BIT Mesra at ImageCLEF 2023: Fusion of Blended Image and Text Features for Medical VQA

Sushmita Upadhyay[1,†], Sanjaya Shankar Tripathy[1,*,†]

[1] *Birla Institute of Technology Mesra, Ranchi, India*

Abstract

This paper describes the blended image and text features fusion for medical Visual Question Answering (VQA), submitted to the MEDVQA-GI task organized by ImageCLEF 2023. The goal of medical VQA is to interpret medical images and generate accurate answers to questions about the image. The proposed network uses a combination of deep and handcrafted methods to extract more abstract high level information and domain specific image feature representation. The question feature embeddings have been captured by the BioBERT model which has been trained specifically on medical datasets. Finally, the multimodal features have been combined together using the Multimodal Factorized High-Order Pooling with co-attention mechanism to generate a fused embedding representation. The unified embedding is then used for classification. MedVQA models can have significant applications in medical diagnosis, treatment planning, decision-making and assisting healthcare professionals.

Keywords

VQA-MED, BioBERT, VGG Network, Blended image features

1. Introduction

Visual Question Answering (VQA) is a challenging multimodal problem which requires answering a text-based question by utilizing information provided in the input image. Like a human, the Artificial Intelligence (AI) machine is expected to intelligently utilize and understand the information conveyed by the contents in the image and find a relation which will help in answering the question. Colonoscopy is a procedure used to detect lesions and diseases in the colon. When VQA is used on colonoscopy images, it enables the retrieval of diverse information by answering various questions about these images. However, the application of VQA in medical images poses greater difficulties compared to general domain images [1]. This is because medical images are generally low resolution images which adds complexity to information retrieval. The creation of medical VQA databases demands expert involvement, making the process time-consuming and costly. Moreover, answering medical questions may require the model to be familiar with specialized medical terminologies potentially requiring training with medical knowledge databases [1].

CLEF 2023: Conference and Labs of the Evaluation Forum, September 18–21, 2023, Thessaloniki, Greece

*Corresponding author.

†These authors contributed equally.

phdec10001.21@bitmesra.ac.in (S. Upadhyay); sstripathy@bitmesra.ac.in (S. S. Tripathy)

0000-0002-8697-2526 (S. S. Tripathy)

Multiple approaches have been investigated to analyze the features of input images and textual questions in VQA. Deep models like VGG16 [2], VGG19 [2], and ResNet [3] have been utilized to generate feature embedding representing the input image effectively. Simultaneously, sequential models like Gated recurrent unit (GRU) [4], Long Short Term Memory (LSTM) [5] or deep models like transformers [6] have been explored to capture the feature representation or embeddings of a question. The individual features need to be combined to generate an effective and discriminative embedding vector. Feature concatenation, element wise multiplication and bilinear pooling are some of the methods to fuse two vectors. Recently, attention based methods have shown improved performance in tasks like machine translation [7], image captioning [8] etc. Multi-modal Factorized Bilinear Pooling [9] (MFB) gives a compact and efficient fused feature embedding and has been utilized for VQA. It has been generalized in Multimodal Factorized High-Order Pooling (MFH) [10]. Overfitting has been avoided by employing Global Average Pooling[11] on the convolution output thereby improving accuracy in [12]. A powerful language model, Bidirectional Encoder Representations from Transformers (BERT) [13], captures embeddings from the input text and has been pre-trained with general domain data. Each embedding captures the meaning and contextual information of the words in the text. Bio-BERT [14] is a variant of BERT which has been pre-trained on medical data and medical texts. This model captures the semantic and context of the medical texts which makes it well suited to extract the domain specific features. Bio-BERT model was used to generate question embeddings in [15] and [16] which were a part of VQA MED 2020 and VQA MED 2021 challenges respectively. To predict answers, both classification and generation methods can be employed. A classifier will predict the most likely answer from a predefined set of labels. Generative models generate answers to the input question using networks like LSTMs, GRUs or Transformers.

The ImageCLEF association has in 2023 [17] organised Medical Visual Question Answering for GI : MEDVQA-GI2023 task [18]. There are three subtasks in this task: VQA, Visual location question answering (VLQA) and Visual question generation (VQG). It has been designed around colonoscopy images which covers the entire Gastrointestinal (GI) tract from mouth to anus. For VQA, the images and textual questions are to be combined to generate answers.

In this work, the VQA subtask has been approached as a classification problem with the unique entries as the class label. The feature embeddings are obtained for both the two modalities: images and text. These features have been fused using Multimodal Factorized High-Order Pooling with co-attention (MFH+CoAtt) [10]. Additionally, image features were extracted by utilizing a blend of deep learning techniques and handcrafted feature extraction methods [19].

2. Methodology

VQA has two inputs, images and text, with unique features. The features derived from the image and text modalities can be individually analyzed to leverage their specific characteristics. These features can then be fused together to obtain a combined and informative feature vector that captures the complementary information from both modalities. In figure 1, the block diagram of the proposed system has been presented.

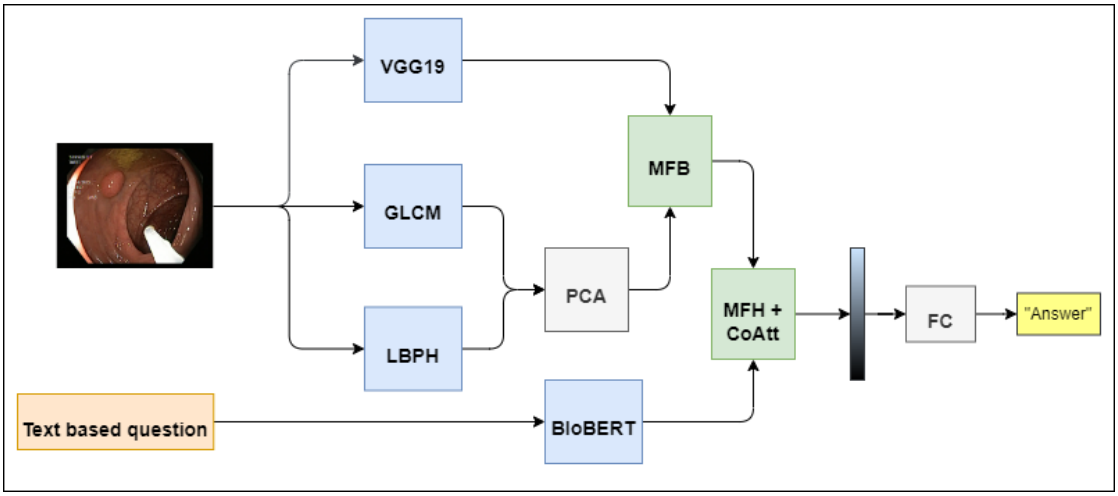

Figure 1: Proposed model block diagram

2.1. Image feature extraction

Colonoscopy images can provide valuable information that can be useful in the detection and diagnosis of any disorder in the GI tract. The presence of abnormalities in a region can be identified by examining the color, texture, and shape variations present in these images. The image features have been extracted by utilizing both deep learning methods and handcrafted methods. This approach utilizes the power of deep models to automatically learn and extract relevant features from the image. Simultaneously, it also incorporates handcrafted techniques that capture domain specific characteristics from the image. The deep features are obtained using the VGG19 convolutional neural network [2]. The top classification layers from the network have been removed, and the features from the last max pooling layers have been considered for further analysis. Global average pooling [11] has been applied to the last max pool output to get a feature vector of length 512. The initial layers of a CNN extract low level features from the input. The subsequent layers extract more abstract high level features that represent complex patterns, shapes, and structures within the input. The handcrafted methods have been employed to extract the features specific to the medical image. Uniform Local Binary Pattern Histogram (LBPH) [20] identifies patterns in an image generating rotation invariant patterns. Uniform patterns primarily correspond to significant micro-textures such as lines, spots, corners, and similar features, and non-uniform patterns represent complex micro-textures like random patterns or network like structures etc. [19]. This operator is performed locally, and the histogram is calculated which represents the distribution of patterns in that region. Finally, all histograms are concatenated to get the final feature. To extract the features, the image has been divided into windows of size 20x20 considering an overlap of 8 pixels. The Gray Level Co-occurrence Matrix (GLCM) [21] is defined by calculating how often a pair of pixels with a specific value occur in the image. This matrix is analyzed to calculate statistical features like contrast, dissimilarity, correlation, energy, homogeneity and Angular Second Moment (ASM). The LBPH and GLCM features are concatenated to get a feature vector of length 3634. Principal

Component Analysis (PCA), which is a dimentionality reduction method, is applied on the resulting features to get the final embedding of length 643. In figure 1 VGG19, GLCM and LBPH blocks extract the respective image features.

2.2. Question feature extraction

In addition to the input image, the other input is a textual question about the image. Like image, the features extracted from the text is also important. The feature representation from the pre-trained BERT [13] model is used for various downstream tasks like text classification, named entity recognition, question answering, by fine tuning. Bio-BERT[14] is a transformer-based architecture with a stack of 12 encoder layers. These layers extract the contextual relations from the input sequence [13]. There are 12 attention heads in every encoder layer to perform the multi-head self-attention mechanism. The encoder layer operates on each token to generate a representation of length 768. The multiple heads operate in parallel, and its purpose is to extract different patterns and dependencies from the input sequence. The attention head outputs are combined and passed through a feed forward network generating the final feature representation of the encoder. In this work, the output of the last layer has been utilized generating a feature embedding of length 768. The Bio-BERT block in figure 1 extracts question features.

2.3. Feature fusion

The image and text feature embeddings individually capture the characteristics and context of the two modalities. The multimodal feature fusion methods aim to integrate these embeddings into a more detailed representation which will help in generating relevant answers for the VQA task. The Bilinear pooling operation generates a high dimensional feature vector by combining the two input feature vectors by taking their outer product. Multimodal Factorized Bilinear (MFB) pooling method [9] expands the individual feature representation to higher dimensional space and then uses factorization to get the compact representation. MFH [10] captures complex interactions between the features generating a compact representation of the feature vector. In VQA, through visual attention mechanism, the model focuses on specific segments of the image which are crucial for generating the answer of the question. In MFH with co-attention [10], along with visual attention, question attention is also performed. Question attention focuses on specific parts of the question.

In this work, the question and images features have been fused using MFH with co-attention. Additionally, the deep learning and handcrafted features are merged using MFB pooling method to get a fused image feature of length 1000. Both methods are shown in figure 1 in blocks MFH+CoAtt and MFB. Finally, the features obtained from the fusion of image and question are passed through a full-connected (FC) layer to classify and generate the answer. By learning the effective fused feature representation, the model can generate an appropriate prediction and classify accurately.

Table 1
Question-wise accuracy score of the model

Question	Accuracy Score
What type of procedure is the image taken from?	0.983488
Have all polyps been removed?	0.881837
Is this finding easy to detect?	0.755934
Is there a green/black box artefact?	0.941692
Is there text?	0.931373
What color is the abnormality?	0.107843
What color is the anatomical landmark?	0.994840
How many findings are present?	0.748710
How many polyps are in the image?	0.863261
How many instruments are in the image?	0.873581
Where in the image is the abnormality?	0.083591
Where in the image is the instrument?	0.738390
Are there any abnormalities in the image?	0.782250
Are there any anatomical landmarks in the image?	0.763674
Are there any instruments in the image?	0.792054
Where in the image is the anatomical landmark?	0.789990
What is the size of the polyp?	0.737874
What type of polyp is present?	0.764706

3. Dataset description

The VQAMed 2023 dataset [18] contains 2000 training images and 1938 test images. The dataset consists of GI tract images spanning from mouth to anus. Specifically, it includes images from various procedures such as gastroscopy, colonoscopy, and capsule endoscopy. A set of 18 questions has been defined, and every image is processed with the question set to generate the corresponding answers for each image. The questions within the dataset cover a wide range of topics, including abnormalities, surgical instruments, and normal findings related to factors like location, color, and more. The training set consists of 36000 unique image-QA pairs. For the testing data, answers are to be predicted for the same predefined set of questions for each image.

4. Experiments

4.1. Model Training

For every image, the answers of 18 predefined questions have been provided for the training set, and the unique entries are treated as a single label. The multimodal image and the corresponding question embeddings are fused to get a descriptive feature vector for the input image question pair. The classification layer utilizes the information encoded in the fused features and aims to predict the most relevant answer. The model has been trained for 30 epochs. L2 regularizer and dropouts have been used to avoid over-fitting, and the batch size is set to 64.

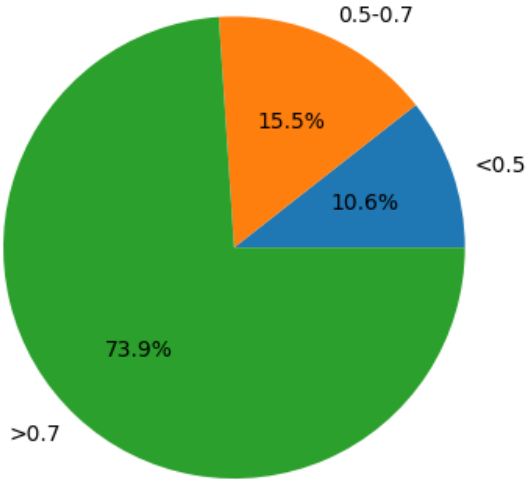

Figure 2: Image-wise accuracy of the model

4.2. Evaluation

The performance of the VQA model in the MEDVQA-GI [18] task was evaluated using the accuracy score metric [18]. The overall accuracy of the proposed model is 75.19%. Additionally, question-wise and image-wise accuracy was also measured. Table 1 shows the question-wise accuracy score, while the image-wise accuracy score is presented graphically in figure 2 by grouping the accuracy scores in the range 0-0.5, 0.5-0.7 and 0.7-1. The question-wise accuracy reveal that the proposed model achieved comparatively high accuracy for the majority of questions. However, there were two specific questions, "What color is the abnormality?" and "Where in the image is the abnormality?" where the model's accuracy decreased noticeably. These questions were about color and location of abnormality and the model needs to be improved further to capture these details about abnormality region. Also, figure 2 suggest that for 73.9% of all the images the image-wise accuracy score was more than 0.7.

5. Conclusion

In this paper, our model submitted for VQA subtask in the MEDVQA-GI challenge of ImageCLEF 2023 was described. The proposed model achieved an overall accuracy of 75.19%. Our model combined deep and domain-specific handcrafted features, to obtained effective feature representation for the input image. In addition, the text features were obtained from Bio-BERT model which is pre-trained on medical images. The fusion of the image and text features using MFH+CoAtt method generated a meaningful embedding for classification. Moving forward, the model's accuracy will be improved by exploring alternative feature extraction methods and

evaluating its performance on additional datasets.

References

[1] Z. Lin, D. Zhang, Q. Tac, D. Shi, G. Haffari, Q. Wu, M. He, Z. Ge, Medical visual question answering: A survey, 2022. arXiv:2111.10056.

[2] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, 2015. arXiv:1409.1556.

[3] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770–778. doi:10.1109/CVPR.2016.90.

[4] K. Cho, B. van Merriënboer, D. Bahdanau, Y. Bengio, On the properties of neural machine translation: Encoder–decoder approaches, in: Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation, Association for Computational Linguistics, Doha, Qatar, 2014, pp. 103–111. URL: https://aclanthology.org/W14-4012. doi:10.3115/v1/W14-4012.

[5] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural Computation 9(8) (1997) 1735–1780.

[6] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, I. Polosukhin, Attention is all you need, in: I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, R. Garnett (Eds.), Advances in Neural Information Processing Systems, volume 30, Curran Associates, Inc., 2017. URL: https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

[7] D. Bahdanau, K. Cho, Y. Bengio, Neural machine translation by jointly learning to align and translate, in: Y. Bengio, Y. LeCun (Eds.), 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings, 2015. URL: http://arxiv.org/abs/1409.0473.

[8] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, Y. Bengio, Show, attend and tell: Neural image caption generation with visual attention, in: F. Bach, D. Blei (Eds.), Proceedings of the 32nd International Conference on Machine Learning, volume 37 of *Proceedings of Machine Learning Research*, PMLR, Lille, France, 2015, pp. 2048–2057. URL: https://proceedings.mlr.press/v37/xuc15.html.

[9] 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 (ICCV), 2017.

[10] 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. URL: https://doi.org/10.1109%2Ftnnls.2018.2817340. doi:10.1109/tnnls.2018.2817340.

[11] M. Lin, Q. Chen, S. Yan, Network in network, 2014. arXiv:1312.4400.

[12] 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.

[13] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, 2019. arXiv:1810.04805.

[14] 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.

[15] B. Jung, L. Gu, T. Harada, bumjunjung at vqa-med 2020: Vqa model based on feature extraction and multi-modal feature fusion, in: CLEF (Working Notes), 2020.

[16] Q. Xiao, X. Zhou, Y. Xiao, K. Zhao, Yunnan university at vqa-med 2021: Pretrained biobert for medical domain visual question answering, in: CLEF (Working Notes), 2021.

[17] B. Ionescu, H. Müller, A. Drăgulinescu, W. Yim, A. Ben Abacha, N. Snider, G. Adams, M. Yetisgen, J. Rückert, A. García Seco de Herrera, C. M. Friedrich, L. Bloch, R. Brüngel, A. Idrissi-Yaghir, H. Schäfer, S. A. Hicks, M. A. Riegler, V. Thambawita, A. Storås, P. Halvorsen, N. Papachrysos, J. Schöler, D. Jha, A. Andrei, A. Radzhabov, I. Coman, V. Kovalev, A. Stan, G. Ioannidis, H. Manguinhas, L. Ştefan, M. G. Constantin, M. Dogariu, J. Deshayes, A. Popescu, Overview of ImageCLEF 2023: Multimedia retrieval in medical, social media and recommender systems applications, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction, Proceedings of the 14th International Conference of the CLEF Association (CLEF 2023), Springer Lecture Notes in Computer Science (LNCS), Thessaloniki, Greece, 2023.

[18] S. A. Hicks, A. Storås, P. Halvorsen, T. de Lange, M. A. Riegler, V. Thambawita, Overview of imageclefmedical 2023 – medical visual question answering for gastrointestinal tract, in: CLEF2023 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, Thessaloniki, Greece, 2023.

[19] D. T. Nguyen, T. D. Pham, N. R. Baek, K. R. Park, Combining deep and handcrafted image features for presentation attack detection in face recognition systems using visible-light camera sensors, Sensors 18 (2018). URL: https://www.mdpi.com/1424-8220/18/3/699. doi:10.3390/s18030699.

[20] T. Ojala, M. Pietikainen, T. Maenpaa, Multiresolution gray-scale and rotation invariant texture classification with local binary patterns, IEEE Transactions on Pattern Analysis and Machine Intelligence 24 (2002) 971–987. doi:10.1109/TPAMI.2002.1017623.

[21] R. M. Haralick, K. Shanmugam, I. Dinstein, Textural features for image classification, IEEE Transactions on Systems, Man, and Cybernetics SMC-3 (1973) 610–621. doi:10.1109/TSMC.1973.4309314.