PMC-CLIP: Contrastive Language-Image Pre-training Using Biomedical Documents

Weixiong Lin[1], Ziheng Zhao[1], Xiaoman Zhang[1,2], Chaoyi Wu[1,2], Ya Zhang[1,2], Yanfeng Wang[1,2], and Weidi Xie[1,2(✉)]

[1] Cooperative Medianet Innovation Center, Shanghai Jiao Tong University, Shanghai, China
[2] Shanghai AI Laboratory, Shanghai, China
{wx_lin,Zhao_Ziheng,xm99sjtu,wtzxxxwcy02,ya_zhang, wangyanfeng,weidi}@sjtu.edu.cn

Abstract. Foundation models trained on large-scale dataset gain a recent surge in CV and NLP. In contrast, development in biomedical domain lags far behind due to data scarcity. To address this issue, we build and release PMC-OA, a biomedical dataset with 1.6M image-caption pairs collected from PubMedCentral's OpenAccess subset, which is 8 times larger than before, PMC-OA covers diverse modalities or diseases, with majority of the image-caption samples aligned at finer-grained level, *i.e.*, subfigure and subcaption. While pretraining a CLIP-style model on PMC-OA, our model named PMC-CLIP outperform previous state-of-the-art models on various downstream tasks, including image-text retrieval on ROCO, MedMNIST image classification, Medical VQA, for example, +8.1% R@10 on image-text retrieval, +3.9% accuracy on image classification.

Keywords: Multimodal Dataset · Vision-Language Pretraining

1 Introduction

In the recent literature, development of foundational models has been the main driving force in artificial intelligence, for example, large language models [2,8,27,30] trained with either autoregressive prediction or masked token inpainting, and computer vision models [19,29] trained by contrasting visual-language features. In contrast, development in the biomedical domain lags far behind due to limitations of data availability from two aspects, (i) the expertise required for annotation, (ii) privacy concerns. This paper presents our preliminary study for constructing a **large-scale, high-quality, image-text** biomedical dataset using publicly available scientific papers, with **minimal manual efforts** involved.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-3-031-43993-3_51.

H. Greenspan et al. (Eds.): MICCAI 2023, LNCS 14227, pp. 525–536, 2023.
https://doi.org/10.1007/978-3-031-43993-3_51

In particular, we crawl figures and corresponding captions from scientific documents on PubMed Central, which is a free full-text archive of biomedical and life sciences journal literature at the U.S. National Institutes of Health's National Library of Medicine (NIH/NLM) [31]. This brings two benefits: (i) the contents in publications are generally well-annotated and examined by experts, (ii) the figures have been well-anonymized and de-identified. In the literature, we are clearly not the first to construct biomedical datasets in such manner, however, existing datasets [28,34,38] suffer from certain limitations in diversity or scale from today's standard. For example, as a pioneering work, ROCO [28] was constructed long time ago with only 81k radiology images. MedICAT [34] contains 217k images, but are mostly consisted of compound figures.

In this work, we tackle the above-mentioned limitations by introducing an automatic pipeline to generate dataset with subfigure-subcaption correspondence from scientific documents, including three major stages: medical figure collection, subfigure separation, subcaption separation & alignment. The final dataset, PMC-OA, consisting of 1.65M image-text pairs (not including samples from ROCO), covers a wide scope of diagnostic procedures and diseases, as shown in Fig. 1 and Fig. 3.

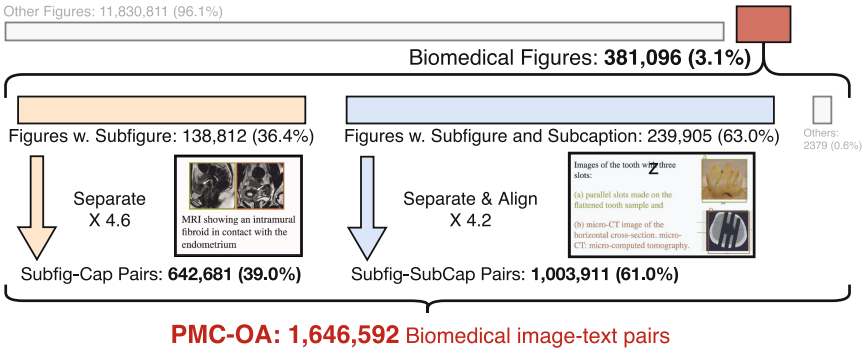


Fig. 1. Statistics over the pipeline and the collected PMC-OA.

Along with the constructed dataset, we train a CLIP-style vision-language model for the biomedical domain, termed as PMC-CLIP. To achieve such a goal, the model is trained on PMC-OA with standard image-text contrastive (ITC) loss, and to encourage the joint interaction of image and text, masked language modeling (MLM) is also applied. We evaluate the pre-trained model on several downstream tasks, including medical image-text retrieval, medical image classification, and medical visual question answering (VQA). PMC-CLIP achieves state-of-the-art performance on various downstream tasks, surpassing previous methods significantly.

Overall, in this paper, we make the following contributions: **First**, we propose an automatic pipeline to construct high-quality image-text biomedical datasets

from scientific papers, and construct an image-caption dataset via the proposed pipeline, named PMC-OA, which is 8× larger than before. With the proposed pipeline, the dataset can be continuously updated. **Second**, we pre-train a vision-language model on the constructed image-caption dataset, termed as PMC-CLIP, to serve as a foundation model for biomedical domain. **Third**, we conduct thorough experiments on various downstream tasks (retrieval, classification, and VQA), and demonstrate state-of-the-art performance. The dataset and pre-trained model will be made available to the community.

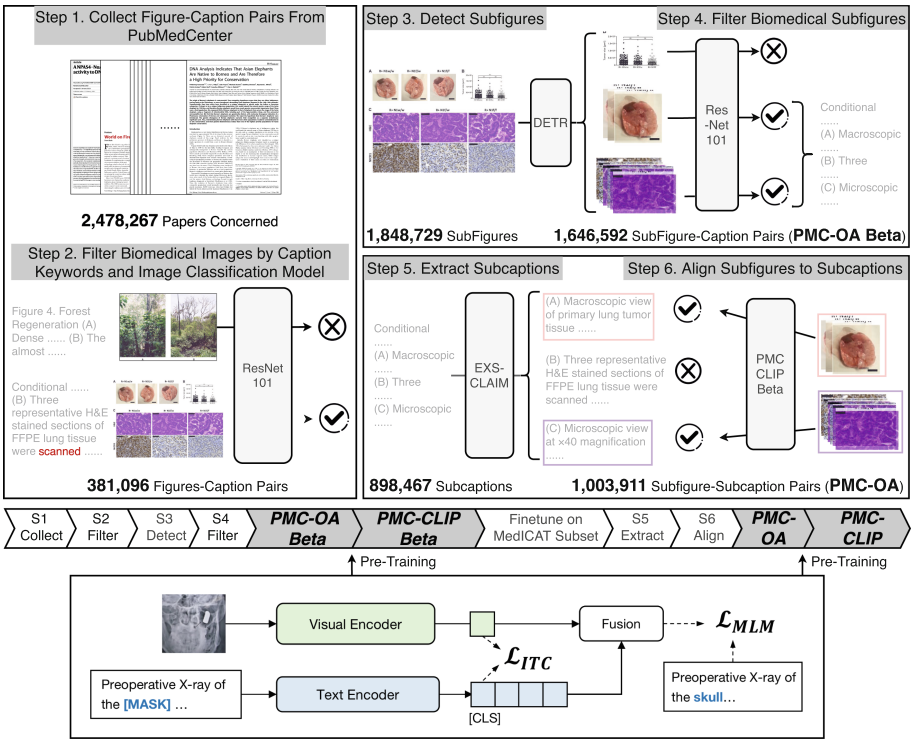


Fig. 2. The proposed pipeline to collect PMC-OA (upper) and the architecture of PMC-CLIP (bottom).

2 The PMC-OA Dataset

In this section, we start by describing the dataset collection procedure in Sect. 2.1, followed by a brief overview of PMC-OA in Sect. 2.2.

2.1 Dataset Collection

In this section, we detail the proposed pipeline to create PMC-OA, a large-scale dataset that contains 1.65M image-text pairs. The whole procedure consists of three major stages: (i) medical figure collection, (ii) subfigure separation, (iii) subcaption separation & alignment, as summarised in Fig. 2.

Medical Figure Collection (Step 1 and 2 in Fig. 2). We first extract figures and captions from PubMedCentral (till 2022-09-16) [31]. 2.4M papers are covered and 12M figure-caption pairs are extracted. To derive medical figures, inspired by MedICat [34], we first filter out the captions without any medical keywords[1] and then use a classification network trained on DocFigure [14] to further pick out the medical figure, ending up with 381K medical figures.

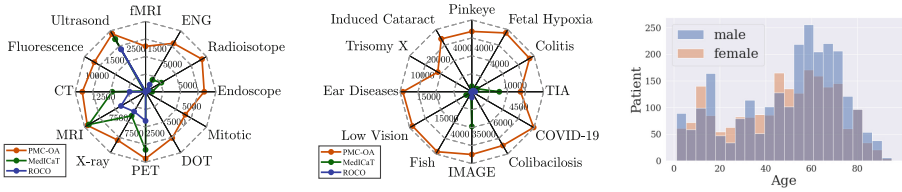
Subfigure Separation (Step 3 and 4 in Fig. 2). We randomly check around 300 figures from previous step, and find that around 80% of figures are compound, *i.e.* multiple pannels. We thus train a subfigure detector on MedICaT subfigure-subcaption subset [34] (MedICatSub) to break the compound figures into subfigures. After separation, to filter out non-medical subfigures missed in the former step, we apply the aforementioned classifier again on the derived subfigures, obtaining 1.6M subfigure-caption pairs. We termed this dataset as **PMC-OA Beta** version.

Subcaption Separation and Alignment (Step 5 and 6 in Fig. 2). To further align subfigure to its corresponding part within the full caption, *i.e.*, subcaption, we need to break the captions into subcaptions first, we apply an off-shelf caption distributor [32]. We pretrain a CLIP-style model (termed as **PMC-CLIP-Beta**, training detail will be described in Sect. 3) on **PMC-OA-Beta**, then finetune it on MedICaTSub for subfigure-subcaption alignment, which achieves alignment accuracy=73% on test set. We finally align 1,003,911 subfigure-subcaption pairs, along with the remaining 642,681 subfigure-caption pairs, we termed this dataset as **PMC-OA**. Note that, we have explicitly removed duplication between our data and ROCO by identify each image-caption pair with paperID and image source link. We consequently pretrain the **PMC-CLIP** on it.

2.2 Dataset Overview

In this section, we provide a brief statistical overview of the collected dataset PMC-OA with UMLS parser [1] from three different perspectives, *i.e.*, diagnostic procedure, diseases and findings, and fairness. *First*, PMC-OA covers a wide range of diagnostic procedures, spanning from common (*CT, MRI*) to rare ones (*mitotic figure*), which is more diverse than before (Fig. 3(a)). *Second*, PMC-OA contains various diseases and findings, and is more up-to-date, covering new emergent diseases like COVID-19 (Fig. 3(b)). And the wide disease coverage in

[1] Follow Class TUI060 "*Diagnostic Procedure*" defined in UMLS [1].



(a) Diagnostic procedure. (b) Disease and findings. (c) Patients' age & gender.

Fig. 3. Statistical overview of PMC-OA.

our dataset supports learning the shared patterns of diseases, promoting accurate auto-diagnosis. *Third*, we also provide the sex-ratio across ages in Fig. 3(c), as we can see PMC-OA is approximately gender-balanced, with 54% males. The fairness on population ensures our dataset sightly suffers from patient characteristic bias, thus providing greater cross-center generalize ability.

Discussion. Compared to pioneering works [28,34] for constructing dataset based on PubMedCentral, our proposed PMC-OA is of larger scale, diversity, and has more accurate alignment: *First*, PMC-OA covers a wider range of papers (2.4M) than ROCO [28](1.8M) and MedICaT [34](131K), and thus enlarge our dataset(1.6M). *Second*, unlike ROCO [28], we maintain the non-radiology images, which makes PMC-OA a more diverse biomedical dataset as shown in Fig. 3. *Third*, to the best of our knowledge, we are the first to integrate subfigures separation, subcaptions separation and the alignment into the data collection pipeline, which explicitly enlarges our dataset (8 times of MedICaT and 20 times of ROCO), while reducing the noise as much as possible.

3 Visual-language Pre-training

With our constructed image-caption dataset, we further train a visual-language model, termed as PMC-CLIP as shown in Fig. 2 (bottom). We describe the architecture first and then introduce the two training objectives separately.

Architecture. Given N image-caption training pairs, *i.e.*, $\mathcal{D} = \{(\mathcal{I}_i, \mathcal{T}_i)|_{i=1}^{N}\}$, where $\mathcal{I}_i \in \mathbb{R}^{H\times W\times C}$ represents images, H, W, C are height, width, channel, and $\mathcal{T}_i$ represents the paired text. We aim to train a CLIP-style visual-language model with an image encoder Φ_{visual} and a text encoder Φ_{text}.

In detail, given a specific image-caption pair $(\mathcal{I}, \mathcal{T})$, we encode it separately with a ResNet-based Φ_{visual} and a BERT-based Φ_{text}, the embedding dimension is denoted as d and the text token length as l:

$$\boldsymbol{v} = \Phi_{\text{visual}}(\mathcal{I}) \in \mathbb{R}^d, \tag{1}$$

$$\boldsymbol{T} = \Phi_{\text{text}}(\mathcal{T}) \in \mathbb{R}^{l\times d}, \boldsymbol{t} = \boldsymbol{T}_0 \in \mathbb{R}^d, \tag{2}$$

where $\boldsymbol{v}$ represents the embedding for the whole image, $\boldsymbol{T}$ refers to the sentence embedding, and $\boldsymbol{t}$ denotes the embedding for [CLS] token.

Image-Text Contrastive Learning (ITC). We implement ITC loss following CLIP [29], that aims to match the corresponding visual and text representations from one sample. In detail, denoting batch size as b, we calculate the softmax-normalized cross-modality dot product similarity between the current visual/text embedding ($\boldsymbol{v}$ / $\boldsymbol{t}$) and all samples within the batch, termed as $p^{\mathrm{i2t}}, p^{\mathrm{t2i}} \in \mathbb{R}^b$, and the final ITC loss is:

$$L_{\mathrm{ITC}} = \mathbb{E}_{(\mathcal{I},\mathcal{T})\sim\mathcal{D}}\big[\mathrm{CE}(y^{\mathrm{i2t}}, p^{\mathrm{i2t}}) + \mathrm{CE}(y^{\mathrm{t2i}}, p^{\mathrm{t2i}})\big], \tag{3}$$

where $y^{\mathrm{i2t}}, y^{\mathrm{t2i}}$ refer to one-hot matching labels, CE refers to InfoNCE loss [25].

Masked Language Modeling (MLM). We implement MLM loss following BERT [7]. The network is trained to reconstruct the masked tokens from context contents and visual cues. We randomly mask the word in texts with a probability of 15% and replace it with a special token '[MASK]'. We concatenate the image embedding $\boldsymbol{v}$ with the text token embeddings $\boldsymbol{T}$, input it into a self-attention transformer-based fusion module Φ_{fusion}, and get the prediction for the masked token at the corresponding position in the output sequence, termed as $p^{\mathrm{mask}} = \Phi_{\mathrm{fusion}}(\boldsymbol{v}, \boldsymbol{T})$. Let y^{mask} denote the ground truth, and the MLM loss is:

$$L_{\mathrm{MLM}} = \mathbb{E}_{(\mathcal{I},\mathcal{T})\sim\mathcal{D}}\big[\mathrm{CE}(y^{\mathrm{mask}}, p^{\mathrm{mask}})\big] \tag{4}$$

Total Training Loss. The final loss is defined as $L = L_{\mathrm{ITC}} + \lambda L_{\mathrm{MLM}}$, where λ is a hyper-parameter deciding the weight of L_{MLM}, set as 0.5 by default.

Disscussion. While we recognize a lot of progress in VLP methodology [13, 35,36], PMC-CLIP is trained in an essential way to demonstrate the potential of the collected PMC-OA, and thus should be orthogonal to these works.

4 Experiment Settings

4.1 Pre-training Datasets

ROCO [28] is a image-caption dataset collected from PubMed [31]. It filters out all the compound or non-radiological images, and consists of 81K samples.

MedICaT [34] extends ROCO to 217K samples (image-caption pairs), however, 75% of its figures are compound ones, *i.e.* one figure with multiple subfigures.

MIMIC-CXR [15] is the largest chest X-ray dataset, containing 377,110 samples (image-report pairs). Each image is paired with a clinical report describing findings from doctors.

PMC-OA contains 1.65M image-text pairs, which we have explicitly conducted deduplication between ROCO.

4.2 Downstream Tasks

Image-Text Retrieval (ITR). ITR contains both image-to-text(I2T) and text-to-image(T2I) retrieval. We train PMC-CLIP on different datasets, and sample 2,000 image-text pairs from ROCO's testset for evaluation, following previous works [4,5,34]. **Note that**, as we have explicitly conducted deduplication, the results thus resemble *zero-shot* evaluation.

Classification. We finetune the model for different downstream tasks that focus on image classification. Spcifically, MedMINIST [37] contains 12 tasks for 2D images, and it covers primary data modalities in biomedical images, including Colon Pathology, Dermatoscope, Retinal OCT, etc.

Visual Question Answering (VQA). We evaluate on the official dataset split of SLAKE [22], and follow previous work's split [24] on VQA-RAD [18], where SLAKE is composed of 642 images and 14,028 questions and VQA-RAD contains 315 images and 3,515 questions. The questions in VQA-RAD and Slake are categorized as close-ended if answer choices are limited, otherwise open-ended. The image and text encoders are initialized from PMC-CLIP and fine-tuned, we refer the reader for more details in supplementary.

4.3 Implementation Details

For the visual and text encoders, we adopt ResNet50 [12] and PubmedBERT [11]. And we use 4 transformer layers for the fusion module. For input data, we resize each image to 224×224. During pre-training, our text encoder is initialized from PubmedBERT, while the vision encoder and fusion module are trained from scratch. We use AdamW [23] optimizer with $lr = 1 \times 10^{-4}$. We train on GeForce RTX 3090 GPUs with batch size 128 for 100 epochs. The first 10 epochs are set for warming up.

5 Result

We conduct experiments to validate our proposed dataset, and the effectiveness of model trained on it. In Sec. 5.1, we first compare with existing large-scale biomedical datasets on the image-text retrieval task to demonstrate the superiority of PMC-OA. In Sect. 5.2, we finetune the model (pre-trained on PMC-OA) across three different downstream tasks, namely, retrieval, classification, and visual question answering. And we also perform a thorough empirical study of the pretraining objectives and the model architectures in Sect. 5.3. **Note that**, for all experiments, we use the default setting: ResNet50 for image encoder, and pre-train with both ITC and MLM objectives, unless specified otherwise.

5.1 PMC-OA surpasses SOTA large-scale biomedical dataset

As shown in Table 1, we pre-train PMC-CLIP on different datasets and evaluate retrieval on ROCO test set. The performance can be largely improved by simply switching to our dataset, confirming the significance of it.

Table 1. Ablation studies on pre-training dataset.

Methods	Pretrain Data	DataSize	I2T			T2I		
			R@1	R@5	R@10	R@1	R@5	R@10
PMC-CLIP	ROCO	81 K	12.30	35.28	46.52	13.36	35.84	47.38
PMC-CLIP	MedICaT	173 K	17.44	41.08	52.72	17.14	40.42	51.71
PMC-CLIP	PMC-OA Beta	1.6 M	30.42	59.11	70.16	27.92	55.99	66.35
PMC-CLIP	PMC-OA	1.6 M	31.41	61.15	71.88	28.02	58.33	69.69

5.2 PMC-CLIP achieves SOTA across downstream tasks

To evaluate the learnt representation in PMC-CLIP, we compare it with several state-of-the-art approaches across various downstream tasks, including image-text retrieval, image classification, and visual question answering.

Image-Text Retrieval. As shown in Table 2, we report a state-of-the-art result on image-text retrieval. On I2T Rank@10, PMC-CLIP outperforms previous state-of-the-art by 8.1%. It is worth mentioning that, the training set of ROCO has been used during pretraining in M3AE [4], ARL [5]. While our dataset does not contain data from ROCO.

Table 2. Zero-shot Image-Text Retrieval on ROCO.

Methods	Pretrain Data	DataSize	I2T			T2I		
			R@1	R@5	R@10	R@1	R@5	R@10
ViLT [16]	COCO [20], VG [17], SBU, GCC	4.1M	11.90	31.90	43.20	9.75	28.95	41.40
METER [9]	COCO, VG, SBU [26], GCC [33]	4.1M	14.45	33.30	45.10	11.30	27.25	39.60
M3AE [4]	ROCO, MedICaT	233 K	19.10	45.60	61.20	19.05	47.75	61.35
ARL [5]	ROCO, MedICaT, CXR	233 K	23.45	50.60	62.05	23.50	49.05	63.00
PMC-CLIP	PMC-OA	1.6 M	31.41	61.15	71.88	28.02	58.33	69.69

Image Classification. To demonstrate the excellent transferability of PMC-CLIP, we validate it on MedMNIST and compare it with SOTA methods *i.e.,* DWT-CV [6] and SADAE [10]. We present the results of 3 of 12 sub-tests here, and the full results can be found in the supplementary material. As shown in Table 3, PMC-CLIP obtains consistently higher results, and it is notable that finetuning from PMC-CLIP achieves significant performance gains compared with training from scratch with ResNet.

Visual Question Answering. VQA requires model to learn finer grain visual and language representations. As Table 4 shows, we surpass SOTA method M3AE in 5 out of 6 results.

Table 3. Classification results on MedMNIST.

Methods	PneumoniaMNIST		BreastMNIST		DermaMNIST	
	AUC↑	ACC↑	AUC↑	ACC↑	AUC↑	ACC↑
ResNet50 [12]	96.20	88.40	86.60	84.20	91.20	73.10
DWT-CV [6]	95.69	88.67	89.77	85.68	91.67	74.75
SADAE [10]	98.30	91.80	91.50	87.80	92.70	75.90
PMC-CLIP	99.02	95.35	94.56	91.35	93.41	79.80

Table 4. VQA results on VQA-RAD and Slake.

Methods	VQA-RAD			Slake		
	Open	Closed	Overall	Open	Closed	Overall
MEVF-BAN [24]	49.20	77.20	66.10	77.80	79.80	78.60
CPRD-BAN [21]	52.50	77.90	67.80	79.50	83.40	81.10
M3AE [4]	67.23	83.46	77.01	80.31	87.82	83.25
PMC-CLIP	67.00	84.00	77.60	81.90	88.00	84.30

5.3 Ablation Study

Training Objectives. We pre-train PMC-CLIP with different objectives (*ITC, MLM*) for ablation studies, and summarize the results in the supplementary material (Table 5 in the supplementary). Here, we present a summary of the observations: *First*, ITC objective is essential for pretraining, and contributes most of the performance. *Second*, MLM using only text context works as a regularization term. *Third*, With incorporation of visual features, the model learns finer grain correlation between image-caption pairs, and achieve the best results.

Data Collection Pipeline. To demonstrate the effectiveness of subfigure-subcaption alignment, we compare PMC-CLIP with the model pretrained on dataset w/o alignment (Table 6(1–3) in the supplementary). The result verify that subfigure-subcaption alignment reduces dataset's noise thus enhance the pretrained model.

Visual Backbone. We have also explored different visual backbones, using the same setting as CLIP [29] (Table 6(4–7) in the supplementary). We observe that all ResNet variants have close performance with RN50, outperforming ViT-B/32, potentially due to the large patch size.

6 Conclusion

In this paper, we present a large-scale dataset in biomedical domain, named PMC-OA, by collecting image-caption pairs from abundant scientific docu-

ments. We train a CLIP-style model on PMC-OA, termed as PMC-CLIP, it achieves SOTA performance across various downstream biomedical tasks, including image-text retrieval, image classification, visual question answering. With the automatic collection pipeline, the dataset can be further expanded, which can be beneficial to the research community, fostering development of foundation models in biomedical domain.

Acknowledgement. This work is supported by the National Key R&D Program of China (No. 2022ZD0160702), STCSM (No. 22511106101, No. 18DZ2270700, No. 21DZ1100100), 111 plan (No. BP0719010), and State Key Laboratory of UHD Video and Audio Production and Presentation.

References

1. Bodenreider, Olivier: The unified medical language system (umls): integrating biomedical terminology. Nucleic Acids Research **32**, D267–D270 (2004)
2. Brown, Tom, et al.: Language models are few-shot learners. Advances in Neural Information Processing Systems **33**, 1877–1901 (2020)
3. Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In Computer Vision-ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16, pages 213–229. Springer, 2020
4. Zhihong Chen, Yuhao Du, Jinpeng Hu, Yang Liu, Guanbin Li, Xiang Wan, and Tsung-Hui Chang. Multi-modal masked autoencoders for medical vision-and-language pre-training. In Medical Image Computing and Computer Assisted Intervention-MICCAI 2022: 25th International Conference, Singapore, September 18–22, 2022, Proceedings, Part V, pages 679–689. Springer, 2022
5. Zhihong Chen, Guanbin Li, and Xiang Wan. Align, reason and learn: Enhancing medical vision-and-language pre-training with knowledge. In Proceedings of the 30th ACM International Conference on Multimedia, pages 5152–5161, 2022
6. Cheng, Jianhong, Kuang, Hulin, Zhao, Qichang, Wang, Yahui, Lei, Xu., Liu, Jin, Wang, Jianxin: Dwt-cv: Dense weight transfer-based cross validation strategy for model selection in biomedical data analysis. Future Generation Computer Systems **135**, 20–29 (2022)
7. Jacob Devlin et al. Bert: Pre-training of deep bidirectional transformers for language understanding. ArXiv preprint ArXiv:1810.04805, 2018
8. Ming Ding et al. Cogview2: Faster and better text-to-image generation via hierarchical transformers. ArXiv preprint ArXiv:2204.14217, 2022
9. Zi-Yi Dou et al. An empirical study of training end-to-end vision-and-language transformers. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 18166–18176, 2022
10. Ge, Xiaolong, et al.: A self-adaptive discriminative autoencoder for medical applications. IEEE Transactions on Circuits and Systems for Video Technology **32**(12), 8875–8886 (2022)
11. Yu, Gu., Tinn, Robert, Cheng, Hao, Lucas, Michael, Usuyama, Naoto, Liu, Xiaodong, Naumann, Tristan, Gao, Jianfeng, Poon, Hoifung: Domain-specific language model pretraining for biomedical natural language processing. ACM Transactions on Computing for Healthcare (HEALTH) **3**(1), 1–23 (2021)

12. Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 770–778, 2016
13. Shih-Cheng Huang, Liyue Shen, Matthew P Lungren, and Serena Yeung. Gloria: A multimodal global-local representation learning framework for label-efficient medical image recognition. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 3942–3951, 2021
14. KV Jobin, Ajoy Mondal, and CV Jawahar. Docfigure: A dataset for scientific document figure classification. In 2019 International Conference on Document Analysis and Recognition Workshops, volume 1, pages 74–79. IEEE, 2019
15. Johnson, Alistair EW., et al.: Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. Scientific Data **6**(1), 317 (2019)
16. Wonjae Kim et al. Vilt: Vision-and-language transformer without convolution or region supervision. In International Conference on Machine Learning, pages 5583–5594. PMLR, 2021
17. Krishna, Ranjay, et al.: Visual genome: Connecting language and vision using crowdsourced dense image annotations. International Journal of Computer Vision **123**, 32–73 (2017)
18. Jason J Lau et al. A dataset of clinically generated visual questions and answers about radiology images. Scientific Data, 5(1), 1–10, 2018
19. Li, Junnan, Selvaraju, Ramprasaath, Gotmare, Akhilesh, Joty, Shafiq, Xiong, Caiming, Hoi, Steven Chu Hong.: Align before fuse: Vision and language representation learning with momentum distillation. Advances in Neural Information Processing Systems **34**, 9694–9705 (2021)
20. Tsung-Yi Lin et al. Microsoft coco: Common objects in context. In Computer Vision-ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6–12, 2014, Proceedings, Part V 13, pages 740–755. Springer, 2014
21. Bo Liu et al. Contrastive pre-training and representation distillation for medical visual question answering based on radiology images. In Medical Image Computing and Computer Assisted Intervention-MICCAI 2021: 24th International Conference, Strasbourg, France, September 27-October 1, 2021, Proceedings, Part II 24, pages 210–220. Springer, 2021
22. Bo Liu et al. 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, 2021
23. Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In International Conference on Learning Representations, 2017
24. Binh D Nguyen et al. Overcoming data limitation in medical visual question answering. In Medical Image Computing and Computer Assisted Intervention-MICCAI 2019, pages 522–530. Springer, 2019
25. Aaron van den Oord et al. Representation learning with contrastive predictive coding. ArXiv preprint ArXiv:1807.03748, 2018
26. Vicente Ordonez et al. Im2text: Describing images using 1 million captioned photographs. Advances in Neural Information Processing Systems, 24, 2011
27. Long Ouyang et al. Training language models to follow instructions with human feedback. ArXiv preprint ArXiv:2203.02155, 2022
28. Obioma Pelka, Sven Koitka, Johannes Rückert, Felix Nensa, and Christoph M Friedrich. Radiology objects in context (roco): a multimodal image dataset. In MICCAI Workshop on Large-scale Annotation of Biomedical Data and Expert Label Synthesis (LABELS) 2018, pages 180–189. Springer, 2018

29. Alec Radford et al. Learning transferable visual models from natural language supervision. In International Conference on Machine Learning, pages 8748–8763. PMLR, 2021
30. Aditya Ramesh et al. Hierarchical text-conditional image generation with clip latents. ArXiv preprint ArXiv:2204.06125, 2022
31. Richard J Roberts. Pubmed central: The genbank of the published literature, 2001
32. Eric Schwenker et al. Exsclaim!-an automated pipeline for the construction of labeled materials imaging datasets from literature. ArXiv preprint ArXiv:2103.10631, 2021
33. Piyush Sharma et al. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 2556–2565, 2018
34. Sanjay Subramanian et al. Medicat: A dataset of medical images, captions, and textual references. In Findings of EMNLP, 2020
35. Zifeng Wang, Zhenbang Wu, Dinesh Agarwal, and Jimeng Sun. Medclip: Contrastive learning from unpaired medical images and text. arXiv preprint arXiv:2210.10163, 2022
36. Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. Medklip: Medical knowledge enhanced language-image pre-training. MedRxiv, pages 2023–01, 2023
37. Yang, Jiancheng, et al.: Medmnist v2-a large-scale lightweight benchmark for 2d and 3d biomedical image classification. Scientific Data **10**(1), 41 (2023)
38. Zheng Yuan, Qiao Jin, Chuanqi Tan, Zhengyun Zhao, Hongyi Yuan, Fei Huang, and Songfang Huang. Ramm: Retrieval-augmented biomedical visual question answering with multi-modal pre-training. arXiv preprint arXiv:2303.00534, 2023