UMass at ImageCLEF Medical Visual Question Answering(Med-VQA) 2018 Task

Yalei Peng[1], Feifan Liu[2], and Max P. Rosen[2]

[1] Worcester Polytechnic Institute, Worcester MA 01609, USA ypeng5@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 2018 Med-VQA task. The goal is to build a system that is able to reason over medial images and questions and generate the corresponding answers. We explored and implemented a co-attention based deep learning framework where residual networks is used to extract visual features from image that interact with the long-short term memory(LSTM) based question representation providing fine-grained contextual information for answer derivation. To efficiently integrate visual features from the image and textual features from the question, we employed Multi-modal Factorized Bilinear(MFB) pooling as well as Multi-modal Factorized High-order(MFH) pooling. In addition, we exploited transfer learning on pre-trained ImageNet model where embedding based topic model(ETM) is applied on the question texts of the training data and the corresponding topic labels are attached to each image for transfer learning. We submitted 3 valid runs for this task, and we found the ETM based transfer learning outperformed other models, achieving the best WBSS score of 0.186, which ranked first among participating groups.

Keywords: Visual Question Answering · Attention Mechanism · LSTM · Residual Nets · Multi-modal Fusion · Topic Analysis

1 Introduction

Given an image and a question in natural language, visual question answering (VQA) system is expected to reason over both visual and textual information to infer the correct answer. It is a challenging task that combines computer vision with natural language processing (NLP) and has received increasing attention. Various kinds of methods, like joint embedding approaches, attention mechanisms and compositional models, have been designed and practiced on this task. Meanwhile, data sets for learning VQA have also been evolving from simple image-QA datasets like COCO-QA to knowledge base-enhanced datasets like Visual Genome.

However, the study of VQA so far is mainly in general domain. There are few

practice of VQA in other domain. With the increasing implementation of artificial intelligence (AI) into medical domain to support clinical decision making and improve patient engagement, the automation of medical image interpretation is becoming more and more desirable. The system is expected to help patients better understand their conditions regarding their available data which can be structured and unstructured, graphical and textual. Also the system, as a opinion machine, may enhance clinicians confidence in interpreting complex medical images. Motivated by this important need for automated image understanding in an advanced question answering manner for clinical domain, ImageCLEF 2018 [1] organized the inaugural edition of the Medical Domain Visual Question Answering(Med-VQA) Task [2].

The implementation of VQA into medical domain is challenging not only because texts and images in medical domain are distinct from those in general domain, but also because the data resources in medical domain are limited compared with those in general domain. Thus, transfer learning from general domain is more promising than directly training from scratch.

Our main contributions in participating this challenge are as follows: First, we explored transfer learning on image channel to extract meaningful features from the medical images, where we present a novel approach of utilizing Embedding-based topic modeling for transfer learning. Second, we implemented co-attention mechanism integrated with Multi-modal Factorized Bilinear Pooling (MFB) and Multi-modal Factorized High-order Pooling (MFH) to medical VQA.

2 Related Work

There Research on VQA has been showing increased interest due to methodological advances in both computer vision and NLP, and the availability of relevant large-scale datasets. The straightforward solution to VQA is the joint embedding method(e.g. [7]), where image and question are represented as global features which are merged to predict the answers. The limitation for this approach is that an image could contain more information than needed to answer a question, which may add noises to the classification model, making it difficult to answer questions pertaining to a specific part of the image. Therefore recent work on VQA explored attention mechanisms(e.g. [3]) to improve the performance by steering the model to specific sections of the input (image and/or question. The main idea is to replace the global image features with fine-grained spatial feature maps so that feature maps can interact with the given question to derive salient features for answer prediction.

Another line of work in VQA focuses on efficient ways for multi-modal feature fusion. A simple approach that has been widely used is linear fusion model, where visual features from image and textual features from question are concatenated or element-wise added. Due to the largely different distributions of two

feature sets, the expressive power of the resulting fused representation is limited in terms of facilitating the final answer prediction. To address this issue, several approaches were proposed, such as Multi-modal Compact Bilinear (MCB) [6], Multi-modal Low-rank Bilinear (MLB) [4], and Multi-modal Factorized Bilinear pooling (MFB) [5]. In our medical VQA system, we integrated the MFB approach for multi-modal feature fusion which was shown to outperform both MCB and MLB in general domain VQA datasets.

3 Methods

Our system consists of four main components: Feature fusion, Co-attention mechanism, Transfer learning, and answer prediction, which are shown in Fig. 1. Specifically, visual context is extracted from the image facilitated by transfer learning, then fused with textual context from the question using co-attention mechanism and feature fusion techniques. Finally, answer is predicted based on the fused multi-modal contextual information.

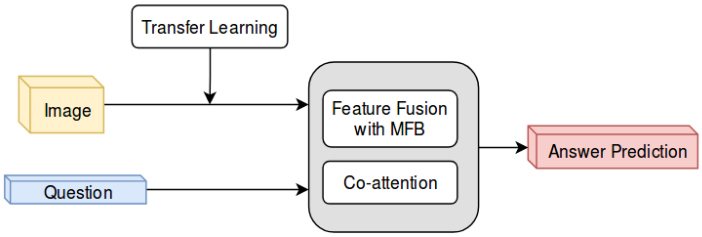

Fig. 1. Our system architecture at MED-VQA.

3.1 Feature Fusion with Multi-modal Factorized Bilinear Pooling

We used MFB pooling method to merge the visual features from image and textual features from question, as it was shown to have dual benefits of compact output features of MLB and robust expressive capacity of MCB. For comparison, we also integrated multi-modal factorized high-order(MFH) pooling which consists of N MFB modules (N is a hyper-parameter).

Each MFB block contains two stages: expand and squeeze. In the expand stage, the textual context and the visual context are transformed into the same dimension by a fully-connected layer respectively for the next element-wise multiplication. Additionally, a dropout layer is next to the element-wise multiplication unit. Then, the fused context is further transformed in squeeze stage which contains sum pooling, power normalization and L2-normalization.

In the MFH module, the output from the dropout layer of the previous MFB block is fed into the next MFB block as additional input as shown in Fig. 2, and the output from multiple MFB blocks are merged together as a final fused feature representation.

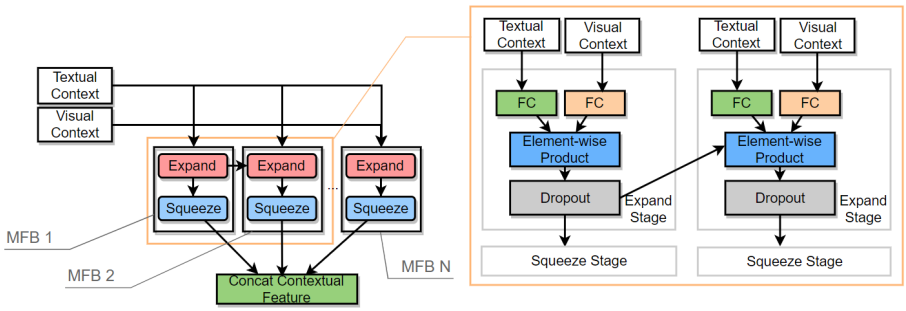

Fig. 2. The high-order MFH model which consists of N MFB blocks [5]

.

3.2 Co-attention with MFB

Similar to [5], we also implemented Co-attention mechanism for MED-VQA. The pre-trained ResNet152 model of ImageNet (excluding the last 3 layers) is used as a image feature extractor, and a LSTM layer is used to encode the question into textual feature vectors. A pre-trained word-embedding (dimension of 200) on wikipedia, pubmed articles and Pittsburgh clinical notes is used as embedding input layer. MFB was used to fuse the the multi-modal features, followed by some feature transformations (e.g., $1 * 1$ convolution and ReLU activation) and softmax normalization to predict the attention weight for each grid location. Based on the attention map, the attentional image features are obtained by the weighted sum of the spatial grid vectors. Multiple attention maps are generated to enhance the learned attention map, and these attention maps are concatenated to output the attentional image features. Next, the final attentional image features are merged with the question features using MFB for downstream answer prediction.

3.3 Transfer Learning to Tune Pretrained ResNet with ETM Labels

ImageNet data are very different from medical images in MED-VQA task, which motivates us to employ transfer learning to adapted pre-trained model to this task. Instead of fine tuning the pre-trained model on the fly, the off-line transfer learning based method can efficiently reduce the training time.

We explored topic analysis to derive semantic label for each image in order to enable transfer learning. The assumption is that the semantics of the question text should match the corresponding image. However, the question text is typically short which is challenging for traditional topic analysis approaches, such as probabilistic latent semantic analysis (PLSA) and latent Dirichlet allocation (LDA), to infer reliable topics as only very limited word co-occurrence information is available in short texts. Embedding-based topic model [8] not only solves the problem of very limited word co-occurrence information by aggregating short texts into long pseudo-texts, but also utilizes a Markov Random Field regularized model that gives correlated words a better chance to be put into the same topic as shown in Fig. 3. First, short texts are merged into long pseudo-texts based on clustering methods using a word embedding pre-trained on a large relevant corpus. Then, embedding-based topic model is applied on the long pseudo-texts to generate latent topics.

Specifically, we applied ETM on question texts of the MED-VQA data, assigning a topic label to each question which can in turn be used as a semantic label for its corresponding image. We then performed transfer learning in a context of image classification, where the parameters of pre-trained residual nets were tuned with the goal of correctly classifying all the images to their corresponding topic labels. The fine-tuned network (removing the last convolution block, fully-connected layer and softmax layer) was used as the static feature extractor in our system architecture.

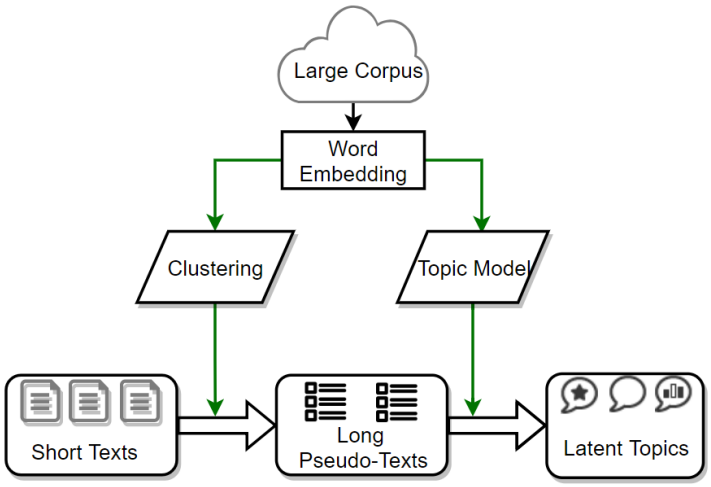

Fig. 3. Embedding based topic model for short texts [8].

3.4 Answer Prediction

The input to answer prediction is the attentional image features from Co-attention, fused with the LSTM based question representational features through MFB. Here we employed a simple multi-label classification method where each unique word in the answer sentence is considered a answer label for the corresponding image-question pair. Based on distribution of all the answer labels, the final answer is generated using sampling method.

4 Experiments

4.1 Data

Statistics of Med-VQA dataset is shown in Table 1. The training, validation and test data splits have 5413, 500 and 500 instances respectively. Both questions and answers are on average longer than those in VQA datasets in general domain. The word-embedding (dimension of 200) ,which was pretained on wikipedia, pubmed articles and Pittsburgh clinical notes, has a good coverage (roughly over 95%) on both question and answer words of each data split. Also, note that the number of images is less than the number of question-answer pairs, which means several question-answer pairs may share a common image. Especially in training dataset shown in Table 1, there are 2278 images which are less than half of the number of question-answer pairs (5413).

Table 1. Statistics of Med-VQA datasets

		Train	Valid	Test
Question	Num	5413	500	500
	Max_length	28	15	14
	Min_length	3	4	4
	Avg_length	9.63	7.38	6.968
	Emd_Coverage	94.99%	96.93%	95.52%
Answer	Num	5413	500	500
	Max_length	26	14	\
	Min_length	1	1	\
	Avg_length	6.03	4.06	\
	Emd_Coverage	95.05%	96.54%	\
Image	Num	2278	324	264

4.2 Preprocessing

Question-Answer Pair Preprocessing on question-answer pairs includes tokenization and lower casing, so that each word can be mapped to its dense representation by looking up pre-trained word embeddings.

Image Although original preprocessing procedure is recommended to better facilitate the transfer learning, we notice that lots of images in medical VQA data set are long shape consisting 2 - 5 sub-images. Therefore, a lot of areas would be cut off, and features would be resized to be too small and blur if the original preprocessing is directly applied. Therefore, we reshape the long images into approximate squares by re-arranging the order of sub-images. Then, the original preprocessing when pre-training the ImageNet ResNet is applied.

4.3 Validation Runs

We experimented with the three co-attention systems with variant settings on feature fusion and transfer learning:(1)ResNet152+MFB which uses MFB for feature fusion and the pre-trained ResNet152 is directly used; (2)ResNet152+MFH which uses MFH for feature fusion and the pre-trained ResNet152 is directly used; (3)ResNet152+ETM+MFH which uses MFH for feature fusion, and the pre-trained ResNet152 is also tuned through transfer learning which is based on ETM topic modeling. In Fig. 4, we shows the performance curves of 3 systems on validation dataset. We can see the MFH based feature fusion constantly outperforms the MFB based method.

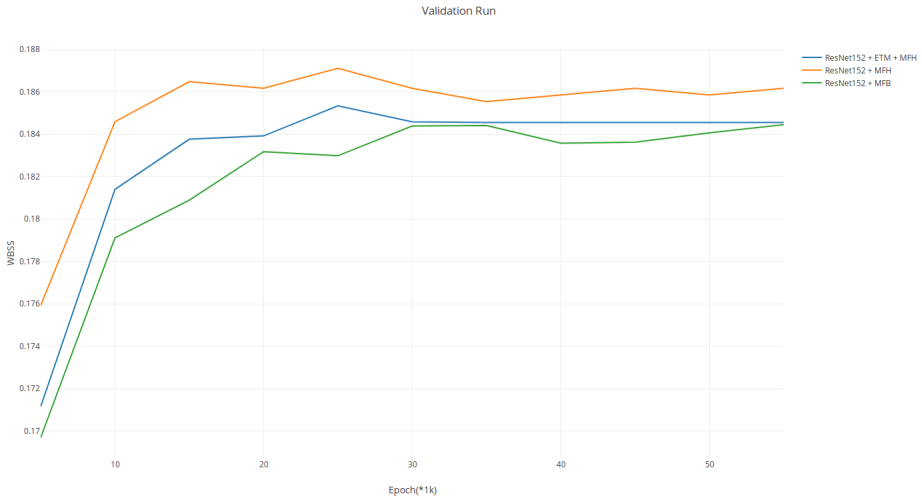

Fig. 4. Validation runs of 3 architectures

4.4 Official Test Runs in ImageCLEF 2018

We submitted 3 valid runs based on the aforementioned system architectures, and the run from "ResNet152+ETM+MFH" achieved the best WBSS score of

0.186, and the run from "ResNet152+MHF" obtained the best BLEU score of 0.162 as shown in Table. 2. Note that the run (ID6091) is not a valid run due to a code error.

Table 2. Summary of submissions in ImageCLEF 2018

Run	WBSS	BLEU	CBSS	Type	Models
6069	0.18616	0.15833	0.02295	automatic	ResNet152 + ETM + MFH
6113	0.18455	0.16159	0.01649	automatic	ResNet152 + MFH
5980	0.18445	0.15966	0.02053	automatic	ResNet152 + MFB

5 conclusions

We experimented with 3 different deep learning architectures for MED-VQA task 2018, where we proposed a novel method for transfer learning using embedding based topic analysis. We found that transfer learning and MFH based feature fusion is helpful on improving the system's performance. Due to time limitation, we didn't model the sequential information in the answer sequence which will be explored in the future to make the answer more natural and readable.

References

1. Bogdan Ionescu, Henning Mller, Mauricio Villegas, Alba Garca Seco de Herrera, Carsten Eickhoff, Vincent Andrearczyk, Yashin Dicente Cid, Vitali Liauchuk, Vassili Kovalev, Sadid A. Hasan, Yuan Ling, Oladimeji Farri, Joey Liu, Matthew Lungren, Duc-Tien Dang-Nguyen, Luca Piras, Michael Riegler, Liting Zhou, Mathias Lux and Cathal Gurrin. Overview of ImageCLEF 2018: Challenges, Datasets and Evaluation, Proceedings of the Ninth International Conference of the CLEF Association (CLEF 2018), 2018.
2. Hasan, S.A., Ling, Y., Farri, O., Liu, J., Lungren, M., Mller, H.: Overview of the ImageCLEF 2018 Medical Domain Visual Question Answering Task. In: CLEF2018 Working Notes. http://ceur-ws.org/, Avignon, France (2018).
3. Ilievski, I., Yan, S., Feng, J.: A Focused Dynamic Attention Model for Visual Question Answering. arXiv:1604.01485 [cs]. (2016).
4. Kim, J.-H., On, K.-W., Lim, W., Kim, J., Ha, J.-W., Zhang, B.-T.: Hadamard Product for Low-rank Bilinear Pooling. arXiv:1610.04325 [cs]. (2016).
5. 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).
6. Fukui, A., Park, D.H., Yang, D., Rohrbach, A., Darrell, T., Rohrbach, M.: Multimodal Compact Bilinear Pooling for Visual Question Answering and Visual Grounding. In: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. pp. 457468. Association for Computational Linguistics, Austin, Texas (2016).
7. Kim, J.-H., Lee, S.-W., Kwak, D.-H., Heo, M.-O., Kim, J., Ha, J.-W., Zhang, B.-T.: Multimodal Residual Learning for Visual QA. arXiv:1606.01455 [cs]. (2016).

8. Qiang, J., Chen, P., Wang, T., Wu, X.: Topic Modeling over Short Texts by Incorporating Word Embeddings. arXiv:1609.08496 [cs]. (2016).