Contrastive Pre-training and Representation Distillation for Medical Visual Question Answering Based on Radiology Images

Bo Liu, Li-Ming Zhan, and Xiao-Ming Wu

Department of Computing, The Hong Kong Polytechnic University, Hong Kong
{csbliu,cslmzhan,csxmwu}@comp.polyu.edu.hk

Abstract. One of the primary challenges facing medical visual question answering (Med-VQA) is the lack of large-scale well-annotated datasets for training. To overcome this challenge, this paper proposes a two-stage pre-training framework by learning transferable feature representations of radiology images and distilling a lightweight visual feature extractor for Med-VQA. Specifically, we 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. Then, we distill the teacher models to a lightweight student model that can be used as a universal visual feature extractor for any Med-VQA system. The lightweight feature extractor can be readily fine-tuned on the training radiology images of any Med-VQA dataset, saving the annotation effort while preventing overfitting to small-scale training data. The effectiveness and advantages of the pre-trained model are demonstrated by extensive experiments with state-of-the-art Med-VQA methods on existing benchmarks. The source code and the pre-training dataset can be downloaded from https://github.com/awenbocc/cprd.

Keywords: Medical visual question answering · Contrastive learning · Representation distillation · Model compression

1 Introduction

Medical visual question answering (Med-VQA) has gained increasing attention over the past few years. Given a medical image and a clinical question about the image, it aims to find the correct answer by analyzing the visual information of the image. Med-VQA technology has great potential in medical and healthcare services. It can be used for computer-assisted diagnosis, intelligent medical guidance, clinical education and training, etc., which can help to significantly improve the quality of medical services and meet the increasing demand of the general public for medical resources.

While recent breakthroughs in image recognition and natural language processing have laid the foundation for the development of Med-VQA systems, the research progress of Med-VQA is impeded by the absence of large-scale well-annotated training datasets. The visual feature extraction module of existing

Med-VQA models usually employs deep architectures and needs to be trained on a large collection of annotated radiology images, which however are often unavailable and costly to collect. To address this issue, a pioneering work [17] proposes mixture of enhanced visual features (MEVF) to pre-train the visual feature extraction module by constructing an auxiliary organ disease classification task on the radiology images of VQA-RAD[13] and observes positive effect. However, this approach cannot be transferred to other datasets, since the auxiliary pre-training task is designed based on the VQA-RAD dataset and requires extra effort for annotation.

In this paper, we tackle the data scarcity challenge by utilizing easily-available unannotated radiology image datasets for pre-training and representation distillation. First, we observe that the radiology images in current Med-VQA benchmarks mainly involve three human body regions – brain, chest, and abdomen, and there are large amounts of open-source unlabelled radiology images available for each region. Therefore, we propose to pre-train a visual feature extraction model (*teacher*) for each region respectively via contrastive learning. Second, to obtain a general and lightweight feature extractor, we distill the three teacher models into a small *student* model by contrastive representation distillation. The distilled model can be readily fine-tuned on any training dataset to facilitate the training of a Med-VQA system, without requiring further annotating process. Moreover, the small size of the distilled model can prevent overfitting to the training data, which typically only contains hundreds of radiology images.

To summarize, our contributions are two-fold. (1) We propose a new pre-training framework that leverages easily-acquired unannotated radiology images to pre-train and distill a general and lightweight visual feature extractor for Med-VQA, which can be easily adapted to small-scale training datasets. (2) We conduct extensive experiments with state-of-the-art Med-VQA methods on two benchmarks VQA-RAD [13] and SLAKE [14] to demonstrate the usefulness and benefits of the pre-trained model.

2 Related Work

Medical Visual Question Answering. Existing Med-VQA methods including [1,21,26] in ImageCLEF-Med challenge [11,2], often employ deep pre-trained architectures such as VGG [22] or ResNet [8] as the visual feature extraction module, which tend to cause overfitting due to limited training data in the Med-VQA domain. To overcome data limitation, MEVF [17] combines convolutional denoising auto-encoder (CDAE) [16] and meta-learning [24] to train a useful initialization for the visual feature extractor. Based on MEVF [17], conditional reasonin (CR) [29] further enhances the reasoning ability of the multimodal feature fusion module. Nevertheless, the pre-training process of MEVF requires additional data annotations on the training images, which requires medical expertise and is laborious and costly.

Contrastive Learning. Contrastive learning aims to learn high-quality feature representations by deriving self-supervision signals. CPC [18] pioneers in using

the InfoNCE loss for contrastive learning on sequential tasks such as text or audio, which has been followed by many recent contrastive learning methods [7,5,6]. MoCo [7] utilizes a queue to efficiently store a large number of negative samples; SimCLR [5] explores the effectiveness of diverse image augmentation combinations; MoCo-v2 [6] takes advantages of both MoCo and SimCLR to enhance representation learning. These unsupervised methods have achieved promising results in learning image representations.
Model Compression. Knowledge distillation is introduced in [4,9] to compress a large model into a smaller one without losing too many generalization abilities, which is achieved by minimizing Kullback–Leibler divergence (KLD) between the probabilistic outputs of the large and the smaller models. A recent work [23] argues that the independence assumption in the KLD loss fails to retain important structural information of the large model and proposes to combine KLD with contrastive representation distillation to achieve better performance.

3 Contrastive Pre-training and Representation Distillation (CPRD)

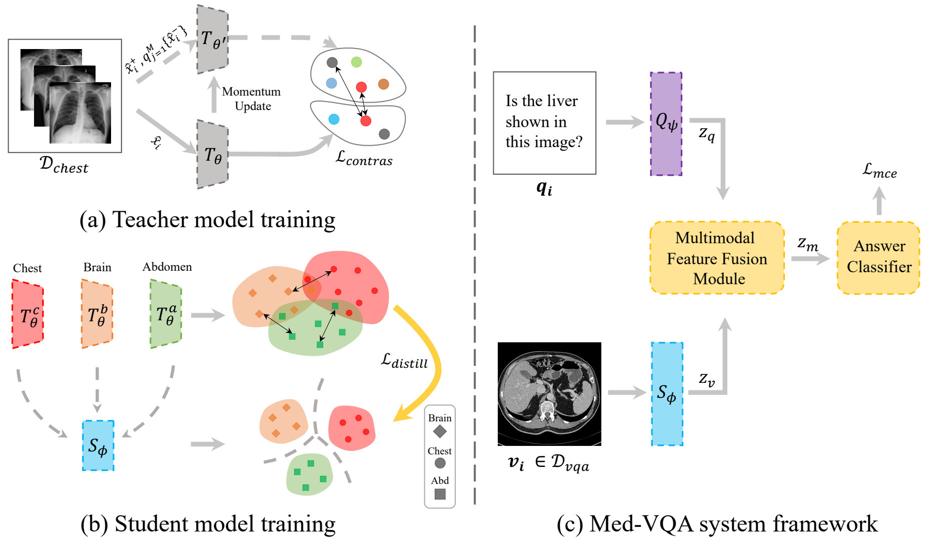
In current Med-VQA benchmarks, the radiology images mainly involve three human body regions: brain, chest, and abdomen. For each region, unlabeled images can be easily obtained from many large-scale open-source datasets. Motivated by this observation, we propose to train three specialized teacher models to focus on different body region respectively and then teach a student model to learn both intra- and inter-region features for Med-VQA, as illustrated in Fig. 1.

3.1 Teachers: Intra-region Contrastive Pre-training

Let $\mathcal{D}_{brain}, \mathcal{D}_{chest}, \mathcal{D}_{abdomen}$ denote the set of radiology images for the three body regions respectively. Radiology images in each region have large diversity in terms of different organs and versatile imaging modalities, e.g., liver MRI, liver CT, and intestine CT in the abdomen region. Therefore, we employ Momentum Contrast [6], a self-supervised contrastive learning method, to train a *Teacher* model for each region with the corresponding dataset $\mathcal{D}_r$ ($r \in \{brain, chest, abdomen\}$) to implicitly model these differences. As shown in Fig. 1 (a), we sample an image x_i and a queue $q = \{x_j^-\}_{j=1}^M$ of M images different from x_i from $\mathcal{D}_r$. Then, data augmentation (such as resize, crop, color distort, and Gaussian blur), denoted as Aug, is applied on all the sampled images and produce:

$$\boldsymbol{\hat{x}_i} = Aug(x_i), \boldsymbol{\hat{x}_i^+} = Aug(x_i),\ \boldsymbol{\hat{q}} = \{\boldsymbol{\hat{x}_j^-} = Aug(x_j^-)\}_{j=1}^M, \tag{1}$$

where $\boldsymbol{\hat{x}_i}$ and $\boldsymbol{\hat{x}_i^+}$ are two different views of x_i, generated by applying random augmentation on x_i twice. An encoder T_θ is used to learn the feature representation of $\hat{x}_i$, i.e., $\boldsymbol{z_i} = T_\theta(\boldsymbol{\hat{x}_i})$. Another momentum encoder $T_{\theta'}$ is used to produce the representations of $\boldsymbol{\hat{x}_i^+}$ and $\boldsymbol{\hat{q}}$, i.e., $\{\boldsymbol{z_i^+}, \boldsymbol{z_1^-}, \boldsymbol{z_2^-}, ..., \boldsymbol{z_M^-}\}$. Since $\boldsymbol{z_i}$ and $\boldsymbol{z_i^+}$ are



(a) Teacher model training

(b) Student model training

(c) Med-VQA system framework

Fig. 1. Our proposed CPRD framework for Med-VQA. (a) Train a teacher model T_θ by self-supervised contrastive learning on the chest region. (b) Distill three teacher models into one student model S_ϕ. (c) Apply the student model S_ϕ for Med-VQA.

the representations of different views of x_i, $\boldsymbol{z_i}$ should be similar to $\boldsymbol{z_i^+}$ but dissimilar to the other M representations in $\hat{\boldsymbol{q}}$. The learning process can be guided by the InfoNCE contrastive loss [18]:

$$\mathcal{L}_{\boldsymbol{z_i},\boldsymbol{z_i^+},\{\boldsymbol{z_j^-}\}} = -\log \frac{exp(\boldsymbol{z_i} \cdot \boldsymbol{z_i^+}/\tau)}{exp(\boldsymbol{z_i} \cdot \boldsymbol{z_i^+}/\tau) + \sum_{j=1}^{M} exp(\boldsymbol{z_i} \cdot \boldsymbol{z_j^-}/\tau)}, \tag{2}$$

where τ is a temperature parameter [25] and $\cdot$ stands for dot product. In practice, the length of the queue q is usually much larger than the mini-batch size, making it costly to update $T_{\theta'}$ by gradient back-propagation. Following [6], we update it in an efficient way: $\theta' \leftarrow \beta\theta' + (1-\beta)\theta$, where β is the momentum coefficient. By optimizing the loss in Eq. (2), we obtain the teacher model T_θ for the region.

3.2 Student: Inter-region Representation Distillation

After obtaining the three teacher models: T_θ^a for $\mathcal{D}_{abdomen}$, T_θ^b for $\mathcal{D}_{brain}$, and T_θ^c for $\mathcal{D}_{chest}$, we design a lightweight *Student* model S_ϕ to distill representations of the teacher models, as shown in Fig. 1 (b). Let $\mathcal{D}_{all} = \{\mathcal{D}_{brain}, \mathcal{D}_{chest}, \mathcal{D}_{abdomen}\}$. Inspired by the idea of contrastive representation distillation [23], for each region $\mathcal{D}_r \in \mathcal{D}_{all}$, for any image $x_i^r \in \mathcal{D}_r$, we randomly sample K images x_j^o ($j = \{1, \ldots, K\}$) from the other two datasets $\mathcal{D}_o = \mathcal{D}_{all} \backslash \mathcal{D}_r$. First, we make the student model inherit knowledge of each teacher by enforcing its representation of x_i^r, $S_\phi(x_i^r)$, to be similar to that of the corresponding teacher model, $T_\theta^r(x_i^r)$, by minimizing the loss function

$$\mathcal{L}_{sim} = -\frac{1}{N}\sum_{r=1}^{3}\sum_{i=1}^{L_r}\log(\frac{e^{T_\theta^r(x_i^r)\cdot S_\phi(x_i^r)/\tau}}{e^{T_\theta^r(x_i^r)\cdot S_\phi(x_i^r)/\tau}+\frac{K}{N}}), \tag{3}$$

where τ is the temperature parameter, L_r is the size of $\mathcal{D}_r$, and N is the size of $\mathcal{D}_{all}$ $(1 < K < N)$. Meanwhile, we enable the student model to acquire the ability to distinguish the three regions by enforcing $S_\phi(x_i^r)$ to be dissimilar to $T_\theta^o(x_j^o)$, the representation of x_j^o (image of other regions) produced by the corresponding teacher model, by minimizing the loss function

$$\mathcal{L}_{dissim} = -\frac{1}{N\times K}\sum_{r=1}^{3}\sum_{i=1}^{L_r}\sum_{j=1}^{K}\log(1-(\frac{e^{T_\theta^o(x_j^o)\cdot S_\phi(x_i^r)/\tau}}{e^{T_\theta^o(x_j^o)\cdot S_\phi(x_i^r)/\tau}+\frac{K}{N}})). \tag{4}$$

Further, we train the student model to produce more discriminative representations by learning to identify the body region R of x_i^r. Note that the images are already grouped by regions in open-source databases so the region labels can be automatically generated. This is achieved by minimizing the classification loss

$$\mathcal{L}_{class} = -\frac{1}{N}\sum_{i=1}^{N}\log P(R=r|WS_\phi(x_i^r)), \tag{5}$$

where W is a linear classification layer, and P is the prediction probability of the target region. Finally, by combining Eqs. (3), (4) and (5), the student model is trained with the loss function

$$\mathcal{L}_{distill} = \alpha(\mathcal{L}_{dissim} + \mathcal{L}_{sim}) + (1-\alpha)\mathcal{L}_{class}, \tag{6}$$

where α is a balancing parameter.

4 Applying CPRD for Med-VQA

The distilled student model can be used as a universal visual feature extractor for any Med-VQA system based on radiology images. Fig. 1 (c) shows a typical Med-VQA pipeline. Given a radiology image v_i and a question q_i as inputs, the student model S_ϕ is applied on v_i to extract the visual features $\boldsymbol{z_v} = S_\phi(v_i)$, and a text encoder (e.g., LSTM [10] network) is used to extract the textual features q_i, i.e., $\boldsymbol{z_q} = Q_\psi(q_i)$. Then, $\boldsymbol{z_v}$ and $\boldsymbol{z_q}$ will be fused by some attention-based module (e.g., BAN [12]) to produce multimodal features $\boldsymbol{z_m}$.

Similar to general VQA, Med-VQA is also formulated as a classification problem [3]: predicting an answer from C fixed candidate answers in the training dataset. Note that there might be multiple correct answers for one question. As such, the multimodal features $\boldsymbol{z_m}$ will be fed to a classifier $\Phi(\cdot)$ (e.g., multilayer perceptron), to predict the probability of each candidate answer. All the model parameters, including those of the visual extractor S_ϕ, the text encoder Q_ψ, the feature fusion module and the classifier, are optimized in an end-to-end manner by minimizing the multi-label cross-entropy loss:

$$\mathcal{L}_{mce} = -\frac{1}{I}\sum_{i=1}^{I}\sum_{c=1}^{C}[l_i^c \log(\sigma^c(\Phi(\boldsymbol{z_m}))) + (1 - l_i^c)\log(1 - \sigma^c(\Phi(\boldsymbol{z_m})))], \qquad (7)$$

where l_i is the multi-hot encoding of the answers for the current (v_i, q_i) pair, σ is the sigmoid function, and I is the size of the training dataset.

5 Experiments

In this section, we extensively evaluate the effectiveness of the visual feature extractor pre-trained by our proposed CPRD framework on the only two available manually-annotated Med-VQA datasets. We experiment with state-of-the-art Med-VQA methods and show that the pre-trained feature extractor can be used to significantly improve their performance.

5.1 Datasets

VQA-RAD [13] consists of 315 radiology images and $3,515$ question-answer pairs. We follow the data splitting in [17]. **SLAKE** [14] is a recently proposed bi-lingual Med-VQA dataset. We use the English version, referred to as SLAKE-EN, which contains 642 radiology images and $7,033$ question-answer pairs. We use the original data splitting. Besides, questions in VQA-RAD and SLAKE are both categorized into "closed-ended" questions whose answers are in limited choices, and "open-ended" questions whose answers are free-form text.

5.2 Experimental Setup

To train the teacher and student models, we randomly sample $22,995$ unlabelled radiology images from open-resource databases[1], including $7,811$ chest X-Rays, $7,592$ abdomen CTs, and $7,592$ brain CTs and MRIs. Our experiments are conducted on a Ubuntu server with 8 NVIDIA TITAN 12GB Xp GPUs. All the hyper-parameters of the teacher and student models are chosen by cross validation via observing the loss in Eq. (2) and Eq. (6).

Teachers. For each region-focused teacher model, we use ResNet-50 to instantiate T_θ and $T_{\theta'}$ (Sec. 3.1) and train for 800 epochs with 4 GPUs for about 7 hours. In each epoch, the mini-batch size is 128, and the queue length M is $1,024$. The temperature parameter τ is set to be 0.2, 0.1, and 0.1 for brain, chest and abdomen respectively. For model optimization, we use SGD optimizer with $1.5e^{-2}$ initial learning rate decayed by cosine schedule.

Student. We use ResNet-8 as the student model (Sec. 3.2) and train for 240 epochs with 1 GPU. We use SGD optimizer to minimize the loss $\mathcal{L}_{distill}$ with

[1] `http://medicaldecathlon.com/`

Table 1. Test accuracy of our method and baselines.

	VQA-RAD [13]			SLAKE-EN [14]		
Models	Overall	Open	Closed	Overall	Open	Closed
MFB fw. [28]	50.6	14.5	74.3	73.3	72.2	75.0
SAN fw. [27]	54.3	31.3	69.5	76.0	74.0	79.1
BAN fw. [12]	58.3	37.4	72.1	76.3	74.6	79.1
MEVF+SAN [17]	64.1	49.2	73.9	76.5	75.3	78.4
MEVF+BAN [17]	66.1	49.2	77.2	78.6	77.8	79.8
CPRD+BAN (ours)	**67.8**	**52.5**	**77.9**	**81.1**	**79.5**	**83.4**
MEVF+BAN+CR [29]	71.6	60.0	79.3	80.0	78.8	82.0
CPRD+BAN+CR (ours)	**72.7**	**61.1**	**80.4**	**82.1**	**81.2**	**83.4**

Table 2. Comparison of different visual modules in test accuracy and model size on VQA-RAD [13]. The number of parameters is calculated on the visual module only.

Visual Modules	Overall(%)	Open(%)	Closed(%)	#Parameters (M)
VGG-16 [22] (ImageNet)	56.8	35.2	71.0	134.8
ResNet-50 [8] (ImageNet)	58.3	37.4	72.1	23.8
MEVF [17]	66.1	49.2	77.2	1.2
ResNet-8 (random init)	63.2	47.2	73.8	0.1
ResNet-8 (our CPRD)	**67.8**	**52.5**	**77.9**	**0.1**

0.05 initial learning rate decayed by cosine schedule. Besides, the queue length K is 8192, the temperature parameter τ is 0.07, and α in Eq. (6) is 0.9.

Med-VQA. After training the student model, we use the weights in the last epoch as initialization and fine-tune the model on a Med-VQA dataset for 100 epochs. We use Adamax optimizer with initial learning rate $2e^{-3}$ for model optimization. Following CR [29], we use accuracy as evaluation metric.

5.3 Comparison with the State-of-the-arts

We use our pre-trained model CPRD as the visual feature extractor, combined with the BAN attention mechanism [12] with or without the CR reasoning module [29] for Med-VQA. To demonstrate the necessity of domain-specific pre-training, we compare with general VQA frameworks including MFB [28], SAN [27], and BAN [12].[2] Further, we compare with MEVF [17], which is the only baseline that uses a small model and pre-trains with medical images.

The results on VQA-RAD [13] and SLAKE [14] are reported in Table 1. For a fair comparison, all methods use a 1024-D LSTM network to extract textual features with word embeddings pre-trained by GloVe [19]. For MFB, SAN and BAN, we use ResNet-50 pre-trianed on ImageNet as the visual feature extractor. The following observations can be made. (1) Our method CPRD+BAN not

[2] MFB, SAN, and BAN stand for the key reasoning module of the respective framework, where the visual and textual modules can be any applicable models.

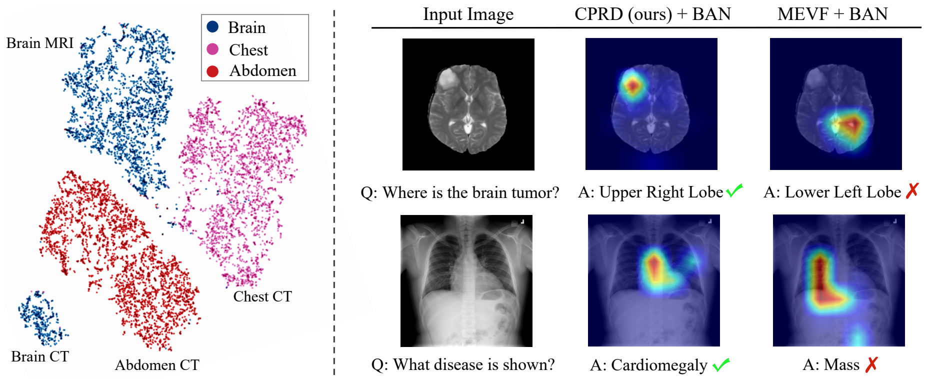


Fig. 2. (Left) t-SNE visualization of the representations learned by the student model; (Right) Grad-CAM maps from the visual modules of Med-VQA methods. ✓ and ✗ indicate the correctness of the answer given by each method.

only improves upon the performance of the strong baseline MEVF+BAN [17], but also achieves state-of-the-art results on the two benchmarks when further incorporating the CR [29] module. (2) Although MEVF+BAN [17] can significantly outperform the base framework BAN [12] on VQA-RAD, its performance gain on SLAKE is less significant ($\sim 2\%$), far lower than the gain brought by our CPRD+BAN ($\sim 5\%$). This demonstrates the generalization ability of our CPRD model on different datasets.

5.4 Ablation Analysis

We conduct an ablation study to analyze the impact of different pre-training strategies for the visual feature extraction module of Med-VQA. The results are summarized in Table 2. Specifically, we use BAN [12] as the multimodal feature fusion module and LSTM as the textual encoder for all methods in this subsection. Compared with the large models (i.e., VGG-16 and ResNet-50) pre-trained on ImageNet, it can be seen that lightweight models (i.e., MEVF and ResNet-8) perform better. Further, ResNet-8 pre-trained by our CPRD achieves better results than with random initialization, and outperforms the strongest baseline MEVF with much fewer parameters. This again demonstrates the effectiveness and advantages of our CPRD model.

5.5 Visualization

The t-SNE [15] visualization of the representations learned by the ResNet-8 student model on the images of $\mathcal{D}_{all}$ (Sec. 3.2) is shown in Fig. 2 (left). It can be clearly seen that the student model learns discriminative representations for different regions. Further, the representations of brain CT and brain MRI are well separated, indicating that the student model also captures the differences among versatile imaging modalities for the same region. To demonstrate the visual evidence used in Med-VQA models for prediction, in Fig. 2 (right), we show the Grad-CAM [20] maps for visual modules based on the final predicted answers of our CPRD+BAN and a strong baseline MEVF+BAN. The first row is about a brain MRI image, and the second is about a chest X-Ray image, both from the test set of the SLAKE [14] dataset. It can be seen that our model can correctly answer the questions by locating the right visual evidence about the questions, which demonstrates the effectiveness of our visual module.

6 Conclusion

In this paper, we have proposed a two-stage pre-training framework to tackle the challenge of data scarcity in the Med-VQA domain. Our framework leverages large amounts of unannotated radiology images to pre-train and distill a lightweight visual feature extractor via contrastive learning and representation distillation. By applying this pre-trained model in current Med-VQA methods, we achieve new state-of-the-art performance on existing benchmarks.

Acknowledgment. This research was supported by the grant of P0030935 (ZVPY) funded by PolyU (UGC).

References

1. Abacha, A.B., Gayen, S., Lau, J.J., Rajaraman, S., Demner-Fushman, D.: NLM at imageclef 2018 visual question answering in the medical domain. In: Working Notes of CLEF 2018 - Conference and Labs of the Evaluation Forum. CEUR Workshop Proceedings, vol. 2125. CEUR-WS.org, Avignon, France (2018)
2. Abacha, A.B., Hasan, S.A., Datla, V.V., Liu, J., Demner-Fushman, D., Müller, H.: Vqa-med: Overview of the medical visual question answering task at imageclef 2019. In: Working Notes of CLEF 2019 - Conference and Labs of the Evaluation Forum. CEUR Workshop Proceedings, vol. 2380. CEUR-WS.org, Lugano, Switzerland (2019)
3. Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C.L., Parikh, D.: Vqa: Visual question answering. In: Proceedings of the IEEE international conference on computer vision. pp. 2425–2433 (2015)
4. Buciluundefined, C., Caruana, R., Niculescu-Mizil, A.: Model compression. In: Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. KDD '06, Association for Computing Machinery, New York, NY, USA (2006)
5. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: International conference on machine learning. pp. 1597–1607. PMLR (2020)
6. Chen, X., Fan, H., Girshick, R., He, K.: Improved baselines with momentum contrastive learning (2020)
7. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning (2020)

8. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
9. Hinton, G., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network (2015)
10. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Computation pp. 1735–1780 (1997)
11. Ionescu, B., Müller, H., Villegas, M., de Herrera, A.G.S., Eickhoff, C., Andrearczyk, V., Cid, Y.D., Liauchuk, V., Kovalev, V., Hasan, S.A., Ling, Y., Farri, O., Liu, J., Lungren, M.P., Dang-Nguyen, D., Piras, L., Riegler, M., Zhou, L., Lux, M., Gurrin, C.: Overview of imageclef 2018: Challenges, datasets and evaluation. In: Experimental IR Meets Multilinguality, Multimodality, and Interaction - 9th International Conference of the CLEF Association, CLEF. Lecture Notes in Computer Science, vol. 11018, pp. 309–334. Springer, Avignon, France (2018)
12. Kim, J., Jun, J., Zhang, B.: Bilinear attention networks. In: Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems, NeurIPS. pp. 1571–1581. NeurIPS, Montréal, Canada (2018)
13. Lau, J.J., Gayen, S., Abacha, A.B., Demner-Fushman, D.: A dataset of clinically generated visual questions and answers about radiology images. Scientific data (2018)
14. Liu, B., Zhan, L.M., Xu, L., Ma, L., Yang, Y., Wu, X.M.: Slake: A semantically-labeled knowledge-enhanced dataset for medical visual question answering (2021)
15. Van der Maaten, L., Hinton, G.: Visualizing data using t-sne. Journal of machine learning research (2008)
16. Masci, J., Meier, U., Ciresan, D.C., Schmidhuber, J.: Stacked convolutional autoencoders for hierarchical feature extraction. In: Artificial Neural Networks and Machine Learning - ICANN 2011 - 21st International Conference on Artificial Neural Networks, Proceedings, Part I. Lecture Notes in Computer Science, vol. 6791, pp. 52–59. Springer, Espoo, Finland (2011)
17. Nguyen, B.D., Do, T., Nguyen, B.X., Do, T., Tjiputra, E., Tran, Q.D.: Overcoming data limitation in medical visual question answering. In: Medical Image Computing and Computer Assisted Intervention - MICCAI 2019 - 22nd International Conference, Part IV. Lecture Notes in Computer Science, vol. 11767, pp. 522–530. Springer, Shenzhen, China (2019)
18. Oord, A.v.d., Li, Y., Vinyals, O.: Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748 (2018)
19. Pennington, J., Socher, R., Manning, C.D.: Glove: Global vectors for word representation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP, A meeting of SIGDAT, a Special Interest Group of the ACL. pp. 1532–1543. ACL, Doha, Qatar (2014)
20. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-cam: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE international conference on computer vision. pp. 618–626 (2017)
21. Shi, L., Liu, F., Rosen, M.P.: Deep multimodal learning for medical visual question answering. In: Working Notes of CLEF 2019 - Conference and Labs of the Evaluation Forum. CEUR Workshop Proceedings, vol. 2380. CEUR-WS.org, Lugano, Switzerland (2019)
22. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)

23. Tian, Y., Krishnan, D., Isola, P.: Contrastive representation distillation. arXiv preprint arXiv:1910.10699 (2019)
24. Vuorio, R., Sun, S., Hu, H., Lim, J.J.: Multimodal model-agnostic meta-learning via task-aware modulation. In: Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems, NeurIPS. pp. 1–12. Vancouver, BC, Canada (2019)
25. Wu, Z., Xiong, Y., Yu, S., Lin, D.: Unsupervised feature learning via non-parametric instance-level discrimination. arXiv preprint arXiv:1805.01978 (2018)
26. Yan, X., Li, L., Xie, C., Xiao, J., Gu, L.: Zhejiang university at imageclef 2019 visual question answering in the medical domain. In: Working Notes of CLEF 2019 - Conference and Labs of the Evaluation Forum. CEUR Workshop Proceedings, vol. 2380. CEUR-WS.org, Lugano, Switzerland (2019)
27. Yang, Z., He, X., Gao, J., Deng, L., Smola, A.J.: Stacked attention networks for image question answering. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition. pp. 21–29. IEEE Computer Society, Las Vegas, NV, USA (2016)
28. Yu, Z., Yu, J., Fan, J., Tao, D.: Multi-modal factorized bilinear pooling with co-attention learning for visual question answering (2017)
29. Zhan, L.M., Liu, B., Fan, L., Chen, J., Wu, X.M.: Medical visual question answering via conditional reasoning. In: Proceedings of the 28th ACM International Conference on Multimedia. MM '20, Association for Computing Machinery, New York, NY, USA (2020)