bumjun_jung at VQA-Med 2020: VQA model based on feature extraction and multi-modal feature fusion

Bumjun Jung[1], Lin Gu[2,1], and Tatsuya Harada[1,2]

[1] The University of Tokyo, Japan
{jung, lingu, harada}@mi.t.u-tokyo.ac.jp
[2] RIKEN AIP, Japan

Abstract. This paper describes the submission of University of Tokyo for Medical Domain Visual Question Answering (VQA-Med) task [3] at ImageCLEF 2020 [11]. The data set for the task mostly consists of Medical Images and Question Answer pair considering the abnormality appeared in the images. We extract visual features by VGG16 network [16] with Global Average Pooling (GAP) [14]. Compared to the model [18] that ranked first in last year's competition that used BERT [6] model to encode semantic features of questions, we used bioBERT model [13], which is a BERT model pre-trained by biomedical textual data. We also apply multi-modal Factorized High-order (MFH) Pooling [20] with co-attention which shows higher performance than Multi-modal Factorized Bilinear (MFB) Pooling [19] used in [18], to fuse two feature modalities. The fused features are then fed to a decoder to predict the answer in a manner of classification. The score of our model is 0.466 in accuracy, 0.502 in BLEU score, and ranked 3rd among all the participating teams in the VQA-Med task [3] at ImageCLEF 2020 [11].

Keywords: Visual Question Answering · Medical Imagery · Global Average Pooling · bioBERT · Multi-modal Factorized High-order Pooling

1 Introduction

With many achievements and rapid progress in the field of Artificial Intelligence (AI) related to Computer Vision (CV) and Natural Language Processing (NLP), recently the AI technology is applied in the medical domain to analyze the pathological images and medical reports. To be specific, it is used to detect abnormalities or symptoms shown in the images or to generate explanations regarding the medical images.

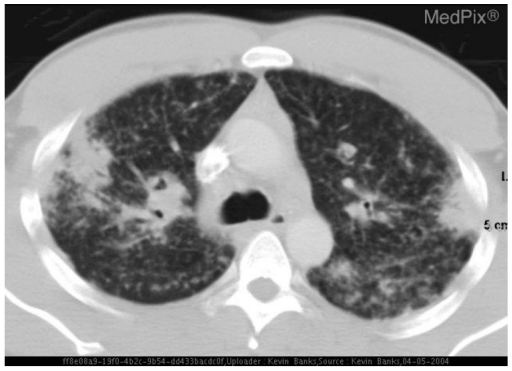
Visual Question Answering (VQA) task involves both CV and NLP techniques to process the data. VQA data set is comprised of both Images and

Question Answer (QA) pairs about the medical images. The images and questions become the inputs to VQA system, whose goal is to predict the answers for the given questions.

Large-scale data sets of VQA for general domain [2], [8] exist and there are many advanced models and techniques that effectively solve the task. With increasing interest in applying AI technology in the medical field, VQA in medical domain is drawing attention due to the importance of supporting the doctors' clinical decision and enhancing the patients' understanding of their conditions from the medical images especially in patient-centered medical care.

VQA for medical domain is a challenging task compared to that of general domain. First, since the cost of collecting valid data is high, valid medical data for training are limited compared to those in general domain such as [2], [8] where hundreds of thousands of images and QA pairs are available. Second, the vocabulary used in QA pairs or medical reports is quite distinct from the language used in daily life.

VQA-Med data set provided by ImageCLEF 2020 consists of 4,000 training set with radiology images and QA pairs, 500 validation set, and 500 test set only with questions without answers. As illustrated in Fig.1, VQA-Med 2020 data set generally asks questions related to the abnormalities shown in the images.



Question: what is the primary abnormality in this image?

Answer: sarcoid

Fig. 1. One example of VQA-Med data set provided by ImageCLEF 2020

The proposed framework in this paper is shown in Fig.2 and can be described as following steps with Fig2: 1. VGG16 network [16] with GAP [14] (Green) is used to extract image features from input image 2. bioBERT model [13] (Blue) is used to capture the semantic of questions and encode it into textual features. 3. Visual features and textual features are fused by fusion mechanism called MFH Pooling [20] (Purple) 4. Co-attention mechanism (Purple) is applied to both

visual and textual features to focus on particular image regions based on the question features and vice versa. 5. Finally, the features, fused by MFH Pooling [20] with co-attention, are fed to the decoder to predict the answer in a manner of classification.

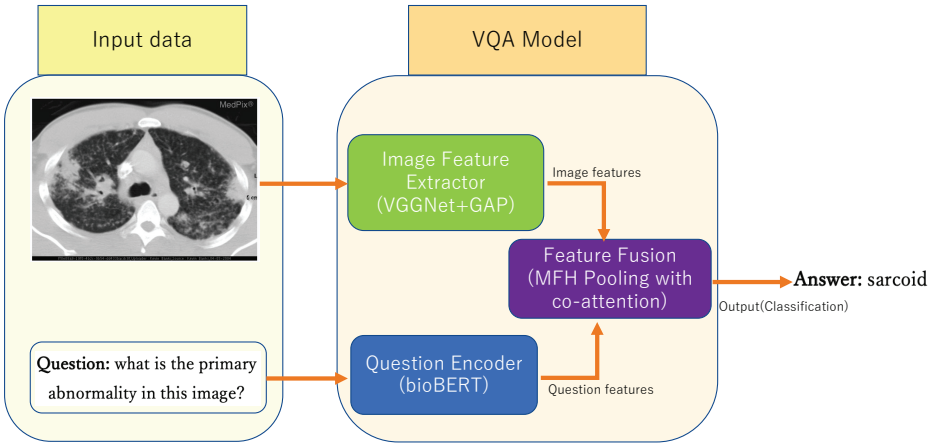


Fig. 2. General pipeline of the proposed framework

The improvements and contributions we made compared to the method [18] comprises of three points: First, for the visual feature extraction, the dimension of extracted feature is reduced to 1472 from 1984 to avoid over-fitting problem while maintaining the quantity of information in extracted features. Second, bioBERT model [13] is used to extract textual features instead of BERT [6] model used in [18]. While bioBERT [13] model has the same network structure as BERT [6] , the data used in pre-training of bioBERT [13] model were biomedical texts which are different from the text data of general domain used in pre-training of BERT [6] . Third, MFH Pooling [20] is used to fuse the visual and textual features which is the advanced version of MFB Pooling [19] used in [18]. MFH Pooling [20] method achieves higher performance than MFB Pooling [19] method.

2 Related Works

There has been much of developments in methods and models used in open domain VQA [2], [8]. For these tasks, deep learning models for image processing based on deep Convolution Neural Networks (CNNs) such as VGGNet [16], ResNet [9] are frequently used to extract image features after pre-trained by large-scale data set in the general domain such as Image net data set [5]. Regarding the question information processing, models for NLP, which are based on Recurrent Neural Networks (RNNs) such as long short-term memory (LSTM)

[10] and grated recurrent units (GRU) [4], are frequently used not only to encode the textual features but also to generate answers as output. Similarlly, NLP models such as BERT [6] pre-trained by large-scale data are applied to extract semantic features from text data.

Attention mechanism and multi-modal feature fusion methods are the important factors of VQA system since VQA is a multidisciplinary task that involves both CV and NLP approaches. Attention mechanisms have been successfully employed in image captioning [17] and NLP models such as BERT [6] also have adopted self-attention transformers in the network structure. Multi-modal feature fusion is essential for VQA task since it combines the information from both modalities to predict the right answer. Fusion techniques have evolved starting from hierarchical co-attention model (Hie+CoAtt) [15] which employs co-attention mechanism using element-wise summations, concatenation, and fully connected layers. Multimodal Compact Bilinear (MCB) pooling [7] computes the outer product between two features to represent every information from the features and this also reduces the computational cost compared to simple outer product calculation. Also, Multi-modal Low-rank Bilinear (MLB) pooling [12] generate output features with lower dimensions and models with fewer parameters compared to MCB Pooling. MFB Pooling [19] method fixed the slow convergence rate of MLB Pooling as well as the part that it is sensitive to the hyper-parameters. MFH Pooling [20] extends the MFB Pooling to a generalized high-order setting to fuse the multi-modal features more effectively.

[18] describes the first rank method of VQA-Med challenge at ImageCLEF 2019 [1] which uses VGG16 network with GAP to extract visual features from input images and BERT model to extract textual features from questions. Those extracted features are fused by MFB Pooling with co-attention method and the fused features are used to predict the answer in a manner of classification.

In addition, bioBERT [13] model is a pre-trained language representation model for the biomedical domain which shares the same network structure with BERT [6]. bioBERT is pre-trained on biomedical domain corpora and it can capture semantic features of biomedical texts such as medical reports more effectively than BERT.

3 Methodology

This section describes the whole pipeline of our VQA model submitted for ImageCLEF 2020 VQA-Med task [3]. As shown in Fig.2, first, from the input image and question the image features and question features are extracted by Image feature extractor and Question encoder. The extracted features are then fused with feature fusion method with co-attention to a classification network for the answer selecting.

3.1 Image feature extractor

In our VQA framework, VGG16 network pre-trained by ImageNet data set [5] is used to extract image features. GAP [14] strategy is applied with VGG16 net-

work to prevent over-fitting problem. The GAP method take the average of last convolution outputs of each layers that have different number of channels. If the input image shape is 224x224x3 as in our model, the output shapes of VGG16 network layers are as follows: 224x224x64, 112x112x128, 56x56x256, 28x28x512, 14x14x512, 7x7x512. The last number of each output shape is the channel size of the convolution layer outputs. After taking the average of the outputs by channel, the extracted features' dimension become the channel size of the layers. Those features are concatenated to form a 1472-dimensional (64+128+256+512+512=1472)[3] vector and it is used as image features and fed to the next network.

3.2 Question encoder

bioBERT [13] is used to extract the semantic features of the given questions. bioBERT is pre-trained with biomedical text and has the same network structure as BERT [6]. bioBERT largely outperforms BERT and previous state-of-the-art models in a variety of biomedical text mining tasks when pre-trained on biomedical corpora. To extract the textual features that can represent the question sentences, we average the last layer of bioBERT-base model to obtain a 768-dimensional question feature vector.

3.3 Feature fusion with co-attention

Fusing multi-modal features is essential and important technique to improve the performance of VQA model. As mentioned in section 2, Multi-modal Factorized High-order (MFH) Pooling [20] method can fuse multi-modal features with less computational cost and improved performance. Co-attention mechanism can help the model to learn the importance of each part in both visual and textual features. It can use the relative information from both modalities to learn which parts of the features are important and to ignore the irrelevant information. We therefore employ the MFH Pooling with co-attention to fuse visual and textual features.

4 Training

Our model is trained for 990 epochs on one Quadro GV100 for about 4 hours. This section describes the detailed process and parameters used in the actual training.

4.1 Train data extension

Besides the data set provided in the ImageCLEF 2020 VQA-Med task [3], we also took advantage of VQA-Med data set of ImageCLEF 2019 [1]. From the

[3] The last convolution layer output that shaped 7x7x512 is precluded when extracting features because it is only the output of Max pooling layer that represents the same information as the former layer output.

data set in [1], only the data comprised of QA pair existing in VQA-Med 2020 data set is used to train the model. 978 pairs in training set and 143 pairs in validation set from [1] are used to extend the VQA-Med 2020 data set.

4.2 Hyper parameters

Hyper parameters are set according to the performance on the validation data set. We used Binary cross-entropy loss as loss function, ADAM optimizer with initial learning rate of 3e-5 and the L1 regularization with co-efficient of 5e-11. MFH Pooling [20] is used with default parameters explained in [20] except for the dropout co-efficient which is set to 0.85 to prevent the over-fitting problem.

5 Evaluation

Two evaluation methods were adopted to VQA-Med 2020 competition, accuracy (strict) and BLEU score. The accuracy measures the ratio of correct prediction and the BLEU score measures the similarity between the real answer and predicted answer. The max validation accuracy of our model was 0.612 and the accuracy transition by training epoch is shown in Fig.3. For the actual training of submitted model, the validation accuracy became 1.0 since the validation data set was also included during the training.

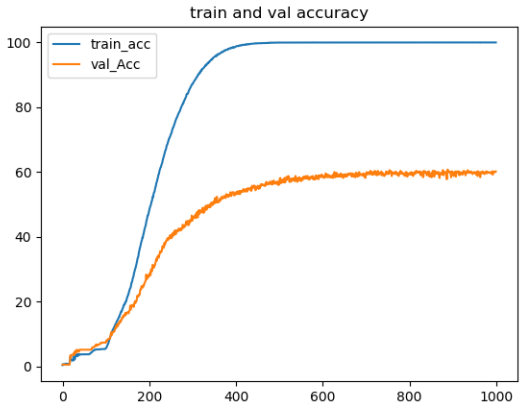


Fig. 3. Accuracy transition by training epoch

Among the 5 valid submissions, the model described in this paper that comprises of VGG16 (with GAP) + bioBERT + MFH Pooling (with co-attention) achieved the accuracy score of 0.466 and BLEU score of 0.502 for the test data set. Our submission took 3rd place in the competition. Fig.4 shows the leaderboard page of VQA-Med competition.

Fig. 4. Leader-board page of the competition. Our team ID is bumjun_jung

6 Conclusion

This paper describes the model submitted in ImageCLEF 2020 VQA-Med challenge. Our model ranked 3rd place and achieved accuracy of 0.466 and BLEU score of 0.502 on test data set. We applied bioBERT [13] model to extract textual features which has stronger performance on encoding biomedical texts compared BERT [6]. Also MFH Pooling [20] is used to fuse the multi-modal features that extends the MFB Pooling [19] to a generalized high-order setting to perform better. For the future work, we will continue to improve the current network and apply it to other data set or tasks.

Acknowledgement

This work was supported by JSPS KAKENHI Grant Number JP20H05556, JST AIP Acceleration Research Grant Number JPMJCR20U3 and JST ACT-X Grant Number JPMJAX190D. We would like to thank Kohei Uehara, Ryohei Shimizu, Dr. Hiroaki Yamane, and Dr. Yusuke Kurose for helpful discussion.

References

1. Abacha, A.B., 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 (Working Notes) (2019)

2. Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Lawrence Zitnick, C., Parikh, D.: Vqa: Visual question answering. In: Proceedings of the IEEE international conference on computer vision. pp. 2425–2433 (2015)
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)
4. Chung, J., Gulcehre, C., Cho, K., Bengio, Y.: Empirical evaluation of gated recurrent neural networks on sequence modeling. arXiv preprint arXiv:1412.3555 (2014)
5. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. pp. 248–255. Ieee (2009)
6. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
7. 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)
8. Goyal, Y., Khot, T., Summers-Stay, D., Batra, D., Parikh, D.: Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 6904–6913 (2017)
9. 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)
10. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural computation **9**(8), 1735–1780 (1997)
11. 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. 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)
12. Kim, J.H., Lee, S.W., Kwak, D., Heo, M.O., Kim, J., Ha, J.W., Zhang, B.T.: Multimodal residual learning for visual qa. In: Advances in neural information processing systems. pp. 361–369 (2016)
13. 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 **36**(4), 1234–1240 (2020)
14. Lin, M., Chen, Q., Yan, S.: Network in network. arXiv preprint arXiv:1312.4400 (2013)
15. Lu, J., Yang, J., Batra, D., Parikh, D.: Hierarchical question-image co-attention for visual question answering. In: Advances in neural information processing systems. pp. 289–297 (2016)
16. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)

17. Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., Zemel, R., Bengio, Y.: Show, attend and tell: Neural image caption generation with visual attention. In: International conference on machine learning. pp. 2048–2057 (2015)
18. Yan, X., Li, L., Xie, C., Xiao, J., Gu, L.: Zhejiang university at imageclef 2019 visual question answering in the medical domain. In: CLEF (Working Notes) (2019)
19. Yu, Z., Yu, J., Fan, J., Tao, D.: Multi-modal factorized bilinear pooling with co-attention learning for visual question answering. In: Proceedings of the IEEE international conference on computer vision. pp. 1821–1830 (2017)
20. Yu, Z., Yu, J., Xiang, C., Fan, J., Tao, D.: Beyond bilinear: Generalized multimodal factorized high-order pooling for visual question answering. IEEE transactions on neural networks and learning systems **29**(12), 5947–5959 (2018)