Lijie at ImageCLEFmed VQA-Med 2021: Attention Model-based Efficient Interaction between Multimodality

Jie Li[1], Shengyan Liu[2]

[1]School of Information Science and Engineering, Yunnan University, Kunming 650091, P.R.China
[2]CSIC 750 proving ground,Yunnan Province, Kunming 650216, P.R.China

Abstract

In this paper, we describe the visual question answering (VQA medicine) task in the medical domain that we submitted on the ImageCLEF 2021 challenge. In terms of semantic feature extraction of question text, we use a more efficient method than BERT, which is processed through the pre-trained BioBERT model on the biomedical data set. Then the image and text features are merged and effectively interacted between multimodality through a more efficient MFH (High-order pooling) and co-attention than MFB (Multimodal factorized bilinear pooling), then we concatenate the various image features from the problem attention. Finally, the text features after multimodal interaction are mapped to the image feature vector space for the second fusion. In this way, the result is obtained by sending it to the fully connected layer and Softmax layer output after two effective fusions. In the ImageCLEF 2021 task, the overall_accuracy of our model is 0.316 and the BLEU is 0.352, ranking sixth among all participating teams this time.

Keywords

Multi-modal Factorized High-order Pooling, BioBERT, Co-attention, Visual Question Answering

1. Introduction

In recent years, artificial intelligence technology (AI) [1] has become more and more mature, especially the rapid development of CV (computer vision) and NLP (natural language processing), so that some difficult tasks have been mentioned again, and it has also collided with all walks of life and produced fierce sparks, and gradually penetrated our daily lives. With the advancement of deep learning [2] algorithms and big data computing power, a medical revolution triggered by artificial intelligence has come quietly. VQA-Med (Visual Question Answering in Medical Domain) is one of the most attractive tasks. That is to say, for many diseases, viewing and analyzing medical images (CT, MRI, Ultrasound) will undoubtedly allow the doctor to inquire about the patient's physical condition clearly and intuitively than asking the patient's feelings. The same is true for the intelligent diagnosis and treatment system. If the questions raised by the patient and the medical images provided by the patient can be combined, it can answer the questions that the patient wants to know more accurately, and even answer some more

CLEF 2021 – Conference and Labs of the Evaluation Forum, September 21–24, 2021, Bucharest, Romania
782097233@qq.com (S. Liu)
0000-0002-6590-6693 (J. Li); 0000-0003-4750-5033 (S. Liu)

complex medical questions.

For clinicians, the medical visual question answering system can enhance their confidence in diagnosing the patient's condition. For patients, the medical visual question answering system can save them a lot of time and money. Instead of having to search for some unverified information on the Internet to understand their condition, patients can learn more accurately what they want to know.

The concept of Visual Question Answering (VQA) first appeared in 2014. VQA also began to develop gradually, then the 2018 ImageCLEF competition first proposed the VQA-Med task [3], and the VQA-Med task of the ImageCLEF competition is still open to university research teams every year and provides corresponding data sets, attracting the participation of a large number of researchers. In 2018, the task was mainly to answer questions about abnormal medical images. There were not many groups that participated at that time, so there were only 5 groups and the method was relatively simple. In addition, the 2018 VQA-Med task is automatically generated from the image caption before being manually checked by a human annotator. The questions and basic answers are variable-length and free-form, which increases the difficulty of answer generation. In VQA-Med in 2019, the classification task is more clarified, using only radiographic images and asking questions from four aspects: image modalities, imaging planes, visual organ systems, and abnormalities that can be detected in images. In the 2020 task, the data set only contains questions about whether or not and what kind of questions. It continues until this year's 2021 VQA-Med competition has not changed to try to get better results.

For the VQA-Med task in ImageCLEF in 2021 [4], Our model is modified by referring to the combined model of MFB and Co-attention structure proposed by Zhou [5] and others in ICCV 2017 which applied to general VQA tasks. Specific steps are as follows: 1. For question text extraction, we use Bio-BERT's pre-trained model on the medical data set to process. 2. For image processing, we use vgg8 [6] for processing, but there is no such complex network as ResNet152 to avoid problems such as excessive training parameters, running delays, and overfitting. 3. MFH is used for efficient fusion during fusion, and the co-attention mechanism is introduced during fusion to improve the effect.

The other parts of this paper are organized as follows. The second section briefly describes the literature review on VQA and VQA-Med. The third section introduces the VQA-Med task and the detailed analysis of its data set. In the fourth part, we introduce the specific methods and principles we used. In the fifth section, we introduce the model and specific steps we used in the experiment. The sixth part introduces the results we submitted. Finally, the paper summarizes and prospects in the seventh part.

2. Related Work

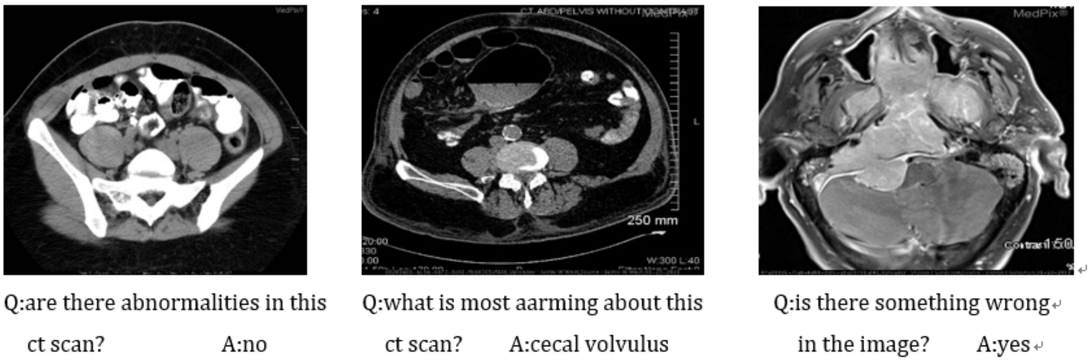
For the development of VQA tasks in the general field, in 2014, Malinowski et al. [7] initially proposed the concept of “open-world" for visual question and answer, and designed a Bayesian model frame model that combines image semantic scene segmentation and text symbol reasoning two methods to realize automatic question and answer of natural language questions. Also using the Bayesian structure model framework is the view put forward by Kushal et al. [8], which transforms open questions into multi-classification problems, such as converting the question “What color is this cat?" into the type of color recognition problem. In fact, most of the initial approach is to use the CNN-RNN framework to process image and text features separately, and almost all images are processed by CNN convolution, and text is processed by RNN, and then the fusion between multi-modality uses the factorization of bilinear pooling, such as MCB (multimodal compact bilinear pooling) [9] and MLB (Multimodal low-rank bilinear pools) [10], or more advanced MFB (Multimodal factorized bilinear pooling) and MFH (High-order pooling) [11]. Of course, there are other fusion methods, such as those that map to the same vector space. The subsequent development is the introduction of the latest models and the tuning process of various algorithms.

For the medical VQA, the development is much slower. The main reason is that the data set in the medical field is much more difficult to obtain than in the general field. Because this requires labeling by professional medical staff and a lot of time to carefully select the quality of the data set, such as the sharpness of the image. Because of the scarcity of data sets, the VQA task has not been rapidly developed in the medical field. The ImageCLEF competition is one of the few organizers that provide data sets in the medical field of VQA. In 2018, Peng et al. [12] proposed a model based on collaborative attention mechanism and MFB feature fusion, and their experimental results achieved first place in the ImageCLEF2018 VQA-Med task. Zhou et al. [13] proposed a model based on Inception-Resnet-v2 [14] and Bi-LSTM [15] and won the second place in the competition. Zachary et al. [16] proposed a model of SAN (two-layer Attention mechanism layer stacking) structure, which ranked third in the competition. In the second year of the ImageCLEF VQA-Med mission, the Zhejiang University team [17] proposed a combination of Bert [18] and MFB. In particular, this model extracts image features from the middle layer of ImageNet pre-trained VGG16, using Bert performs word embedding on the text to extract features and then uses MFB to perform feature fusion, and the results of the experiment won first place in the ImageCLEF2019 VQA-Med [19] and MFB task. The ImageCLEF2020 VQA-Med [20] task competition has just come to an end. It can be seen from the literature that researchers have made certain innovations to the traditional VQA depth model. The University of Adelaide team [21] proposed Skeleton-based Sentence Mapping (SSM) combined with a knowledge reasoning model and won first place in the competition. They combined the knowledge reasoning method into VQA-Med for the first time. Medical VQA is equivalent to start development based on the general VQA field, and the model draws on the methods used in the general VQA development process, but its limitation is that the lack and difference of data sets have led to many method limitations.

In the 2021 ImageCLEF VQA-Med competition and drawing on the methods used in previous competitions, we also made improvements and innovations. In data processing, image enhancement methods such as ZCA (whitening technology image enhancement model) [22] are also introduced to make feature extraction richer to get better output.

3. Task and Dataset

Compared with ImageCLEF VQA-Med 2020, the data set is not much different. Last year’s data set consisted of 4,000 medical images and 4000 question-answer (QA) pairs in the training set, 500 medical images and 500 QA pairs in the validation set, and 500 questions and 500 medical



Q:are there abnormalities in this ct scan? A:no

Q:what is most aarming about this ct scan? A:cecal volvulus

Q:is there something wrong in the image? A:yes

Figure 1: Three forms in ImageCLEF VQA-Med 2021 data set

images in the test set. This year's training set contains 4000 radiological images and related question and answer (QA) pairs. The verification set contains 500 radiological images and related question (QA) pairs. The test set contains 500 radiological images and related question pairs. In addition, the data set was classified into four categories (modal, plane, organ system, and anomaly) in 2019.

To improve the accuracy of the experiment we also used last year's data set as an extended data set for training, but erased the four classification information labels, and the 2020 data set is added to the training set for training. In addition, before training this year's data set separately, image enhancement techniques such as ZCA (whitening technology image enhancement model) were used for image enhancement. The specific images and question-and-answer pairs (QA) in the ImageCLEF VQA-Med 2021 data set [23] are shown in Figure 1.

4. Methods

4.1. Image extraction feature

We first preprocessed the image and used the whitening technology image enhancement model and the adaptive histogram equalization method to limit the contrast to effectively enhance the details of the medical image. While subtly enhancing the contrast of medical images, it also plays a role in suppressing noise. Then a simplified version of the VGG8 model based on the VGG16 model pre-trained on the ImageNet data set [24] was used to extract image features. Because networks such as VGG16 or ResNet50 are too large and the amount of calculation is too large, and the use of such large networks to extract image feature extraction is too redundant and wasteful of resources. The actual experiment also proved that after the previous preprocessing, as long as the small network can achieve the same extraction effect as these large network models, it can effectively avoid overfitting and shorten the training time. So we reduced the original 13-layer convolutional layer of VGG16 to 5 layers and reduced the number of nodes in the following 3 layers of fully connected layers to 128.

4.2. Feature extraction and coding aspects of question text

We used BioBERT [25] which is better than BERT to extract the semantic features of the problem. Since BERT performed well in the ImageCLEF VQA-Med competitions in previous years, we also continued to use the pre-trained model for semantic feature extraction. BioBERT is pre-trained in biomedical text. The network structure is the same as BERT. It inherits almost all the advantages of BERT, and its performance in various biomedical text mining tasks is much better than BERT and previous advanced models. We only need to modify the last layer to make it average to more effectively represent the text features of the question sentence.

4.3. Feature fusion

Feature fusion is the same as image feature extraction and question text feature extraction, which is the key point of whether the VQA task can perform well. To make the interaction between different modalities more effective, we use the MFH which is more efficient than the previous MFB. Because in the dimensionality reduction operation before multiplying between multimodal matrices, MFH can be converted to a more suitable dimension for more effective fusion. At the same time, co-attention is introduced to achieve the characteristics of the problem text and pay more attention to the feature area of the image to improve the effect. Using question text features to capture and attention image-specific area features, and a total of two effective MFH fusions have achieved more accurate regional feature extraction.

5. Experiment

In the ImageCLEF 2021 VQA competition, the model we used is shown in Figure 2. Among them, in terms of extracting image features, we use VGG8 which is a simplified VGG16 network. Because with limited resources, through the number of image data sets after data enhancement, VGG8 is effective enough for extracting image features. Compared with large-scale networks (such as ResNet50), the effect gap is not too huge but the speed is greatly improved. In addition to improving speed, it also prevents overfitting. The text features are pre-trained on BioBERT. After the combined model of MFB and Co-attention fusion is performed, the image is weighted, and then concatenating is performed. The next step is to re-extract features after performing attention operations on the original image. Here we set 4 sets of attention values, because too many extraction groups will ignore the relationship between the information in the image, and too little will make it impossible to better extract the important features of the image So, in the end, we chose 4 groups which is the best grouping.

After having a better interaction between image features and question text features and giving important features greater weight, once again, the two features are fused and output, here we no longer use the MFH module for fusion. Because the previous 4 sets of features and multiple pieces of training have made the interaction between the modalities sufficient, the image feature information is mapped to the text feature information vector space for fusion, which can effectively reduce the amount of calculation and save resources. Finally, through FC Layer and Softmax layer output.

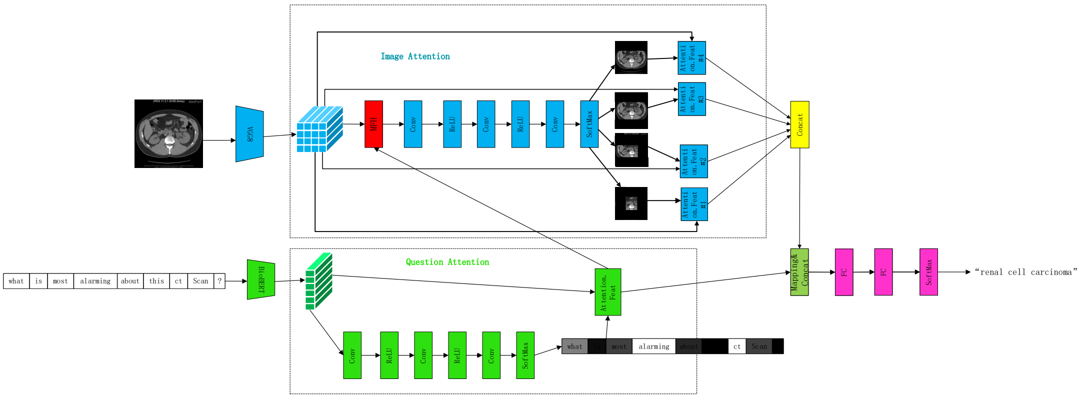


Figure 2: The model we used in the ImageCLEFmed VQA-Med 2021 competition

In the experiment, the loss function we used is the binary cross-entropy loss function, the optimizer is Adam, and the learning rate is 1e-5.

6. Results And Summary

6.1. Results

In the ImageCLEFmed VQA-Med 2021 competition, the overall _accuracy and BLEU are used as the evaluation indicators for the final submission results ranking display. That is the proportion of correct predictions, the similarity between the real answer and the predicted answer. Figure 3 shows the change curve of accuracy and loss during our training. The final results after we submitted were 0.316 and 0.352 respectively, ranking 6th among valid submitters. Figure 4 shows the ranking page of this ImageCLEF VQA-Med 2021 medical competition.

6.2. Summary

Due to the limitations of hardware resources and other conditions, the main idea of this experiment is to obtain the best results with the least resource cost, so the smallest possible modified version of VGG8 is used to extract image features. The corresponding remedy is to use image enhancement. And set an appropriate number of extraction groups in the subsequent collaborative attention mechanism to improve the accuracy. After experimenting with the VGG8 model, we also tried to use larger networks such as VGG16 and ResNet50 to extract features. The accuracy rate has indeed improved, but the time has been much longer and the space overhead has also been much higher, so in the end, I chose a more cost-effective small network, which I originally designed to achieve. If you only start with high-precision considerations, it is better to use large-scale network training when there is ample time. Table 1 shows the comparison of the accuracy results of the tried several models on the validation set.

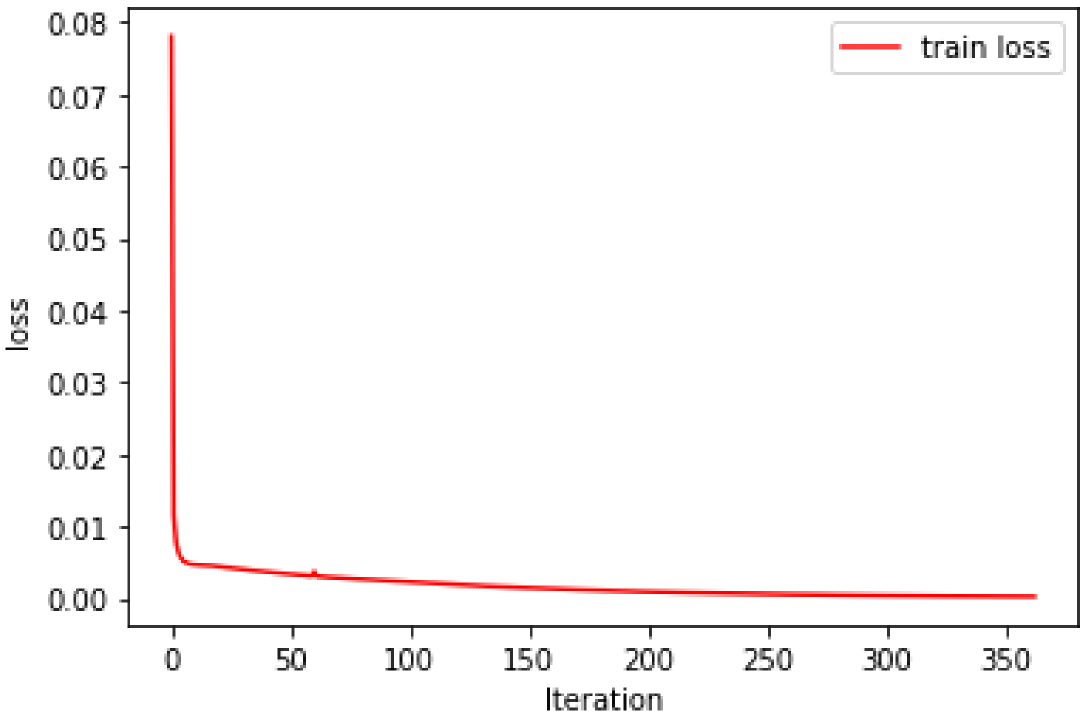


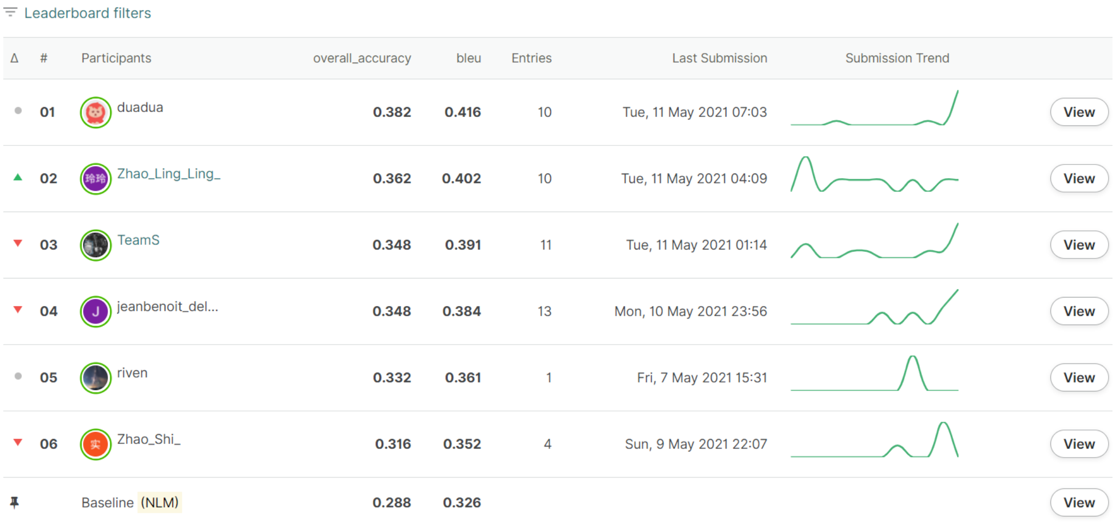
Figure 3: The change curve of accuray and loss during training

Table 1
Comparison of several models on the validation set

Model	Accuray on the validationSet
VGG16+BioBERT+Co-Attention+MFB	0.66
ResNet50+BioBERT+Co-Attention+MFH	0.69
VGG8+BioBERT+Co-Attention+MFH+ZCA	0.62

7. Perspectives For Future work

VQA technology mainly includes the solution of three problems: the extraction of image features, the extraction of problem text features, and the effective fusion of multi-modal features. The effectiveness of these three areas directly affects the quality of the results. In this experiment, for the extraction and characterization of text features, we used Bio-Bert's pre-training weights on the biomedical data set, used the VGG8 model for image feature extraction, and used efficient MFH for fusion. Considering the limited resources, in the image feature extraction, VGG8 with a small number of layers is used, and in the second fusion, a faster mapping method is used instead of matrix multiplication MFH and other methods. In other words, that is to reduce the final result score to save more resources and time. In addition, the image pre-training model



Leaderboard filters

Δ	#	Participants	overall_accuracy	bleu	Entries	Last Submission	Submission Trend	
●	01	duadua	0.382	0.416	10	Tue, 11 May 2021 07:03		View
▲	02	Zhao_Ling_Ling_	0.362	0.402	10	Tue, 11 May 2021 04:09		View
▼	03	TeamS	0.348	0.391	11	Tue, 11 May 2021 01:14		View
▼	04	jeanbenoit_del...	0.348	0.384	13	Mon, 10 May 2021 23:56		View
●	05	riven	0.332	0.361	1	Fri, 7 May 2021 15:31		View
▼	06	Zhao_Shi_	0.316	0.352	4	Sun, 9 May 2021 22:07		View
📌		Baseline (NLM)	0.288	0.326				View

Figure 4: 2021 VQA-Med leaderboard

is not pre-trained on a large medical data set and the training time is too long, which leads to insufficient training times, and the parameters and models are not adjusted to the best, which leads to the result It was not optimal.

In addition, although we used the attention mechanism to align the question text with the corresponding area of the image, after all, there is no refined feature mark for reference, and the result is inevitably bad. In future work, we plan to use VisualBert [26], ImageBert [27], and the Transformer structure model to achieve better performances. Try to migrate from a data set marked with position coordinates, and introduce the method of target detection to make the alignment of text and image more accurate, making the effect better.

References

[1] A. Saffiotti, An ai view of the treatment of uncertainty, The Knowledge Engineering Review 2 (1987) 75–97.

[2] I. Goodfellow, Y. Bengio, A. Courville, Y. Bengio, Deep learning, volume 1, MIT press Cambridge, 2016.

[3] S. A. Hasan, Y. Ling, O. Farri, J. Liu, H. Müller, M. P. Lungren, Overview of imageclef 2018 medical domain visual question answering task., in: CLEF (Working Notes), 2018.

[4] 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 Meets

Multilinguality, Multimodality, and Interaction, Proceedings of the 12th International Conference of the CLEF Association (CLEF 2021), LNCS Lecture Notes in Computer Science, Springer, Bucharest, Romania, 2021.

[5] Z. Yu, J. Yu, J. Fan, D. Tao, Multi-modal factorized bilinear pooling with co-attention learning for visual question answering, in: Proceedings of the IEEE international conference on computer vision, 2017, pp. 1821–1830.

[6] V. Liauchuk, Imageclef 2019: Projection-based ct image analysis for tb severity scoring and ct report generation., in: CLEF (Working Notes), 2019.

[7] M. Malinowski, M. Fritz, A multi-world approach to question answering about real-world scenes based on uncertain input, arXiv preprint arXiv:1410.0210 (2014).

[8] K. Kafle, C. Kanan, Answer-type prediction for visual question answering, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 4976–4984.

[9] A. Fukui, D. H. Park, D. Yang, A. Rohrbach, T. Darrell, M. Rohrbach, Multimodal compact bilinear pooling for visual question answering and visual grounding, arXiv preprint arXiv:1606.01847 (2016).

[10] J.-H. Kim, K.-W. On, W. Lim, J. Kim, J.-W. Ha, B.-T. Zhang, Hadamard product for low-rank bilinear pooling, arXiv preprint arXiv:1610.04325 (2016).

[11] Z. Yu, J. Yu, C. Xiang, J. Fan, D. Tao, Beyond bilinear: Generalized multimodal factorized high-order pooling for visual question answering, IEEE transactions on neural networks and learning systems 29 (2018) 5947–5959.

[12] Y. Peng, F. Liu, M. P. Rosen, Umass at imageclef medical visual question answering (med-vqa) 2018 task., in: CLEF (Working Notes), 2018.

[13] Y. Zhou, X. Kang, F. Ren, Employing inception-resnet-v2 and bi-lstm for medical domain visual question answering., in: CLEF (Working Notes), 2018.

[14] C. Szegedy, S. Ioffe, V. Vanhoucke, A. Alemi, Inception-v4, inception-resnet and the impact of residual connections on learning, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 31, 2017.

[15] M. Schuster, K. K. Paliwal, Bidirectional recurrent neural networks, IEEE transactions on Signal Processing 45 (1997) 2673–2681.

[16] A. B. Abacha, S. Gayen, J. J. Lau, S. Rajaraman, D. Demner-Fushman, Nlm at imageclef 2018 visual question answering in the medical domain., in: CLEF (Working Notes), 2018.

[17] X. Yan, L. Li, C. Xie, J. Xiao, L. Gu, Zhejiang university at imageclef 2019 visual question answering in the medical domain., in: CLEF (Working Notes), 2019.

[18] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, arXiv preprint arXiv:1810.04805 (2018).

[19] A. B. Abacha, S. A. Hasan, V. V. Datla, J. Liu, D. Demner-Fushman, H. Müller, Vqa-med: Overview of the medical visual question answering task at imageclef 2019., in: CLEF (Working Notes), 2019.

[20] A. B. Abacha, V. V. Datla, S. A. Hasan, D. Demner-Fushman, H. Müller, Overview of the vqa-med task at imageclef 2020: Visual question answering and generation in the medical domain, CLEF 2020 Working Notes (2020) 22–25.

[21] Z. Liao, Q. Wu, C. Shen, A. van den Hengel, J. Verjans, Aiml at vqa-med 2020: Knowledge inference via a skeleton-based sentence mapping approach for medical domain visual question answering, CLEF, 2020.

[22] H. K. Verma, S. Sindhu Ramachandran, Harendrakv at vqa-med 2020: Sequential vqa with attention for medical visual question answering (2020).

[23] 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.

[24] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, Imagenet: A large-scale hierarchical image database, in: 2009 IEEE conference on computer vision and pattern recognition, Ieee, 2009, pp. 248–255.

[25] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, J. Kang, Biobert: a pre-trained biomedical language representation model for biomedical text mining, Bioinformatics 36 (2020) 1234–1240.

[26] L. H. Li, M. Yatskar, D. Yin, C.-J. Hsieh, K.-W. Chang, Visualbert: A simple and performant baseline for vision and language, arXiv preprint arXiv:1908.03557 (2019).

[27] D. Qi, L. Su, J. Song, E. Cui, T. Bharti, A. Sacheti, Imagebert: Cross-modal pre-training with large-scale weak-supervised image-text data, arXiv preprint arXiv:2001.07966 (2020).