, Alba García Seco de Herrera, Narciso Garcia, Ergina Kavallieratou, Carlos Roberto del Blanco, Carlos Cuevas Rodríguez, Nikos Vasillopoulos, Konstantinos Karampidis, Jon Chamberlain, Adrian Clark, and Antonio Campello. ImageCLEF 2019: Multimedia retrieval in medicine, lifelogging, security and nature. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, Proceedings of the 10th International Conference of the CLEF

Association (CLEF 2019), Lugano, Switzerland, September 9-12 2019. LNCS Lecture Notes in Computer Science, Springer.

9. Jin-Hwa Kim, Sang-Woo Lee, Dong-Hyun Kwak, Min-Oh Heo, Jeonghee Kim, Jung-Woo Ha, and Byoung-Tak Zhang. Multimodal residual learning for visual qa. *neural information processing systems*, pages 361–369, 2016.

10. Min Lin, Qiang Chen, and Shuicheng Yan. Network in network. *arXiv preprint arXiv:1312.4400*, 2013.

11. Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. Hierarchical co-attention for visual question answering. *arXiv preprint arXiv:1606.00061*, 2016.

12. Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

13. Jasdeep Singh, Vincent Ying, and Alex Nutkiewicz. Attention on attention: Architectures for visual question answering (vqa). *arXiv preprint arXiv:1803.07724*, 2018.

14. Damien Teney, Peter Anderson, Xiaodong He, and Anton van den Hengel. Tips and tricks for visual question answering: Learnings from the 2017 challenge. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4223–4232, 2018.

15. Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *neural information processing systems*, pages 5998–6008, 2017.

16. Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. *international conference on machine learning*, pages 2048–2057, 2015.

17. Zichao Yang, Xiaodong He, Jianfeng Gao, Li Deng, and Alexander J. Smola. Stacked attention networks for image question answering. pages 21–29, 2016.

18. Zhou Yu, Jun Yu, Jianping Fan, and Dacheng Tao. Multi-modal factorized bilinear pooling with co-attention learning for visual question answering. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 1839–1848, 2017.