PATHOLOGICAL VISUAL QUESTION ANSWERING

Xuehai He*
University of California San Diego
x5he@ucsd.edu

Zhuo Cai*
Tsinghua University
caiz17@mails.tsinghua.edu.cn

Wenlan Wei
Wuhan University
wwl999@whu.edu.cn

Yichen Zhang
University of California San Diego
yiz037@eng.ucsd.edu

Luntian Mou
Beijing University of Technology
ltmou@pku.edu.cn

Eric Xing
Carnegie Mellon University
epxing@cs.cmu.edu

Pengtao Xie
University of California San Diego
pengtaoxie2008@gmail.com

ABSTRACT

Is it possible to develop an "AI Pathologist" to pass the board-certified examination of the American Board of Pathology (ABP)? To build such a system, three challenges need to be addressed. First, we need to create a visual question answering (VQA) dataset where the AI agent is presented with a pathology image together with a question and is asked to give the correct answer. Due to privacy concerns, pathology images are usually not publicly available. Besides, only well-trained pathologists can understand pathology images, but they barely have time to help create datasets for AI research. The second challenge is: due to the fact that it is difficult to hire highly experienced pathologists to create pathology visual questions and answers, the resulting pathology VQA dataset may contain errors such as some questions may not be relevant to the image or the answers are not given correctly. Training pathology VQA models using these noisy or even erroneous data will lead to problematic models that cannot generalize well on unseen images. The third challenge is: the medical concepts and knowledge covered in pathology question-answer (QA) pairs are very diverse while the number of QA pairs available for modeling training is limited. How to learn effective representations of diverse medical concepts based on limited data is technically demanding. In this paper, we aim to address these three challenges. To our best knowledge, our work represents the first one addressing the pathology VQA problem. To deal with the issue that a publicly available pathology VQA dataset is lacking, we create PathVQA, a VQA dataset with 32,795 questions asked from 4,998 pathology images. The questions in PathVQA are similar to those in the ABP tests. To our best knowledge, this is the first dataset for pathology VQA. To address the second challenge, we propose a learning-by-ignoring approach which automatically identifies training examples that have bad-quality and remove them from the training dataset. To address the third challenge, we propose to use cross-modal self-supervised learning to learn powerful visual and textual representations jointly. We perform experiments on our created PathVQA dataset and the results demonstrate the effectiveness of our proposed learning-by-ignoring method and cross-modal self-supervised learning methods.

*equal contribution

1 INTRODUCTION

Pathology studies the causes and effects of diseases or injuries. It underpins every aspect of patient care, such as diagnostic testing, providing treatment advice, preventing diseases using cutting-edge genetic technologies, to name a few. Medical professionals practicing pathology are called pathologists, who examine bodies and body tissues. To become a board-certificated pathologist in the US, a medical professional needs to pass a certification examination organized by the American Board of Pathology (ABP), which is a very challenging task. We are interested in asking: whether an artificial intelligence (AI) system can be developed to pass the ABP examination? It is an important step towards achieving AI-aided clinical decision support and clinical education. Among the ABP test questions, one major type is to understand the pathology images. Given a pathology image and a question, the examinees are asked to select a correct answer. Such a problem is called visual question answering (VQA) (Antol et al., 2015) in the AI community. VQA is an interdisciplinary research problem that has drawn extensive attention recently. Given an image (e.g., an image showing a dog is chasing a ball) and a question asked about the visual content of the image (e.g., "what is the dog chasing?"), VQA aims to develop AI algorithms to infer the correct answer (e.g., "ball"). VQA requires a deep comprehension of both images and textual questions, as well as the relationship between visual objects and textual entities, which is technically very demanding.

To train an AI system to perform VQA on pathology images and pass the ABP test, we first need to collect a dataset containing questions similar to those in the ABP test. ABP provides some sample questions, but they are too few to be useful for training data-driven models. Some commercial institutes provide a larger number of practice questions, but they are very expensive to buy and they cannot be shared with the public due to copyright issues. One possible way to create pathology VQA dataset is to leverage crowdsourcing, which is used successfully for creating general domain VQA datasets (Malinowski & Fritz, 2014; Antol et al., 2015; Ren et al., 2015a; Johnson et al., 2017; Goyal et al., 2017). However, it is much more challenging to build medical VQA datasets than general domain VQA datasets via crowdsourcing. First, medical images such as pathology images are highly domain-specific, which can only be interpreted by well-educated medical professionals. It is very difficult and expensive to hire medical professionals to help create medical VQA datasets. Second, to create a VQA dataset, one first needs to collect an image dataset. While images in the general domain are pervasive, medical images are very difficult to obtain due to privacy concerns.

To address these challenges, we resort to pathology textbooks, especially those that are freely accessible online, as well as online digital libraries. These textbooks contain a lot of pathology images, covering the entire domain of pathology. Each image has a caption describing pathological findings in the image. The caption is carefully worded and clinically precise. We extract images and captions from the textbooks and online digital libraries. Given these images, question-answer pairs are created based on image captions. These QA pairs are verified by medical professionals to ensure clinical meaningfulness and correctness. In the end, we create a pathology VQA dataset called PathVQA, which contains 32,795 questions asked from 4,998 pathology images. To our best knowledge, this is the first dataset for pathology VQA.

Given the pathology VQA dataset, the next step is to develop a pathology VQA system, which is also very challenging, due to the following reasons. First, while we have tried our best to ensure the clinical correctness of the PathVQA dataset, it may still contain noises and errors that can only be identified by very experienced pathologists who unfortunately do not have time to do so for all the data examples in PathVQA. To address this problem, we propose a learning-by-ignoring method which can automatically identify bad-quality data (errors, noises, outliers, etc.) and remove them from the training set. The learning-by-ignoring strategy analyzes the collection of training examples holistically and determines which ones should be ignored. The likelihood of ignoring each training example is learned by maximizing the performance on the validation set in a bi-level optimization framework. The second challenge is: the medical concepts involved in PathVQA are very diverse while the number of question-answer pairs available for training is limited. Learning effective representations of these diverse medical concepts using limited data is technically difficult. Poorly learned representations lead to inferior VQA performance. To address the second challenge, we propose cross-modal self-supervised learning approaches to pretrain the representation learning modules in VQA models for obtaining effective visual and textual embeddings. Self-supervised learning (SSL) (Gidaris et al., 2018b; Zhang et al., 2016a; Pathak et al., 2016b) is an unsupervised learning approach which creates auxiliary tasks on input data without using human-provided labels and learns data representations by solving these auxiliary tasks. We create two types of cross-modal

SSL tasks: 1) given an image and a question, judge whether this question is asked from this image; 2) given an image and an answer, judge whether this answer is relevant to this image. We also conduct a single-modal SSL on question-answer pairs: we pretrain the text encoder by predicting answers only based on the questions without considering the input images. Experiments on our developed PathVQA dataset demonstrates the effectiveness of our proposed methods.

The major contributions of this paper are as follows:

- We create a pathology visual question answering dataset – PathVQA, to foster the research of medical VQA. To our best knowledge, this is the first dataset for pathology VQA.
- We propose a learning-by-ignoring approach which automatically identifies problematic training examples and removes them from the training set. Our method performs data ignoring by maximizing the validation performance end-to-end.
- We propose cross-modal self-supervised learning (SSL) approaches to learn better image encoders and text encoders in VQA models. Three SSL strategies are studied, including 1) predicting whether an image and a question match, 2) predicting whether an image and an answer match, and 3) predicting answers solely based on questions.
- On our PathVQA dataset, we demonstrate the effectiveness of our proposed learning-by-ignoring and cross-modal SSL methods in detecting noisy training examples and learning powerful visual-textual representations.

2 Related Works

2.1 Medical VQA Datasets

To our best knowledge, there are two existing datasets for medical visual question answering. The VQA-Med (Abacha et al., 2019) dataset is created on 4,200 radiology images and has 15,292 question-answer pairs. Most of the questions are in multiple-choice (MC) style and can be answered by multi-way classifiers. This makes the difficulty of this dataset significantly lower. VQA-RAD (Lau et al., 2018) is a manually-crafted dataset where questions and answers are given by clinicians on radiology images. It has 3515 questions of 11 types. Our dataset differs from VQA-Med and VQA-RAD in two-fold. First, ours is about pathology while VQA-Med and VQA-RAD (Lau et al., 2018) are both about radiology. Second, our dataset is a truly challenging QA dataset where most of the questions are open-ended while in VQA-Med and VQA-RAD the majority of questions have a fixed number of candidate answers and can be answered by multi-way classification. Besides, the number of questions in our dataset is much larger than that in VQA-Med and VQA-RAD.

2.2 Self-supervised Learning

Self-supervised learning (SSL) has been widely studied to learn better representations of images and texts. SSL learns useful features automatically by constructing a loss from a pretext task without much demand for human annotations. It purely uses the input data to create auxiliary tasks and enables deep networks to learn effective latent features by solving these auxiliary tasks. Various strategies have been proposed to construct auxiliary tasks, based on temporal correspondence (Li et al., 2019b; Wang et al., 2019a), cross-modal consistency (Wang et al., 2019b), etc. In computer vision, examples of auxiliary tasks include rotation prediction (Gidaris et al., 2018a), image inpainting (Pathak et al., 2016a), automatic colorization (Zhang et al., 2016b), instance discrimination (Wu et al., 2018), to name a few. In SSL for natural language processing, examples of auxiliary tasks include next-word prediction in the GPT model (Radford et al., 2019), next sentence prediction, masked word prediction in the BERT model (Devlin et al., 2018), and so on.

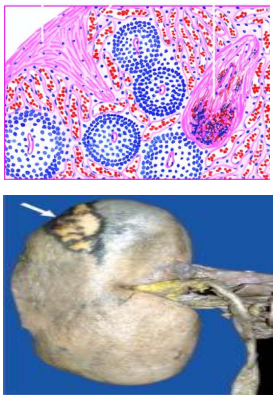


Figure 1: Two exemplar images with generated questions. Both images have three types of questions: "what", "where", and "yes/no".

Cross-modal self-supervised learning has been studied as well, which learns representations for data with multiple modalities by solving cross-modal auxiliary tasks. VisualBERT (Li et al., 2019a) learns representations for images and texts by implicitly aligning elements of an input text and regions in an associated input image with self-attention. Two visually-grounded language model objectives are used for pretraining VisualBERT on image caption data.

VideoBERT (Sun et al., 2019) performs vector quantization on video frames to get visual tokens, and then trains masked language models on the concatenation of visual tokens and text tokens. Chung et al. (2020) proposes to learn cross-modal joint embeddings using self-supervised learning for cross-modal retrieval. CVLP (Shi et al., 2020) proposes an unbiased contrastive visual-linguistic pretraining approach, which constructs a visual self-supervised loss based on contrastive learning. LXMERT (Tan & Bansal, 2019) designs five pretraining tasks: masked language modeling, feature regression, label classification, cross-modal matching, and image question answering, to pretrain a large Transformer model. ViLBERT (Lu et al., 2019) proposes to pretrain a vision-and-language BERT model through masked multi-modal modeling and multi-modal alignment prediction tasks and then transfer the model to visual question answering tasks.

2.3 Data selection and data reweighting

A number of approaches have been proposed for data selection. Matrix column subset selection (Deshpande & Rademacher, 2010; Boutsidis et al., 2009) aims to select a subset of data examples that can best reconstruct the entire dataset. Similarly, coreset selection (Bachem et al., 2017) chooses representative training examples in a way that models trained on the selected examples have comparable performance with those trained on all training examples. These methods perform data selection and model training separately. As a result, the validation performance of the model cannot be used to guide data selection. Ren et al. (2018) propose a meta learning method to learn the weights of training examples by performing a meta gradient descent step on the weights of the current mini-batch of examples. Shu et al. (2019) propose a method which can adaptively learn an explicit weighting function directly from data. Different from these works, our learning-by-ignoring method is based on a bi-level optimization framework which can flexibly select data elements with various granularity, such as pixels, images, bags of instances, etc., in a unified way.

Table 1: Statistics of the PathVQA dataset

	Max	Avg	Min
# questions per image	14	6.6	1
# words per question	28	9.5	3
# words per answer	10	2.5	1

3 The PathVQA Dataset

The PathVQA dataset consists of 32,795 question-answer pairs generated from 1,670 pathology images collected from two pathology textbooks: "Textbook of Pathology" (Muir et al., 1941) and "Basic Pathology" (Robbins et al., 1981), and 3,328 pathology images collected from the PEIR[1] digital library. Figure 1 shows some examples. On average, each image has 6.6 questions. The maximum and minimum number of questions for a single image is 14 and 1 respectively. The average number of words per question and per answer is 9.5 and 2.5 respectively. Table 1 summarizes these statistics. There are eight different categories of questions: what, where, when, whose, how, why, how much/how many, and yes/no. Table 2 shows the number of questions and percentage of each category. The questions in the first 7 categories are open-ended: 16,466 in total and accounting for 50.2% of all questions. The rest are close-ended "yes/no" questions. The questions cover various aspects of visual contents, including color, location, appearance, shape, etc. Such clinical diversity poses great challenges for AI models to solve this pathology VQA problem.

Table 2: Frequency of questions in different categories

Question type	Total number and percentage
Yes/No	16,329 (49.8%)
What	13,401 (40.9%)
Where	2,157 (6.6%)
How	595 (1.8%)
How much/many	139 (0.4%)
Why	114 (0.3%)
When	51 (0.2%)
Whose	9 (0.1%)

[1] http://peir.path.uab.edu/library/index.php?/category/2

4 METHODS

In this section, we propose a learning-by-ignoring approach for automatically identifying and removing problematic training examples to avoid distorting the model by these bad-quality examples. We also propose several cross-modal self-supervised learning methods to learn effective visual and textual representations. These proposed methods can be applied to any VQA method. In this work, we choose two well-established and state-of-the-art VQA methods to perform the study while noting that other VQA methods are applicable as well.

4.1 LEARNING TO IGNORE

To automatically identify and remove bad-quality examples from the training data to avoid distorting the model by them, we propose a learning-by-ignoring (LBI) approach, where a data example is taken as the input and a corresponding ignoring variable $a \in [0, 1]$ is learned to indicate how likely this example should be ignored. For the loss L defined on each training example, we multiply it with the ignoring variable. If a is close to zero, then L is close to zero and this data example does not contribute to model training. We learn these ignoring variables using the following formulation:

$$\begin{aligned} \min_A \quad & \sum_{i=1}^{N^{(\text{val})}} L(d_i^{(\text{val})}; W^*(A)) \\ \text{s.t.} \quad & W^*(A) = \text{argmin}_W \sum_{i=1}^{N^{(\text{tr})}} a_i L(d_i^{(\text{tr})}; W) \end{aligned} \tag{1}$$

where $A = \{a_i\}_{i=1}^{N^{(\text{tr})}}$. W denotes the weights of the VQA model. $L(d_i^{(\text{tr})}; W)$ is the training loss defined on the training example $d_i^{(\text{tr})}$. $a_i \in [0, 1]$ is an ignoring variable indicating how likely $d_i^{(\text{tr})}$ should be ignored. Given the weighted training loss $\sum_{i=1}^{N^{(\text{tr})}} a_i L(d_i^{(\text{tr})}; W)$, we learn the VQA model weights W by minimizing this loss and get the optimal weights W^*. Note that W^* is a function of A. When A changes, the ignoring variables changes and the weighted training loss changes. The optimal model trained by minimizing the weighted training loss changes accordingly. Given the trained VQA model $W^*(A)$, we measure its loss on the validation dataset $\sum_{i=1}^{N^{(\text{val})}} L(d_i^{(\text{val})}; W^*(A))$. We assume all validation examples are double-checked by humans and have good quality. The validation loss is a function of A. We learn A by minimizing this validation loss, i.e., finding the optimal ignoring variables to remove bad-quality training examples so that the model trained on the remaining good-quality examples achieves the best performance on the validation set. Figure 2 illustrates the idea. In our PathVQA dataset, the number of training data examples is not very large (tens of thousands), we can directly learn an ignoring variable for each data example. In other applications, if there are millions of training examples, learning millions of ignoring variables may not be a good choice. Under such circumstances, we can use a neural network (called ignoring network) to parameterize the ignoring variable, where the input of the network is a feature representation of the data example and the output of the network is an ignoring variable. The ignoring network and the VQA model can share the same encoder used for representation learning.

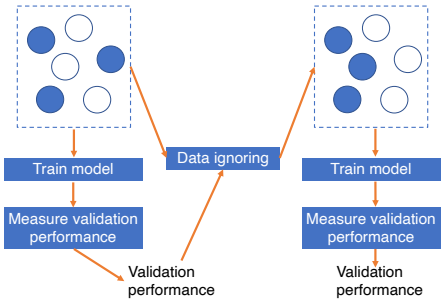


Figure 2: Learning by ignoring.

Algorithm 1 Algorithm for learning-by-ignoring

while not converged **do**
1. Update ignoring variables A by descending $\nabla_A L_{val}\left(W - \xi \nabla_W L_{train}(W, A)\right)$
2. Update weights W by descending $\nabla_W L_{train}(W, A)$

end while

The algorithm of learning-by-ignoring is shown in Algorithm 1. Similar to Liu et al. (2018), we approximate $W^*(A)$ using one step of gradient descent update of W: $W^*(A) = W - \xi \nabla_W L_{train}(W, A)$ where $L_{train}(W, A) = \sum_{i=1}^{N^{(\text{tr})}} a_i L(d_i^{(\text{tr})}; W)$. Then we plug this approximation into the validation loss: $L_{val}\left(W - \xi \nabla_W L_{train}(W, A)\right)$, and update A by performing gradient descent on the approximated validation loss. The update of W and A are performed alternatively until convergence.

4.2 SELF-SUPERVISED LEARNING ON PATHVQA

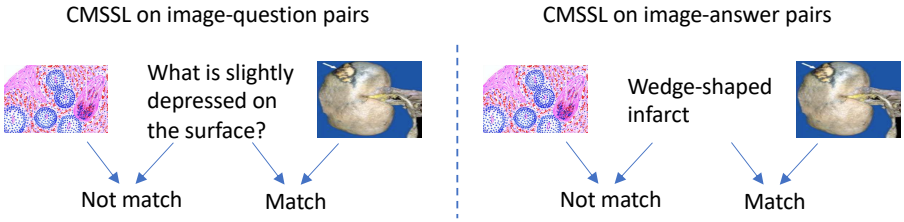


Figure 3: Cross-modal self-supervised learning.

To learn powerful visual and textual representations on limited data, we develop cross-modal self-supervised learning (CMSSL) approaches. Given a pair of data (x, y), where x and y are from different modalities and could be an image, a question, or an answer, we define a CMSSL task on (x, y) which judges whether x and y are from the same training example. We learn the image encoder and text encoder by solving this task. Figure 3 illustrates the idea. We define this task on two types of pairs: image-question and image-answer. For image-question CMSSL, given an image and a question, if the question is asked based on the content of the image, then they are labeled as a match. Otherwise, they are labeled as not a match. The image encoder and question encoder are learned jointly to predict whether an image and a question match with each other. Such a task can help to learn the correspondence between image regions and words in the question. The encoders learned in CMSSL are used to initialize the encoders in VQA models, which are then finetuned on the PathVQA dataset. We also perform CMSSL on image-answer pairs: given an image and an answer, judge whether this answer is relevant to the image. We use the question encoder to encode the answer. In addition, we perform single modal SSL on question-answer pairs (SSL-QA), to capture the semantic correspondence between words in questions and answers. In SSL-QA, we first pretrain the question encoder by predicting answers solely based on questions themselves without using image information, then finetune the question encoder together with other modules in the full VQA task involving images. Given CMSSL on image-question pairs, CMSSL on image-answer pairs, and SSL-QA, we perform them simultaneously in a multi-task learning framework which minimizes the weighted sum of losses of the three SSL tasks.

4.3 VQA MODELS

We evaluate the effectiveness of learning-by-ignoring and cross-modal self-supervised learning on two VQA models.

- **Method 1**: In Tan & Bansal (2019), a large-scale Transformer (Vaswani et al., 2017) model is built that consists of three encoders: an object relationship encoder, a language encoder, and a cross-modal encoder. The three encoders are built mostly based on two kinds of attention layers — self-attention layers and cross-attention layers. The object relationship encoder and the language encoder are both single-modality encoders. Each layer of them contains a self-attention sub-layer and a feed-forward sub-layer, where the feed-forward sub-layer is composed of two fully-connected sub-layers. A cross-modal encoder is proposed to learn the connections between vision and language. Each layer of it consists of two self-attention sub-layers, one bi-directional cross-attention sub-layer, and two feed-forward sub-layers.
- **Method 2**: The method proposed in Kim et al. (2018) uses a Gated Recurrent Unit (GRU) (Cho et al., 2014) recurrent network and a Faster R-CNN (Ren et al., 2015b) network to embed the question and the image. It extends the idea of co-attention into bilinear attention which considers every pair of multimodal channels. It learns bilinear attention distributions using the bilinear attention networks (BAN) and uses low rank approximation techniques to approximate the bilinear interaction between question embeddings and image embeddings. It also proposes residual learning of attention which keeps the size of intermediate features constant.

5 EXPERIMENT

5.1 EXPERIMENTAL SETTINGS

Data split We partition the images in the PathVQA dataset along with the associated questions into a training set, validation set, and testing set with a ratio of about 3:1:1. In the PathVQA dataset, the frequencies of question

Table 3: Statistics of the data split

	Training set	Validation set	Test set
# images	3,021	987	990
# QA pairs	19,755	6,279	6,761

Table 4: Accuracy (%), BLEU-n (%), and F1 (%) achieved by different methods. We denote cross-modal SSL on image-question pairs and image-answer pairs as CMSSL-IQ, CMSSL-IA, and denote single-modal SSL on question-answer pairs as SSL-QA

Method	Accuracy	BLEU-1	BLEU-2	BLEU-3	F1
Method 1 without image	49.2	50.2	2.8	1.2	9.5
Method 1	57.6	57.4	3.1	1.3	9.9
Method 1 with ignoring	58.5	58.9	3.5	2.0	10.2
Method 1 with CMSSL-IQ	58.7	59.0	3.5	2.1	11.0
Method 1 with CMSSL-IA	58.6	58.9	3.4	2.0	10.3
Method 1 with SSL-QA	58.7	59.0	3.5	2.1	11.2
Method 1 with joint pretraining	59.3	59.2	4.7	2.8	11.6
Method 1 with joint pretraining+ignoring	60.1	59.9	5.1	3.2	12.2
Method 2 without image	46.2	46.5	1.0	0.0	0.8
Method 2	55.1	56.2	3.2	1.2	8.4
Method 2 with ignoring	56.3	57.4	3.5	1.8	9.6
Method 2 with CMSSL-IQ	55.9	57.1	3.4	1.4	9.2
Method 2 with CMSSL-IA	55.9	57.1	3.5	1.5	9.2
Method 2 with SSL-QA	57.6	58.8	4.1	1.5	10.8
Method 2 with joint pretraining	57.7	59.1	4.2	2.2	10.9
Method 2 with joint pretraining+ignoring	58.4	59.5	4.4	2.6	11.2

categories are imbalanced. Because of this, during the partition process, we perform sampling to ensure the frequencies of these categories in each set to be consistent. There are 19,755 question-answer pairs in the training set, 6,279 in the validation set, and 6,761 in the testing set. For all the data examples in the validation set and test set, senior radiologists helped to carefully examine them to ensure they are clinically correct. The training set was not examined by senior radiologists. The statistics are summarized in Table 3.

Implementation details We basically follow the original model configurations used in Tan & Bansal (2019), Kim et al. (2018), and Yang et al. (2016). Data augmentation is applied to the images, including shifting, scaling, and shearing. From questions and answers in the PathVQA dataset, we create a vocabulary of 4,631 words that have the highest frequencies. In Method 1, we use the default hyperparameter settings in Tan & Bansal (2019). For the text encoder, the hidden size was set to 768. The image features were extracted from the outputs of the Faster-RCNN network, which is pretrained on BCCD[2] – a medical dataset containing blood cells photos, as well as on Visual Genome (Krishna et al., 2017). The initial learning rate was set to 5e-5 with the Adam (Kingma & Ba, 2014a) optimizer used. The batch size was set to 256. The model was trained for 200 epochs. In the cross-modal SSL pretraining on Method 1, we train a linear classifier with a dimension of 1,280 to judge whether one modality of data (image, question, answer) matches with another. In Method 2, words in questions and answers are represented using GloVe (Pennington et al., 2014) vectors pretrained on general-domain corpora such as Wikipedia, Twitter, etc. The image features are extracted from the outputs of the Faster-RCNN network pretrained on BCCD and Visual Genome. Given an image and a question, the model outputs an answer from a predefined set of answers. The dropout (Krizhevsky et al., 2012) rate for the linear mapping was set to 0.2 while for the classifier it was set to 0.5. The initial learning rate was set to 0.005 with the Adamax optimizer (Kingma & Ba, 2014b) used. The batch size was set to 512. The model was trained for 200 epochs. In the cross-modal SSL pretraining on Method 2, similar to that on Method 1, we train a linear classifier with a dimension of 1,280 to predict whether two modalities of data match or not. For learning-by-ignoring, we update ignoring variables using the Adam optimizer, with an initial learning rate of 0.01. We perform training for 120 epochs in Method 1 with ignoring and for 180 epochs in Method 2 with ignoring.

Evaluation metrics We perform evaluation using three metrics: 1) accuracy (Malinowski & Fritz, 2014) which measures the percentage of inferred answers that match exactly with the ground-truth using string matching; only exact matches are considered as correct; 2) macro-averaged F1 (Goutte & Gaussier, 2005), which measures the average overlap between the predicted answers and ground-truth, where the answers are treated as bag of tokens; 3) BLEU (Papineni et al., 2002), which measures the similarity of predicted answers and ground-truth by matching n-grams.

[2]https://public.roboflow.ai/object-detection/bccd

Table 6: Accuracy (%) on open-ended questions of different types

Method	Question types				
	What	Where	How	How much/many	Why
Method 1 without image	0.08	0.39	0.16	0.41	0.50
Method 1	0.22	0.73	0.12	0.45	0.50
Method 1 with ignoring	0.24	0.76	0.15	0.45	0.64
Method 1 with CMSSL-IQ	0.24	0.73	0.13	0.45	0.59
Method 1 with CMSSL-IA	0.24	0.74	0.13	0.45	0.59
Method 1 with SSL-QA	0.26	0.78	0.15	0.50	0.64
Method 1 with joint pretraining	0.29	0.79	0.16	0.50	0.68
Method 1 with joint pretraining+ignoring	**0.32**	**0.81**	**0.16**	**0.56**	**0.68**
Method 2 without image	0.05	0.29	0.00	0.00	0.00
Method 2	0.18	0.64	0.11	0.36	0.32
Method 2 with ignoring	0.24	0.72	0.12	0.41	0.41
Method 2 with CMSSL-IQ	0.20	0.71	0.12	0.36	0.50
Method 2 with CMSSL-IA	0.20	0.72	0.11	0.41	0.45
Method 2 with SSL-QA	0.20	0.71	0.11	0.36	0.45
Method 2 with joint pretraining	0.21	0.72	0.12	0.45	0.55
Method 2 with joint pretraining+ignoring	**0.24**	**0.72**	**0.14**	**0.45**	**0.59**

5.2 RESULTS

Table 4 shows the VQA performance achieved by different methods. From this table, we make the following observations. **First**, for both Method 1 and Method 2, applying learning-by-ignoring (LBI) improves the performance. This demonstrates the effectiveness of LBI in improving the generalization ability of trained VQA models. LBI learns to identify and remove noisy and erroneous training data examples, which can avoid the model to be distorted by such bad-quality examples. **Second**, for both Method 1 and 2, applying cross-modal SSL (CMSSL) methods including CMSSL-IQ and CMSSL-IA improves the performance, which demonstrates the effectiveness of CMSSL. CMSSL uses auxiliary tasks, including judging whether an image matches with a question and judging whether an image matches with an answer, to learn semantic correspondence between image regions and words in questions/answers, which can improve the effectiveness of visual and textual representations for accurate VQA. **Third**, using SSL-QA improves VQA performance of Method 1 and 2. SSL-QA learns the correspondence between words in questions and words in answers, which can better extract semantic representations of questions and answers. **Fourth**, joint pretraining which performs CMSSL-IQ, CMSSL-IA, and SSL-QA jointly achieves better performance than performing the three SSL tasks individually, for both Method 1 and 2. This is because letting the model solve several SSL tasks simultaneously is more challenging, which encourages the model to learn more powerful textual and visual representations. **Fifth**, applying both joint pretraining and learning-by-ignoring achieves the best performance in Method 1 and 2. Table 6 shows the accuracy scores achieved on open-ended questions belonging to the following categories: what, where, how, how much/how many, and why respectively by different methods. Table 5 shows the accuracy on yes/no questions. Similar to the observations made from Table 4, the results in Table 6 and Table 5 also demonstrate that learning-by-ignoring and cross-modal SSL both help to improve VQA performance. In Table 6, all methods perform the best on "where" questions. This is because it is relatively easy to recognize image regions of interest for "where" questions, which helps the model to give the correct answer. In Table 5, all methods perform much better than random guesses (where the accuracy is 50%). This indicates that our PathVQA dataset is clinically meaningful, which allows VQA models to be learnable.

Table 5: Accuracy (%) on "yes/no" questions

Method	Accuracy
Method 1 without image	85.1
Method 1	86.1
Method 1 with ignoring	86.4
Method 1 with CMSSL-IQ	86.2
Method 1 with CMSSL-IA	86.4
Method 1 with SSL-QA	86.2
Method 1 with joint pretraining	86.8
Method 1 with joint pretraining+ignoring	87.1
Method 2 without image	84.5
Method 2	85.7
Method 2 with ignoring	86.4
Method 2 with CMSSL-IQ	86.4
Method 2 with CMSSL-IA	86.4
Method 2 with SSL-QA	86.8
Method 2 with joint pretraining	86.6
Method 2 with joint pretraining+ignoring	87.2

One may suspect how much information in images are used during the inference of the answers? Could it be possible that the models simply learn the correlations between questions and answers and ignore the images? In light of these concerns, we perform studies where the images are not fed into VQA models and only questions are used as inputs for inferring answers. Table 4 shows the results of not using images ("Method 1/2 without image"). As can be seen, for both Method

1 and 2, ignoring images leads to substantial degradation of performance. This shows that images in our dataset provide valuable information for VQA and PathVQA is a meaningful VQA dataset. The models trained on our datasets are not degenerated to simply capture the correlation between questions and answers.

6 CONCLUSION

In this paper, towards the goal of developing AI systems to pass the board-certificated examinations of the American Board of Pathology and fostering research in medical visual question answering, we build a pathology VQA dataset – PathVQA – that contains 32,795 question-answer pairs of 8 categories, generated from 4,998 images. Majority of questions in our dataset are open-ended, posing great challenges for the medical VQA research. Our dataset is publicly available. To address the challenges that the training data may contain errors and the effective representations of pathology images and questions are difficult to learn on limited data, we propose a learning-by-ignoring approach to automatically identify and remove problematic training examples and develop cross-modal self-supervised learning approaches to learn visual and textual representations effectively. The experiments on our collected PathVQA dataset demonstrate the effectiveness of our proposed methods.

REFERENCES

Asma Ben Abacha, Sadid A Hasan, Vivek V Datla, Joey Liu, Dina Demner-Fushman, and Henning Müller. Vqa-med: Overview of the medical visual question answering task at imageclef 2019. In *CLEF2019 Working Notes.*, 2019.

Stanislaw Antol, C Lawrence Zitnick, and Devi Parikh. Zero-shot learning via visual abstraction. In *ECCV*, 2014.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *ICCV*, 2015.

Olivier Bachem, Mario Lucic, and Andreas Krause. Practical coreset constructions for machine learning. *arXiv preprint arXiv:1703.06476*, 2017.

Christos Boutsidis, Michael W Mahoney, and Petros Drineas. An improved approximation algorithm for the column subset selection problem. In *Proceedings of the twentieth annual ACM-SIAM symposium on Discrete algorithms*, pp. 968–977. SIAM, 2009.

Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*, 2014.

Soo-Whan Chung, Joon Son Chung, and Hong Goo Kang. Perfect match: Self-supervised embeddings for cross-modal retrieval. *IEEE Journal of Selected Topics in Signal Processing*, 2020.

Amit Deshpande and Luis Rademacher. Efficient volume sampling for row/column subset selection. In *2010 ieee 51st annual symposium on foundations of computer science*, pp. 329–338. IEEE, 2010.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations. *arXiv preprint arXiv:1803.07728*, 2018a.

Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations. *arXiv preprint arXiv:1803.07728*, 2018b.

Cyril Goutte and Eric Gaussier. A probabilistic interpretation of precision, recall and f-score, with implication for evaluation. In *European Conference on Information Retrieval*, 2005.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *CVPR*, 2017.

Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *CVPR*, 2017.

Aniruddha Kembhavi, Minjoon Seo, Dustin Schwenk, Jonghyun Choi, Ali Farhadi, and Hannaneh Hajishirzi. Are you smarter than a sixth grader? textbook question answering for multimodal machine comprehension. In *CVPR*, 2017.

Jin-Hwa Kim, Jaehyun Jun, and Byoung-Tak Zhang. Bilinear attention networks. In *NIPS*, 2018.

Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *International Conference on Learning Representations*, 12 2014a.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014b.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 2017.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, 2012.

Jason J Lau, Soumya Gayen, Asma Ben Abacha, and Dina Demner-Fushman. A dataset of clinically generated visual questions and answers about radiology images. *Scientific data*, 2018.

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*, 2019a.

Xueting Li, Sifei Liu, Shalini De Mello, Xiaolong Wang, Jan Kautz, and Ming-Hsuan Yang. Joint-task self-supervised learning for temporal correspondence. In *Advances in Neural Information Processing Systems*, pp. 317–327, 2019b.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014.

Hanxiao Liu, Karen Simonyan, and Yiming Yang. Darts: Differentiable architecture search. *arXiv preprint arXiv:1806.09055*, 2018.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Advances in Neural Information Processing Systems*, pp. 13–23, 2019.

Mateusz Malinowski and Mario Fritz. A multi-world approach to question answering about real-world scenes based on uncertain input. In *NIPS*, 2014.

Robert Muir et al. Text-book of pathology. *Text-Book of Pathology.*, (Fifth Edition), 1941.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *ACL*, 2002.

Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros. Context encoders: Feature learning by inpainting. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2536–2544, 2016a.

Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros. Context encoders: Feature learning by inpainting. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2536–2544, 2016b.

Jeffrey Pennington, Richard Socher, and Christopher Manning. Glove: Global vectors for word representation. In *EMNLP*, 2014.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8):9, 2019.

Mengye Ren, Ryan Kiros, and Richard Zemel. Image question answering: A visual semantic embedding model and a new dataset. *NIPS*, 2015a.

Mengye Ren, Wenyuan Zeng, Bin Yang, and Raquel Urtasun. Learning to reweight examples for robust deep learning. *arXiv preprint arXiv:1803.09050*, 2018.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pp. 91–99, 2015b.

Stanley L Robbins, Marcia Angell, and Vinay Kumar. *Basic pathology*. WB Saunders, 1981.

Lei Shi, Kai Shuang, Shijie Geng, Peng Su, Zhengkai Jiang, Peng Gao, Zuohui Fu, Gerard de Melo, and Sen Su. Contrastive visual-linguistic pretraining. *arXiv preprint arXiv:2007.13135*, 2020.

Jun Shu, Qi Xie, Lixuan Yi, Qian Zhao, Sanping Zhou, Zongben Xu, and Deyu Meng. Meta-weight-net: Learning an explicit mapping for sample weighting. In *Advances in Neural Information Processing Systems*, pp. 1919–1930, 2019.

Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *ECCV*, 2012.

Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. Videobert: A joint model for video and language representation learning. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 7464–7473, 2019.

Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490*, 2019.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pp. 5998–6008, 2017.

Xiaolong Wang, Allan Jabri, and Alexei A Efros. Learning correspondence from the cycle-consistency of time. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2566–2576, 2019a.

Xin Wang, Qiuyuan Huang, Asli Celikyilmaz, Jianfeng Gao, Dinghan Shen, Yuan-Fang Wang, William Yang Wang, and Lei Zhang. Reinforced cross-modal matching and self-supervised imitation learning for vision-language navigation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6629–6638, 2019b.

Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3733–3742, 2018.

Zichao Yang, Xiaodong He, Jianfeng Gao, Li Deng, and Alex Smola. Stacked attention networks for image question answering. In *CVPR*, 2016.

Richard Zhang, Phillip Isola, and Alexei A Efros. Colorful image colorization. In *European conference on computer vision*, pp. 649–666. Springer, 2016a.

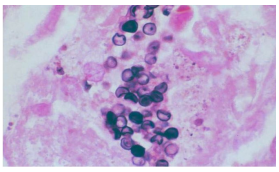
Richard Zhang, Phillip Isola, and Alexei A Efros. Colorful image colorization. In *European conference on computer vision*, pp. 649–666. Springer, 2016b.

C Lawrence Zitnick and Devi Parikh. Bringing semantics into focus using visual abstraction. In *CVPR*, 2013.

A APPENDIX

A.1 EXAMPLE OF ABP TEST QUESTIONS

An example of ABP test questions is shown in Figure 4.



Q: What are these GMS-stained organisms?
A1: *Blastomyces dermatitidis.*
A2: *Cryptococcus neoformans.*
A3: *Pneumocystis jiroveci.*
A4: trophozoites of *Entamoeba histolytica.*
A5: yeasts of *Candida* species.

Figure 4: An example of ABP test questions.

A.2 COMPARISON OF EXISTING VQA DATASETS

The comparison of existing VQA datasets is shown in Table 7. Table 7 presents a comparison of different VQA datasets. The first five datasets are in the general domain while the last three are in the medical domain. Not surprisingly, the size of general-domain datasets (including the number of images and question-answer pairs) is much larger than that of medical datasets since general-domain images are much more available publicly and there are many qualified human annotators to generate QA pairs on general images. Our dataset is larger than the two medical datasets: VQA-Med and VQA-RAD, and majority of questions in our dataset are open-ended while majority of questions in VQA-Med and VQA-RAD are in multiple-choices style.

Table 7: Comparison of VQA datasets

	Domain	# images	# QA pairs	Answer type
DAQUAR	General	1,449	12,468	Open
VQA	General	204K	614K	Open/MC
VQA v2	General	204K	1.1M	Open/MC
COCO-QA	General	123K	118K	Open/MC
CLEVR	General	100K	999K	Open
VQA-Med	Medical	4,200	15,292	Open/MC
VQA-RAD	Medical	315	3,515	Open/MC
Ours	Medical	4,998	32,795	Open

A.3 NUMBER OF QUESTIONS IN DIFFERENT CATEGORIES FOR TRAINING, VALIDATION, AND TEST SET

For our data split, the number of questions in different categories in each set is shown in Table 8.

Table 8: Number of questions in different categories in each set

Dataset	Question types					
	What	Where	How	How much/many	Why	Yes/No
Training set	8083	1316	366	62	71	9804
Validation set	2565	409	108	21	21	3135
Testing set	2753	432	121	18	22	3390

A.4 DERIVATION OF GRADIENT IN ALGORITHM 1

In Algorithm 1, there are two steps. In the first step, we update ignoring variables A by descending $\nabla_A L_{val}\left(W - \xi \nabla_W L_{train}(W, A)\right)$, where we approximate W^* using one step gradient descent

update of W:

$$W^* \approx W - \xi \nabla_W L_{train}(W, A),$$

where ξ is the learning rate.

We compute $\nabla_A L_{val}(W^*)$ as follows:

$$\nabla_A L_{val}(W^*) \tag{2a}$$
$$\approx \nabla_A L_{val}(W - \xi \nabla_W L_{train}(W, A)) \tag{2b}$$
$$= -\xi \nabla^2_{A,W} L_{train}(W, A) \nabla_{W^*} L_{val}(W^*) \tag{2c}$$
$$\approx -\xi \frac{\nabla_A L_{train}(W^+, A) - \nabla_A L_{train}(W^-, A)}{2\epsilon}, \tag{2d}$$

where ϵ is a small scalar and

$$W^{\pm} = W \pm \epsilon \nabla_{W^*} L_{val}(W^*).$$

A.5 Additional Related Works

A number of visual question answering datasets have been developed in the general domain. DAQUAR (Malinowski & Fritz, 2014) is built on top of the NYU-Depth V2 dataset (Silberman et al., 2012) which contains RGBD images of indoor scenes. DAQUAR consists of (1) synthetic question-answer pairs that are automatically generated based on textual templates and (2) human-created question-answer pairs produced by five annotators. The VQA dataset (Antol et al., 2015) is developed on real images in MS COCO (Lin et al., 2014) and abstract scene images in (Antol et al., 2014; Zitnick & Parikh, 2013). The question-answer pairs are created by human annotators who are encouraged to ask "interesting" and "diverse" questions. VQA v2 (Goyal et al., 2017) is extended from the VQA (Antol et al., 2015) dataset to achieve more balance between visual and textual information, by collecting complementary images in a way that each question is associated with a pair of similar images with different answers. In the COCO-QA (Ren et al., 2015a) dataset, the question-answer pairs are automatically generated from image captions based on syntactic parsing and linguistic rules. CLEVR (Johnson et al., 2017; Kembhavi et al., 2017) is a dataset developed on rendered images of spatially related objects (including cube, sphere, and cylinder) with different sizes, materials, and colors. The locations and attributes of objects are annotated for each image. The questions are automatically generated from the annotations.