PubMedCLIP: How Much Does CLIP Benefit Visual Question Answering in the Medical Domain?

Sedigheh Eslami, Christoph Meinel, Gerard de Melo

Hasso Plattner Institute / University of Potsdam

{sedigheh.eslami, christoph.meinel, gerard.demelo}@hpi.de

Abstract

Contrastive Language–Image Pre-training (CLIP) has shown remarkable success in learning with cross-modal supervision from extensive amounts of image–text pairs collected online. Thus far, the effectiveness of CLIP has been investigated primarily in general-domain multimodal problems. In this work, we evaluate the effectiveness of CLIP for the task of Medical Visual Question Answering (MedVQA). We present PubMedCLIP, a fine-tuned version of CLIP for the medical domain based on PubMed articles. Our experiments conducted on two MedVQA benchmark datasets illustrate that PubMedCLIP achieves superior results improving the overall accuracy up to 3% in comparison to the state-of-the-art Model-Agnostic Meta-Learning (MAML) networks pre-trained only on visual data. The PubMedCLIP model with different back-ends, the source code for pre-training them and reproducing our MedVQA pipeline is publicly available at https://github.com/sarahESL/PubMedCLIP.

1 Introduction

Medical visual question answering (MedVQA) seeks answers to natural language questions about a given medical image. The development of MedVQA has considerable potential to benefit healthcare systems, as it may aid clinicians in interpreting medical images and obtaining more accurate diagnoses by consulting a second opinion. Thus, it has become a very active area of research, with competitive benchmarks and yearly competitions (Abacha et al., 2021). Yet, visual question answering in the medical domain in particular remains non-trivial, as we suffer from a general lack of large balanced training data, in part due to privacy concerns. To solve the multimodal task of MedVQA, a system must understand both medical images and textual questions and infer the associations between them sufficiently well to produce a correct answer (Antol et al., 2015). Thus, the success of these solutions is tied to the effectiveness of their visual and question encoders. Current approaches for MedVQA adopt deep artificial neural network encoders to interpret the image and the question. Previous studies in MedVQA (Nguyen et al., 2019; Zhan et al., 2020; Pan et al., 2021; Gong et al., 2022) commonly exploit the Mixture of Enhanced Visual Features (MEVF) model (Nguyen et al., 2019) as their visual encoder to overcome data limitations. However, MEVF is custom-tailored for the particular challenges encountered in the VQA-RAD (Lau et al., 2018) dataset, i.e., specifically designed for the organs present in this dataset, limiting its generalizability to other settings.

In non-medical settings, recent work (Su et al., 2019; Zhang et al., 2020; Cho et al., 2021; Wang et al., 2021; Radford et al., 2021; Yu et al., 2022) has shown improvements of visual encoders when learning from multimodal image–text pairs in comparison to learning from just visual images. Among these approaches, the contrastive pre-training of language–image data in OpenAI's CLIP (Radford et al., 2021) has been particularly prominent. CLIP is trained using a vast number of image–text pairs acquired from the Internet with close to zero additional human annotation. We argue that this is particularly promising for the medical domain, since data annotation requires expert medical knowledge, making it expensive and time-consuming. Following CLIP, we investigate to what extent learning from publicly available medical image–text pairs without any further annotation can aid in the MedVQA task. To this end, we use image–text pairs obtained from PubMed articles to train a new version of CLIP called PubMedCLIP. We then examine the outcomes when incorporating PubMedCLIP into state-of-the-art MedVQA methods, investigating whether CLIP benefits MedVQA.

To the best of our knowledge, this is the first study introducing a PubMed-optimized CLIP and

Findings of the Association for Computational Linguistics: EACL 2023, pages 1181–1193

assessing the effectiveness of its visual and textual encoders for VQA. Unlike prior work on MedVQA, PubMedCLIP is trained using medical images from a diverse range of body regions and is not restricted to only a few organs. We conduct extensive experiments on two MedVQA benchmark datasets and employ diverse back-end visual encoders in PubMedCLIP. Our experiments show that using PubMedCLIP as a pre-trained visual encoder improves previous models by up to 3%. Our experiments further reveal question type distributional differences in the two MedVQA benchmark datasets that have not been imparted in previous work and cause different back-end visual encoders in PubMedCLIP to exhibit different behavior on these datasets.

2 Related Work

Shen et al. (2021) showed the benefits of CLIP for general-domain visual question answering. However, MedVQA approaches generally need to be able to learn from small amounts of training data and be able to pick up fine-granular details such as subtle medical abnormalities. Recent MedVQA approaches typically employ deep pre-trained neural encoders and consist of four main components: a visual encoder, question encoder, attention-based fusion of vision and text features, and an answer classifier (Nguyen et al., 2019; Vu et al., 2020; Zhan et al., 2020; Pan et al., 2021; Liu et al., 2021a; Gong et al., 2022). Skip-thought vectors, LSTM, and GRU recurrent neural networks have been popular question encoders in prior work. Due to the lack of diversity in the semantics of the questions in the ImageCLEF VQA-Med 2021 Challenge (Abacha et al., 2021), the winning teams (Gong et al., 2021; Eslami et al., 2021) were able to treat MedVQA as a multi-class image classification task, without any need to encode and interpret the questions. Bilinear attention networks (Kim et al., 2018), multimodal compact bilinear pooling (Fukui et al., 2016), stacked attention networks (Yang et al., 2016), and element-wise production are popular as multimodal pooling approaches in MedVQA. With regard to the visual encoder, the winning teams in the ImageCLEF VQA-Med Challenges (Abacha et al., 2020, 2021) often fine-tune an ensemble of pre-trained VGG (Simonyan and Zisserman, 2014) and various ResNet (Lei et al., 2018) encoders. A notable number of papers (Nguyen et al., 2019; Zhan et al., 2020; Pan et al., 2021; Gong et al., 2022) employ the Mixture of Enhanced Visual Features (MEVF; Nguyen et al. 2019) in order to overcome image data limitations. MEVF consists of two modules: 1. the pre-trained meta-learning module, which uses Model-Agnostic Meta-Learning (MAML; Finn et al. 2017) with the objective of solving a k-shot n-way classification problem with the abnormality status of chest, abdomen, and brain organs as classes, 2. the Convolutional Denoising Autoencoder (CDAE; Masci et al. 2011) module in order to have a robust visual encoder for noisy medical images. The pre-training of MEVF is custom-tailored for the particular organs that are present in the VQA-RAD (Lau et al., 2018) dataset, i.e., chest, brain, abdomen. Another study (Do et al., 2021) similarly trained multiple meta-models confined to these three body regions, combined with a scoring mechanism to select the n most robust and accurate encoders and concatenate their outputs to represent the visual features. Liu et al. (2021a) also restricted the objective of their visual encoding to chest, brain, and abdomen, and pre-trained three separate visual encoder teacher models for these respective body regions. They distilled the three teacher models into a smaller student model by contrastive representation distillation. As opposed to previous work, which learns from just visual data, we design an alternative encoder, PubMedCLIP, which not only uses natural language as supervision for visual representation learning, but also learns features in medical images of various modalities and diverse body organs, and hence, is not limited to only a few body regions.

3 PubMedCLIP

Our first step is to fine-tune the original general-domain CLIP using medical image–text pairs. We refer to the fine-tuned version as PubMedCLIP. Figure 1 (A) shows an overview of the training procedure for PubMedCLIP. Texts and images are encoded separately using CLIP, which we denote by $\mathbf{e}_\text{t} \in \mathbb{R}^{b\times d}$, $\mathbf{e}_\text{v} \in \mathbb{R}^{b\times d}$, respectively, for a batch of size b. For each image–text pair, a label $y \in \mathbb{R}$ represents the correspondence of the pairing of image and text. The cosine similarities between text and image features are computed to represent the respective visual and textual logits $\hat{y}_\text{v}$, $\hat{y}_\text{t}$, i.e.,

$$\hat{y}_\text{v} = \frac{\mathbf{e}_\text{v}^\intercal\, \mathbf{e}_\text{t}}{\|\mathbf{e}_\text{v}\|\,\|\mathbf{e}_\text{t}\|}, \qquad \hat{y}_\text{t} = \frac{\mathbf{e}_\text{t}^\intercal\, \mathbf{e}_\text{v}}{\|\mathbf{e}_\text{v}\|\,\|\mathbf{e}_\text{t}\|}. \quad (1)$$

As formulated in Eq. 2, a weighted sum of the vision and language loss values is computed to

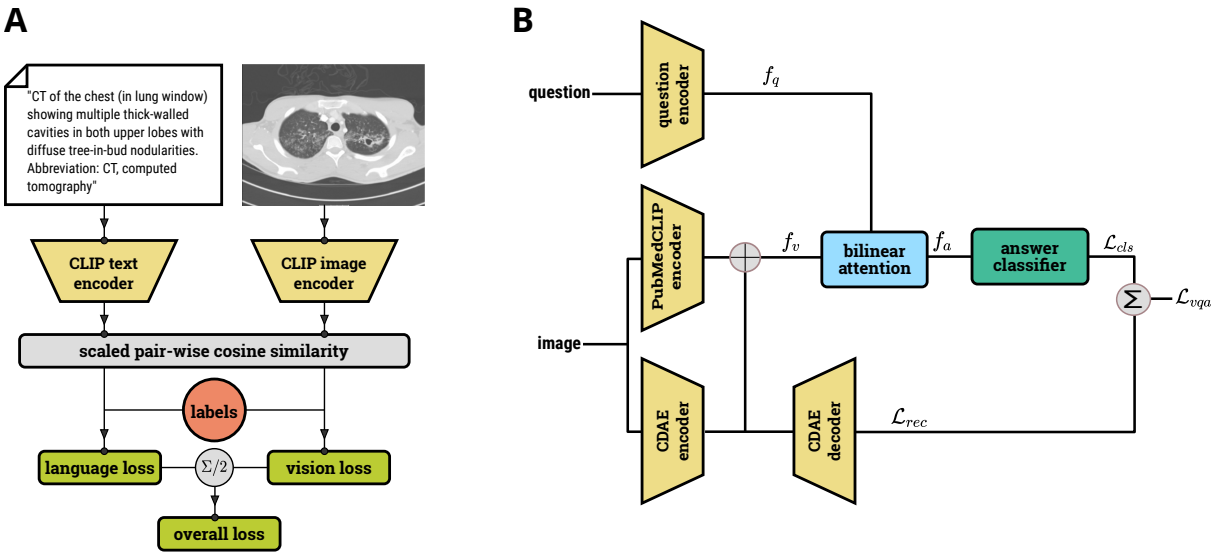

Figure 1: (A) Overview of how PubMedCLIP is pre-trained. (B) Schematic of MedVQA backbone with PubMedCLIP pre-trained visual encoder.

represent the overall loss. $Y \in \mathbb{R}^b$ denotes the set of labels y for a total of b image–text pairs in the batch. In this work, we use the cross-entropy loss

$$\mathcal{L} = \lambda H(\hat{y}_{\text{v}}, Y) + (1 - \lambda) H(\hat{y}_{\text{t}}, Y). \quad (2)$$

Following CLIP, we set $\lambda = 0.5$ to obtain the average of vision and language losses.

For training PubMedCLIP, we drew on the Radiology Objects in COntext (ROCO) dataset (Pelka et al., 2018). Previous work (Rajpurkar et al., 2017; Wang et al., 2017; Irvin et al., 2019; Johnson et al., 2019a) also proposes large-scale multimodal datasets in the medical domain. However, they include images of only one imaging modality, i.e., X-ray, for a very limited number of body regions. In contrast, ROCO includes over $80K$ samples of diverse imaging modalities such as ultrasound, X-rays, PET scans, CT scans, MRI, angiography, from various human body regions, e.g., head, neck, spine, chest, abdomen, hand, foot, knee, and pelvis. Learning visual representations of diverse organs with various imaging modalities is valuable for a MedVQA system, as it is expected to interpret images given such diversities. The image–text pairs in ROCO stem from PubMed articles. The texts are taken from the relatively short captions (average length of 20 words) associated with images in the articles, which provide rich explanatory information about the content of images. In this work, the training and validation data splits from the original paper (Pelka et al., 2018) were used to train PubMedCLIP, with ViT-B/32 Vision Transformer (Dosovitskiy et al., 2021), ResNet RN-50 (He et al., 2016), and RN-50x4 visual encoder back-ends. With respect to the maximum text length accepted by CLIP, which is 76, we trimmed any longer captions, while zero-padding shorter ones. PubMedCLIP was trained for 50 epochs with a batch size of 64, and Adam optimization (Kingma and Ba, 2014) with a learning rate of 10^{-5}. The trained models, source code as well as further implementation details are available online at https://github.com/sarahESL/PubMedCLIP.

Figures 2 and 3 show PCA visualizations of the caption and image embeddings, respectively, for the ROCO validation set. Comparing CLIP and PubMedCLIP embeddings, PubMedCLIP appears to obtain more semantic-aware visual and textual features with regard to body locations. For instance, looking at chest, abdomen, and head body locations, the corresponding embeddings form clusters for PubMedCLIP. However, the original CLIP embeddings are scattered without much separation.[1]

4 PubMedCLIP for MedVQA

Given a MedVQA training dataset represented as $T = \{(v_i, q_i, a_i)\}_{i=1}^{D}$ of size D, where v_i is a medical image, q_i is the corresponding natural language question, and a_i is natural language answer, our goal is to learn to emit correct answer a_i given

[1]In Appendix A, we provide more information on our approach for proxy-labeling the unannotated captions from the ROCO dataset. The proxy-labels have been merely used for the purpose of visualisations in this paper.

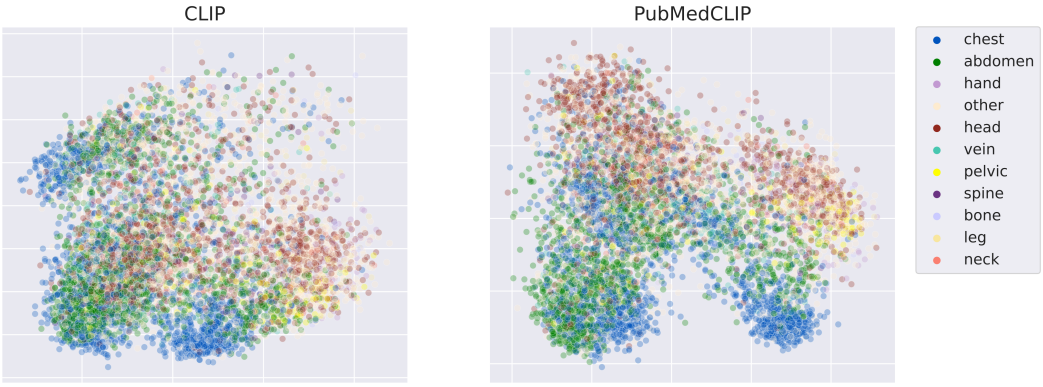

Figure 2: PCA visualizations of image embeddings.

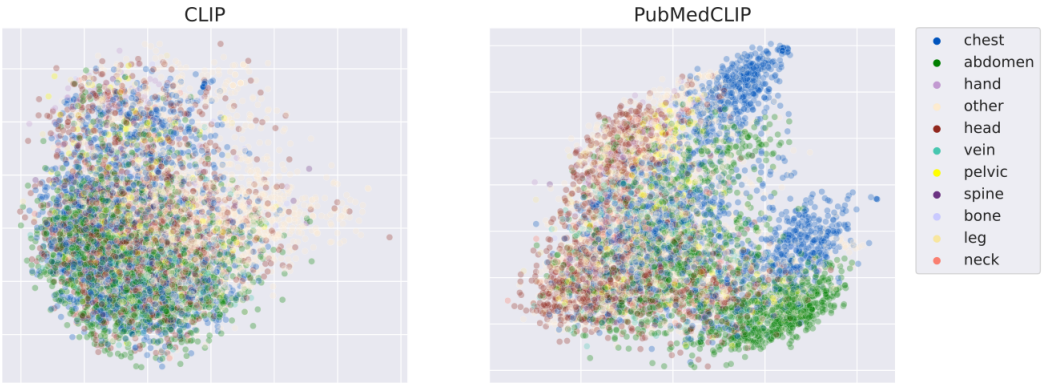

Figure 3: PCA visualizations of text embeddings.

image–question pair (v_i, q_i). For this, we assume appropriate encoding functions to obtain $\mathbf{f}_v \in \mathbb{R}^n$ as an n-dimensional vector encoding for image v_i and the sequence embedding of $\mathbf{f}_q \in \mathbb{R}^{m \times l}$ for the question q_i with length l. We then cast MedVQA as a multi-label classification function $F : \mathbb{R}^n \times \mathbb{R}^{m \times l} \longrightarrow \{0, 1\}^{|A|}$ where A is the overall set of possible answers and $F(\mathbf{f}_v, \mathbf{f}_q) = \mathbf{a}_i$ for the one-hot encoded answer $\mathbf{a}_i$.

Our goal is to investigate the effect of employing PubMedCLIP as the pre-trained visual encoder in MedVQA. To this end, we considered two prominent MedVQA methods, MEVF (Zhan et al., 2020) and QCR (Nguyen et al., 2019), that adopt MAML as their pre-trained visual encoder and GloVe word embeddings followed by a Recurrent Neural Network (RNN) as their question encoder. We substitute the pre-trained MAML module in MEVF and QCR with the pre-trained visual encoder from PubMedCLIP. A schematic architecture of our pipeline is shown in Figure 1 (B). The representative visual feature $\mathbf{f}_v$ in this solution is the concatenation of the outputs of the PubMedCLIP network and the CDAE encoder. The objective of CDAE's encoder is to robustly encode the noisy version v_i' of an image v_i while the decoder learns to reconstruct the original non-noisy images. Denoting the reconstructed image as v_i^{rec}, Equation 3 defines the image reconstruction loss of CDAE as the mean squared error.

$$\mathcal{L}_{\text{rec}} = \|v_i - v_i^{\text{rec}}\|^2 \tag{3}$$

The multimodal pooling mechanism for combining $\mathbf{f}_v$ and $\mathbf{f}_q$ is BAN (Kim et al., 2018) to obtain the answer feature vector $\mathbf{f}_a$, as illustrated in Figure 1 (B). For answer prediction, which is a classification task in our case, a sigmoid layer preceding a binary cross-entropy loss is utilized in order to allow multiple correct answers per question. Eq. 4 formulates the answer classification loss function.

$$\begin{aligned}\mathcal{L}_{\text{cls}} = -\frac{1}{D}\sum_{i=1}^{D}\sum_{c=1}^{A} a_{i,c}\log(\hat{a}_{i,c}) \\ +(1 - a_{i,c})\log(1 - \hat{a}_{i,c})\end{aligned} \tag{4}$$

Here, $\hat{a}_{i,c} = \sigma(\mathcal{M}(\mathbf{f}_a))$, where σ represents the sigmoid function. Following BAN (Kim et al., 2018), the answer classifier $\mathcal{M}$ is a two-layer feed-forward network with ReLU activation.

The objective of MedVQA is to simultaneously minimize the error of answer classification and image reconstruction, denoted as:

$$\mathcal{L}_{\text{vqa}} = \mathcal{L}_{\text{cls}} + \mathcal{L}_{\text{rec}}. \quad (5)$$

5 Experiments

5.1 Datasets and Setup

We conducted our experiments using two well-known MedVQA datasets:

1. **VQA-RAD** (Lau et al., 2018) consists of 315 images and 3,515 English language question–answer pairs. Following previous work, we adopt the data split proposed in MEVF (Nguyen et al., 2019). We notice that all the images in the test dataset are also present in the training set. However, the set of question–answer pairs for these images in the test set are unseen in the training set.

2. The **SLAKE** (Liu et al., 2021b) dataset consists of English and Chinese questions. In this work, we utilize the English subset of the dataset, comprising 642 images and more than 7,000 question–answer pairs. Using the original data split, we observe that in contrast to VQA-RAD, all the images in the test set of SLAKE are unseen in the training set.

To ensure a fair comparison, our experiments followed the same setups used in the original MEVF and QCR studies. MEVF was trained for 20 epochs, QCR for 200, both with Adam optimization. When using PubMedCLIP as either the pre-trained visual encoder or the text encoder, we set the learning rate to 1×10^{-3} and 2×10^{-3} and the batch size to 16 and 32 in QCR and MEVF, respectively. All implementations are based on the PyTorch framework (Paszke et al., 2019). We ran the original MEVF and QCR on our machine and report the results here to have a fair comparison. Due to the non-deterministic behaviour of the cuDNN library in CUDA convolution operations (Pham et al., 2020), we observed non-deterministic results in different runs of the original MEVF and QCR. For a more robust comparison, we repeated all experiments 10 times and report the average accuracy scores.

5.2 Results and Analysis

The results of our experiments using PubMedCLIP's visual encoder are given in Table 1. In order to see the effectiveness of PubMedCLIP in comparison to the general domain CLIP, we also report the results when using CLIP. We provide the overall accuracy along with the accuracy of answering only open-ended or closed-ended questions.

When using CLIP and PubMedCLIP as the pre-trained visual encoder only, it is observed that the results of both the MEVF and QCR approaches improve. Furthermore, PubMedCLIP yields an absolute improvement of up to 1% in comparison with the original CLIP. On the VQA-RAD dataset, PubMedCLIP with the ResNet-50 backend achieves the best results, improving the overall accuracy of MEVF up to 6% and for QCR up to 3%. Results on the SLAKE dataset show that PubMedCLIP with ViT-B/32 Vision Transformer encoder back-end attains the best accuracy. It enhances MEVF by up to 3% and QCR up to 2%. We witness the same trend of improvement among overall, open-ended, and closed-ended accuracy scores.

In Figure 4, a comparison of image embeddings when using MAML as apposed to PubMedCLIP's visual encoder is shown using PCA analysis for the VQA-RAD dataset. We find that in contrast to the MAML encoder, PubMedCLIP's visual encoding results in organ-aware visual embeddings i.e., images of head, chest, and abdomen form more coherent and distinct clusters.

In Table 2, we compare the performance of PubMedCLIP with the recent state-of-the-art models in MedVQA. All the models use BAN as the fusion mechanism. In Table 2, PubMedCLIP refers to using PubMedCLIP as the pre-trained visual encoder in QCR. The comparison shows that PubMedCLIP achieves the best results on open-ended, closed-ended, and overall accuracies.

Behavior of visual encoder back-ends. The fact that PubMedCLIP with ResNet-50 back-end achieves the best results for VQA-RAD, while PubMedCLIP with ViT performs best on the SLAKE dataset points us to underlying differences in the question type distribution in these datasets. As Figure 5 shows, the majority of the questions in the VQA-RAD ask about the presence of an abnormality in the images. This requires the visual encoder to detect local features and local abnormalities. Thus, the CNN-based ResNet model with better visual localization outperforms the Vision Transformer. However, on SLAKE, the majority of questions are of the type "organ", asking which organ is present in the image. For such cases, the

MedVQA Model	Question Encoder	Visual Encoder	VQA-RAD Accuracy			SLAKE Accuracy		
			Open	Closed	Overall	Open	Closed	Overall
MEVF	GloVe+RNN	MAML + AE (*)	42.1%	73.2%	60.8%	74.1%	77.5%	75.5%
		CLIP-ViT-B + AE	50.8%	75%	65.4%	75.8%	80.5%	77.7%
		CLIP-RN50 + AE	47%	77.4%	65.4%	75.7%	79.6%	77.2%
		CLIP-RN50x4 + AE	46.8%	76.6%	64.8%	75.9%	79.1%	77.2%
		PubMedCLIP-ViT-B + AE	48.9%	76.7%	65.5%	**76.5%**	**80.4%**	**78%**
		PubMedCLIP-RN50 + AE	**48.6%**	**78.1%**	**66.5%**	76.2%	79.9%	77.6%
		PubMedCLIP-RN50x4 + AE	47.1%	77.8%	65.6%	76.6%	79.1%	77.6%
QCR	GloVe+RNN	MAML + AE (+)	56%	77.9%	69.2%	76.8%	80.6%	78.3%
		CLIP-ViT-B + AE	57.6%	79.5%	70.7%	78.6%	81%	79.5%
		CLIP-RN50 + AE	58.3%	80%	71.3%	78.2%	81.5%	79.7%
		CLIP-RN50x4 + AE	59.9%	79.4%	71.3%	77.6%	80.5%	78.7%
		PubMedCLIP-ViT-B + AE	58.4%	79.5%	71.1%	**78.4%**	**82.5%**	**80.1%**
		PubMedCLIP-RN50 + AE	**60.1%**	**80%**	**72.1%**	77.8%	81.4%	79.3%
		PubMedCLIP-RN50x4 + AE	60%	79.7%	71.8%	77.7%	81.3%	79.1%

Table 1: Accuracy scores on VQA-RAD and SLAKE datasets. (*) denotes the original MEVF (Nguyen et al., 2019) and (+) denotes the original QCR (Zhan et al., 2020). Bold numbers represent the rows that achieved best overall accuracy. Light cyan, yellow, and green highlight correspond to the results when using MAML, CLIP and PubMedCLIP as the visual encoder only, respectively.

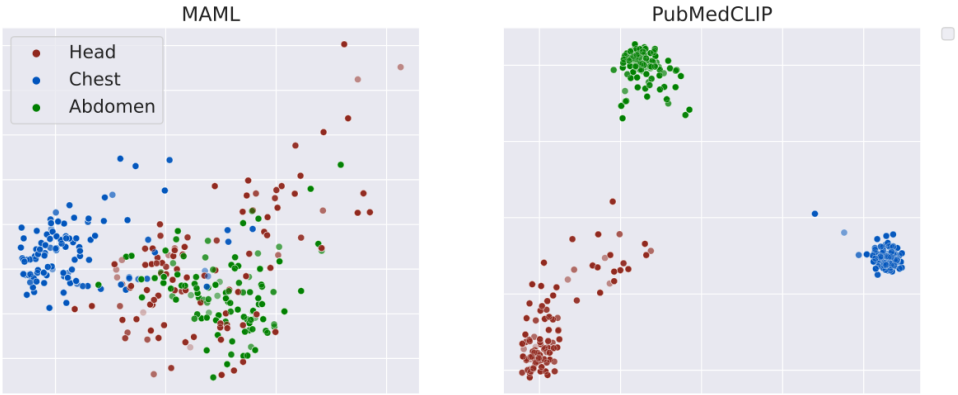

Figure 4: PCA visualizations of MAML and PubMedCLIP image embeddings for VQA-RAD dataset.

MedVQA Model	VQA-RAD Accuracy			SLAKE Accuracy		
	Open	Closed	Overall	Open	Closed	Overall
MEVF (Nguyen et al., 2019)	42.1%	73.2%	60.8%	74.1%	77.5%	75.5%
QCR (Zhan et al., 2020)	56%	77.9%	69.2%	76.8%	80.6%	78.3%
MMQ (Do et al., 2021)	53.7%	75.8%	67%	—	—	—
VQAMix (Gong et al., 2022)	56.6%	79.6%	70.4%	—	—	—
PubMedCLIP + BAN (ours)	**60.1%**	**80%**	**72.1%**	**78.4%**	**82.5%**	**80.1%**

Table 2: Comparison of PubMedCLIP with state-of-the-art MedVQA models. Results for the SLAKE dataset are not reported in the MMQ and VQAMix papers.

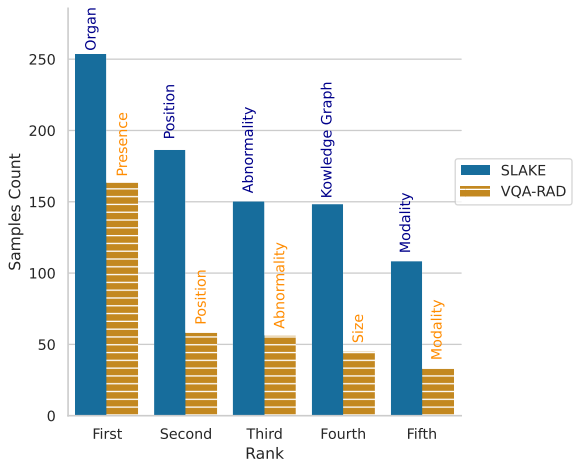

Figure 5: Distribution of top 5 most frequent question types in VQA-RAD and SLAKE.

visual encoder needs to be able to acquire a holistic overall understanding of the image and thus capture long-range dependencies of image patches. Vision Transformers indeed are capable of accounting for such features (Yu et al., 2021), and hence perform better on the SLAKE dataset.

PubMedCLIP as the text encoder. We expanded our experiments to investigate the effects of PubMedCLIP's text encoder in MedVQA. To this end, we replaced the question encoder in the MEVF model with PubMedCLIP's text encoder, i.e., instead of using GloVe word embeddings and an RNN network to model the question, we use PubMedCLIP's text tokenizer and encoder, which receives the question q_i with l words and outputs a sequence-level embedding $\mathbf{f}_q \in \mathbb{R}^m$. Note that the size of image and text embeddings when using PubMedCLIP is equal. The results of our experiments in Table 3 suggest that invoking PubMedCLIP to encode questions in MedVQA is not as successful as using it for images. Furthermore, Table 3 shows that using both the visual and textual encoders of PubMedCLIP achieves absolute improvements of up to 5% in comparison to the original MEVF model. However, the best results are achieved with PubMedCLIP as the visual encoder together with GloVe+RNN for encoding questions.

In order to have a better understanding of the PubMedCLIP's text encoder, a PCA visualization of the question embeddings is provided in Figure 6. The top row shows the embeddings when annotated according to their respective body location and the bottom row depicts them when labeled with question types, i.e., whether the question asks

Visual Encoder	Question Encoder	VQA-RAD Accuracy		
		Open	Closed	Overall
MAML	GloVe+RNN(*)	42.1%	73.2%	60.8%
	PubMedCLIP	26.5%	72.9%	54.3%
PubMedCLIP	GloVe+RNN	**48.6%**	**78.1%**	**66.5%**
	PubMedCLIP	48%	77.4%	65.6%

Table 3: Accuracy of PubMedCLIP as text encoder in the MEVF model. (*) denotes the original MEVF.

about the *Presence* of abnormality, *Position* of abnormality, type of *Abnormality*, etc. For having a comprehensible analysis, we visualize the top five frequent question types shown in Figure 5. Observations from Figure 6 suggest that PubMedCLIP's text encoder emits organ-aware textual embeddings in contrast to GloVe+RNN. However, PubMedCLIP does not separate embeddings based on the question type, while GloVe+RNN results in better question type clusters. These findings suggest that question type awareness when encoding questions might be more beneficial than organ awareness for the MedVQA task. Based on our experiments, exploiting PubMedCLIP as the visual encoder in the QCR model is the most effective solution.

Furthermore, we sampled a few questions from the VQA-RAD test set and compared their pairwise cosine similarities when using GloVe+RNN versus PubMedCLIP encoding. We seek to examine the power of PubMedCLIP text encoder in identifying semantic differences. Figure 7 reports the cosine similarities when using PubMedCLIP in contrast to GloVe+RNN embeddings. As can be seen, when using PubMedCLIP text encoder, different questions about "lung abnormality" and "image plane" are equally similar to the "rib fracture" question, i.e., 0.77, and the encoder does not distinguish them. However, the cosine similarities are more intuitive when using GloVe+RNN. For instance, questions "Is there a rib fracture?" and "Describe the lung abnormalities?" have a small similarity of 0.27, while questions "Which plane is this image taken?" and "What is the plane of this image?" have a high similarity of 0.86.

In addition, it is observed that PubMedCLIP generally results in embeddings that are highly close to each other, with cosine similarities of more than 0.7 for different questions on disparate topics. In contrast, similarities of GloVe+RNN encoding are spread in the range of $[-0.09, 1]$, meaning that these embeddings are scattered over the m-

Figure 6: PCA of question embeddings. (Top) Labeled with body locations. (Bottom) Labeled with question types.

dimensional embedding space. We conclude that GloVe+RNN distinguishes the semantics of questions more effectively in comparison to PubMedCLIP's text encoder for the MedVQA task.

CLIP versus PubMedCLIP. In order to better see the impact of fine-tuning PubMedCLIP, we additionally looked into the intermediate task of image–text matching using nearest neighbors vector retrieval. Considering that the pre-training objective in CLIP and PubMedCLIP is to minimize the cosine distance between paired image and text embeddings while maximizing this distance for non-paired image–text combinations, we argue that with a rich representation learning model, a nearest neighbor approach using the cosine distance metric should be fairly successful in retrieving matching image–text pairs. We randomly selected a subset of $D' = 10,000$ samples from the ROCO training data and used them to compare the outcomes of image–text matching in the medical domain. We exploit the text encoder as well as the visual encoder in CLIP and PubMedCLIP. Using Faiss (Johnson et al., 2019b) for vector retrieval, we investigated KNN with $K = 1$ on batches of size b. For each batch, the objective was to find the closest encoded text for a given encoded image, using the cosine distance metric. The evaluation metric for this setting is the overall accuracy of image–text matching over all batches:

$$\text{acc} = \frac{\sum_{i=1}^{S} \#\text{ correct matches in batch } i}{D'}, \quad (6)$$

V-L encoder	Batch size	ViT-B/32	RN50	RN50x4
CLIP	8	58.1%	49.1%	57.7%
	16	44%	36.1%	45.1%
	32	21.6%	25.5%	33.1%
PubMedCLIP	8	**93.1%**	**89.2%**	**92.2%**
	16	87.6%	81.1%	85.7%
	32	80.1%	70.6%	76.2%

Table 4: Accuracy scores of image-text matching using CLIP and PubMedCLIP vision–language encoders.

where $S = \lceil \frac{D'}{b} \rceil$. Table 4 summarizes the results for batch sizes of 8, 16, and 32. PubMedCLIP achieves over 40% improvement in comparison to CLIP across all batch sizes, with the ViT-B/32 back-end achieving the best results. This shows the effectiveness of our fine-tuning in PubMedCLIP.

Comparison of qualitative examples. In Figure 8, examples from the VQA-RAD and SLAKE datasets are provided that illustrate the performance of the original MEVF and QCR in comparison with PubMedCLIP, used here as either the visual or question encoder for QCR. PubMedCLIP_TE_VE, PubMedCLIP_TE and PubMedCLIP_VE refer to the scenarios of PubMedCLIP as both visual and textual encoders, as textual encoder only, and as the visual encoder only, respectively.

We find that the MEVF model often has difficulties discerning which organ is depicted in the image. For instance, regardless of the asked ques-

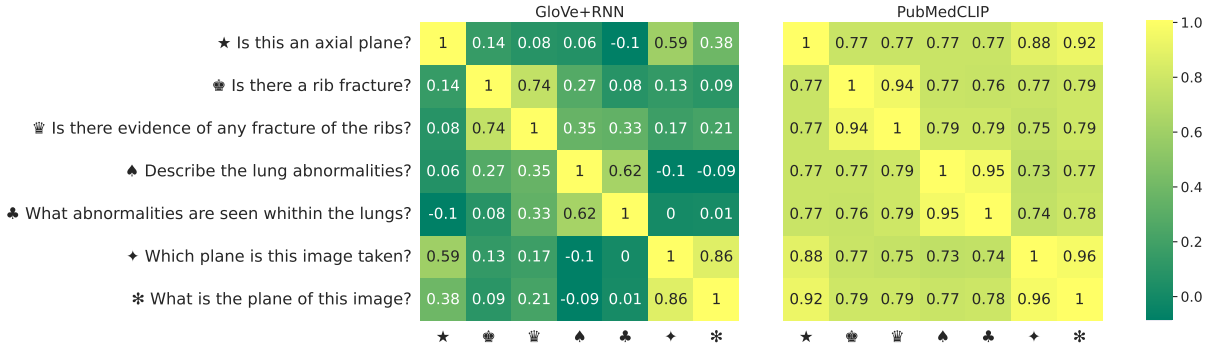

Figure 7: Pair-wise cosine similarities of questions from VQA-RAD encoded with GloVe+RNN compared with PubMedCLIP. Each question is associated with a symbol and represented only by the symbol on the horizontal axis.

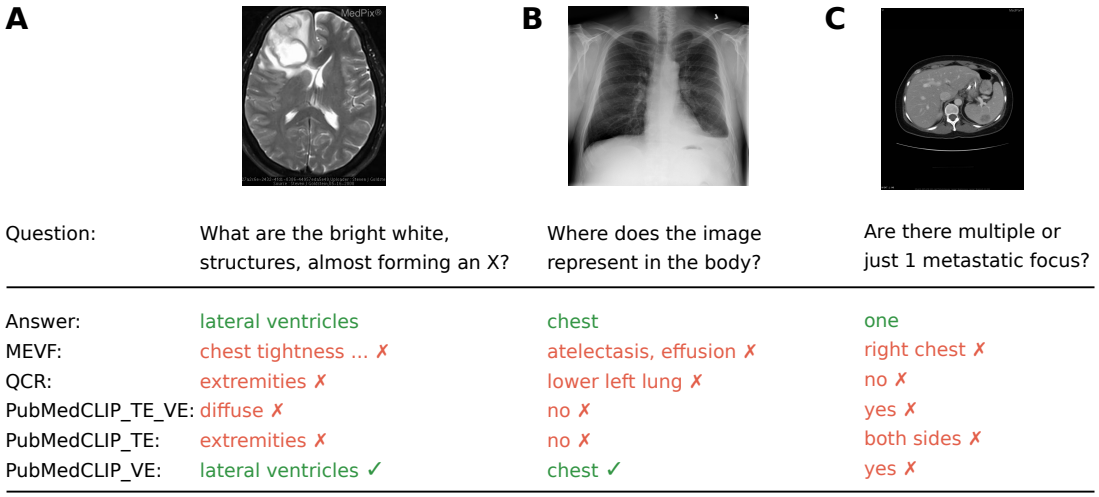

Figure 8: (A) Example from VQA-RAD dataset. (B) Example from SLAKE dataset. (C) Example from VQA-RAD dataset that all models fail to answer correctly.

tion in Figure 8 (A), MEVF provides an answer related to the chest region, while the image is of the brain. This behaviour is also seen in Figure 8 (B) and 8 (C). From this perspective, QCR appears to be providing answers that are at least relevant to the given image. As Figure 8 (B) shows, the answer provided by QCR is related to the chest X-ray, although it is not a correct answer. Furthermore, it is observed that when PubMedCLIP is used as the question encoder, the model has difficulties providing the correct answers and often misinterprets open-ended questions as close-ended. In contrast, PubMedCLIP as the visual encoder successfully yields the correct answers.

Figure 8 (C) shows an example from the VQA-RAD that all models fail to answer correctly. MEVF again provides irrelevant answers about body organs not present in the image. QCR and PubMedCLIP misinterpret the question as a yes/no one. In spite of this, the fact that PubMedCLIP_VE answers with “yes" may illustrate that it has at least detected the “one" metastatic focus in the image. In comparison, QCR answers with “no", showing its troubles in interpreting the image and recognizing the metastatic focus. Figure 8 (C) reveals that these models still have shortcomings in understanding questions and correctly relating them to the images.

6 Conclusion

This work introduces PubMedCLIP, a pre-trained vision–language encoder for the medical domain trained via contrastive learning of medical image–caption pairs from PubMed articles. We demonstrated that PubMedCLIP results in organ-aware vision and language embeddings and evaluated its effectiveness for the task of MedVQA in comprehensive experiments across two heterogeneous MedVQA benchmarks. While PubMedCLIP’s text encoder is found to be less powerful for MedVQA, we showed that PubMedCLIP’s visual encoder outperforms previously used pre-trained visual encoders by up to 3%, leading to state-of-the-art results.

Limitations

Although we envision that in the long term, MedVQA systems can be sufficiently successful and trustworthy to aid medical practitioners towards better interpreting medical images and providing better healthcare, we emphasize that the development of these systems is still in its infancy stage and they are not yet ready for fully automated and unsupervised use in real-world clinical settings. Despite the notable improvement of accuracy in MedVQA brought by PubMedCLIP, further evaluations of these models from the vantage points of scalability, trustworthiness, explainability, and generalizability are required before they can be deployed for sensitive clinical tasks. In future work, we plan to perform further analysis of these models using explainable AI techniques such as Grad-CAM visualizations to assess the regions of focus within the image from the class activation maps. Furthermore, due to a lack of suitable data to train large-scale models for other languages, our current experiments are limited to English language MedVQA, so different findings may be observed for typologically different languages. By releasing PubMedCLIP, we hope to enable further research investigating these aspects as well as its effectiveness in other use cases, e.g., image classification for medical diagnosis and radiology report generation.

Discussions on Ethics

As remarked above, MedVQA models are still in their early stages of development and have limitations that should be considered before being used in any real-world scenarios.

Acknowledgements

The authors acknowledge the financial support by the German Federal Ministry for Education and Research (BMBF) within the project »KI-Servicezentrum Berlin Brandenburg« 01IS22092.

References

Asma Ben Abacha, Vivek V Datla, Sadid A Hasan, Dina Demner-Fushman, and Henning Müller. 2020. Overview of the VQA-Med task at ImageCLEF 2020: Visual question answering and generation in the medical domain. In *CLEF (Working Notes)*.

Asma Ben Abacha, Mourad Sarrouti, Dina Demner-Fushman, Sadid A Hasan, and Henning Müller. 2021. Overview of the VQA-Med task at ImageCLEF 2021: Visual question answering and generation in the medical domain. In *CLEF (Working Notes)*.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. VQA: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433.

Jaemin Cho, Jie Lei, Hao Tan, and Mohit Bansal. 2021. Unifying vision-and-language tasks via text generation. In *International Conference on Machine Learning*, pages 1931–1942. PMLR.

Tuong Do, Binh X Nguyen, Erman Tjiputra, Minh Tran, Quang D Tran, and Anh Nguyen. 2021. Multiple meta-model quantifying for medical visual question answering. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 64–74. Springer.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*.

Sedigheh Eslami, Gerard de Melo, and Christoph Meinel. 2021. TeamS at VQA-Med 2021: BBN-Orchestra for long-tailed medical visual question answering. In *Working Notes of CLEF 2021*, number 2936 in CEUR Workshop Proceedings, pages 1211–1217. CEUR-WS.

Chelsea Finn, Pieter Abbeel, and Sergey Levine. 2017. Model-agnostic meta-learning for fast adaptation of deep networks. In *International Conference on Machine Learning*, pages 1126–1135. PMLR.

Akira Fukui, Dong Huk Park, Daylen Yang, Anna Rohrbach, Trevor Darrell, and Marcus Rohrbach. 2016. Multimodal compact bilinear pooling for visual question answering and visual grounding. In *Conference on Empirical Methods in Natural Language Processing*, pages 457–468. ACL.

Haifan Gong, Guanqi Chen, Mingzhi Mao, Zhen Li, and Guanbin Li. 2022. VQAMix: Conditional triplet mixup for medical visual question answering. *IEEE Trans. on Medical Imaging*.

Haifan Gong, Ricong Huang, Guanqi Chen, and Guanbin Li. 2021. SYSU-HCP at VQA-Med 2021: A data-centric model with efficient training methodology for medical visual question answering. In *CLEF 2021 Working Notes*, volume 201. CEUR-WS.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778.

Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, et al. 2019. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 590–597.

Alistair EW Johnson, Tom J Pollard, Seth J Berkowitz, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Roger G Mark, and Steven Horng. 2019a. MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific data*, 6(1):1–8.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2019b. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 7(3):535–547.

Jin-Hwa Kim, Jaehyun Jun, and Byoung-Tak Zhang. 2018. Bilinear attention networks. *Advances in neural information processing systems*, 31.

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

Jason J Lau, Soumya Gayen, Asma Ben Abacha, and Dina Demner-Fushman. 2018. A dataset of clinically generated visual questions and answers about radiology images. *Scientific data*, 5(1):1–10.

Li Lei, Haogang Zhu, Yuxin Gong, and Qian Cheng. 2018. A deep residual networks classification algorithm of fetal heart CT images. In *2018 IEEE international conference on imaging systems and techniques (IST)*, pages 1–4. IEEE.

Bo Liu, Li-Ming Zhan, and Xiao-Ming Wu. 2021a. Contrastive pre-training and representation distillation for medical visual question answering based on radiology images. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 210–220. Springer.

Bo Liu, Li-Ming Zhan, Li Xu, Lin Ma, Yan Yang, and Xiao-Ming Wu. 2021b. SLAKE: A semantically-labeled knowledge-enhanced dataset for medical visual question answering. In *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, pages 1650–1654. IEEE.

Jonathan Masci, Ueli Meier, Dan Cireşan, and Jürgen Schmidhuber. 2011. Stacked convolutional auto-encoders for hierarchical feature extraction. In *International conference on artificial neural networks*, pages 52–59. Springer.

Binh D Nguyen, Thanh-Toan Do, Binh X Nguyen, Tuong Do, Erman Tjiputra, and Quang D Tran. 2019. Overcoming data limitation in medical visual question answering. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 522–530. Springer.

Haiwei Pan, Shuning He, Kejia Zhang, Bo Qu, Chunling Chen, and Kun Shi. 2021. MuVAM: A multi-view attention-based model for medical visual question answering. *arXiv preprint arXiv:2107.03216*.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32.

Obioma Pelka, Sven Koitka, Johannes Rückert, Felix Nensa, and Christoph M Friedrich. 2018. Radiology Objects in COntext (ROCO): a multimodal image dataset. In *Intravascular Imaging and Computer Assisted Stenting and Large-Scale Annotation of Biomedical Data and Expert Label Synthesis*, pages 180–189. Springer.

Hung Viet Pham, Shangshu Qian, Jiannan Wang, Thibaud Lutellier, Jonathan Rosenthal, Lin Tan, Yaoliang Yu, and Nachiappan Nagappan. 2020. Problems and opportunities in training deep learning software systems: an analysis of variance. In *Proceedings of the 35th IEEE/ACM International Conference on Automated Software Engineering*, pages 771–783.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR.

Pranav Rajpurkar, Jeremy Irvin, Aarti Bagul, Daisy Ding, Tony Duan, Hershel Mehta, Brandon Yang, Kaylie Zhu, Dillon Laird, Robyn L Ball, et al. 2017. Mura: Large dataset for abnormality detection in musculoskeletal radiographs. *arXiv preprint arXiv:1712.06957*.

Sheng Shen, Liunian Harold Li, Hao Tan, Mohit Bansal, Anna Rohrbach, Kai-Wei Chang, Zhewei Yao, and Kurt Keutzer. 2021. How much can CLIP benefit vision-and-language tasks? In *International Conference on Learning Representations*.

Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556.

Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. 2019. VL-BERT: Pre-training of generic visual-linguistic representations. In *International Conference on Learning Representations*.

Minh H Vu, Tommy Löfstedt, Tufve Nyholm, and Raphael Sznitman. 2020. A question-centric model for visual question answering in medical imaging. *IEEE transactions on Medical Imaging*, 39(9):2856–2868.

Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M Summers. 2017. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2097–2106.

Zirui Wang, Jiahui Yu, Adams Wei Yu, Zihang Dai, Yulia Tsvetkov, and Yuan Cao. 2021. SimVLM: Simple visual language model pretraining with weak supervision. In *International Conference on Learning Representations*.

Zichao Yang, Xiaodong He, Jianfeng Gao, Li Deng, and Alex Smola. 2016. Stacked attention networks for image question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 21–29.

Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. 2022. CoCa: Contrastive captioners are image-text foundation models. *arXiv preprint arXiv:2205.01917*.

Qihang Yu, Yingda Xia, Yutong Bai, Yongyi Lu, Alan L Yuille, and Wei Shen. 2021. Glance-and-gaze vision transformer. *Advances in Neural Information Processing Systems*, 34:12992–13003.

Li-Ming Zhan, Bo Liu, Lu Fan, Jiaxin Chen, and Xiao-Ming Wu. 2020. Medical visual question answering via conditional reasoning. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 2345–2354.

Yuhao Zhang, Hang Jiang, Yasuhide Miura, Christopher D Manning, and Curtis P Langlotz. 2020. Contrastive learning of medical visual representations from paired images and text. *arXiv preprint arXiv:2010.00747*.

A Proxy-labeling ROCO dataset for visualization purposes

In order to have a better analysis of the PCA visualizations when comparing CLIP and PubMedCLIP encodings, we created proxy body location labels by identifying organ-specific keywords in ROCO captions. The complete list of keywords used for each body location is provided in Listing 1. Furthermore, the distribution of these proxy labels in the ROCO validation dataset is shown in Figure 9.

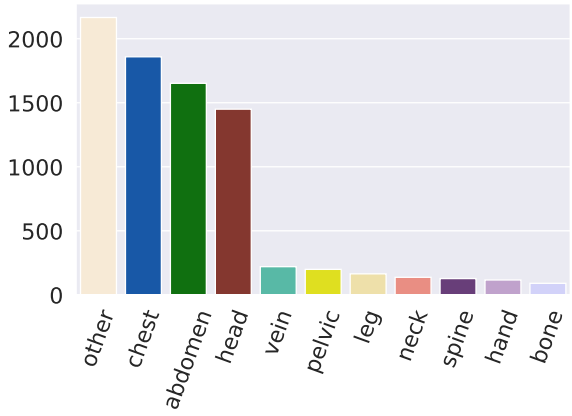

Figure 9: Distribution of proxy labels in ROCO.

```
chest = ['breast', 'lung', 'rib', 'thoracotomy', 'pulmonary', 'mediastinal',
         'bronchus', 'bronchoscopic', 'bronchiectasis', 'bronchial',
         'tuberculosis', 'heart', 'ventricle', 'myocardial', 'valve',
         'thorax', 'thoracic', 'echocardiogram', 'echocardiography',
         'angioplasty', 'diaphragm', 'coronary', 'cardiac', 'coronaries',
         'thoracique', 'chest', 'mitral annulus', 'empyema']
######################
abdomen = ['gastro-oesophageal', 'gastrointestinal', 'gastric',
           'abdomen', 'abdomenal', 'abdominal', 'bowel', 'colon', 'liver',
           'kidney', 'renal', 'stomach', 'ventral', 'esophagus', 'pancreas',
           'pancreatic', 'pancreatitis', 'hernia', 'bladder', 'gallstones',
           'gallbladder', 'spleen', 'splenic', 'appendi', 'intestine',
           'duodenum', 'ileum', 'jejunum', 'rectum', 'ovary', 'uterus',
           'vagina', 'cervix', 'pregnancy', 'cervical', 'prostate', 'penis',
           'testicle', 'testis', 'testicular', 'urethrogram', 'urethra',
           'ureteral', 'ureter', 'peritoneum']
######################
head = ['head', 'skullbase', 'skull', 'zygoma', 'parieto-occipital',
        'parietooccipital', 'parieto occipital', 'cerebellar', 'cerebellum',
        'brain', 'caudate nucleus', 'caudate', 'ear', 'auditory canal',
        'facial', 'eye', 'sinus', 'gland', 'temporal lobe', 'frontal lobe',
        'frontal bone', 'parietal bone', 'parietal lobe', 'occipital lobe',
        'lymph', 'nose', 'nasal', 'mouth', 'tongue', 'cheek', 'jaw',
        'root canal', 'tooth', 'teeth', 'obturation', 'periapical', 'premolars',
        'dental', 'parotid', 'orthopantomograph', 'orthopantomogram',
        'myelinolysis']
######################
neck = ['neck', 'throat', 'theroid', 'thyroid', 'carotid']
######################
spine = ['foraminal', 'spine', 'disk', 'disc', 'spinal', 'lumbosacral',
         'thoracic spine', 'lubmar']
######################
pelvic = ['pelvic', 'pelvis', 'hip', 'perineum', 'iliac', 'gluteal']
######################
hand = ['arm', 'shoulder', 'elbow', 'wrist', 'hand', 'nail', 'finger',
        'humerus', 'thumb']
######################
leg = ['tibias', 'leg', 'thigh', 'foot', 'feet', 'talus', 'toe', 'knee',
       'calcaneus', 'fibula', 'femur', 'femoral', 'femural', 'prosthesis',
       'prostheses', 'limb']
######################
vein = ['vein', 'vessel', 'vascular', 'artery', 'angioplasty', 'angiography',
        'artial',  'aorta', 'aortogram']
######################
bone = ['bone']
```

Listing 1: Proxy-label keywords