Deep Multimodal Learning for Medical Visual Question Answering

Lei Shi[1], Feifan Liu[2§], Max P. Rosen[2]

[1] Worcester Polytechnic Institute, Worcester MA 01609, USA
lshi@wpi.edu
[2] University of Massachusetts Medical School, Worcester MA 01655, USA
feifan.liu@umassmed.edu, max.rosen@umassmemorial.org

Abstract. This paper describes the participation of the University of Massachusetts Medical School in the ImageCLEF 2019 Med-VQA task. The goal is to predict the answers given the medical images and the questions. The categories of the questions are provided for the training and validation datasets. We implemented long-short-term memory (LSTM) for question textual feature extraction and transfer learning followed by the co-attention mechanism for image feature extraction. Due to the provided category information, we implemented the SVM model the predict the question category which is used as another feature for our system. In addition, we applied the embedding based topic model (ETM) to generate question topic distribution as one more feature for our system. To efficiently integrate different types of features, we employed the multi-modal factorized high-order pooling (MFH). For the answer prediction, we developed a two-channel framework to handle different categories of questions through single-label classification and multi-label classification respectively. We submitted 3 valid runs, and the best system achieved the accuracy of 0.566 and the BLEU score of 0.593, ranking the 5th place among 17 participating groups.

Keywords: Visual Question Answering, Transfer Learning, ETM, Multi-modal Fusion.

1 Introduction

Given an image and a natural language question about the image, visual question answering (VQA) task is to provide an accurate natural language answer. This task combines computer vision (CV) and natural language processing (NLP). One of the challenges for VQA task is how to fuse different types of features. Various methods, like LinearSum and Multi-modal Factorized Bilinear Pooling (MFB), have been designed and practiced on the VQA task.

§ Corresponding author.

A lot of studies of VQA task are in the general domain. With increasing implementations of deep learning to support clinical decision making and improve patient engagement, some studies begin to focus on the VQA task in the medical domain. ImageCLEF 2019 [1] organized the inaugural edition of the Medical Domain Visual Question Answering (Med-VQA) Task [2]. Given a medical image with a clinically relevant question, the system is tasked with answering the question based on the visual image content. The dataset of this year is different from the last year. The categories of the questions are provided. For the questions in the first three categories, there are a limited number of answer candidates. And the answers to the questions in the last category are narrative.

In this work, we introduced the question category information and the question topic distribution as two additional information during the information fusion process. To develop an integrated system that is able to handle all four categories of questions, we developed two-channel structures for the answer prediction. One channel is to classify the image-question pairs into the close set of answer candidates. The other channel is to generate a narrative answer given an image-question pair.

2 System description

Our system consists of 6 components: transfer learning for image feature extraction, LSTM for question textual feature extraction, other features (including question category information and question topic distribution), co-attention mechanism, MFH for feature fusion, and answer generation. Fig. 1 shows the architecture of our system.

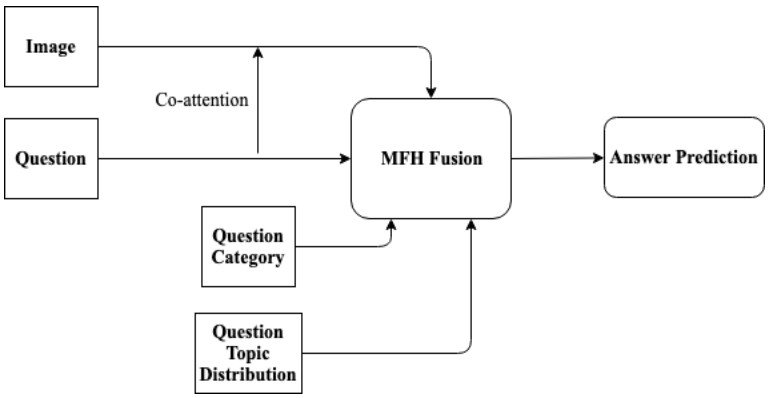

Fig. 1. Our system architecture at Med-VQA

2.1 Question Processing

A pre-trained biomedical word embedding (dimension of 200) is used as the embedding layer. After the word embedding layer, a bidirectional LSTM network is used to extract

textual features of the question. During the training process, the embedding of the "unknown" token is first initialized randomly and then learned. The textual features are transformed to predict the attention weight of different grid locations, which generates the attentional features of the question.

2.2 Image Processing

We applied transfer learning to extract image features. The pre-trained ResNet-152 model of ImageNet (excluding the last 2 layers, pooling layer and fully-connected layer) is the image feature extractor. The parameters of the last 2 convolutional blocks of ResNet-152 model are fine-tuned during the training process. Then we applied the co-attention mechanism to generate the attentional features of the image.

2.3 Question Topic Distribution

ETM [3] is applied to generate several topics from the questions. 10 topics are generated by applying ETM on the questions in the training dataset. Each question is assigned a vector of topic distribution according to the frequencies of topics' words appearing in the question. The topic distribution is used as another input feature of the MFH.

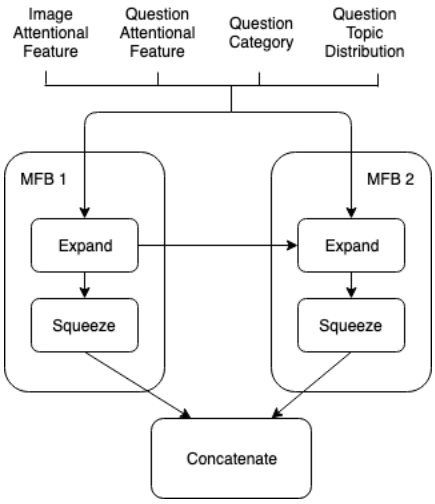

Fig. 2. Feature fusion with MFH

2.4 Question Categorization

According to the instruction of the 2019 Med-VQA, the questions are from 4 categories. SVM is used to classify the category of the question. We applied TF-IDF based unigram vectorization to extract textual features, and trained a support vector machines (SVM) model using the questions from the training dataset. The accuracy of the SVM model

on validation dataset is 100%, which shows that the language used in different categories of questions are relatively unique and less ambiguous. The category information of the question is used as an additional input of the MFH.

2.5 Feature Fusion

MFH [4] contains multiple dependent MFB blocks. The output from the expand stage of the previous MFB block is fed into the next MFB block as additional input, and the output from multiple MFB blocks are merged together as a final fused feature representation.

We applied a 2-block MFH model to fuse 4 types of features including image attentional features, question attentional features, question topic distribution, and question category, which is shown in Fig. 2.

2.6 Answer Prediction

According to the instruction of 2019 Med-VQA, for the questions of the first 3 categories, the corresponding answers are in a limited number of certain candidates. We regarded this case as a single-label classification task. On the other hand, the questions of "abnormality" category are the narrative type. This case is regarded as a multi-label classification task. So, we built two-channel structures which are shown in Fig. 3. One is for the single-label classification task and the other one is for the multi-label classification task. For multi-label classification, each unique word in the answer sentence is considered an answer label for the corresponding image-question pair. Based on the distribution of all the answer labels, the narrative answer is generated using the sampling method.

Therefore, both a classification result and a distribution of words in the answer are predicted through our system for each pair of image-question. If the classification result is one of the certain candidates of the answers, the final answer is that candidate. Otherwise, the final answer is a combination of words generated by the sampling method.

The loss function of our system is an integration of two loss functions. We applied the cross-entropy loss function for the single-label classification structure and Kullback–Leibler divergence loss function for the multi-label classification structure. Given an image-question pair, the loss L is calculated as follows:

$$L = (1 - A) * CrossEntropyLoss + A * KLDivLoss \quad (1)$$

where A is 1 if the predicted category of the question is "Abnormality" otherwise 0.

3 Experiments

We experimented with 4 settings of the pre-trained ResNet-152 model on ImageNet: (1) Res-2 is using the pre-trained ResNet-152 model excluding the last 2 layers (pooling

layer and fully-connected layer); (2) Res-3 is using the pre-trained ResNet-152 model excluding the last 3 layers (last residual block, pooling layer and fully-connected layer); (3) Res-2-tunable is using the pre-trained ResNet-152 model excluding the last 2 layers, and the last residual block is fine-tuned during the training process of our system; (4) ETM-Res-2 is using the pre-trained ResNet-152 model excluding the last 2 layers. We used the topics of the questions generated by the ETM model to label the corresponding images and fine-tuned this ResNet-152 model.

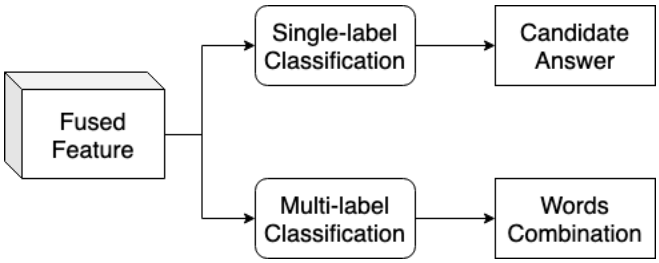

Fig. 3. Two-channel structures for answer prediction

A pre-trained word-embedding (dimension of 200) on PubMed and the clinical notes from MIMIC-III Clinical Database is used as the word embedding layer. We experimented with 2 settings to handle "unknown" token in the questions: (1) Fixed-Unknown is using a fixed 0 vector for "unknown" token; (2) Learned-Unknown is initializing a random vector for "unknown" token and this vector is trained during the training process of our system.

Table 1. Summary of experiments on the validation dataset

Answer Max Length	Word Embedding	Image Feature Extractor	Classification Accuracy	BLEU Score
10	Fixed-Unknown	Res-2	0.575	0.473
10	Fixed-Unknown	Res-3	0.591	0.602
9	Fixed-Unknown	ETM-Res-2	0.558	0.457
6	Learned-Unknown	Res-2-tunable	0.594	0.626

Table 2. Summary of submissions in ImageCLEF 2019

Answer Max Length	Word Embedding	Image Feature Extractor	Accuracy	BLEU Score
9	Fixed-Unknown	ETM-Res-2	0.018	0.039
10	Fixed-Unknown	Res-3	0.48	0.509
6	Learned-Unknown	Res-2-tunable	0.566	0.593

3.1 Performance on Validation Dataset

Table 1 shows the performance of different settings on the validation dataset. We can see that ETM based transfer learning is not as helpful as shown in [5]. It is partially because the questions in 2019 data are less diverse and the topic labels derived from ETM are not distinctive enough to finetune the pre-trained ImageNet model during image classification training. We also found that visual features from different layers of Residual networks perform differently (Res-2 vs. Res-3 in Table 1), which suggests that combining them together may help improve the system performance.

3.2 Official Test Runs in ImageCLEF 2019

We submitted 3 runs on the test dataset and the settings and results are shown in Table 1. The accuracy of our best system is 0.566 and the BLEU score is 0.593 on the test dataset, ranking fifth place. The team ranking first place of this competition obtained an accuracy of 0.624 and a BLEU score of 0.644.

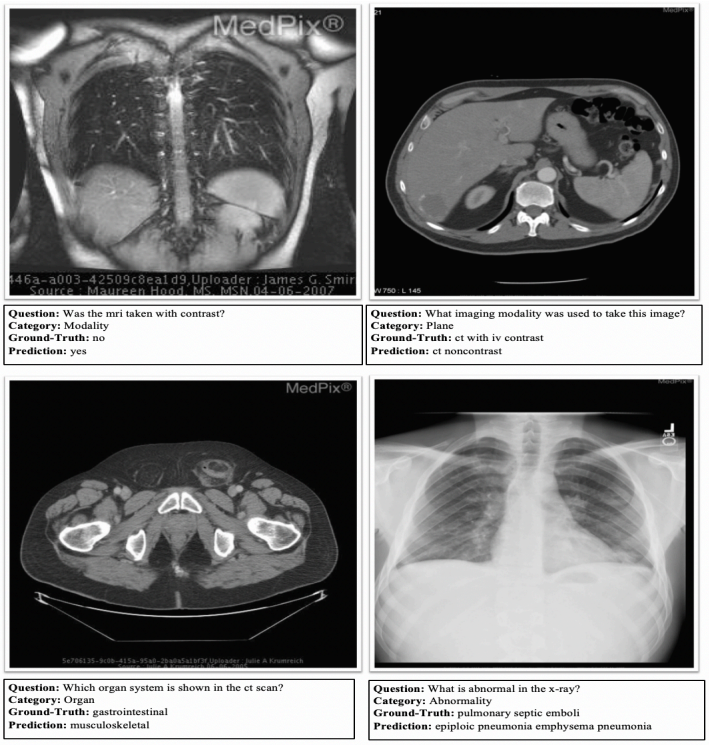

Fig. 4. Examples of poor predictions

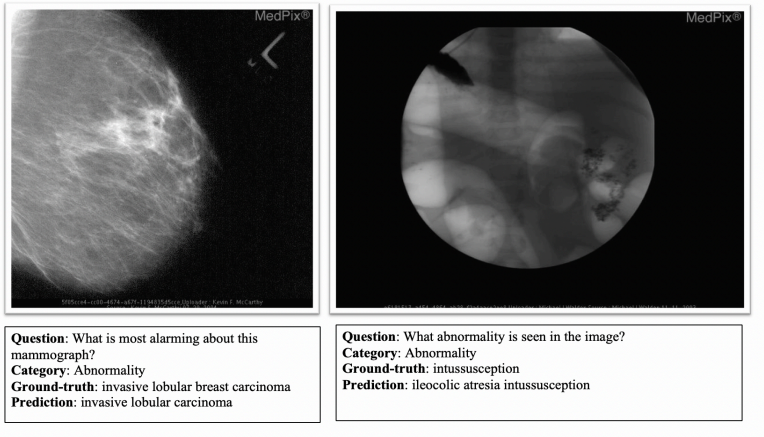

Fig. 5. Examples of good predictions

3.3 Examples of System Outputs on Validation Dataset

Fig. 4 and Fig. 5 show some examples of poor predictions and good predictions that our best system makes on the validation dataset. More analysis is needed to investigate the system's performance on different question categories and identify different error patterns to inform future improvements.

4 Conclusion

We experimented with three different settings of deep learning structures for MED-VQA task 2019, where we introduced two more types of features and we constructed two-channel structures for answer prediction. Due to the limited time, we did not implement Bidirectional Encoder Representations from Transformers (BERT) [6] which will be explored in the future to extract question textual features.

5 Acknowledgments

We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan Xp GPU used for this research.

References

1. Bogdan Ionescu, Henning Müller, Renaud Péteri, Yashin Dicente Cid, Vitali Liauchuk, Vassili Kovalev, Dzmitri Klimuk, Aleh Tarasau, Asma Ben Abacha, Sadid A. Hasan, Vivek V. Datla, Joey Liu, Dina Demner-Fushman, Duc-Tien Dang-Nguyen, Luca Piras, Michael Riegler, Minh-Triet Tranand Mathias Lux, Cathal

Gurrin, Obioma Pelka, Christoph M. Friedrich, Alba García Seco de Herrera, Narciso Garcia, Ergina Kavallieratou, Carlos Roberto del Blanco, Carlos CuevasRodríguez, Nikos Vasillopoulos, Konstantinos Karampidis, Jon Chamberlain, Adrian Clark, Antonio Campello: ImageCLEF 2019: Multimedia Retrieval in Medicine, Lifelogging, Security and Nature. In: Experimental IR Meets Multilinguality, Multimodality, and Interaction. Lecture Notes in Computer Science, Springer, Lugano, Switzerland (2019).
2. Ben Abacha, A., Hasan, S.A., V. Datla, V., Liu, J., Demner-Fushman, D., Müller, H.: VQA-Med: Overview of the Medical Visual Question Answering Task at ImageCLEF 2019. In: CLEF2019 Working Notes. CEUR Workshop Proceedings (http://ceur-ws.org/) Vol. 2380, ISSN 1613-0073, Lugano, Switzerland (2019).
3. Qiang, J., Chen, P., Wang, T., Wu, X.: Topic Modeling over Short Texts by Incorporating Word Embeddings. arXiv:1609.08496 [cs]. (2016).
4. Yu, Z., Yu, J., Fan, J., Tao, D.: Multi-modal Factorized Bilinear Pooling with Co-Attention Learning for Visual Question Answering. arXiv:1708.01471 [cs]. (2017).
5. Peng, Y., Liu, F., Max P. Rosen: UMass at ImageCLEF Medical Visual Question Answering(Med-VQA) 2018 Task. In: CLEF (2018).
6. Devlin, J., Chang, M.-W., Lee, K., Toutanova, K.: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv:1810.04805 [cs]. (2018).