Diffusion-based Visual Representation Learning for Medical Question Answering

Dexin Bian BIANDEXIN@BUPT.EDU.CN
Xiaoru Wang WXR@BUPT.EDU.CN
Meifang Li LIMF@BUPT.EDU.CN
Beijing University of Posts and Telecommunications, 10 Xitucheng RD, Haidian DIST, Beijing, China

Editors: Berrin Yanıkoğlu and Wray Buntine

Abstract

Medical visual question answering (Med-VQA) aims to correctly answer the medical question based on the given image. One of the major challenges is the scarcity of large professional labeled datasets for training, which poses obstacles to feature extraction, especially for medical images. To overcome it, we propose a method to learn transferable visual representation based on conditional denoising diffusion probabilistic model(conditional DDPM).Specifically, we collate a large amount of unlabeled radiological images and train a conditional DDPM with the paradigm of auto-encoder to obtain a model which can extract high-level semantic information from medical images.The pre-trained model can be used as a well initialized visual feature extractor and can be easily adapt to any Med-VQA systems. We build our Med-VQA system follow the state-of-the-art Med-VQA architecture and replace the visual extractor with our pre-trained model.Our proposal method outperforms the state-of-the-art Med-VQA method on VQA-RAD and achieves comparable result on SLAKE.

Keywords: Medical Visual Question Answering, Denoising Diffusion Probabilistic Model, Representation Learning

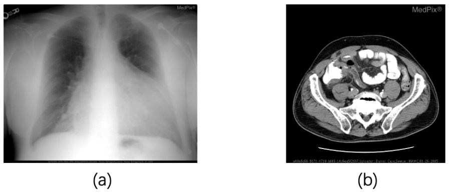
1. Introduction

[1]

Med-VQA system takes medical images and clinical questions as input and then give correct answers to the questions by combining the visual information in the image. Med-VQA has a broad application prospect in the field of healthcare. It can assist doctors in diagnosis, help patients learn more health information, assist clinical teaching, and can also be integrated into the dialogue AI platform to provide intelligent medical consultation services.

Although breakthroughs in computer vision and natural language processing have laid a foundation for the research of Med-VQA, the scarcity of professional Med-VQA datasets is still a huge bottleneck for the development of Med-VQA, especially the amount of medical image is even more scarce. The existing manually annotated datasets such as VQA-RAD (Lau et al. (2018)) has 3515 QA pairs but only 315 medical images, while SLAKE (Liu

[1] Our code is available at https://github.com/lluviosac/Diffusion-Based-MedVQA.



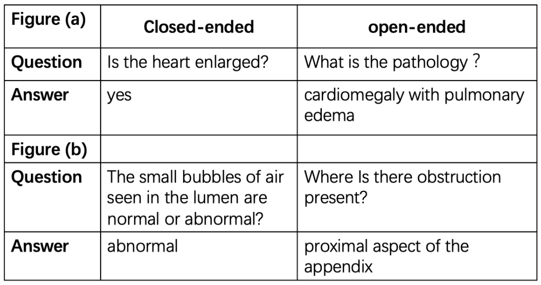


Figure (a)	Closed-ended	open-ended
Question	Is the heart enlarged?	What is the pathology ?
Answer	yes	cardiomegaly with pulmonary edema
Figure (b)		
Question	The small bubbles of air seen in the lumen are normal or abnormal?	Where Is there obstruction present?
Answer	abnormal	proximal aspect of the appendix

Figure 1: An example of Med-VQA, an image may have multiple QA pairs.There are two types of questions, closed-end questions which the answers are restricted to a few specific answers, and open-end questions which the answers are free-form.

et al. (2021b)) has 14000 QA pairs but 642 medical images. It is difficult to train an image encoder which could extract high-quality and high-level semantic visual features solely on these Med-VQA datasets from scratch. Due to high costs and privacy issues, there is no large-scale labeled dataset like ImageNet (Russakovsky et al. (2014)) in the medical field, which makes it impossible to train deep networks from scratch. Some solutions directly transfered deep neural networks pre-trained on ImageNet(Russakovsky et al. (2014)) but get poor performance because of the huge differences between medical images and ordinary images. Nguyen et al. (2019) is the first study to address this issue, proposed to use Model-Agnostic Meta-Learning (MAML)(Finn et al. (2017)) for weight initialization of image encoder. However, the task was specifically designed for VQA-RAD (Lau et al. (2018)) dataset which made it cannot be easily applied to other datasets. Khare et al. (2021) proposed to learn medical image and text semantic representations using Masked Language Modeling (MLM) with image features as the pretext task on ROCO——a large medical image caption dataset(Pelka et al. (2018)). Eslami et al. (2021) fine-tuned CLIP(Radford et al. (2021)) on ROCO(Pelka et al. (2018)). These methods have made great gains in general vision-language field, however, it has little gain in Med-VQA due to the lack of high-quality image text data. Liu et al. (2021a) leverage large amounts of unlabeled radiology images to train three teacher models for the body regions of brain, chest, and abdomen respectively via contrastive learning and then distill the teacher models to a student model that can be used as visual feature extractor for Med-VQA system. However, the latent space of its encoding is still limited to the number of teacher models.

In this paper, we first build a medical image dataset which contains 89946 images by collecting and preprocessing medical image datasets from different medical relevant tasks (image segmentation, image reconstruction, etc.). The dataset contains medical images of multiple modes (CT, MRI, etc.), multiple human parts and organs (brain, chest, abdomen, joints, hands and feet, etc.), multiple planes (sagittal plane, coronal plane and cross section), normal or pathological changes.The dataset can effectively alleviate the problem of the scarcity of medical image of Med-VQA datasets, which is conducive to learning a better representation space, and it can be simply and effectively migrated to the existing or future Med-VQA datasets due to its diversity. We first train an unconditional diffusion probabilistic model (DPM) on this dataset to learn the hidden variables of the medical image, but these hidden variables are the result of the noise in the original space, lacking the high-level semantic information of the data. Therefore, we use an encoder to map the image into a latent variable, and then the DPM take the latent variable and stochastic noise as input to reconstruct the image. We use this encoder as a general medical image encoder, which can be directly fine-tuned on any Med-VQA datasets.

To summarize, our contributions are: (1) We collate a large medical image dataset which contains image of multimodality, multiple parts and organs, and multiple planes. (2) We propose a method to pretrain a general medical visual feature extractor for Med-VQA which is based on conditional DPM and combined with auto encoder paradigm. (3) We conduct extensive experiments with the state-of-the-art Med-VQA methods on VQA-RAD(Lau et al. (2018))and SLAKE(Liu et al. (2021b)) to demonstrate the effectiveness of the model.

2. Related Work

In this section, we briefly review the research progress of Med-VQA and introduce the denoising diffusion probabilistic model(DDPM).

2.1. Medical Visual Question Answering

Med-VQA task was first proposed in the ImageCLEF 2018 challenge(Abacha et al. (2018)). The common framework of Med-VQA follows the framework of general VQA, which is divided into three steps: extracting visual features from images, extracting text feature from questions, and combining visual and text feature to predict answers.

A few works directly applied state-of-the-art VQA model to the medical field, which usually using a pre-trained deep neural network such as ResNet or VGGNet to extract visual feature, and a RNN-based neural networks to extract question feature.For feature fusion, Abacha et al. (2018); Nguyen et al. (2019); Zhan et al. (2020) used concatenation, stacked attention network(Yang et al. (2015)), compact bilinear pooling(Fukui et al. (2016)), and bilinear attention network(Kim et al. (2018)) to fuse the two modal features. However, the direct application was not effective due to the huge data difference between the medical domain and the general domain.

One branch of works focuses on improving the extraction of medical image features. Nguyen et al. (2019) proposed the mixture of enhanced visual features(MEVF), which consists of an unsupervised auto-encoder and an image encoder trained by Model-Agnostic Meta-Learning(Finn et al. (2017)) as a visual feature extraction module. Nguyen et al. (2019) extracted new labels and designed the meta-learning task for VQA-RAD to learn the initialization of the image encoder. This method outperforms direct transfer and reduces the parameter scale, but requires additional data labeling for the training dataset. Ren and Zhou (2020); Khare et al. (2021) used transformer(Vaswani et al. (2017)) to do self-supervised pre-training on the medical image-caption dataset ROCO(Pelka et al. (2018)).

The image feature which encoded by different layers of CNN are fed into a transformer as visual tokens with text tokens. They leveraged the self-attention to align image regions and text tokens. Eslami et al. (2021) fine-tuned the CLIP(Radford et al. (2021)) model as an image encoder on the ROCO (Pelka et al. (2018)) with little improvement as well. The approach which has achieved great success in the general vision-language domain with rich unlabeled image-text pairs got little gain in Med-VQA because of the lack of high-quality image-text data and negative false pairs in the medical domain. Liu et al. (2021a) proposed to learn a general medical image encoder by contrastive learning and distillation. They used momentum contrast to train three teacher models which represent brain, chest, abdomen respectively and then distilled three teacher models into a lightweight student model.

Another branch of works focuses on high-level reasoning capability, such as Shi et al. (2019) extracted topic representations by using embedding-based topic model and SVM, Zhan et al. (2020) proposed a question-conditional reasoning module to guide the importance selection of multi-modal fusion features, and learning different reasoning skills for different types of questions.

In this work, we focus on improving the image feature extraction while follow the reasoning module of Zhan et al. (2020).

2.2. Denoising Diffusion Probabilistic Models

The main idea of the denoising diffusion Probabilistic model (DDPM, Ho et al. (2020)) is to add gaussian noise to corrupts the data distribution during the forward process, and then learn to recover the data distribution during the reverse process. Ho et al. (2020) proposed to apply UNet(Ronneberger et al. (2015)) to learn a function $\epsilon_\theta(x_t, t)$ which takes noisy image x_t and time t to predict the noise and trained it through $||\epsilon - \epsilon_\theta||$.Song et al. (2020a) proposed Denoising Diffusion Implicit Model (DDIM). DDIM can reverse on the subsequence of the complete timestep sequence to reconstruct x_0 from x_T which accelerates the sampling while DDPM needs to reverse on the complete sequence of $T \sim 0$.

Diffusion Probabilistic Models(DPMs) are unconditional generative model, which can only sample randomly and cannot control the output of the model. Recently the conditional DPMs(Rombach et al. (2021); Dhariwal and Nichol (2021); Song et al. (2020b)) have achieved great success in the field of image generation and high resolution. There are two ways to insert "condition" into the DPM, one is called "Classifier-Guidance" and the other is "Classifier-Free". Classifier-Guidance introduces an additional classifier to an unconditional pre-trained DPM to guide the generation. For example, Dhariwal and Nichol (2021); Song et al. (2020b) introduced a classifier $p_{\theta,\phi}(y|x_t)$ which predict the class label y to guide the generation through its gradient. Classifier-Free(Rombach et al. (2021)) is to directly add condition embedding to the training phase of DPM, the model ϵ_θ naturally contains the conditional information when sampling.

3. Method

In this paper, we focus on enhancing the visual feature extraction from medical image. Firstly, as the existing Med-VQA(Lau et al. (2018); Liu et al. (2021b)) benchmarks contain normal/abnormal radiological image with different modalities, different planes, multiple body parts and organs. Therefore, we collect and preprocess medical image datasets from many medical relevant task which will be used for image encoder's pretraining. Secondly, we train an unconditional DPM on our dataset from scratch. Then, as the Fig. 2 shows, we train a semantic encoder which maps the image to a semantic vector z as conditional variable and a gradient estimator attached to the pre-trained DPM which makes it a conditional DPM. Finally, we apply the semantic encoder to the Med-VQA system.

3.1. Diffusion-based Visual Representation Learning

We adopt the "Classifier-Guidance" way to train conditional DPM, which usually attach an additional classifier $p_{\theta,\phi}(y|x_t)$ to the unconditional DPM. The classifier predicts the noised x_t's original label y and the gradient $\nabla_{x_t} p_{\theta,\phi}(y|x_t)$ will guide the DPM to sample towards specific class y.

$$p_{\theta,\phi}(x_{t-1}|x_t, y) = \mathcal{N}(\mu_\theta(x_t) + \sum_\theta \nabla_{x_t} logp_\phi(y|x_t), \sum_\theta) \tag{1}$$

As the Eq. 1, compare to the unconditional DPM, the classifier-guidance way just shifts the mean with $\sum_\theta \nabla_{x_t} p_{\theta,\phi}(y|x_t)$ while sampling. Song et al. (2020b) used stochastic differential equations and score matching tricks to obtain:

$$\hat{\epsilon}_{\theta,\phi}(x_t, t) = \epsilon_\theta(x_t, t) - \sqrt{1 - \overline{\alpha}_t} \nabla_{x_t} \log p_\phi(y|x_t) \tag{2}$$

which replace $\epsilon_\theta(x_t, t)$ in unconditional sampling, we can get a more general sampling process:

$$p(x_{t-1}|x_t, y) = \mathcal{N}(\sqrt{\frac{\overline{\alpha}_{t-1}}{\overline{\alpha}_t}} x_t + (\sqrt{1 - \overline{\alpha}_{t-1} - \sigma_t^2} - \sqrt{\frac{\overline{\alpha}_{t-1}(1 - \overline{\alpha}_t)}{\overline{\alpha}_t}}) \hat{\epsilon}_{\theta,\phi}(x_t, t), \sigma_t^2) \tag{3}$$

when σ_t=0, it is the sampling process of DDIM(Song et al. (2020a)).

According to Eq. 1, the conditional reverse process guided by the gradient of the classifier has an additional shift term at the mean compared to the unconditional one, which can help the reverse process to reconstruct the lost class information in the samples. The unconditional DPM receives a noise x_T and the data generated by gradual denoising is not deterministic. It can ensure that the generated data is consistent with the data distribution of the training dataset, but the specific content of the generation is not controllable. However, conditional DPM can control the result of generation through conditional embedding, that is, in the reverse process, introducing a prior condition y about x_0 can make up for the information gap lost in the forward process, and the more information y contains about x_0, the more gap can be made up. From this aspect, we adopt a model to encode the original image x_0 as $z = Enc_\psi(x_0)$. The DPM will take z as the conditional embedding and the random noise x_T together in reverse process to reconstruct the original image. Therefore, our goal is to let Enc_ψ learn a semantically rich latent space, so that the encoded z contains richer and more discriminative semantic information of medical images which can apply to the Med-VQA system. In our experiment, Enc_ψ only consists of several stacked convolutional layers and a linear layer.

3.2. Gradient Estimator for Unsupervised Representation Learning

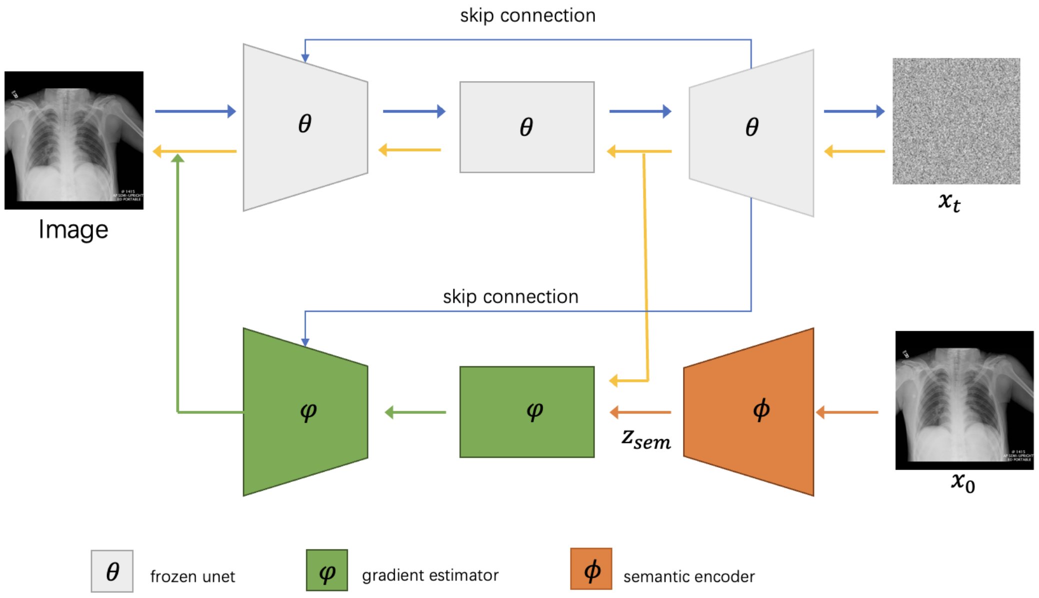


Figure 2: The framework of conditional DPM, contains a semantic encoder ϕ(orange) which map the origin image x into semantic encoding z_{sem}, a gradient estimator ψ(green) and a pre-trained unconditional DPM θ(gray).During the phase of training encoder, DPM model will be frozen, z_{sem} will capture the high-leve semantic information of x_0, the gradient estimator will receive conditional encoding z_{sem}, time embedding and noise to approximate gradient and then guide DPM to reconstruct x_0.

As our dataset is collated from different medical image tasks without consistent and strict annotation. Instead of explicitly introduce a classifier, we add a module to the DPM network as a gradient estimator G_ϕ to approximate the gradient. As Fig. 2 shows, G_ϕ takes noised image x_t, conditional embedding z and time embedding t as input and approximate $\nabla_{x_t} p_{\theta,\phi}(y|x_t)$. The reverse process of the conditional DPM can be expressed as:

$$p_{\theta,\phi}(x_{t-1}|x_t,z) = \mathcal{N}(x_{t-1};\mu_\theta(x_t,t) + \sum_\theta(x_t,t)\cdot G_\phi(x_t,z,t), \sum_\theta(x_t,t)) \tag{4}$$

We train an unconditional DPM θ from scratch on the radiology image dataset introduced previous section, and then freeze θ, optimize the encoder and conditional control module where the training objective is

$$\mathcal{L}(\psi,\phi) = \mathbb{E}_{x_0,t,\epsilon}[\lambda_t||\epsilon - \epsilon_\theta(x_t,t) + \sqrt{\frac{\overline{\alpha}_t(1-\overline{\alpha}_t)}{\overline{\beta}_t}} \cdot \sum_\theta(x_t,t)\cdot G_\phi(x)t, E_\psi(x_0),t)||^2] \tag{5}$$

where $x_t = \sqrt{\overline{\alpha}_t}x_0 + \sqrt{1-\overline{\alpha}_t}\epsilon$, $\sum_\theta = \frac{1-\overline{\alpha}_{t-1}}{1-\overline{\alpha}_t}\beta_t\mathbf{I}$

Song and Ermon (2019) pointed that different stages in the process should not be equally important. For example, with the generation of images, the same images will become more and more similar while two different images will gradually become different. Therefore, we follow the weighting scheme proposed in Zhang et al. (2023), define $\lambda_t = (\frac{1}{1+SNR(t)}^{1-\gamma}) \cdot (\frac{SNR(t)}{SNR(t)+1})^{\gamma}$ to educe the weight of early and late stages, where $SNR(t) = \frac{\overline{\alpha}_t}{1-\overline{\alpha}_t}$.

In our experiment, we choose a network similar to Unet(Ronneberger et al. (2015)) as the gradient estimator. In order to make full use of the knowledge learned by the pre-trained DPM, we reuse the encoder part of the pre-trained DPM and directly use it to encode x_t. Only the middle blocks and output blocks of Unet are added. Skip connections is still used between the new decoder and the reused encoder.

3.3. Apply Diffusion Encoder for Med-VQA

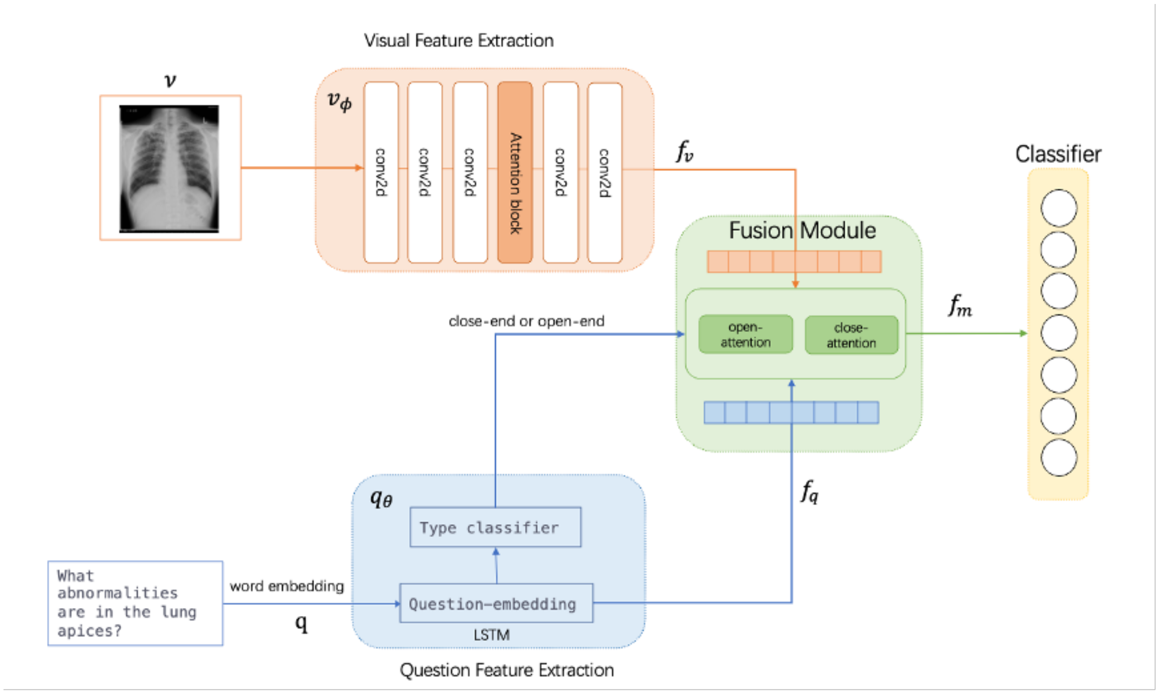


Figure 3: The framework of the Med-VQA system. We follow the architecture of Zhan et al. (2020), and replace the orange block (visual extraction block) with the model we pretrain based on conditional DDPM. Visual feature and question features are extracted separately. The fusion model which fuses visual feature and question feature has two independent attention blocks, one for open-end question, the other for close-end question. The question feature extraction block will classify the type of question, and feed the question embedding and visual embedding into corresponding attention block. The final classifier will receive the fusion feature f_m and give the final answer.

As Fig. 3 shows, we directly adopt the Enc_{ψ} we trained before to the Med-VQA system as a general visual feature extractor. Given a radiological image and a medical question pair (v_i, q_i). The image v_i will be fed into Enc_{ψ} to obtain image feature $f_v = Enc_{\psi}(v_i)$. The question q_i will be encoded by an encoder such as LSTM, GRU to obtain text feature $f_q = Q_{\psi}(v_i)$. Then f_v and f_q will be fused by an attention-based fusion module such as BAN(Kim et al. (2018)), SAN(Yang et al. (2015))to get multimodal feature f_m. Med-VQA is defined as a classification task to predict the correct answer from the candidate answer set C, so f_m will be fed into a classifier to get the correct answer. We compute the cross-entropy loss to optimize the model end-to-end.

4. Experiment

In this section, we conduct experiments to evaluate the effectiveness of our pre-trained encoder on three Med-VQA benchmarks. We follow the reasoning module of QCR(Zhan et al. (2020)) and the experiment results show that our visual encoder can not only extract high-level semantics features and thus improve the performance, but also can be directly applied to different Med-VQA systems.

4.1. Dataset for Pretraining

For training a more general and professional medical image encoder, we collect radiology images from multiple datasets of different medical-related tasks, preprocess them using python tool libraries. In the end, we obtain a large 2D radiology image dataset (89946 images) which contains different imaging modalities (MRI, CT, etc.), different planes (coronal, sagittal plane, horizontal, etc.), multiple body parts and organs (brain, chest, abdomen, bones, joints, etc.), normal or diseased examples. Table 1 shows the composition of the pre-training dataset.

Fig. 4 shows the t-SNE visualization of representation of images from the test set of VQA-RAD and SLAKE.We divide these examples into different classes through organs.(a) shows the representation obtained by auto-encoder trained on the MedVQA dataset (VQA-RAD or SLAKE), (b) shows the representation obtained by auto-encoder trained on the pre-training dataset we collated. By comparing (a) and (b), the model trained on our pre-training dataset makes the representation of different class more dispersed, while the same class more clustered, which indicates that the rich and diverse medical image has brought significant gains to medical image feature extraction.

4.2. Med-VQA Dataset

- RAD-VQA(Lau et al. (2018)) is a professional Med-VQA dataset proposed in 2018. It contains 315 medical images of head, chest, and abdomen in different modalities and 3515 corresponding QA pairs.

- SLAKE(Liu et al. (2021b)) is a bilingual Med-VQA dataset proposed in 2021. We selected the English version of the data, which contains 642 medical images and 7033 QA pairs. The questions were asked by radiologists and the answers were marked using a pre-designed template.

dataset	organ	modality	size	original task
OASIS(Marcus et al. (2007, 2010))	brain	MRI T1	7281	classification
MSD Brain Tumour(Antonelli et al. (2022))	brain	MRI	12105	segmentation
MSD Lung Tumour(Antonelli et al. (2022))	lung	CT	1890	segmentation
MSD Heart(Antonelli et al. (2022))	left atrium	MRI	200	segmentation
CheXpert(Irvin et al. (2019))	chest	CT	24820	classification
ChestX-ray8(Wang et al. (2017))	Chest	CT	16868	classification
CHAOS(Kavur et al. (2021))	live	CT,MRI	400	segmentation
Sliver07(van Ginneken (2019))	liver	CT	400	segmentation
kits19(Heller et al. (2020))	kidney	CT	584	segmentation
MSD Pancreas(Antonelli et al. (2022))	pancreas	CT	602	segmentation
MSD Spleen(Antonelli et al. (2022))	spleen	CT	409	segmentation
MSD HepaticVessel(Antonelli et al. (2022))	hepatic vessel	CT	556	segmentation
MURA(Rajpurkar et al. (2017))	bone	X-ray	18736	classification
MRNet(Bien et al. (2018))	knee	MRI	5982	classification
LERA(LER)	bone	X-ray	1203	classification
total	-	-	89946	-

Table 1: Overview of our pretraining dataset

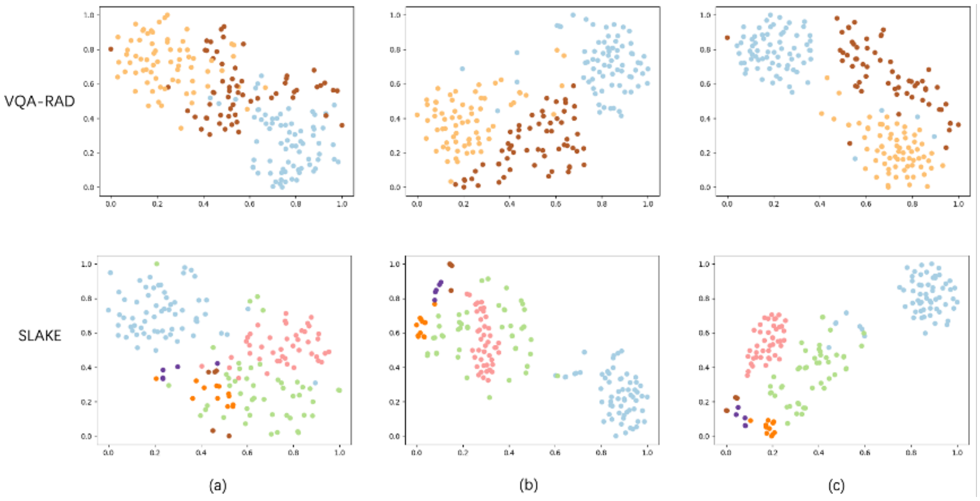


Figure 4: t-SNE visualization of representation of images from the test set.(a)the auto-encoder trained on MedVQA dataset (b)the auto-encoder trained on our dataset (c) condtional DPM trained on our dataset

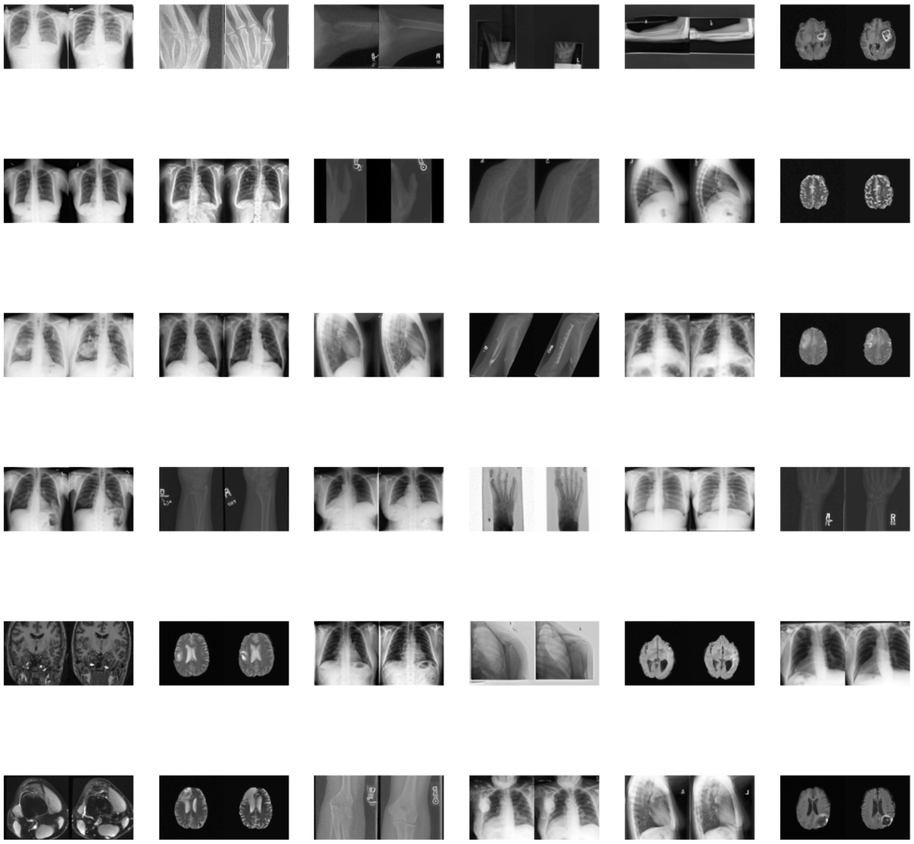


Figure 5: The left of each pair is the original image, and the right is the result of conditional DPM sampling.

4.3. Training Setup

Firstly, we train an unconditional DPM on the pre-training dataset to follow Ho et al. (2020),so that the DPM has the ability to generate medical images. Then we freeze the pre-trained DPM, add the semantic encoder and the gradient estimator to train conditional DPM. We train all models on 1 RTX3090 GPU. We resize all images to 128 × 128 and apply random horizontal flips. We use Unet(Ronneberger et al. (2015)) as the backbone of the DPM and set T to 1000. We set batch size 32, learning rate 0.0001, and the variance of the process to increase linearly from $\beta_1 = 10^{-4}$ to $\beta_T = 0.02$.

As Fig. 5 shows, we use conditional DPM to sample in the DDIM(Song et al. (2020a)) way. We can see that the conditional embedding encoded by semantic encoder can effectively guide the DPM to reconstruct the original image when reverse sampling which indicates that the semantic encoder has indeed learned high-level image semantic information.Compare with (b) and (c) in Fig. 4, the representation generated by conditional DDPM shows more distinct boundary between different classes (especially on SLAKE), which indicates that the model is more capable of distinguishing similar medical images.

VQA-RAD(Lau et al. (2018))	Closed-end	Open-end	Overall
Our	**81.2%**	63.3%	**74.1%**
CPCR(Liu et al. (2023))	80.4%	60.5%	72.5%
QCR(Zhan et al. (2020))	60.0%	79.3%	71.6%
MMBERT(Khare et al. (2021))	63.1%	77.9%	72.0%
CPRD(Liu et al. (2021a))	61.1%	**80.4%**	72.7%
PubMedCLIP(Eslami et al. (2021))	60.1%	80%	72%
MEVF(Nguyen et al. (2019))	49.2%	77.2%	66.1%
SLAKE(Liu et al. (2021b))	Closed-end	Open-end	Overall
Our	**84.6%**	77.5%	80.3%
CPCR(Liu et al. (2023))	84.1%	80.5%	**81.9%**
QCR(Zhan et al. (2020))	83.2%	75.8%	78.7%
CPRD(Liu et al. (2021a))	83.4%	**81.2%**	82.1%
PubMedCLIP(Eslami et al. (2021))	82.5%	78.4%	80.1%
MEVF(Nguyen et al. (2019))	77.5%	74.1%	75.5%

Table 2: Results of our method and others on VQA-RAD and SLAKE.We report the accuracy of closed-end,open-end and overall on VQA-RAD and SLAKE

As our Med-VQA framework shown in Fig. 3, for medical visual feature, we adopt the pre-trained diffusion encoder, for question feature, we follow QCR(Zhan et al. (2020)) to use Glove(Pennington et al. (2014)) to initialize word embedding and use GRU with 1024 hidden dim to extract the semantic information in the question. For feature fusion, we use BAN(Kim et al. (2018)). The classifier will receive the fused feature and predict the answer among the candidates.We use Adam(Kingma and Ba (2014)) optimizer and set the learning rate to 0.0005, except for the visual encoder, which is set to 0.0002.

4.4. Compare with the SOTA

We use accuracy to evaluate the performance of Med-VQA system. In the following, we briefly introduce the baselines. MEVF(Nguyen et al. (2019)) used Modal- Model-Agnostic Meta-Learning(Finn et al. (2017)) to initialize the visual encoder and combines with the output of DAE to enhance the visual feature. QCR(Zhan et al. (2020)), proposed base on MEVF, enhanced the attention to the question feature and reselected the importance of feature after feature fusion. CPRD(Liu et al. (2021a)) used contrastive learning and distillation method to pre-train the image encoder. PubMedCLIP(Eslami et al. (2021)) used medical image-caption pairs to fine-tune CLIP to obtain the visual encoder. MMBERT(Khare et al. (2021)) used a transformer to receive the image token and the question token, and uses the transformer's self-attention to align the text and the image.

Table. 2 shows the results on our method and other SOTA method on the VQA-RAD(Lau et al. (2018)) and SLAKE(Liu et al. (2021b)). On VQA-RAD, our solution outperforms all the existing solution on both open-end and closed-end questions, and improves by 1.4% accuracy compared to the SOTA method overall.On SLAKE dataset, our solution

	diffusion-based encoder			auto-encoder		
	Closed-end	Open-end	Overall	Closed-end	Open-end	Overall
VQA-RAD	81.2%	63.3%	74.1%	79.8%	53.9%	69.4%
SLAKE	84.6%	77.5%	80.4%	82.9%	76.9%	79.3%

Table 3: Ablation study on the diffusion-based encoder.We report the accuracy of closed-end, open-end and overall on VQA-RAD and SLAKE.

also outperforms other methods on closed-end question and outperforms MEVF(Nguyen et al. (2019)) and QCR(Zhan et al. (2020)) on open-end question.

4.5. Ablation Analysis

We conduct an ablation experiment to demonstrate the effectiveness of the diffusion-based encoder. We use the same settings in our method and train an auto-encoder on the pre-training dataset with the goal of data reconstruction. Then We adopt it to the Med-VQA system to compare with the diffusion-based encoder.

As shown in Table. 3 we observe that the diffusion-based visual encoder is highly effective. Compared to the auto-encoder, the diffusion-based visual encoder brings 9.4%, 1.4%, and 4.7% gains for open-end, closed-end and overall on VQA-RAD and brings 0.6%, 1.7%, and 0.9% gains respectively on SLAKE. In particular, we notice that the gains on the open-end of RAD dataset is remarkable, which indicates that the diffusion-based visual encoder can learn rich, effective, high-level medical visual feature. The superior performance verifies that the diffusion-based visual encoder benefits the Med-VQA system to achieve higher accuracy.

5. Conclusion

In this paper, we propose a method to pretrain a general medical image encoder for Med-VQA system, which could extract high-level semantic information in medical image via adding conditional embedding to the DPM. At the same time, we collate a large medical image dataset for pre-training. We follow the framework of QCR but replace the visual encoder with our diffusion-based encoder. Our Med-VQA system achieve the SOTA performance on VQA-RAD and comparable performance on SLAKE.The experiment results show that our visual encoder can be directly and easily applied to existing and future Med-VQA datasets.

Acknowledgments

This work is supported by by National Natural Science Foundation of China No.61976025.

References

Lera- lower extremity radiographs. URL https://aimi.stanford.edu/lera-lower-extremity-radiographs.

Asma Ben Abacha, Soumya Gayen, Jason J. Lau, Sivaramakrishnan Rajaraman, and Dina Demner-Fushman. Nlm at imageclef 2018 visual question answering in the medical domain. In *Conference and Labs of the Evaluation Forum*, 2018.

Michela Antonelli, Annika Reinke, Spyridon Bakas, Keyvan Farahani, Annette Kopp-Schneider, Bennett A. Landman, Geert Litjens, Bjoern Menze, Olaf Ronneberger, and Ronald M. Summers. The medical segmentation decathlon. *Nature Communications*, 13(1), jul 2022. doi: 10.1038/s41467-022-30695-9. URL https://doi.org/10.1038%2Fs41467-022-30695-9.

Nicholas Bien, Pranav Rajpurkar, Robyn L. Ball, Jeremy A. Irvin, Allison Park, Erik Jones, Michael D. Bereket, Bhavik N. Patel, Kristen W. Yeom, and Katie S. Shpanskaya. Deep-learning-assisted diagnosis for knee magnetic resonance imaging: Development and retrospective validation of mrnet. *PLoS Medicine*, 15, 2018.

Prafulla Dhariwal and Alex Nichol. Diffusion models beat gans on image synthesis. *CoRR*, abs/2105.05233, 2021. URL https://arxiv.org/abs/2105.05233.

Sedigheh Eslami, Gerard de Melo, and Christoph Meinel. Does CLIP benefit visual question answering in the medical domain as much as it does in the general domain? *CoRR*, abs/2112.13906, 2021. URL https://arxiv.org/abs/2112.13906.

Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. *CoRR*, abs/1703.03400, 2017. URL http://arxiv.org/abs/1703.03400.

Akira Fukui, Dong Huk Park, Daylen Yang, Anna Rohrbach, Trevor Darrell, and Marcus Rohrbach. Multimodal compact bilinear pooling for visual question answering and visual grounding. *CoRR*, abs/1606.01847, 2016. URL http://arxiv.org/abs/1606.01847.

Nicholas Heller, Fabian Isensee, Klaus H Maier-Hein, Xiaoshuai Hou, Chunmei Xie, Fengyi Li, Yang Nan, Guangrui Mu, Zhiyong Lin, Miofei Han, et al. The state of the art in kidney and kidney tumor segmentation in contrast-enhanced ct imaging: Results of the kits19 challenge. *Medical Image Analysis*, page 101821, 2020.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *CoRR*, abs/2006.11239, 2020. URL https://arxiv.org/abs/2006.11239.

Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn L. Ball, Katie S. Shpanskaya, Jayne Seekins, David A. Mong, Safwan S. Halabi, Jesse K. Sandberg, Ricky Jones, David B. Larson, Curtis P. Langlotz, Bhavik N. Patel, Matthew P. Lungren, and Andrew Y. Ng. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. *CoRR*, abs/1901.07031, 2019. URL http://arxiv.org/abs/1901.07031.

A. Emre Kavur, N. Sinem Gezer, Mustafa Barış, Sinem Aslan, Pierre-Henri Conze, Vladimir Groza, Duc Duy Pham, Soumick Chatterjee, Philipp Ernst, Savaş Özkan, and Bora Baydar. CHAOS Challenge - combined (CT-MR) healthy abdominal organ segmentation. *Medical Image Analysis*, 69:101950, April 2021. ISSN 1361-8415. doi: https://doi.org/10.1016/j.media.2020.101950. URL http://www.sciencedirect.com/science/article/pii/S1361841520303145.

Yash Khare, Viraj Bagal, Minesh Mathew, Adithi Devi, U. Deva Priyakumar, and C. V. Jawahar. MMBERT: multimodal BERT pretraining for improved medical VQA. *CoRR*, abs/2104.01394, 2021. URL https://arxiv.org/abs/2104.01394.

Jin-Hwa Kim, Jaehyun Jun, and Byoung-Tak Zhang. Bilinear attention networks. *CoRR*, abs/1805.07932, 2018. URL http://arxiv.org/abs/1805.07932.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2014.

Jason J. Lau, Soumya Gayen, D. L. Demner, and Asma Ben Abacha. Visual question answering in radiology (vqa-rad). 2018.

Bo Liu, Li-Ming Zhan, and Xiao-Ming Wu. Contrastive pre-training and representation distillation for medical visual question answering based on radiology images. In Marleen de Bruijne, Philippe C. Cattin, Stéphane Cotin, Nicolas Padoy, Stefanie Speidel, Yefeng Zheng, and Caroline Essert, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021*, pages 210–220, Cham, 2021a. Springer International Publishing. ISBN 978-3-030-87196-3.

Bo Liu, Li-Ming Zhan, Li Xu, Lin Ma, Yan Yang, and Xiao-Ming Wu. SLAKE: A semantically-labeled knowledge-enhanced dataset for medical visual question answering. *CoRR*, abs/2102.09542, 2021b. URL https://arxiv.org/abs/2102.09542.

Bo Liu, Li-Ming Zhan, Li Xu, and Xiao-Ming Wu. Medical visual question answering via conditional reasoning and contrastive learning. *IEEE Transactions on Medical Imaging*, 42(5):1532–1545, 2023. doi: 10.1109/TMI.2022.3232411.

Daniel S. Marcus, Tracy H. Wang, Jamie Parker, John G. Csernansky, John C. Morris, and Randy L. Buckner. Open access series of imaging studies (oasis): Cross-sectional mri data in young, middle aged, nondemented, and demented older adults. *Journal of Cognitive Neuroscience*, 19:1498–1507, 2007.

Daniel S. Marcus, Anthony F. Fotenos, John G. Csernansky, John C. Morris, and Randy L. Buckner. Open access series of imaging studies: Longitudinal mri data in nondemented and demented older adults. *Journal of Cognitive Neuroscience*, 22:2677–2684, 2010.

Binh D. Nguyen, Thanh-Toan Do, Binh X. Nguyen, Tuong Do, Erman Tjiputra, and Quang D. Tran. Overcoming data limitation in medical visual question answering. *CoRR*, abs/1909.11867, 2019. URL http://arxiv.org/abs/1909.11867.

Obioma Pelka, Sven Koitka, Johannes Rückert, Felix Nensa, and C. Friedrich. Radiology objects in context (roco): A multimodal image dataset. In *CVII-STENT/LABELS@MICCAI*, 2018.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, 2014. URL http://www.aclweb.org/anthology/D14-1162.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. *CoRR*, abs/2103.00020, 2021. URL https://arxiv.org/abs/2103.00020.

P. Rajpurkar, J. Irvin, A. Bagul, D. Ding, T. Duan, H. Mehta, B. Yang, K. Zhu, D. Laird, and R. L. Ball. Mura: Large dataset for abnormality detection in musculoskeletal radiographs. 2017.

Fuji Ren and Yangyang Zhou. Cgmvqa: A new classification and generative model for medical visual question answering. *IEEE Access*, 8:50626–50636, 2020. doi: 10.1109/ACCESS.2020.2980024.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. *CoRR*, abs/2112.10752, 2021. URL https://arxiv.org/abs/2112.10752.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. *CoRR*, abs/1505.04597, 2015. URL http://arxiv.org/abs/1505.04597.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael S. Bernstein, Alexander C. Berg, and Li Fei-Fei. Imagenet large scale visual recognition challenge. *CoRR*, abs/1409.0575, 2014. URL http://arxiv.org/abs/1409.0575.

Lei Shi, Feifan Liu, and Max P. Rosen. Deep multimodal learning for medical visual question answering. In *Conference and Labs of the Evaluation Forum*, 2019.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *CoRR*, abs/2010.02502, 2020a. URL https://arxiv.org/abs/2010.02502.

Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. *CoRR*, abs/1907.05600, 2019. URL http://arxiv.org/abs/1907.05600.

Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *CoRR*, abs/2011.13456, 2020b. URL https://arxiv.org/abs/2011.13456.

Bram van Ginneken. Sliver07, March 2019. URL https://doi.org/10.5281/zenodo.2597575.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *CoRR*, abs/1706.03762, 2017. URL http://arxiv.org/abs/1706.03762.

Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M. Summers. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. *CoRR*, abs/1705.02315, 2017. URL http://arxiv.org/abs/1705.02315.

Zichao Yang, Xiaodong He, Jianfeng Gao, Li Deng, and Alexander J. Smola. Stacked attention networks for image question answering. *CoRR*, abs/1511.02274, 2015. URL http://arxiv.org/abs/1511.02274.

Li-Ming Zhan, Bo Liu, Lu Fan, Jiaxin Chen, and Xiao-Ming Wu. Medical visual question answering via conditional reasoning. pages 2345–2354, 10 2020. doi: 10.1145/3394171.3413761.

Zijian Zhang, Zhou Zhao, and Zhijie Lin. Unsupervised representation learning from pre-trained diffusion probabilistic models, 2023.