# Multiple Meta-model Quantifying for Medical Visual Question Answering

Tuong Do[1], Binh X. Nguyen[1], Erman Tjiputra[1], Minh Tran[1],
Quang D. Tran[1], and Anh Nguyen[2]

[1] AIOZ, Singapore
{tuong.khanh-long.do,binh.xuan.nguyen,erman.tjiputra,
minh.quang.tran,quang.tran}@aioz.io
[2] University of Liverpool, UK
anh.nguyen@liverpool.ac.uk

**Abstract.** Transfer learning is an important step to extract meaningful features and overcome the data limitation in the medical Visual Question Answering (VQA) task. However, most of the existing medical VQA methods rely on external data for transfer learning, while the meta-data within the dataset is not fully utilized. In this paper, we present a new multiple meta-model quantifying method that effectively learns meta-annotation and leverages meaningful features to the medical VQA task. Our proposed method is designed to increase meta-data by auto-annotation, deal with noisy labels, and output meta-models which provide robust features for medical VQA tasks. Extensively experimental results on two public medical VQA datasets show that our approach achieves superior accuracy in comparison with other state-of-the-art methods, while does not require external data to train meta-models. Source code available at: https://github.com/aioz-ai/MICCAI21_MMQ.

**Keywords:** visual question answering· meta learning.

## 1 Introduction

A medical Visual Question Answering (VQA) system can provide meaningful references for both doctors and patients during the treatment process. Extracting image features is one of the most important steps in a medical VQA framework which outputs essential information to predict answers. Transfer learning, in which the pretrained deep learning models [36,9,24,13,12] that are trained on the large scale labeled dataset such as ImageNet [32], is a popular way to initialize the feature extraction process. However, due to the difference in visual concepts between ImageNet images and medical images, finetuning process is not sufficient [26]. Recently, Model Agnostic Meta-Learning [6] (MAML) has been introduced to overcome the aforementioned problem by learning meta-weights that quickly adapt to visual concepts. However, MAML is heavily impacted by the meta-annotation phase for all images in the medical dataset [26]. Different from normal images, transfer learning in medical images is more challenging due to: *(i)* noisy labels may occur when labeling images in an unsupervised manner; *(ii)* high-level semantic labels cause uncertainty during learning; and *(iii)* difficulty in scaling up the process to all unlabeled images in medical datasets.

In this paper, we introduce a new Multiple Meta-model Quantifying (MMQ) process to address these aforementioned problems in MAML. Intuitively MMQ is designed to: *(i)* effectively increase meta-data by auto-annotation; *(ii)* deal with the noisy labels in the training phase by leveraging the uncertainty of predicted scores during the meta-agnostic process; and *(iii)* output meta-models which contain robust features for down-stream medical VQA task. Note that, compared with the recent approach for meta-learning in medical VQA [26], our proposed MMQ does not take advantage of additional out-of-dataset images, while achieves superior accuracy in two challenging medical VQA datasets.

## 2 Literature Review

**Medical Visual Question Answering** Based on the development of VQA in general images, the medical VQA task inherits similar techniques and achieves certain achievements [2,18,28,1,45,19]. Specifically, the attention mechanisms such as MCB [7], SAN [43], BAN [15], or CTI [5] are applied in [28,1,45,26,41] to learn joint representation between medical visual information and questions. Additionally, in [18,45,28,17], the authors take advantage of transfer learning for extracting medical image features. Recently, approaches which directly solve different aspects of medical VQA are introduced, including reasoning [17,44], diagnose model behavior [40], multi-modal fusion [35], dedicated framework design [20,8], and generative model for dealing with abnormality questions [31].

**Meta-learning** Traditional machine learning algorithms, specifically deep learning-based approaches, require a large-scale labeled training set [21,25,3,23,4]. Therefore, meta-learning [42,34,11,14], which targets to deal with the problem of data limitation when learning new tasks, is applied broadly. There are three common approaches to meta-learning, namely model-based [33,22], metric-based [16,39,38,37], and optimization-based [30,6,27]. A notable optimization-based work, MAML [6], helps to learn a meta-model then quickly adapt it to other tasks. The authors in [26] used MAML to overcome the data limitation problem in medical VQA. However, their work required the use of external data during the training.

## 3 Methodology

### 3.1 Method overview

Our approach comprises two parts: our proposed multiple meta-model quantifying (MMQ - Figure 1) and a VQA framework for integrating meta-models outputted from MMQ (Figure 2). MMQ addresses the meta-annotation problem by outputting multiple meta-models. These models are expected to robust to each other and have high accuracy during the inference phase of model-agnostic tasks. The VQA framework aims to leverage different features extracted from candidate meta-models and then generates predicted answers.

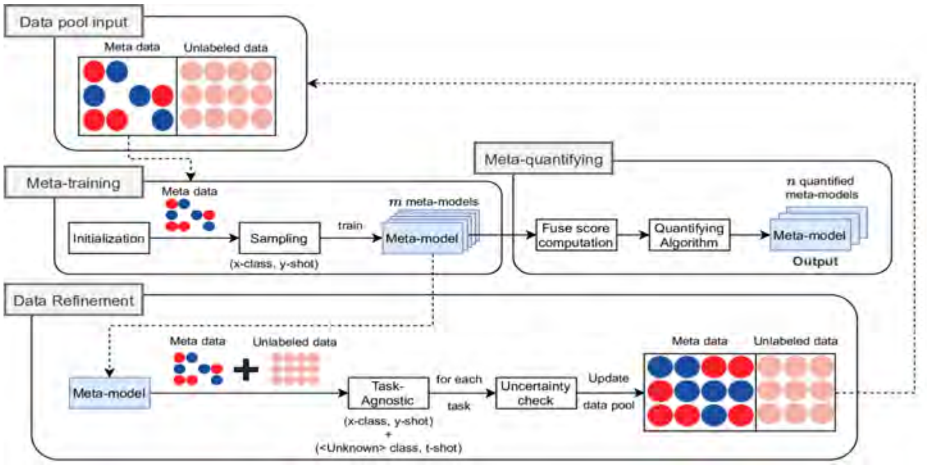

**Fig. 1.** Multiple Meta-model Quantifying in medical VQA. Dotted lines denote looping steps, the number of loop equals to $m$ required meta-models.

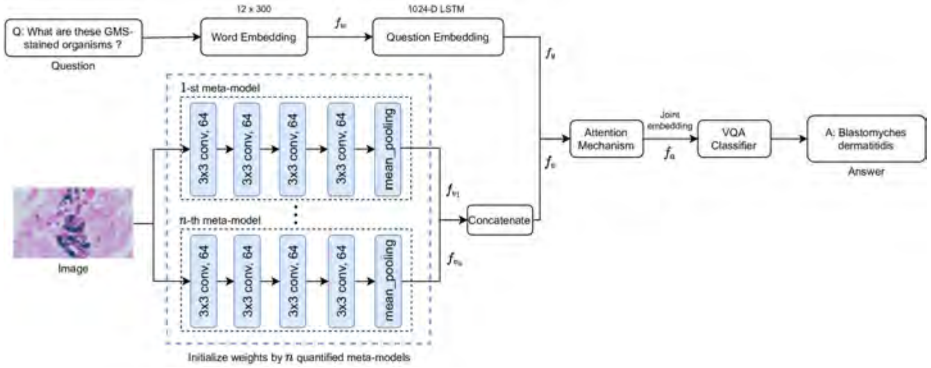

**Fig. 2.** Our VQA framework is designed to integrate robust image features extracted from multiple meta-models outputted from MMQ.

### 3.2 Multiple meta-model quantifying

Multiple meta-model quantifying (Figure 1) contains three modules: *(i)* **Meta-training** which trains a specific meta-model for extracting image features used in medical VQA task by following MAML [6]; *(ii)* **Data refinement** which increases the training data by auto-annotation and deal with the noisy label by leveraging the uncertainty of predicted scores; and *(iii)* **Meta-quantifying** which selects meta-models whose robust to each others and have high accuracy during inference phase of model-agnostic tasks.

**Algorithm 1:** Model-Agnostic for data refinement

```
Input: $\rho(\mathcal{T})$ distribution over tasks; data pool $\mathcal{D}$; meta-model weights $\theta$
Output: Updated data pool $\mathcal{D}'$
Sample batch of tasks $\mathcal{T}_i \sim \rho(\mathcal{T})$
Establish list $A$ with contains list of (Score $S$, Label $L$) of each sample in data
  pool $D$. $(S, L)$ is from the predicted process of Classifier $\mathcal{C}$ of each task $\mathcal{T}_i$.
Set $\alpha$ and $\beta$ be uncertainty checking threshold.
For all task $\mathcal{T}_i$ do
    For all image $I_k$ in $\mathcal{T}_i$ batch do
        $(S_k^i,\ L_k^i) \leftarrow \mathcal{C}_i(\mathcal{T}_i,\ \theta,\ I_k)$. Where $\mathcal{C}_i$ is the $i$-th classifier of task $\mathcal{T}_i$.
        Append $(S_k^i,\ L_k^i)$ into $A[I_k]$.
Establish new version of Meta data split $\mathcal{M}'$ and new version of Unlabeled
  data split $\mathcal{U}'$ of $\mathcal{D}$
For all element $A[I_j]$ in list $A$ do
    If $A[I_j]$ in Meta data split $\mathcal{M}$ of $\mathcal{D}$
        If $\exists A[I_j]\{S\} < \alpha$ and $A[I_j]\{L\}$ is $A[I_j]\{GT_j\}$. Where $GT_j$ is the
          ground-truth label of $I_j$.
            Append $(I_j, A[I_j]\{L\})$ into $\mathcal{U}'$
            Remove $(I_j, A[I_j]\{L\})$ from $\mathcal{M}$
    If $A[I_j]$ in Unlabeled data split $\mathcal{U}$ of $\mathcal{D}$
        If $\exists A[I_j]\{S\} > \beta$
            Append $argmax_{A[I_j]\{S\}}(I_j, A[I_j]\{L\})$ into $\mathcal{M}'$
$\mathcal{U}_f = \mathcal{U}$ - $\mathcal{M}' + \mathcal{U}'$. Where $\mathcal{U}_f$ is the updated Unlabeled data split of $\mathcal{D}'$.
$\mathcal{M}_f = \mathcal{M}$ - $\mathcal{U}' + \mathcal{M}'$. Where $\mathcal{M}_f$ is the updated Meta data split of $\mathcal{D}'$.
return $\mathcal{M}_f$,$\mathcal{U}_f$ of $\mathcal{D}'$
```

**Meta-training** We generally follow MAML [6] to do meta-training. Let $f_\theta$ be the classification meta-model. Hence, $\theta$ represents the parameters of $f_\theta$ while $\{\theta'_0, \theta'_1, ...\theta'_x\}$ is the adapting parameters list of classification models for $x$ given tasks $\mathcal{T}_i$ and their associated dataset $\{\mathcal{D}_i^{tr}, \mathcal{D}_i^{val}\}$. Specifically, for each iteration, $x$ tasks are sampled with $y$ examples of each task. Then we calculate the gradient descent $\nabla_\theta L_{\mathcal{T}_i}(f_\theta(\mathcal{D}_i^{tr}))$ of the classification loss $L_{\mathcal{T}_i}$ and update the corresponding adapting parameters as follow.

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

At the end of each iteration, the meta-model parameters $\theta$ are updated throughout validation sets of all tasks sampled to learn the generalized features as:

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

Unlike MAML [6] where only one meta-model is selected, we develop the following refinement and meta-quantifying steps to select high-quality meta-models for transfer learning to the medical VQA framework later.

**Data refinement** After finishing the meta-training phase, the weights of the meta-models are used for refining the dataset. The module aims to expand the meta-data pool for meta-training and removes samples that are expected to be hard-to-learn or have noisy labels (See Algorithm 1 for more details).

**Meta-quantifying** This module aims to identify candidate meta-models that are useful for the medical VQA task. A candidate model $\theta$ should achieve high performance during the validating process and its features distinct from other features from other candidate models.

To achieve these goals, we design a fuse score $S_F$ as described in (3).

$$S_F = \gamma S_P + (1-\gamma)\sum_{t=1}^{m} 1 - Cosine\,(F_c, F_t)\,\forall F_c \neq F_t \tag{3}$$

where $S_P$ is the predicted score of the current meta-model over ground-truth label; $F_c$ is the feature extracted from the aforementioned meta-model that needs to compute the score; $F_t$ is the feature extracted from $t$-th model of the list of meta-model $\Theta$; Cosine is using for similarity checking between two features.

Since the predicted score $S_P$ at the ground-truth label and diverse score are co-variables, therefore the fuse score $S_F$ is also covariate with both aforementioned scores. This means that the larger $S_F$ is, the higher chance of the model to be selected for the VQA task. Algorithm 2 describes our meta-quantifying algorithm in details.

### 3.3 Integrate quantified meta-models to medical VQA framework

To leverage robust features extracted from quantified meta-models, we introduce a VQA framework as in Figure 2. Specifically, each input question is trimmed to a 12-word sentence and then zero-padded if its length is less than 12. Each word is represented by a 300-D GloVe word embedding [29]. The word embedding is fed into a 1024-D LSTM to produce the question embedding $f_q$.

Each input image is passed through $n$ quantified meta-models got from the meta-quantifying module, which produce $n$ vectors. These vectors are concatenated to form an enhanced image feature, denoted as $f_v$ in Figure 2. Since this vector contains multiple features extracted from different high-performed meta-models and each model has different views, the VQA framework is expected to be less affected by the bias problem. Image feature $f_v$ and question embedding $f_q$ are fed into an attention mechanism (BAN [15] or SAN [43]) 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 [18]). To train the proposed model, we use a Cross Entropy loss for the answer classification task. The whole VQA framework is then fine-tuned in an end-to-end manner.

**Algorithm 2:** Meta-quantifying algorithm

```
Input: Data pool $\mathcal{D}_T$; list of meta-model $\Theta \in [\theta_0, \theta_1, ..., \theta_m]$ where $m$ denotes
        the number of candidate meta-models; number of quantified model $n$.
Output: List of Quantified meta-models $\Theta_n \in [\theta_0, \theta_1, ..., \theta_n]$. $n < m$.
For all $n$ meta-models, sample batch of tasks $\mathcal{T}_i \sim \rho(\mathcal{T})$
Establish list $A$ with contains list of (Score $S_P$, Feature $F$) of each sample in
  quantify data pool $\mathcal{D}_T$. $(S_P, F)$ is got from the predicted process of Classifier
  $\mathcal{C}$ of each task $\mathcal{T}_i$. $S_P$ is the predicted score at ground-truth label.
Set $\gamma$ be effectiveness - robustness balancing hyper-parameter.
Establish Fuse Score list $\mathcal{L}_{S_F}$ for all meta-model in $\Theta$.
For all task $\mathcal{T}_i$ do
    For all image $I_k$ in $\mathcal{T}_i$ batch do
        For all meta-model $\Theta_t$ in $\Theta$ do
            $(S^i_k,\ F^i_k)^{\Theta_t} \leftarrow \mathcal{C}_i(\mathcal{T}_i,\ \theta,\ I_k)$. Where $\mathcal{C}_i$ is the $i$-th classifier of task $\mathcal{T}_i$.
            Append $(S^i_k,\ F^i_k)^{\Theta_t}$ into $A[I_k]$.
        For all meta-model $\Theta_t$ in $\Theta$ do
            For all $A[I_k]$ do
                $S^{\Theta_t}_F \leftarrow$ Compute fuse score using Equation (3).
                $\mathcal{L}^{\Theta_t}_{S_F} += S^{\Theta_t}_F$.
$\mathcal{L}_{S_F} \leftarrow$ Sort $\mathcal{L}_{S_F}$ decreasingly along with corresponding $\theta$.
return $\Theta_n \leftarrow n$-first meta-models selected from $\mathcal{L}_{S_F}$.
```

# 4 Experiments

## 4.1 Dataset

We use the VQA-RAD [18] and PathVQA [10] in our experiments. The VQA-RAD [18] dataset contains 315 images and 3,515 corresponding questions. Each image is associated with more than one question. The PathVQA [10] dataset consists of 32,799 question-answer pairs generated from 1,670 pathology images collected from two pathology textbooks, and 3,328 pathology images collected from the PEIR digital library.

## 4.2 Experimental details

**Meta-training.** Similar to [26], we first create the meta-annotation for training MAML. For the VQA-RAD dataset, we re-use the meta-annotation created by [26]. Note that we do not use their extra collected data in our experiment. For the PathVQA dataset, we create the meta-annotation by categorizing all training images into 31 classes based on body parts, types of images, and organs.

For every iteration of MAML training, 5 tasks are sampled per iteration in RAD-VQA while in PathVQA, this value is 4 instead. For each task, in RAD-VQA, we randomly select 3 classes from 9 classes while in PathVQA we select 5 classes from 31 aforementioned classes. For each class, in RAD-VQA, 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. In PathVQA, the same process is applied with 20 random selected images, 5 of them are used for updating task models and the remains are used for updating meta-model.

**Data refinement.** The meta-model outputted from the meta-training step is then used for updating the data pool through the algorithm described in Section 3.2. The refined data pool is then leveraged as the input for the meta-training step to output another meta-model. This loop is applied by a maximum of 7 times to output up to 7 different meta-models.

**Meta-quantifying.** All meta-models got from the previous step are passed through the Algorithm 2 to quantify their effectiveness. A maximum of 4 models which have high performance is applied to VQA training.

**VQA training.** After selecting candidate meta-models from the meta quantifying module, we use their trained weights to initialize the image feature extraction component in the VQA framework. We then finetune the whole VQA model using the VQA training set. The output vector of each meta-model is set to 32-D in PathVQA and 64-D in VQA-RAD dataset. We use 50% of meta-annotated images for training meta-models. The effect of meta-annotated images can be found in our supplementary material.

**Baselines.** We compare our MMQ results with recent methods in medical VQA: MAML [6], MEVF [26], stacked attention network (StAN) with VGG-16 [10], and bilinear attention network (BiAN) with Faster-RCNN [10]. Two attentions methods SAN [43] and BAN [15] are used in MAML, MEVF, and MMQ. Note that, StAN and BiAN only use pretrained models from the ImageNet dataset, MEVF [26] uses extra collected data to train their meta-model, while our MMQ relies solely on the images from the dataset. For the question feature extraction, all baselines and our method use the same pretrained models (i.e., Glove [29]) and then finetuning on VQA-RAD or PathVQA dataset.

### 4.3 Results

Table 1 presents comparative results between different methods. The results show that our MMQ significantly outperforms other meta-learning methods by a large margin. Besides, the gain in performance of MMQ is stable with different attention mechanisms (BAN [15] or SAN [43]) in the VQA task. It worth noting that, compared with the most recent state-of-the-art method MEVF [26], we outperform 5.3% in free-form questions of the PathVQA dataset and 9.8% in the Open-ended questions of the VQA-RAD dataset, respectively. Moreover, no out-of-dataset images are used in MMQ for learning meta-models. The results imply that our proposed MMQ learns essential representative information from the input images and leverage effectively the features from meta-models to deal with challenging questions in medical VQA datasets.

### 4.4 Ablation study

Table 2 presents our MMQ accuracy in PathVQA dataset when applying $m$ times refining data and $n$ quantified meta-models. The results show that, by

**Table 1.** Performance comparison on VQA-RAD and Path-VQA test set. (*) indicates methods used pre-trained model on ImageNet dataset. We refine data 5 times ($m = 5$) and use 3 meta-models ($n = 3$) in our MMQ.

| **Reference Methods** | **Attention Method** | **PathVQA** | | | **VQA-RAD** | | |
|---|---|---|---|---|---|---|---|
| | | ***Free-form*** | ***Yes/ No*** | ***Over-all*** | ***Open-ended*** | ***Close-ended*** | ***Over-all*** |
| StAN [10](*) | SAN | 1.6 | 59.4 | 30.5 | 24.2 | 57.2 | 44.2 |
| BiAN [10](*) | BAN | 2.9 | 68.2 | 35.6 | 28.4 | 67.9 | 52.3 |
| MAML [6] | SAN | 5.4 | 75.3 | 40.5 | 38.2 | 69.7 | 57.1 |
| | BAN | 5.9 | 79.5 | 42.9 | 40.1 | 72.4 | 59.6 |
| MEVF [26] | SAN | 6.0 | 81.0 | 43.6 | 40.7 | 74.1 | 60.7 |
| | BAN | 8.1 | 81.4 | 44.8 | 43.9 | 75.1 | 62.7 |
| **MMQ (ours)** | SAN | 11.2 | 82.7 | 47.1 | 46.3 | 75.7 | 64.0 |
| | BAN | **13.4** | **84.0** | **48.8** | **53.7** | **75.8** | **67.0** |

using only 1 quantified meta-model outputted from our MMQ, we significantly outperform both MAML and MEVF baselines. This confirms the effectiveness of the proposed MMQ for dealing with the limitation of meta-annotation in medical VQA, i.e., noisy labels and scalability. Besides, leveraging more quantified meta-models also further improves the overall performance.

We note that the improvements of our MMQ are more significant on free-form questions over yes/no questions. This observation implies that the free-form questions/answers which are more challenging and need more information from input images benefits more from our proposed method.

Table 2 also shows that increasing the number of refinement steps and the number of quantified meta-models can improve the overall result, but the gain is smaller after each loop. The training time also increases when the number of meta-models is set higher. However, our testing time and the total number of parameters are only slightly higher than MAML [6] and MEVF [26]. Based on the empirical results, we recommend applying 5 times refinement with a maximum of 3 quantified meta-models to balance the trade-off between the accuracy performance and the computational cost.

## 5 Conclusion

In this paper, we proposed a new multiple meta-model quantifying method to effectively leverage meta-annotation and deal with noisy labels in the medical VQA task. The extensively experimental results show that our proposed method outperforms the recent state-of-the-art meta-learning based methods by a large margin in both PathVQA and VQA-RAD datasets. Our implementation and trained models will be released for reproducibility.

**Table 2.** The effectiveness of our MMQ under $m$ times refining data and $n$ quantified meta-models on PathVQA test set. BAN is used as the attention method.

| **Methods** | **m** | **n** | **Free-form** | **Yes/No** | **Overall** | **Train time (hours)** | **Test time (s/sample)** | **#Paras (M)** |
|---|---|---|---|---|---|---|---|---|
| MAML [6] | - | - | 5.9 | 79.5 | 42.9 | 2.1 | 0.007 | 27.2 |
| MEVF [26] | - | - | 8.1 | 81.4 | 44.8 | 2.5 | 0.008 | 27.9 |
| **MMQ (ours)** | 3 | 1 | 10.1 | 82.1 | 46.2 | 5.8 | 0.008 | 27.8 |
| | 4 | 2 | 12.0 | 83.0 | 47.6 | 7.3 | 0.009 | 28.1 |
| | 5 | 3 | 13.4 | 84.0 | 48.8 | 8.9 | 0.010 | 28.3 |
| | 7 | 4 | 13.6 | 84.0 | 48.8 | 12.1 | 0.011 | 28.5 |

## 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. Abacha, A.B., Hasan, S.A., Datla, V.V., Liu, J., Demner-Fushman, D., Müller, H.: Vqa-med: Overview of the medical visual question answering task at imageclef 2019. In: CLEF (Working Notes) (2019)
3. 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)
4. Chi, W., Dagnino, G., Kwok, T.M., Nguyen, A., Kundrat, D., Abdelaziz, E., Riga, C., Bicknell, C., Yang, G.Z.: Collaborative robot-assisted endovascular catheterization with generative adversarial imitation learning. In: ICRA (2020)
5. Do, T., Do, T.T., Tran, H., Tjiputra, E., Tran, Q.D.: Compact trilinear interaction for visual question answering. In: ICCV (2019)
6. Finn, C., Abbeel, P., Levine, S.: Model-agnostic meta-learning for fast adaptation of deep networks. In: ICML (2017)
7. 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: EMNLP (2016)
8. Gupta, D., Suman, S., Ekbal, A.: Hierarchical deep multi-modal network for medical visual question answering. Expert Systems with Applications (2021)
9. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR (2016)
10. He, X., Zhang, Y., Mou, L., Xing, E., Xie, P.: Pathvqa: 30000+ questions for medical visual question answering. arXiv preprint arXiv:2003.10286 (2020)
11. Hsu, K., Levine, S., Finn, C.: Unsupervised learning via meta-learning. In: ICLR (2019)
12. Huang, B., Tsai, Y.Y., Cartucho, J., Vyas, K., Tuch, D., Giannarou, S., Elson, D.S.: Tracking and visualization of the sensing area for a tethered laparoscopic gamma probe. International Journal of Computer Assisted Radiology and Surgery (2020)
13. Huang, B., Zheng, J.Q., Nguyen, A., Tuch, D., Vyas, K., Giannarou, S., Elson, D.: Self-supervised generative adversarial network for depth estimation in laparoscopic images. In: MICCAI (2021)

14. Khodadadeh, S., B¨ol¨oni, L., Shah, M.: Unsupervised meta-learning for few-shot image classification. In: NIPS (2019)
15. Kim, J.H., Jun, J., Zhang, B.T.: Bilinear attention networks. In: NIPS (2018)
16. Koch, G., Zemel, R., Salakhutdinov, R.: Siamese neural networks for one-shot image recognition. In: ICML Deep Learning Workshop (2015)
17. Kornuta, T., Rajan, D., Shivade, C., Asseman, A., Ozcan, A.S.: Leveraging medical visual question answering with supporting facts. arXiv:1905.12008 (2019)
18. 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)
19. Liu, S., Ding, H., Zhou, X.: Shengyan at vqa-med 2020: An encoder-decoder model for medical domain visual question answering task. CLEF (2020)
20. Lubna, A., Kalady, S., Lijiya, A.: Mobvqa: A modality based medical image visual question answering system. In: TENCON (2019)
21. Maicas, G., Bradley, A.P., Nascimento, J.C., Reid, I., Carneiro, G.: Training medical image analysis systems like radiologists. In: MICCAI (2018)
22. Munkhdalai, T., Yu, H.: Meta networks. In: ICML (2017)
23. Nguyen, A.: Scene understanding for autonomous manipulation with deep learning. arXiv preprint arXiv:1903.09761 (2019)
24. Nguyen, A., Kundrat, D., Dagnino, G., Chi, W., Abdelaziz, E., Guo, Y., Ma, Y., Kwok, T., Riga, C., Yang, G.Z.: End-to-end real-time catheter segmentation with optical flow-guided warping during endovascular intervention. In: ICRA (2020)
25. Nguyen, A., Nguyen, N., Tran, K., Tjiputra, E., Tran, Q.: Autonomous navigation in complex environments with deep multimodal fusion network. In: IROS (2020)
26. Nguyen, B.D., Do, T.T., Nguyen, B.X., Do, T., Tjiputra, E., Tran, Q.D.: Overcoming data limitation in medical visual question answering. In: MICCAI (2019)
27. Nichol, A., Achiam, J., Schulman, J.: On first-order meta-learning algorithms. arXiv preprint arXiv:1803.02999 (2018)
28. Peng, Y., Liu, F., Rosen, M.P.: Umass at imageclef medical visual question answering (med-vqa) 2018 task. CEUR Workshop Proceedings (2018)
29. Pennington, J., Socher, R., Manning, C.D.: Glove: Global vectors for word representation. In: EMNLP (2014)
30. Ravi, S., Larochelle, H.: Optimization as a model for few-shot learning. In: ICLR (2017)
31. Ren, F., Zhou, Y.: Cgmvqa: a new classification and generative model for medical visual question answering. IEEE Access (2020)
32. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al.: Imagenet large scale visual recognition challenge. IJCV (2015)
33. Santoro, A., Bartunov, S., Botvinick, M., Wierstra, D., Lillicrap, T.: Meta-learning with memory-augmented neural networks. In: ICML (2016)
34. Schmidhuber, J.: Evolutionary Principles in Self-referential Learning. (1987)
35. Shi, L., Liu, F., Rosen, M.P.: Deep multimodal learning for medical visual question answering. In: CLEF (Working Notes) (2019)
36. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: ICLR (2015)
37. Snell, J., Swersky, K., Zemel, R.S.: Prototypical networks for few-shot learning. In: NIPS (2017)
38. Sung, F., Yang, Y., Zhang, L., Xiang, T., Torr, P.H., Hospedales, T.M.: Learning to compare: Relation network for few-shot learning. In: CVPR (2018)
39. Vinyals, O., Blundell, C., Lillicrap, T., Kavukcuoglu, K., Wierstra, D.: Matching networks for one shot learning. In: NIPS (2016)

40. Vu, M.H., Löfstedt, T., Nyholm, T., Sznitman, R.: A question-centric model for visual question answering in medical imaging. IEEE TMI (2020)
41. Vu, M., Sznitman, R., Nyholm, T., Löfstedt, T.: Ensemble of streamlined bilinear visual question answering models for the imageclef 2019 challenge in the medical domain. In: Conference and Labs of the Evaluation Forum (2019)
42. Wang, Y.X., Hebert, M.: Learning from small sample sets by combining unsupervised meta-training with cnns. In: NIPS (2016)
43. Yang, Z., He, X., Gao, J., Deng, L., Smola, A.J.: Stacked attention networks for image question answering. In: CVPR (2016)
44. Zhan, L.M., Liu, B., Fan, L., Chen, J., Wu, X.M.: Medical visual question answering via conditional reasoning. In: ACM International Conference on Multimedia (2020)
45. Zhou, Y., Kang, X., Ren, F.: Employing Inception-Resnet-v2 and Bi-LSTM for medical domain visual question answering. CEUR Workshop Proceedings (2018)

# Supplementary material: Multiple Meta-model Quantifying for Medical Visual Question Answering

Tuong Do[1], Binh X. Nguyen[1], Erman Tjiputra[1], Minh Tran[1],
Quang D. Tran[1], and Anh Nguyen[2]

[1] AIOZ, Singapore
{tuong.khanh-long.do,binh.xuan.nguyen,erman.tjiputra,
minh.quang.tran,quang.tran}@aioz.io
[2] University of Liverpool, UK
anh.nguyen@liverpool.ac.uk

**Abstract.** In this supplementary material, we provide further analysis of MMQ to verify its effectiveness for the medical VQA task. In particular, we illustrate the network structure to extract features from meta-models. We clarify the general setup and the details of the meta-annotation step in the PathVQA dataset. We also analyze the effect of different amounts of meta-annotated images for training meta-models. The experiments are conducted using the PathVQA and VQA-RAD datasets.

## 1 Network to extract features from meta-models

Figure 1 shows the network to extract features from the meta-model in our VQA framework. It consists of four $3 \times 3$ 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.

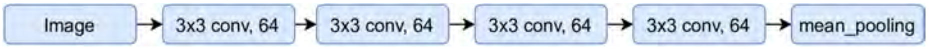

**Fig. 1.** Feature extraction network from meta-models.

## 2 General setup

The image size is set at $84 \times 84$. The proposed MMQ is implemented using PyTorch. The experiments are conducted on a single NVIDIA 1080Ti with 11GB RAM. In all MMQ experiment setups, $\alpha, \beta$, and $\gamma$ are equalled to $0.01, 0.001$, and $0.5$, respectively. There is no fine-tuning process for these hyper-parameters since we use the default $\alpha, \beta$ values in MAML plus balanced initial $\gamma$ for training and still achieve good results.

## 3 Meta-annotation details for PathVQA dataset

For the PathVQA dataset, we create the meta-annotation by manually categorizing all training images into 31 classes based on body parts, types of images, and organs. These classes are: *dense cell*, *drawing*, *process tree*, *x-ray mouth*, *x-ray ribs*, *RGB bone*, *RGB brain*, *RGB endocrine*, *RGB heart*, *RGB intestine*, *RGB kidney*, *RGB liver*, *RGB lung*, *RGB mouth*, *RGB skull*, *RGB uterus*, *RBG arms*, *RGB baby*, *RGB head*, *RGB skin*, *sparse cell*, *chart*, *x-ray legs*, *RGB prostate*, *RGB spleen*, *RBG body*, *RGB legs*, *x-ray arms*, *RBG oral*, *RGB pancreas*, *RGB penis*.

These meta-labels, our source code and trained models will be release for reproducibility.

## 4 Meta-data vs. Unlabeled data in MMQ

Table 1 illustrates the performance of MMQ with different amounts of meta-annotated images for training meta models. Since our data refinement module (See Algorithm 1 in our paper) expands current meta-data as well as removes uncertainty samples simultaneously, keeping the balance in the number of samples between meta-data and unlabeled data at the initial step is worthy. Empirical results also imply that the initial data balance between two data pools greatly increases the effectiveness of our proposed MMQ.

**Table 1.** Performance (%) comparison on VQA-RAD and Path-VQA test set when using MMQ with different amount of meta-annotated images for training meta-models. These results reported after 5 times refining data and 3 quantified meta-models are picked up.

| **% annotated data** | **Attention Mechanism** | **PathVQA** | | | **VQA-RAD** | | |
|---|---|---|---|---|---|---|---|
| | | ***Free-form*** | ***Yes/ No*** | ***Over-all*** | ***Open-ended*** | ***Close-ended*** | ***Over-all*** |
| **25 %** | SAN | 7.2 | 82.8 | 45.1 | 44.7 | 72.4 | 61.4 |
| | BAN | 9.8 | 83.1 | 46.5 | 48.8 | 74.6 | 64.3 |
| **50%** | SAN | 11.2 | 82.7 | 47.1 | 46.3 | 75.7 | 64.0 |
| | BAN | 13.4 | 84.0 | 48.8 | 53.7 | 75.8 | 67.0 |
| **75%** | SAN | 10.1 | 82.7 | 46.5 | 46.3 | 74.6 | 63.3 |
| | BAN | 12.9 | 83.3 | 48.2 | 48.8 | 78.8 | 66.2 |
| **100%** | SAN | 10.6 | 82.8 | 46.8 | 47.2 | 74.6 | 63.6 |
| | BAN | 13.6 | 83.2 | 48.5 | 48 | 78.9 | 66.6 |

## 5 Visualization for data refinement

For visualization, Figure 2 shows the images removed from the meta-annotated dataset due to their high uncertainty score. Additionally, Figure 3 illustrates the images and their labels which are annotated automatically by passing unlabeled image data through the 1-st refinement step. These illustrations indicate that our MMQ successfully extend dataset by labelling meta-data automatically as well as remove samples with noisy labels.

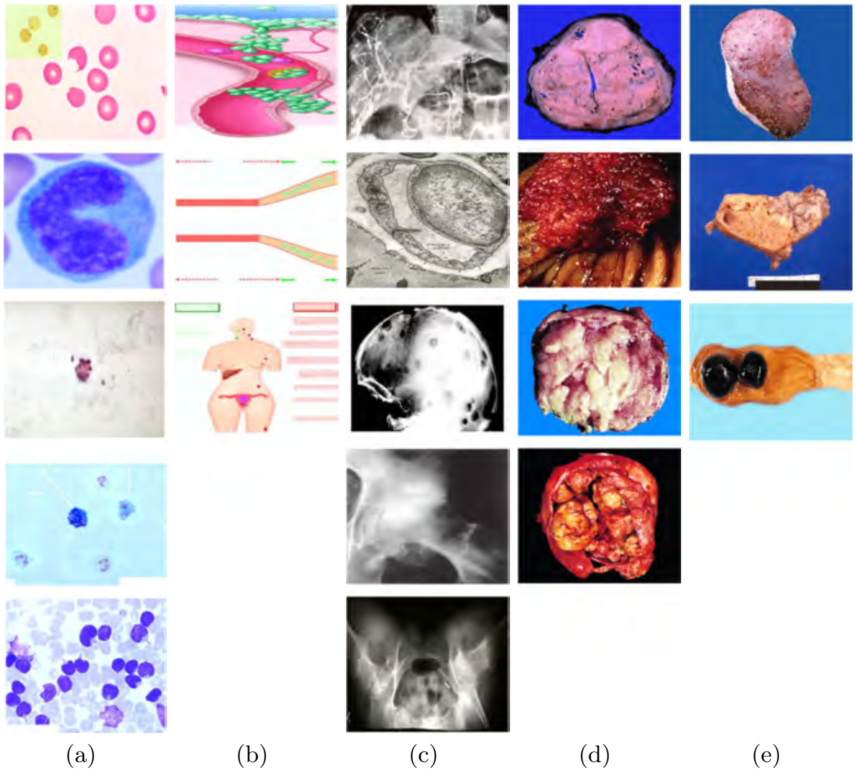

**Fig. 2.** The visualization images from meta-annotated data which are removed during the first refinement step, i.e., caused by their high uncertainty scores . Their labels are: **(a)** Dense Cell, **(b)** Process Tree, **(c)** X-ray Mouth, **(d)** RBG Brain, and **(e)** RBG Kidney, consequently. Best viewed in color.

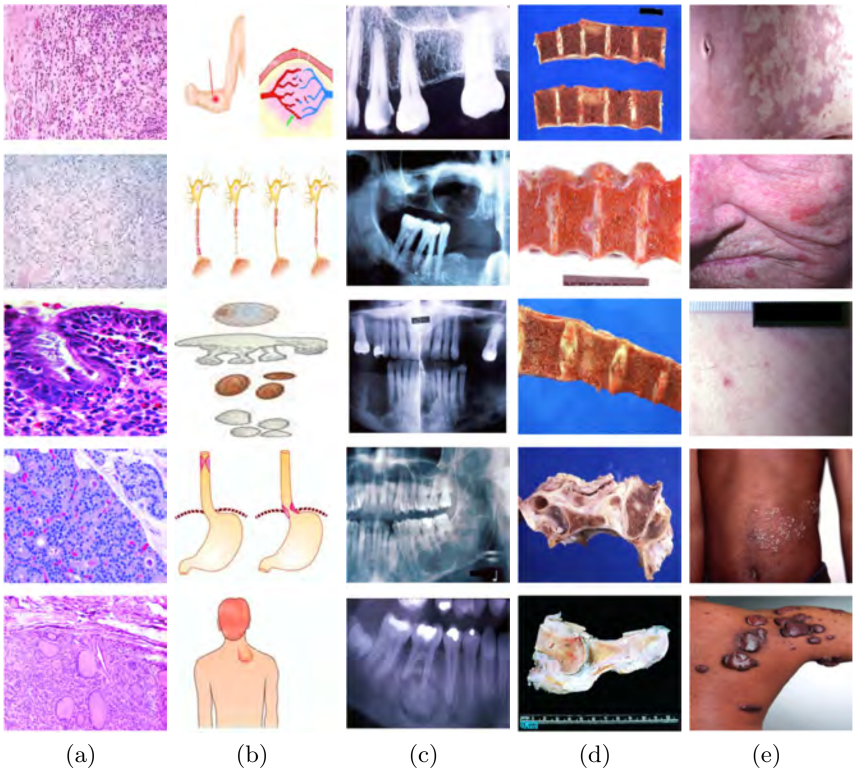

**Fig. 3.** The visualization images from unlabeled data which are chosen to add into meta-annotated data during the first refinement step. Their labels are: **(a)** Dense Cell, **(b)** Drawing, **(c)** X-ray Mouth, **(d)** RBG Bone, and **(e)** RBG Skin, consequently. Best viewed in color.