

Overcoming Data Limitation in Medical Visual Question Answering

Binh D. Nguyen¹, Thanh-Toan Do², Binh X. Nguyen¹, Tuong Do¹,
Erman Tjiputra¹, and Quang D. Tran¹

¹ AIOZ Pte Ltd, Singapore
{binh.duc.nguyen, binh.xuan.nguyen, tuong.khanh-long.do,
erman.tjiputra, quang.tran}@aioz.io

² University of Liverpool
thanh-toan.do@liverpool.ac.uk

Abstract. Traditional approaches for Visual Question Answering (VQA) require large amount of labeled data for training. Unfortunately, such large scale data is usually not available for medical domain. In this paper, we propose a novel medical VQA framework that overcomes the labeled data limitation. The proposed framework explores the use of the unsupervised Denoising Auto-Encoder (DAE) and the supervised Meta-Learning. The advantage of DAE is to leverage the large amount of unlabeled images while the advantage of Meta-Learning is to learn meta-weights that quickly adapt to VQA problem with limited labeled data. By leveraging the advantages of these techniques, it allows the proposed framework to be efficiently trained using a small labeled training set. The experimental results show that the proposed method significantly outperforms the state-of-the-art medical VQA. The source code is available at <https://github.com/aioz-ai/MICCAI19-MedVQA>.

Keywords: Visual Question Answering · Auto-Encoder · Meta-Learning.

1 Introduction

Visual Question Answering (VQA) aims to provide a correct answer to a given question such that the answer is consistent with the visual content of a given image. In medical domain, VQA could benefit both doctors and patients. For example, doctors could use answers provided by VQA system as support materials in decision making, while patients could ask VQA questions related to their medical images for better understanding their health. However, one major problem with medical VQA is the lack of large scale labeled training data which usually requires huge efforts to build. The first attempt for building the dataset for medical VQA is by ImageCLEF-Med [6]. In particular, in [6], images were automatically captured from PubMed Central articles. The questions and answers were automatically generated from corresponding captions of images. By that construction, the data has high noisy level, i.e., the dataset includes many images that are not useful for direct patient care and it also contains

questions that do not make any sense. Recently, in [10], the authors released the first manually constructed dataset VQA-RAD for medical VQA. Unfortunately, it contains only 315 images, which prevents to directly apply the powerful deep learning models for the VQA problem. One may think about the use of transfer learning in which the pretrained deep learning models [18,7] that are trained on the large scale labeled dataset such as ImageNet [16] are used for finetuning on the medical VQA. However, due to difference in visual concepts between ImageNet images and medical images, finetuning with very few medical images is not sufficient, which is confirmed by our experiments in Section 4. Therefore it is necessary to develop a new VQA framework that can improve the accuracy while still only needs a small labeled training data.

The motivation for our approach to overcome the data limitation of medical VQA comes from two observations. Firstly, we observe that there are large scale unlabeled medical images available. These images are from same domain with medical VQA images. Hence if we train an unsupervised deep learning model using these unlabeled images, the trained weights may be easier to be adapted to the medical VQA problem than the pretrained weights on ImageNet images. Another observation is that although the labeled dataset VQA-RAD in [10] is primarily designed for VQA, by spending a little effort, we can extract the new class labels³ for that dataset. The new class labels allow us to apply the recent meta-learning technique [4] for learning meta-weights, that can be quickly adapted to the VQA problem later.

From these two observations, we propose a novel medical VQA framework as presented in Figure 1, in which the Model-Agnostic Meta-Learning (MAML) [4] and the Convolutional Denoising Auto-Encoder (CDAE) [12] are used to initialize the model weights for the image feature extraction.

2 Literature Review

Medical Visual Question Answering. Most approaches for medical VQA [10,13,1,20] are to directly apply the state-of-the-art general VQA models to medical domain. The 2018 ImageCLEF-Med challenge [6] provides a good overview about the approaches and their results. Typically, in [13,1,20], the authors use the state-of-the-art attention mechanisms in general VQA (e.g., MCB [5], SAN [19]) to learn a joint representation between an image and a question. Note that in the mentioned approaches, the models pretrained on ImageNet such as VGG [18] or ResNet [7] are directly finetuned on medical VQA images for image feature extraction. However, directly finetuning those models on medical VQA images is not effective due to the limited medical VQA data.

Meta-learning. Traditional machine learning algorithms, especially deep learning based approaches, require large scale labeled training set when learning a new task, even when the model is pretrained on other classification problems [11,2]. Contrasting with traditional machine learning algorithms, meta-learning approach [17] targets to deal with the problem of data limitation when learning new

³ The descriptions of new defined classes are presented in Section 4.1.

Algorithm 1 Overview of the meta-training procedure

```

1: procedure META-TRAIN( $\mathcal{D}$ , model  $f_\theta$ )
2:   Initialize model parameters  $\theta$ 
3:   for  $h = 1$  to  $H$  do                                      $\triangleright$  Meta-update Loop
4:     Create meta-batch of tasks  $\{\mathcal{T}_1, \mathcal{T}_2, \dots, \mathcal{T}_m\}$ 
5:     for each task  $\mathcal{T}_i$  do
6:       Sample data  $\{\mathcal{D}_i^{tr}, \mathcal{D}_i^{val}\}$  of task  $\mathcal{T}_i$ 
7:       Update task models with Eq. (1) using samples from  $\mathcal{D}_i^{tr}$ 
8:       Update meta-model  $\theta$  with Eq. (2) using  $\{\mathcal{D}_1^{val}, \mathcal{D}_2^{val}, \dots, \mathcal{D}_m^{val}\}$ 

```

tasks. Recently, in [4] the authors proposed a new approach for meta-learning, i.e., Model-Agnostic Meta-Learning (MAML), which helps to learn a meta-model (e.g. network weights) from current tasks that is broadly suitable for many tasks. Hence, the model can be quickly adapted to new tasks that have a small number of training images.

Denoising Auto-Encoder. In medical domain, the lack of labeled data makes training process become inefficiency. Thus, unlabeled data, which is easy to achieve, is encouraged to use for training. Auto-Encoder [15,12], which helps to extract high-level features without any label information, is a typical solution to take the advantage of unlabeled data. Besides, medical images such as MRI, CT, X-ray may contain various degree of noises, which might happen during transmission and acquisition [8]. Hence, it requires a feature extraction model that is robust to noise, i.e., it can still extract useful information from the noisy input image. In this work, to leverage the benefit of large scale unlabeled datasets and also to make the model robust to the noise in input images, we propose to use the Convolutional Denoising Auto-Encoder (CDAE) [12] as one of image feature extraction components in our framework.

3 Methodology

The proposed medical VQA framework is presented in Figure 1. In our framework, the image feature extraction component is initialized by pretrained weights from MAML and CDAE. After that, the VQA framework will be finetuned in an end-to-end manner on the medical VQA data. In the following sections, we detail the architectures of MAML, CDAE, and our framework.

3.1 Model-Agnostic Meta-Learning – MAML

The MAML classification model is represented by a parametrized function f_θ with meta-parameters θ . When adapting to a new task \mathcal{T}_i , the model’s parameters θ become θ'_i . Let $\mathcal{D} = \{x_i, y_i\}_{i=1}^N$ be the dataset for training MAML. N is the number of samples. $\{x_i, y_i\}$ is a pair of image (x_i) and its class label (y_i). A task in MAML is defined as a “**k**-shot **n**-way” classification problem. The dataset for each task is defined as $\mathcal{D}' = \{x'_i, y'_i\}_{i=1}^{N'}$; samples in \mathcal{D}' come from

n different classes which are a subset of classes in \mathcal{D} . The task dataset \mathcal{D}' is split equally into two sets \mathcal{D}^{tr} – training set and \mathcal{D}^{val} – validation set; in \mathcal{D}^{tr} , each class contains k training images. The training procedure is described in the Algorithm 1. In each iteration h , m tasks are generated forming a meta-batch for MAML training. For each task \mathcal{T}_i , the corresponding adapted parameters θ'_i are calculated as follows

$$\theta'_i = \theta - \alpha \nabla_{\theta} L_{\mathcal{T}_i}(f_{\theta}(\mathcal{D}_i^{tr})) \quad (1)$$

where $L_{\mathcal{T}_i}$ is the classification loss of task i . After all adapted parameters of m tasks are calculated, the meta-model’s parameters θ are updated via *stochastic gradient descent* (SGD) as follows

$$\theta \leftarrow \theta - \beta \nabla_{\theta} \sum_{\mathcal{T}_i} L_{\mathcal{T}_i}(f_{\theta'_i}(\mathcal{D}_i^{val})) \quad (2)$$

We follow [4] to design MAML. It consists of four 3×3 convolutional layers with stride 2 and is ended with a mean pooling layer; each convolutional layer has 64 filters and is followed by a ReLu layer. The detail training of MAML is presented in Section 4.2. After training, the weights of the meta-model are used for finetuning in the VQA framework as presented in Figure 1.

3.2 Convolutional Denoising Auto-Encoder – CDAE

The encoder maps an image x' , which is the noisy version of the original image x , to a latent representation z which retains useful amount of information. The decoder transforms z to the output y . The training algorithm aims to minimize the reconstruction error between y and the original image x as follows

$$L_{rec} = \|x - y\|_2^2 \quad (3)$$

In our design, the encoder is a stack of convolutional layers; each of them is followed by a max pooling layer. The decoder is a stack of deconvolutional and convolutional layers. The noisy version x' is achieved by adding Gaussian noise to the original image x . The detail training of CDAE is presented in Section 4.2. After training, the trained weights of both encoder and decoder are used for finetuning in the VQA framework as presented in Figure 1.

3.3 The proposed Medical VQA framework

VQA detail. Each input question is trimmed to a 12-word sentence. The question is zero-padded in case its length is less than 12. Each word is represented by a 600-D vector which is a concatenation of the 300-D GloVe word embedding [14] and the augmenting embedding from the VQA-RAD training data [9]. The word embedding is fed into a 1024-D LSTM in order to produce the question embedding, denoted as f_q in Figure 1. Each input image is passed through the Mixture of Enhanced Visual Features (MEVF) component, which produces two

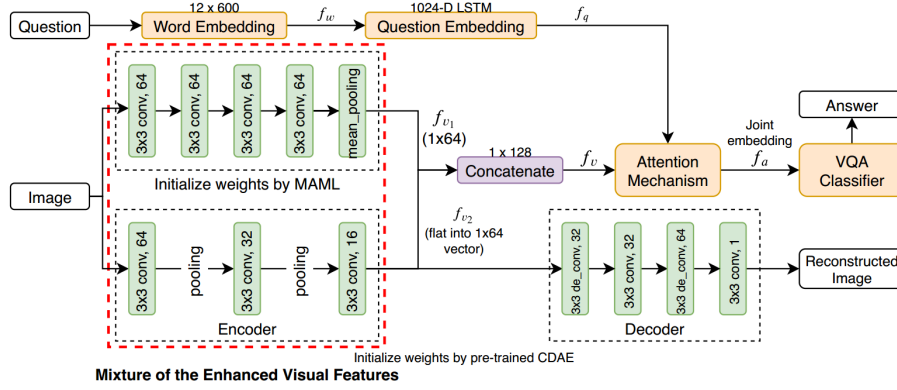


Fig. 1. The proposed medical VQA. The image feature extraction is denoted as “Mixture of Enhanced Visual Features (MEVF)” and is marked with the red dashed box. The weights of MEVF are initialized by MAML and CDAE. Best view in color.

64-D vectors f_{v1} and f_{v2} . Those vectors are concatenated to form an 128-D enhanced image feature, denoted as f_v in Figure 1.

Image feature f_v and question embedding f_q are fed into an attention mechanism (BAN [9] or SAN [19]) to produce a joint representation f_a . This feature f_a is used as input for a multi-class classifier (over the set of predefined answer classes [10]). To train the proposed model, we introduce a multi-task loss function to incorporate the effectiveness of the CDAE to VQA. Formally, our loss function is defined as follows

$$L = \alpha_1 L_{vqa} + \alpha_2 L_{rec} \quad (4)$$

where L_{vqa} is a Cross Entropy loss for VQA classification and L_{rec} stands for the reconstruction loss of CDAE (Eq. 3). The whole VQA model is finetuned in an end-to-end manner using VQA-RAD dataset as presented in Section 4.2.

4 Experiments

4.1 Dataset

The VQA-RAD [10] dataset contains 315 images and 3,515 corresponding questions. Each image is associated with more than one question. The questions are divided into 11 categories which are “Abnormality”, “Attribute”, “Color”, “Count”, “Modality”, “Organ”, “Other”, “Plane”, “Positional reasoning”, “Object/Condition Presence”, “Size”. We use exactly the same training set and test set described in [10]. The test set contains 451 questions and the rest is for training. The questions can be close-ended questions, i.e. the questions in which the answers are “yes/no” and other limited choices, or open-ended questions, i.e., the questions do not have a limited structure and could have multiple correct answers. The dataset has 458 answers. The VQA is posed as a classification over the set of answers.

4.2 Training MAML, CDAE, and the whole VQA framework

MAML. We create the dataset for training MAML by *manually reviewing* around three thousand question-answer pairs from the training set of VQA-RAD dataset. In our annotation process, images are split into three parts based on its *body part* labels (head, chest, abdomen). Images from each body part are further divided into three subcategories based on the interpretation from the question-answer pairs corresponding to the images. These subcategories are: (1) *normal* images in which no pathology is found; (2) *abnormal present* images in which there are the existence of fluid, air, mass, or tumor; (3) *abnormal organ* images in which the organs are large in size or in wrong position. Thus, all the images are categorized into 9 classes: *head normal*, *head abnormal present*, *head abnormal organ*, *chest normal*, *chest abnormal organ*, *chest abnormal present*, *abdominal normal*, *abdominal abnormal organ*, and *abdominal abnormal present*. For every iteration of MAML training (line 3 in Alg. 1), 5 tasks are sampled per iteration. For each task, we randomly select 3 classes (from 9 classes). For each class, we randomly select 6 images in which 3 images are used for updating task models and the remaining 3 images are used for updating meta-model.

CDAE. To train CDAE, we collect 11,779 unlabeled images available online which are brain MRI images [3], chest X-ray images⁴ and CT abdominal images⁵. The dataset is split into train set with 9,423 images and test set with 2,356 images. We use Gaussian noise to corrupt the input images before feeding them to the encoder.

VQA. After training MAML and CDAE, we use their trained weights to initialize the MEVF image feature extraction component in the VQA framework. We then finetune the whole VQA model using the training set of VQA-RAD dataset. In order to make a fair comparison to [10], we evaluate our framework on 300 free-form questions of the test set. The proposed framework is implemented using PyTorch. The experiments are conducted on a single NVIDIA 1080Ti with 11GB RAM. The VQA accuracy is computed as the percentage of the total correct answers over the number of testing questions.

4.3 Ablation Study

We evaluate the effectiveness of different image feature extraction methods in the VQA model when using only MAML, using only CDAE, and their combination MEVF. For each extraction method, we present results when training the VQA model using only VQA-RAD training set (i.e. *from scratch*) or when pretraining as described in Section 4.2 and then finetuning using VQA-RAD training set (i.e. *finetuning*). We also present the results when the pretrained VGG model (on ImageNet) is finetuned on VQA-RAD for image feature extraction.

Table 1 presents VQA accuracy in both VQA-RAD open-ended and close-ended questions on the test set. The results show that for both MAML and

⁴ <https://www.kaggle.com/paultimothymooney/chest-xray-pneumonia>

⁵ <https://www.synapse.org/#!/Synapse:syn3193805/wiki/217753>

Table 1. VQA results on VQA-RAD test set. All reference methods differ at the image feature extraction component. Other components are similar. The Stacked Attention Network (SAN) [19] is used as the attention mechanism in all methods.

Reference methods	VQA accuracy (%)	
	Open-ended	Close-ended
VGG-16 (finetuning)[10]	24.2	57.2
MAML (from scratch)	6.5	68.6
MAML(finetuning)	38.2	69.7
CDAE (from scratch)	13.8	69.2
CDAE (finetuning)	36.7	70.8
MEVF (from scratch)	15.4	70.8
MEVF (finetuning)	40.7	74.1

Table 2. Performance comparison on VQA-RAD test set. The results of SAN framework (fw.) and MCB framework (fw.) are cited from the paper [10].

	SAN fw.[10,19] (baseline)	MCB fw.[10,5] (baseline)	BAN fw.[9] (baseline)	SAN + proposal	BAN + proposal
Open-ended	24.2	25.4	27.6	40.7	43.9
Close-ended	57.2	60.6	66.5	74.1	75.1

CDAE, by firstly pretraining as described in Section 4.2, then finetuning, the finetuning significantly improves the performance over the training from scratch using only VQA-RAD. In addition, the results also show that our pretraining and finetuning of MAML and CDAE give better performance than the finetuning of VGG-16 which is pretrained on the ImageNet dataset. Our proposed image feature extraction MEVF which leverages both pretrained weights of MAML and CDAE, then finetuning them give the best performance. This confirms the effectiveness of the proposed MEVF for dealing with the limitation of labeled training data for medical VQA.

The results also show that for all MAML, CDAE, and MEVF, the accuracy on close-ended questions (CEQ) are higher than those on the open-ended questions (OEQ). Furthermore, the improvements of the finetuning over the training from scratch are more significant on OEQ. We found that OEQ are usually difficult to answer than CEQ, i.e., OEQ mainly ask about the detail description and require long answers, while CEQ mainly ask about the confirmation (i.e., “yes/no”) and usually have short answers. That observation implies that the description answers which need more information from input images take more benefits from the proposed image feature extraction.

4.4 Comparison with the state of the art

We compare our framework (Figure 1) with the baselines in [10]. In [10], the authors report the results when applying the general VQA frameworks, i.e., SAN

framework [19], MCB framework [5]⁶ and finetuning on VQA-RAD dataset. We also report another strong baseline when finetuning the state-of-the-art BAN framework [9]. For our framework, we report results when using SAN [19] or BAN [9] as the attention mechanisms, although other attention mechanisms are straightforward to use in our framework.

Table 2 presents comparative results between methods. Note that for the image feature extraction, the baselines use the pretrained models (VGG or ResNet) that have been trained on ImageNet and then finetune on the VQA-RAD dataset. For the question feature extraction, all baselines and our framework use the same pretrained models (i.e., Glove [14]) and finetuning on VQA-RAD. The results show that when BAN or SAN is used as the attention mechanism in our framework, it significantly outperforms the baseline frameworks BAN [9] and SAN [10,19]. Our best setting, i.e. the one with BAN as the attention, achieves the state-of-the-art results and it significantly outperforms the best baseline framework BAN [9], i.e., the improvements are 16.3% and 8.6% on open-ended and close-ended VQA, respectively.

5 Conclusion

In this paper, we proposed a novel medical VQA framework that leverages the meta-learning MAML and denoising auto-encoder CDAE for image feature extraction in order to overcome the limitation of labeled training data. Specifically, CDAE helps to leverage information from the large scale unlabeled images, while MAML helps to learn meta-weights that can be quickly adapted to the VQA problem. We establish new state-of-the-art results on VQA-RAD dataset for both close-ended and open-ended questions.

References

1. Abacha, A.B., Gayen, S., Lau, J.J., Rajaraman, S., Demner-Fushman, D.: NLM at ImageCLEF 2018 visual question answering in the medical domain. CEUR Workshop Proceedings (2018)
2. Bar, Y., Diamant, I., Wolf, L., Greenspan, H.: Deep learning with non-medical training used for chest pathology identification. In: Medical Imaging: Computer-Aided Diagnosis (2015)
3. Clark, K., Vendt, B., Smith, K., Freymann, J., et.al.: The Cancer Imaging Archive (TCIA): maintaining and operating a public information repository. Journal of Digital Imaging pp. 1045–1057 (2013)
4. Finn, C., Abbeel, P., Levine, S.: Model-agnostic meta-learning for fast adaptation of deep networks. In: ICML (2017)
5. Fukui, A., Park, D.H., Yang, D., Rohrbach, A., Darrell, T., Rohrbach, M.: Multi-modal compact bilinear pooling for visual question answering and visual grounding. In: EMNLP (2016)

⁶ Those frameworks are completed VQA models in which the core components in those frameworks are SAN and MCB attentions. We refer the reader to the corresponding papers [19,5] for the detail of those models.

6. Hasan, S.A., Ling, Y., Farri, O., Liu, J., Lungren, M., Müller, H.: Overview of the ImageCLEF 2018 medical domain visual question answering task. CEUR Workshop Proceedings (2018)
7. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR (2016)
8. Jifara, W., Jiang, F., Rho, S., Cheng, M., Liu, S.: Medical image denoising using convolutional neural network: a residual learning approach. The Journal of Supercomputing pp. 1–15 (2017)
9. Kim, J.H., Jun, J., Zhang, B.T.: Bilinear attention networks. In: NIPS (2018)
10. Lau, J.J., Gayen, S., Abacha, A.B., Demner-Fushman, D.: A dataset of clinically generated visual questions and answers about radiology images. Nature (2018)
11. Maicas, G., Bradley, A.P., Nascimento, J.C., Reid, I., Carneiro, G.: Training medical image analysis systems like radiologists. In: MICCAI (2018)
12. Masci, J., Meier, U., Cireşan, D., Schmidhuber, J.: Stacked convolutional auto-encoders for hierarchical feature extraction. In: ICANN (2011)
13. Peng, Y., Liu, F., Rosen, M.P.: Umass at imageclef medical visual question answering (med-vqa) 2018 task. CEUR Workshop Proceedings (2018)
14. Pennington, J., Socher, R., Manning, C.D.: Glove: Global vectors for word representation. In: EMNLP (2014)
15. Rumelhart, D.E., Hinton, G.E., Williams, R.J.: Learning internal representations by error propagation. Tech. rep. (1985)
16. Russakovsky, O., Deng, J., Su, H., et.al.: Imagenet large scale visual recognition challenge. IJCV pp. 211–252 (2015)
17. Schmidhuber, J.: Evolutionary Principles in Self-referential Learning. (1987)
18. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: ICLR (2015)
19. Yang, Z., He, X., Gao, J., Deng, L., Smola, A.J.: Stacked attention networks for image question answering. In: CVPR (2016)
20. Zhou, Y., Kang, X., Ren, F.: Employing Inception-Resnet-v2 and Bi-LSTM for medical domain visual question answering. CEUR Workshop Proceedings (2018)