Large-Scale Domain-Specific Pretraining for Biomedical Vision-Language Processing

Preprint · March 2023

DOI: 10.48550/arXiv.2303.00915

Cliff Wong
Microsoft
39 PUBLICATIONS **4,764** CITATIONS

Matthew Lungren
Stanford University
253 PUBLICATIONS **21,467** CITATIONS

Large-Scale Domain-Specific Pretraining for Biomedical Vision-Language Processing

Sheng Zhang*, **Yanbo Xu***, **Naoto Usuyama***, **Jaspreet Bagga**, **Robert Tinn**, **Sam Preston**, **Rajesh Rao**, **Mu Wei**, **Naveen Valluri**, **Cliff Wong**, **Matthew P. Lungren**, **Tristan Naumann**, **and Hoifung Poon**
Microsoft Research

Abstract

Contrastive pretraining on parallel image-text data has attained great success in vision-language processing (VLP), as exemplified by CLIP and related methods. However, prior explorations tend to focus on general domains in the web. Biomedical images and text are rather different, but publicly available datasets are small and skew toward chest X-ray, thus severely limiting progress. In this paper, we conducted by far the largest study on biomedical VLP, using 15 million figure-caption pairs extracted from biomedical research articles in PubMed Central. Our dataset (PMC-15M) is two orders of magnitude larger than existing biomedical image-text datasets such as MIMIC-CXR, and spans a diverse range of biomedical images. The standard CLIP method is suboptimal for the biomedical domain. We propose BiomedCLIP with domain-specific adaptations tailored to biomedical VLP. We conducted extensive experiments and ablation studies on standard biomedical imaging tasks from retrieval to classification to visual question-answering (VQA). BiomedCLIP established new state of the art in a wide range of standard datasets, substantially outperformed prior VLP approaches. Surprisingly, BiomedCLIP even outperformed radiology-specific state-of-the-art models such as BioViL on radiology-specific tasks such as RSNA pneumonia detection, thus highlighting the utility in large-scale pretraining across all biomedical image types. We release our models at aka.ms/biomedclip to facilitate future research in biomedical VLP.

1 Introduction

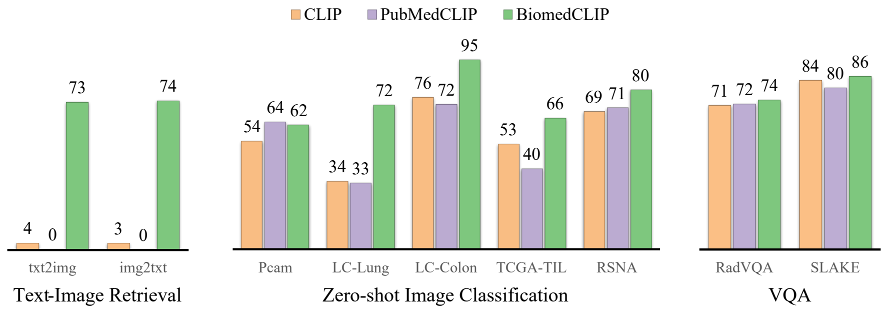


Figure 1: BiomedCLIP significantly outperforms CLIP (Radford et al., 2021) and PubMedCLIP (Eslami et al., 2021) across various biomedical vision-language datasets.

Parallel image-text data abounds in general domains, such as web images and captions. Contrastive pretraining has proven to be an effective approach to leverage such parallel data for self-supervised vision-language learning, as illustrated by CLIP (Radford et al., 2021) and related methods (Jia et al., 2021; inter alia). By learning to map corresponding image and text pairs closer

*These authors contributed equally to this work.

and non-corresponding ones further apart, the resulted foundation models can significantly improve downstream vision-language tasks such as cross-modal retrieval (Lin et al., 2014), image classification (Deng et al., 2009), and visual question answering (VQA) (Antol et al., 2015).

While successful in general domains, such pretrained models are ill suited for biomedical applications, as biomedical images and text are drastically different from standard web content. Domain-specific pretraining of large language models has been shown to help biomedical NLP applications (Lee et al., 2020; Huang et al., 2019; Gu et al., 2021; Luo et al., 2022). In biomedical vision-language processing (VLP), however, progress is hindered by limited availability of parallel data. MIMIC-CXR (Johnson et al., 2019) is the largest public image-text dataset in biomedicine, but it comprises entirely chest X-ray images and reports, and only has 228k pairs.

In this paper, we conducted a large-scale study on domain-specific pretraining for biomedical VLP, by curating figure-caption pairs from biomedical research articles in PubMed Central. This yields a dataset with 15 million biomedical image-text pairs (PMC-15M), which is two orders of magnitude larger than MIMIC-CXR and covers a diverse range of image types (Table 1 and Figure 3 to 5).

We find that the standard CLIP method is suboptimal for the biomedical domain, due to its large difference from the web and general domains. We conducted a thorough study on potential domain-specific adaptations (e.g., encoders and batch size) to identify best practice for biomedical VLP. Based on the results, we propose BiomedCLIP, which substantially outperforms alternative approaches. We conducted extensive experiments on standard biomedical imaging tasks such as retrieval, classification, and visual question-answering (VQA). BiomedCLIP established new state of the art in a wide range of standard datasets (shown in Figure 1). Interestingly, with large-scale pretraining across diverse biomedical image types, BiomedCLIP even outperformed radiology-specific state-of-the-art models such as BioViL (Boecking et al., 2022) on radiology-specific tasks such as RSNA pneumonia detection (Shih et al., 2019). To facilitate future research in biomedical VLP, we release our BiomedCLIP models at aka.ms/biomedclip.

2 Methods

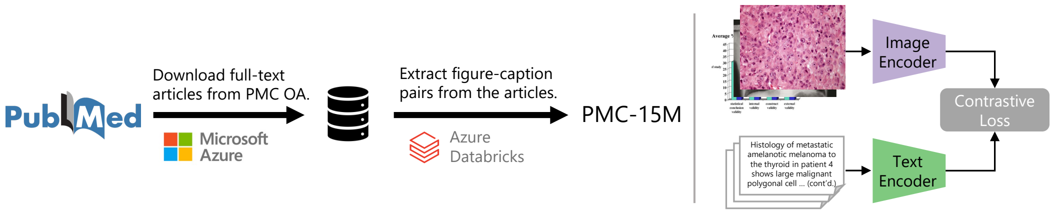

PubMed
Download full-text articles from PMC OA.
Microsoft Azure
Extract figure-caption pairs from the articles.
Azure Databricks
PMC-15M
Image Encoder
Contrastive Loss
Histology of metastatic amelanotic melanoma to the thyroid in patient 4 shows large malignant polygonal cell ... (cont'd.)
Text Encoder

Figure 2: Overview of PMC-15M creation pipeline (left) and BiomedCLIP pretraining (right).

2.1 PMC-15M: A large parallel image-text dataset for biomedicine

Motivation Pretraining on parallel image-text data is key to the success of general-domain vision-language models such as CLIP (Radford et al., 2021), DALLE-E (Ramesh et al., 2021) and Stable Diffusion (Rombach et al., 2022). Therefore, increasing efforts have been made in creating large datasets from mining web images and captions (Sharma et al., 2018; Changpinyo et al., 2021; Srinivasan et al., 2021; Schuhmann et al., 2022). In the biomedical domain, however, parallel image-text datasets are still relatively small, ranging between 15k-228k pairs, e.g., MIMIC-CXR (Johnson et al., 2019), CheXpert (Irvin et al., 2019), ARCH (Gamper & Rajpoot, 2021), and ROCO (Pelka et al., 2018). Additionally, these datasets skew predominantly towards chest X-ray. Prior studies in the general domains have demonstrated the advantage in pretraining on diverse, large-scale datasets (Radford et al., 2021). To facilitate large-scale vision-language pretraining in biomedicine, we created to our knowledge the largest biomedical image-text dataset by mining PubMed Central (PMC) articles. PubMed is a comprehensive repository of biomedical research papers. Prior use of PubMed focuses on leveraging text for pretraining biomedical language models (e.g., PubMedBERT (Gu et al., 2021), BioGPT (Luo et al., 2022)). Here, we instead leverage the abundant figure-caption pairs in PMC full-text articles for vision-language pretraining.

Data Creation PubMed Central contains 4.4 million publicly available full-text articles (as of June 15 2022). We downloaded and extracted compressed directories with complete article packages[1]. Each article is represented as a package of XML, PDF, media, and supplementary materials. We extracted figure files and the matching captions, along with PMID and PMCID of the provenance articles. This yields a dataset PMC-15M with 15 million figure-caption pairs from over 3 million articles.

Statistics Table 1 provides summary statistics of PMC-15M, with training/dev/test splits for pretraining and held-out evaluation. As shown in Figure 3, most captions are much longer than the default max length (77) used by the standard CLIP method . Figures are also much larger than the standard image size (224) as used in general domains.

Articles	Avg. figures per article	Avg. figure size (px)	Avg. caption length (token)	Figure-caption pairs	Training	Dev	Test
3,298,780	4.6	582×702	110	15,282,336	13.9M	13.6k	726k

Table 1: PMC-15M statistics.

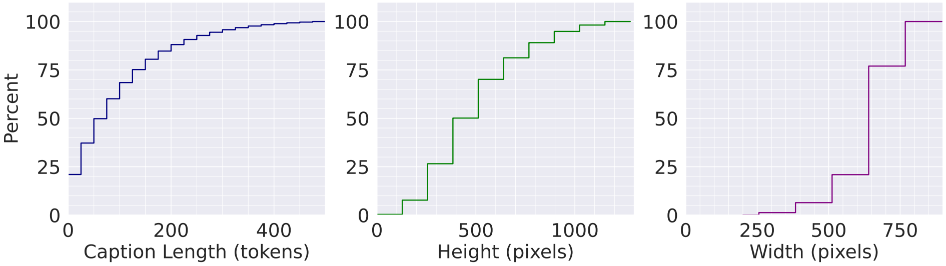


Figure 3: PMC-15M histograms of caption length and figure size.

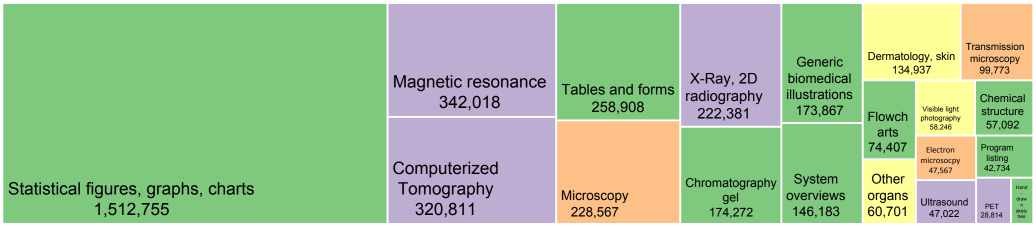


Figure 4: Estimate of PMC-15M figure type distribution based on the image taxonomy in García Seco de Herrera et al. (2015). Block size is proportional to keyword frequency for the corresponding image class.

Diversity To probe the diversity and coverage of image classes in PMC-15M, we generated a word cloud of the captions: Figure 5. As expected, the word distribution is quite distinct compared to general domains. For a more precise estimate of the figure distribution, we used the taxonomy introduced by García Seco de Herrera et al. (2015) with manually assigned keywords for each figure type, and estimated the frequency of each figure type based on the sum of frequencies of its keywords. The numbers likely overcount real class frequency, but are close enough for a ball-park estimate. Figure 4 shows the

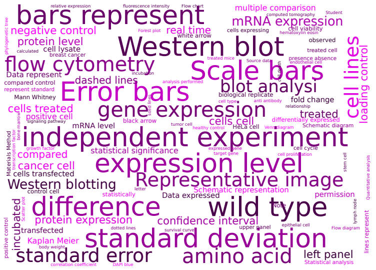


Figure 5: Word cloud of PMC-15M.

[1] https://www.ncbi.nlm.nih.gov/pmc/tools/ftp/#indart

top 20 figure types in PMC-15M. Images in PMC-15M
are extremely diverse, ranging from generic biomedical illustration (e.g., statistical figures, graphs, charts, and tables and forms) to radiography (e.g., magnetic resonance, computerized tomography, and X-ray) to microscopy (e.g., transmission microscopy, and electron microscopy), among others.

2.2 BIOMEDCLIP: LARGE-SCALE VISION-LANGUAGE PRETRAINING FOR BIOMEDICINE

Background We give a brief overview of the CLIP pretraining approach (Radford et al., 2021). Given a batch of N (image, text) pairs, CLIP learns a multimodal embedding space by jointly training an image encoder and a text encoder to maximize the cosine similarity between the image and text embeddings of the N pairs in the batch while minimizing the cosine similarity of the embeddings of the other $N^2 - N$ non-pairs. Concretely, CLIP minimizes the InfoNCE loss (Oord et al., 2018), i.e., a symmetric cross entropy loss over these similarity scores:

$$\mathcal{L} = -\frac{1}{2N}\left(\sum_{i=1}^{N} \log \frac{e^{\cos(\boldsymbol{I}_i, \boldsymbol{T}_i)/\tau}}{\sum_{j=1}^{N} e^{\cos(\boldsymbol{I}_i, \boldsymbol{T}_j)/\tau}} + \sum_{i=1}^{N} \log \frac{e^{\cos(\boldsymbol{I}_i, \boldsymbol{T}_i)/\tau}}{\sum_{j=1}^{N} e^{\cos(\boldsymbol{I}_j, \boldsymbol{T}_i)/\tau}}\right) \tag{1}$$

where τ is a learnable temperature parameter, directly optimized during training as a log-parameterized multiplicative scalar; $\boldsymbol{I}_i$ and $\boldsymbol{T}_i$ are embeddings for the i-th image and text, produced by a linear projection layer on top of the image encoder and text encoder. Rather than initializing with pretrained weights, CLIP trains the image encoder and text encoder from scratch. For the image encoder, CLIP considers two different architectures, ResNet-50 (He et al., 2016) and Vision Transformer (ViT; Dosovitskiy et al., 2020). The text encoder is effectively GPT-2 (Radford et al., 2019) based on transformer (Vaswani et al., 2017).

Adapting CLIP for the biomedical domain Biomedical text and images are drastically different from the web data used in CLIP pretraining. We find that the standard CLIP settings are suboptimal for biomedical vision-language pretraining. We thus conducted a systematic study of potential adaptations and identified a series of domain-specific adaptations for the biomedical domain. We used the optimization loss and cross-modal retrieval results on the validation set to guide our initial exploration and report detailed ablation studies in the Evaluations section.

On the text side, we replace a blank-slate GPT-2 with a pretrained language model more suited for biomedicine. Specifically, we initialize with PubMedBERT, which shows substantial gains from domain-specific pretraining (Gu et al., 2021). Correspondingly, for the tokenizer, we replace Byte-Pair Encoding (BPE; Sennrich et al., 2016) with WordPiece (Kudo & Richardson, 2018), which uses unigram-based likelihood rather than shattering all words to characters and greedily forming larger tokens based on frequency. The original CLIP uses a context of 77 tokens, but biomedical text is typically longer, as shown in Figure 3. We thus increase the context size to 256, which covers 90% of PMC captions. Table 2 shows that both modifications bring substantial improvements over the original CLIP model on the validation set.

				img2txt (%)	txt2img (%)
text encoder	vocab	context length	loss (↓)	R@1(↑)	R@1(↑)
GPT	50k general domain	77	0.6626	64.53	63.56
PubMedBERT	30k domain specific	77	0.5776	69.03	67.41
PubMedBERT	30k domain specific	256	**0.4807**	**73.50**	**72.26**

Table 2: Improvements from text-side domain-specific adaptations, as measured on the PMC-15M validation set.

On the image side, we first evaluated Vision Transformer (ViT) across different scales, ranging from ViT-Small, ViT-Medium, to ViT-Base. The suffix "/16" in the ViT model names refers to the patch size of 16×16 pixels i.e., the input images are divided into patches of this size, and fed through the transformer blocks. As shown in Table 3, we found that larger ViT results in better performance, confirming the importance of model scalability on our new dataset PMC-15M. We used the largest one (ViT-B/16) in all subsequent experiments. Biomedical image understanding often requires fine-grained visual features (Zhang et al., 2020). We conducted a series of experiments to explore the impact of image resolution; see Table 4. By increasing image resolution from 224×224 pixels to

336×336 pixels, we observe significant gains in validation results. But this also leads to doubling of pretraining time. By applying random dropout to 50% of the patches (Li et al., 2022b), we restore the pretraining speed while still attaining decent performance gain, especially with the addition of one epoch of unmasked tuning (out of 8 epochs), which helps close the distribution gap caused by patch dropout. When training longer (40 epochs), we observed that pretraining with patch dropout yields better results than the version without it. This could be attributed to the regularization effect, which encourages the models to focus on smaller details and sub-figures.

				img2txt (%)	txt2img (%)
encoder	trainable params	hidden dim	loss (↓)	R@1(↑)	R@1(↑)
ViT-S/16	22M	384	0.5342	69.45	68.02
ViT-M/16	39M	512	0.5063	71.85	70.22
ViT-B/16	86M	768	**0.4807**	**73.50**	**72.26**

Table 3: Validation performance for various ViT models (Small, Medium, Base). All experiments use PubMedBERT to initialize the text encoder with the maximal context length of 256.

			img2txt (%)	txt2img (%)
image size	training time	loss (↓)	R@1(↑)	R@1(↑)
224px	1.00x	0.4807	73.50	72.26
336px	2.14x	0.4250	**77.12**	**76.24**
336px w. 50% patch dropout	1.19x	0.4415	75.86	74.79
+ unmasked tuning	+ δ	**0.4137**	76.38	75.22

Table 4: Pretraining time and validation performance with image-side domain-specific adaptations. All experiments use ViT-B/16 as the image encoder and PubMedBERT to initialize the text encoder, with the maximal context length of 256.

Finally, we investigated the impact of batch size. We increase batch size by gradient accumulation (Cui et al., 2022), which caches embeddings of each sub-iteration and computes the gradient until the batch size is reached. In Table 5, we studied two batch schedules: (1) training with a constant batch size of 4k for 40 epochs, and (2) training with a batch size of 4k for the first 8 epochs followed by a batch size of 64k for the remaining 32 epochs. We find that instead of using a large batch size at the beginning, starting from a smaller batch size and then gradually increasing the batch size attains the best trade-off of learning speed and stabilization.

	image-to-text retrieval (%)	text-to-image retrieval (%)
batch size	R@1(↑)	R@1(↑)
4k→4k	83.98	82.71
4k→64k	**85.76**	**83.89**

Table 5: Validation performance with constant batch size of 4k for all 40 epochs vs linearly increasing batch size from 4k to 64k in the first 8 epochs.

Putting it all together We pretrained a series of BiomedCLIP models pretrained on PMC-15M using the optimal batch schedule above and compare them with general-domain CLIP models (Radford et al., 2021). As Table 6 shows, large-scale pretraining or continual pretraining on PMC-15M is always helpful, and the best results are generally attained using biomedical pretrained language model (PubMedBERT), large vision transformer, and higher image resolution.

Implementation Our implementation is based on OpenCLIP (Ilharco et al., 2021), an open source software adapted for large-scale distributed training with contrastive image-text supervision. The pretraining experiments were conducted with up to 16 NVIDIA A100 GPUs or 16 NVIDIA V100 GPUs, via PyTorch DDP (Li et al., 2020; Paszke et al., 2019). The hyperparameters are reported in Appendix A. To reduce memory consumption, we enable gradient checkpointing and automatic mixed precision (AMP) with datatype of bfloat16 (whenever supported by the hardware). In addition, we use a sharding contrastive loss (Cherti et al., 2022), which achieves identical gradients to

model	config	data	img2txt (%) R@1(↑)	txt2img (%) R@1(↑)
OpenAI CLIP	ResNet-50-224-GPT/77	*WIT-400M*	10.31	10.38
OpenAI CLIP	ViT-B/16-224-GPT/77	*WIT-400M*	11.82	11.65
BiomedCLIP	ResNet-50-224-GPT/77	*WIT-400M* → PMC-15M	85.65	84.14
BiomedCLIP	ViT-B/16-224-GPT/77	*WIT-400M* → PMC-15M	87.50	86.95
BiomedCLIP	ViT-B/16-224-PMB/256	PMC-15M	87.32	86.66
BiomedCLIP	ViT-B/16-336-PMB/256	PMC-15M	89.14	87.60
BiomedCLIP	ViT-B/16-448-PMB/256	PMC-15M	**89.36**	**88.44**

Table 6: Comparison of BiomedCLIP and general-domain CLIP models on validation performance. "*WIT-400M* → PMC-15M" indicates initialization with OpenAI CLIP weights that were pretrained on WIT-400M (Radford et al., 2021), followed by continual pretraining on PMC-15M. "PMB/256" denotes PubMedBERT with the maximal context length of 256.

InfoNCE (Oord et al., 2018) and reduces memory usage by eliminating redundant computations and only computing the similarities of locally relevant features on each GPU.

3 Evaluations

Task	**Dataset**	**Metric**	**Description**	**Data size**		
				Train	Dev	Test
Cross-Modal Retrieval	PMC-15M	R@k	Given textual description (caption), retrieve the corresponding image, or vice versa.	13.9M	13.6k	726k
Image Classification	PCam	Accuracy	Binary classification on whether a histopathology image of lymph node contains metastatic tumor tissue.	262,144	32,769	32,769
	LC25000 (Lung)	Accuracy	Ternary classification (benign, adenocarcinoma, squamous cell carcinoma) on histopathology images of lung tissue.	-	-	15,000
	LC25000 (Colon)	Accuracy	Binary classification (benign, adenocarcinoma) on histopathology images of colon tissue.	-	-	10,000
	TCGA-TIL	AUCROC	Binary classification on whether lung H&E whole-slide image patches show adenocarcinoma.	-	-	2,480
	RSNA	Accuracy	Binary classification on whether chest X-rays show pneumonia.	18,678	4,003	9,069
VQA	VQA-RAD	Accuracy	Answer clinician questions about radiology images.	3,064	-	451
	SLAKE	Accuracy	Answer clinician questions about radiology images (X-rays, CTs, MRIs).	9,849	2,109	2,070

Table 7: Tasks, datasets, and evaluation metrics used in our biomedical vision-language processing study. The numbers in train, dev, and test represent numbers of image-caption pairs in cross-modal retrieval, and numbers of images in image classification and visual question-answering (VQA). $R@k$ is the recall score among top k results.

3.1 Evaluating Biomedical Vision-Language Models

The overarching objective of pretraining is to improve performance across a broad spectrum of downstream applications. In general domains, comprehensive benchmarks like ELEVATER (Li et al., 2022a) have spurred rapid advances in vision-language pretraining by facilitating direct comparison among pretrained models. In contrast, prior work on biomedical vision-language pretraining tends to use different tasks and datasets for downstream evaluation, e.g., Zhang et al. (2020); Wang et al. (2022); Eslami et al. (2021). To facilitate evaluations of biomedical vision-language pretraining and expedite progress in biomedical vision-language processing, we compile a collection of eight standard vision-language datasets spanning key downstream tasks: image-to-text and text-to-image retrieval, image classification, and visual question answering. Table 7 provides an overview of datasets used in our study.

3.2 Cross-Modal Retrieval

In cross-modal retrieval, we evaluate the efficacy in retrieving the corresponding image from caption (text-to-image retrieval) or vice versa (image-to-text retrieval). The retrieval tasks mirror various image search and text generation in real-world applications, and can be evaluated automatically in held-out image-text pairs. We use a held-out test set from PMC-15M comprising 725,739 PMC figure-caption pairs. pretraining of 725,739 PubMed image-caption pairs. To evaluate the retrieval performance, we follow prior work (Schuhmann et al., 2022) to precompute image and text embeddings and perform approximate nearest neighbour search. See Table 8 for results. General-domain CLIP model performs poorly in the biomedical domain, whereas BiomedCLIP attains remarkably high retrieval accuracy. Out of over 700 thousand candidates, BiomedCLIP's top-5 results contain the correct one over 90% of times, and its top-1 results are correct over 70% of times. To our surprise, the supposedly in-domain model of PubMedCLIP (Eslami et al., 2021) performs even worse than CLIP. On hind sight, the reasons are simple. Despite being named after PubMed, PubMedCLIP only used radiology image/text pairs in continual pretraining of CLIP, which account for a small fraction of images in the biomedical literature (a rough estimate from Figure 4 suggests only about 1% in PMC-15M). Moreover, continual pretraining on a small dataset without extra care (e.g., augmenting with pretraining objective) may lead to catastrophic forgetting (McCloskey & Cohen, 1989). In contrast, with large-scale in-domain pretraining on PMC-15M, even our continual pretrained model (BiomedCLIP ViT-B/16-224-GPT/77) performs very well.

		text-to-image (%)			image-to-text (%)		
model	config	R@1	R@5	R@10	R@1	R@5	R@10
CLIP	ViT-B/16-224-GPT/77	3.50	6.71	8.42	3.43	6.74	8.50
PubMedCLIP	ViT-B/32-224-GPT/77	0.27	0.70	1.00	0.21	0.59	0.86
BiomedCLIP	ViT-B/16-224-GPT/77	71.20	90.44	93.10	69.34	88.34	90.89
BiomedCLIP	ViT-B/16-448-PMB/256	**73.38**	**90.93**	**93.63**	**74.04**	**90.86**	**93.37**

Table 8: Cross-modal retrieval results on PMC-15M test set.

Case Study To understand how BiomedCLIP outperforms general-domain CLIP in biomedical cross-modal retrieval, we show three random examples in Figure 6. In each example, we show the top-4 image retrieval results given the text prompt, with the correct answer shown in a gold box. General CLIP can find images matching common keywords such as "chest X-ray", but have trouble differentiating subtle semantics such as "pleural effusion", "spindle shaped cells", or even important biomedical image categories like Arterial Spin Labeling (ASL). By contrast, BiomedCLIP recognizes not only high-level categories, but also details like "a large pleural effusion on the right". E.g., images in the first row are completely correct according to the caption, except the right-bottom one, which still looks quite alike others. In the second row, BiomedCLIP is able to find the correct answer and additional H&E images with spindle-shaped cells, as described in the prompt, unlike general CLIP.

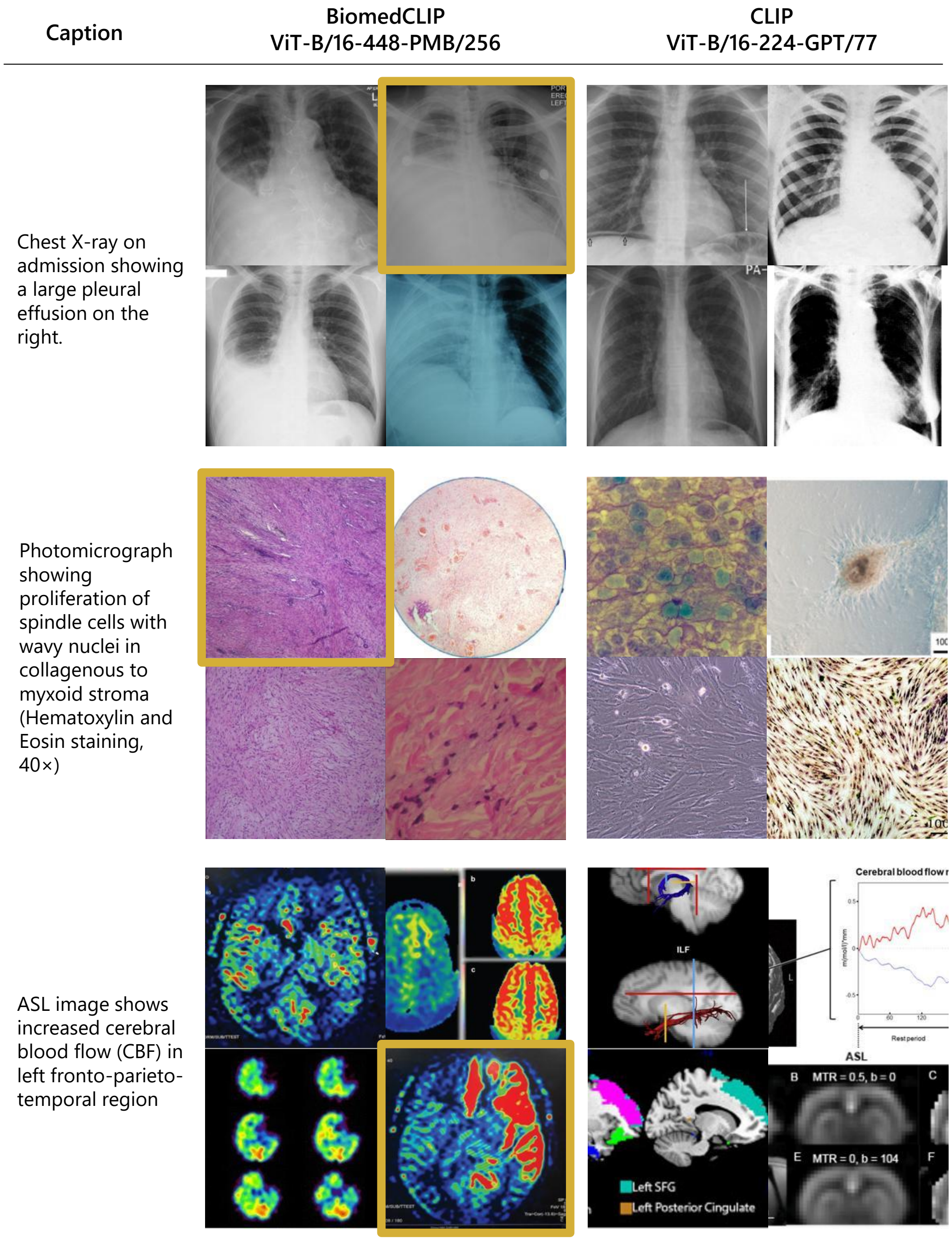

Caption
BiomedCLIP
ViT-B/16-448-PMB/256
CLIP
ViT-B/16-224-GPT/77
Chest X-ray on admission showing a large pleural effusion on the right.
PA
Photomicrograph showing proliferation of spindle cells with wavy nuclei in collagenous to myxoid stroma (Hematoxylin and Eosin staining, 40×)
ASL image shows increased cerebral blood flow (CBF) in left fronto-parieto-temporal region
Cerebral blood flow
ILF
ASL
B MTR = 0.5, b = 0
C
E MTR = 0, b = 104
F
Left SFG
Left Posterior Cingulate

Figure 6: Examples comparing BiomedCLIP and general-domain CLIP on text-to-image retrieval for PMC captions (top-4 predictions). Gold box signifies the corresponding figure for the caption.

3.3 Image Classification

We use the evaluation toolkit ELEVATER (Li et al., 2022a) to facilitate our experiments on image classification. It is an easy-to-use toolkit that can efficiently adapt pretrained vision-language models and automatically tune hyper-parameters. It supports zero-shot, few-shot and full-shot evaluations, with linear probing and full model fine-tuning available for the latter two settings. It also contains twenty image classification datasets collected from various domains, including a biomedical one PatchCamelyon, which we use in our experiments. In addition, we evaluated on three standard biomedical imaging benchmarks LC25000, TCGA-TIL and RSNA. An overview of the datasets can be found in Table 7. See below for details.

Datasets **PatchCamelyon** (PCam) (Veeling et al., 2018) contains $327,680$ color images (96×96px), which were taken from histophathology scans of lymph node sections. The images have been assigned a binary label indicating whether or not they contain metastatic tissue. **LC25000** (Borkowski et al., 2019) contains $25,000$ histopathology images (768×768px). These images were generated by augmentation from a collection of HIPAA-compliant, validated sources originally comprising 750 images of lung tissue (250 benign, 250 adenocarcinomas, and 250 squamous cell carcinomas) and 500 images of colon tissue (250 benign and 250 adenocarcinomas). The dataset is divided into five classes: lung benign tissue, lung adenocarcinoma, lung squamous cell carcinoma, colon adenocarcinoma, and colon benign tissue, with each class containing 5,000 images. **TCGA-TIL** (Saltz et al., 2018a;b) contains $2,480$ image patches (500×500px) that were partitioned from H&E whole-slide images from The Cancer Genome Atlas (TCGA) (Clark et al., 2013) lung adenocarcinoma (LUAD) cases (5.9% of the patches are labelled as LUAD). **RSNA** Pneumonia (Shih et al., 2019) contains about $30,000$ frontal-view chest radiographs collected from the National Institutes of Health's public database of chest X-rays. It contains binary labels classifying pneumonia against normal cases.

Zero-shot settings We evaluate zero-shot performance on BiomedCLIP models, along with three baselines CLIP, MedCLIP and PubMedCLIP. MedCLIP (Wang et al., 2022) extends the pretraining to include large unpaired images and texts through contrastive learning. It uses the pretrained BioClinicalBERT and Swin Transformer (Liu et al., 2021b) as the backbone text encoder and visual encoder respectively, and fine-tunes on MIMIC-CXR and CheXpert datasets. PubMedCLIP (Eslami et al., 2021) fine-tunes CLIP on the Radiology Objects in COntext (ROCO) dataset Pelka et al. (2018), which consists of 80K radiology image-text pairs drawn from PubMed articles. All the models are adapted into ELEVATER for evaluation. Table 9 summarizes the zero-shot results.

model	config	PCam	LC25000 (Lung)	LC25000 (Colon)	TCGA-TIL	RSNA	mean
CLIP	ViT-B/16-224-GPT/77	54.02	33.97	75.86	52.56	68.81	57.04
MedCLIP*	Swin-T/4-224-BioB/77	52.72	33.19	52.53	39.89	66.96	48.90
PubMedCLIP	ViT-B/32-224-GPT/77	**63.58**	32.77	72.39	40.27	70.67	55.94
BiomedCLIP	ViT-B/16-224-GPT/77	61.09	62.75	76.10	67.88	77.10	68.98
BiomedCLIP	ViT-B/16-224-PMB/256	62.16	**72.17**	**94.65**	65.58	**79.72**	**74.86**
BiomedCLIP	ViT-B/16-448-PMB/256	61.68	47.96	70.66	**69.25**	78.81	65.67

Table 9: Zero-shot image classification. AUCROC (%) for TCGA-TIL; accuracy (%) for others. *MedCLIP was evaluated on a sub-sample of RSNA in its original paper but is evaluated on the full dataset here for head-to-head comparison with other methods.

BiomedCLIP models exhibit superior performance on zero-shot classification, where BiomedCLIP "ViT-B/16-224-PMB/256" achieves the highest overall accuracy (mean of the scores) across all the benchmarks. Interestingly, BiomedCLIP with larger image size 448 does not appear to help in general, especially on LC25000. This is noticeably different from other tasks, where increasing image resolution typically helps. We leave it to future work to study the root cause for this curious performance drop.

model	pretraining data	zero-shot	1%-shot	10%-shot	100%-shot
CLIP	WIT-400M	68.80	-	-	-
MedCLIP	MIMIC-CXR + CheXpert	66.96	-	-	-
PubMedCLIP	ROCO	70.70	-	-	-
GLoRIA	CheXpert	70.00	72.00	78.00	79.00
BioViL	MIMIC-CXR	73.20	80.50	81.20	82.20
BiomedCLIP	PMC-15M	**79.72**	**80.75**	**82.95**	**83.33**

Table 10: Zero-shot and fine-tuned (linear probing) accuracy on RSNA pneumonia image classification dataset. GLoRIA and BioViL results are from the respective papers.

Supervised settings We evaluate few-shot/full-shot performance on the standard radiology benchmark RSNA by linear probing the models with 1%, 10% and 100% of training data, respectively.

See Table 10. Interestingly, by pretraining on diverse data across all biomedical image classes, BiomedCLIP even outperformed the state-of-the-art radiology-specific BioViL model (Boecking et al., 2022) on this radiology benchmark. Furthermore, it is noticeable that BiomedCLIP already outperforms fully supervised BioViL using only 10% of labeled data. As shown in Figure 4 (figure type estimate by keyword frequency, likely resulting in overcount), the radiology-related images in BiomedCLIP are no more than that in MIMIC-CXR used in BioViL pretraining, and the image-text pairs are likely to be much noisier. So it is unlikely that the superior performance of BiomedCLIP on RSNA stems from more radiology-specific pretraining. Instead, the overall large-scale pretraining, even though across other image types, might have helped pretrained a more robust image encoder.

3.4 Medical Visual Question Answering (VQA)

We utilize the METER (Dou et al., 2022) framework to facilitate our experiments on visual question answering (VQA). It formulates the VQA task as a classification task. The core module of METER is a transformer-based *co-attention* multimodal fusion module that produces cross-modal representations over the image and text encodings, which are then fed to a classifier for predicting the final answer. We compare BiomedCLIP with general-domain CLIP, MAML (Model-Agnostic Meta-Learning) network pre-trained only on visual data, and the state-of-the-art PubMedCLIP. All three models were fine-tuned for VQA tasks using the QCR (Question answering via Conditional Reasoning) framework (Zhan et al., 2020) that alternatively uses a MLP-based attention networks with conditional reasoning as the fusion module. We evaluated the models on two standard datasets below.

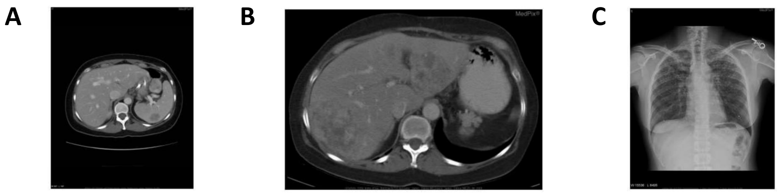
Datasets **Radiology VQA** (VQA-RAD) (Lau et al., 2018) consists of 315 radiology images and $3,515$ question-answer pairs that were manually constructed by clinicians. Images in the test set are also present in training set but the question-answer pairs do not overlap. **SLAKE** (English only) (Liu et al., 2021a) consists of 642 radiology images and over $7,000$ question-answer pairs annotated by experienced physicians. It covers more human body parts than VQA-RAD and does not have common images between the training and test sets. See Table 7 for details.

Table 11 presents our evaluation results on the two datasets. Results for MAML, CLIP and PubMedCLIP are from Eslami et al. (2021). We report the overall accuracy as well as respective accuracies for open-ended questions and closed-ended questions. Again, BiomedCLIP "ViT-B/16-448-PMB/256" exhibits superior performance compared to the radiology-specific state-of-the-art PubMedCLIP on both radiology benchmarks, with about three points and five points increases in overall accuracy respectively. BiomedCLIP's gains are particularly pronounced for open-ended questions in VQA-RAD and for all question types in SLAKE.

	fine-tuning	VQA-RAD			SLAKE		
model	framework	open	closed	overall	open	closed	overall
MAML	QCR	56.0	77.9	69.2	76.8	80.6	78.3
CLIP	QCR	59.9	79.4	71.3	78.6	81.0	79.5
PubMedCLIP	QCR	60.1	**80.0**	72.1	78.4	82.5	80.1
BiomedCLIP	METER	**67.6**	79.8	**75.2**	**82.5**	**89.7**	**85.4**

Table 11: Test accuracy for visual question answering on VQA-RAD and SLAKE. Results of MAML, CLIP and PubMedCLIP are from the PubMedCLIP paper (Eslami et al., 2021).

Case Study We evaluated BiomedCLIP on examples reported in the PubMedCLIP paper (Eslami et al., 2021), where all prior state-of-the-art models (including PubMedCLIP) failed to answer correctly. See Figure 7. For example B, prior models fail to return the right answer, whereas for example C, their answers indicate that they even fail to understand what the question is about. BiomedCLIP nails both answers perfectly. For example A, MEVF misidentifies the body part as displayed in the image, whereas QCR and PubMedCLIP misinterpret the question as a binary one (yes/no). While BiomedCLIP didn't get the correct answer either, it correctly identified the relevant organ as presented in the image.



Question:	Are there multiple or just 1 metastatic focus?	What are the hyperdensities on the periphery of the image?	What is the biological sex of the patient?
Answer:	one	ribs	female
MEVF:	right chest ✗	storage of urine ✗	inflammation … ✗
QCR:	no ✗	intestine ✗	treat brain diseases … ✗
PubMedCLIP:	yes ✗	spinal cord ✗	nodule ✗
BiomedCLIP:	right lobe of liver ✗	ribs ✓	female ✓

Figure 7: Examples from VQA-RAD where all previous VQA models (including the previous state-of-the-art PubMedCLIP) failed to produce the correct answer, as reported in the PubMedCLIP paper (Eslami et al., 2021).

4 LIMITATIONS

While BiomedCLIP shows clear benefits of large-scale domain-specific pretraining for biomedical vision-language processing, there are several limitations in our current method: 1) Compound figures are prevalent in scientific literature. Our data pipeline does not impose special treatment for compound figures. Splitting them into sub-figures can largely increase the data size and potentially lead to better vision-language representations. How to split the corresponding captions is also challenging. 2) Besides captions, the context of in-line references can also be naturally paired with the corresponding figures to create additionally training signals. Our current data pipeline leaves it untouched. 3) Due to the compute constraints, the largest vision encoder we use is ViT-B, which is still relatively smaller compared with ViT-L, ViT-H and ViT-G. The input image size 448 is also bounded by the compute. As shown in Figure 3, figures whose raw size is below 448 only account for less 25% of the entire PMC-15M. 4) We observe a performance gap between pretraining and several downstream image classification tasks (e.g., PCam, LC25000, and RSNA). The best BiomedCLIP model "ViT-B/16-448-PMB/256" does not perform best on these tasks. BiomedCLIP models that use smaller image size or are pretrained shorter epochs show better performance. This is because PubMed articles are often curated and contain images within a larger study to optimize the relevant finding, and thus its distribution would be expected to skew toward less common pathologies than would be seen in a typical medical setting.

5 RELATED WORK

Our work is most relevant to vision-language representation learning. In the general domain, Joulin et al. (2016); Li et al. (2017); Sariyildiz et al. (2020); Desai & Johnson (2021) learn a good visual representations by predicting the captions from images, but they are limited to to small datasets such as Flickr (Joulin et al., 2016; Li et al., 2017) and COCO Captions (Sariyildiz et al., 2020; Desai & Johnson, 2021), and the learned models do not produce vision-language representations needed for tasks like cross-modal retrieval. Frome et al. (2013); Faghri et al. (2017) are the first to propose the deep visual-semantic embedding models, followed by improved versions that leverage object detectors, dense feature maps, or multi-attention layers (Socher et al., 2014; Karpathy et al., 2014; Nam et al., 2017; Kiros et al., 2018; Li et al., 2019; inter alia). Recent advances focus on cross-modal attention layers, e.g.,Liu et al. (2019); Tan & Bansal (2019); Lu et al. (2019); Chen et al. (2020). While they show superior performance, they are much slower and impractical to train on large-scale data. Radford et al. (2021); Jia et al. (2021) demonstrate that the simple pretraining task of matching captions with images in a constrastive learning setting is an efficient and scalable way to learn the SOTA vision-language representations from large-scale web data.

In the biomedical domain, most studies of image-text pretraining focus on chest X-ray (CXR) with limited amounts of training data. ConVIRT (Zhang et al., 2020) pioneers the use of naturally occurring pairing of medical images and text data for self-supervision and demonstrates the potential of contrastive loss in pretraining. Their image encoders benefit downstream CXR classification and retrieval tasks. Their text encoder inherits a general-domain vocabulary, which leads to frequent encounters of out-of-vocabulary words when processing medical text. While the word-piece tokenization mitigates this issue, common biomedical terms are often shattered into pieces, leading to a suboptimal performance (Gu et al., 2021). GLoRIA (Huang et al., 2021) uses the same general-domain vocabulary and extends ConVIRT by jointly learning multimodal global and local representations of medical images via contrasting attention weighted image regions with words in the paired reports. LoVT (Müller et al., 2022) proposes a similar pretraining approach that aligns local representations of image regions and report sentences. Liao et al. (2021) learn multimodal representations by maximizing the mutual information between local features of medical images and text. PubMedCLIP (Eslami et al., 2021) fine-tunes the the original CLIP on 80K radiology image-caption pairs from the ROCO dataset (Pelka et al., 2018), which is collected from PubMed Central (Roberts, 2001), but has limited scale and diversity due to heavy filtering and manual revision. PubMedCLIP runs evaluations only on medical visual question answering. Wang et al. (2021) propose a transformer-based framework for mix-up image-text pretraining, which uses masked vision/language modeling for image-only or text-only data and uses binary cross entropy for paired image-text data. They demonstrate the benefits of adopting pretrained models in three CXR applications, i.e., classification, retrieval and image regeneration. MedCLIP (Wang et al., 2022) similarly extends contrastive learning to cover image-only and text-only data. They additionally introduce medical knowledge to alleviate false negatives. MedAug (Vu et al., 2021) leverages patient metadata to select positive image pairs that go beyond augmentations of the same image. BioViL (Boecking et al., 2022) improves contrastive learning in self-supervised vision-language processing with principled textual semantic modeling. It achieves the state of the art not only in radiology natural language inference but also in a range of CXR benchmarks. Iyer et al. (2022) show self-supervised multimodal pretraining on CXR data consistently outperforms ImageNet-pretrained models for CXR interpretation. We refer readers to Heiliger et al. (2022); Huang et al. (2020) for a systematic review of multimodal learning in radiology.

6 Conclusion

We present to our knowledge the largest study on biomedical vision-language pretraining using 15 million figure-caption pairs mined from PubMed Central full-text articles. Our pretraining data is at least two orders of magnitude larger than prior datasets, spanning an extremely diverse range of biomedical images. We conducted a systematic study on domain-specific adaptations for the biomedical domain and propose BiomedCLIP for biomedical vision-language processing. In extensive experiments on eight standard biomedical datasets, BiomedCLIP establishes new state of the art on various tasks such as cross-modal retrieval, image classification, and visual question answering. Future directions include: further improvement on pretraining and fine-tuning; multi-modal generation; real-world applications such as image search, digital pathology, multi-modal fusion for precision medicine.

Acknowledgements

We give warm thanks to Jianwei Yang & Chunyuan Li for helpful advice on pretraining/evaluation.

References

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

Benedikt Boecking, Naoto Usuyama, Shruthi Bannur, Daniel C. Castro, Anton Schwaighofer, Stephanie Hyland, Maria Wetscherek, Tristan Naumann, Aditya Nori, Javier Alvarez-Valle, Hoifung Poon, and Ozan Oktay. Making the most of text semantics to improve biomedical vision–

language processing. In *European Conference on Computer Vision (ECCV)*, pp. 1–21. Springer, 2022.

Andrew A Borkowski, Marilyn M Bui, L Brannon Thomas, Catherine P Wilson, Lauren A DeLand, and Stephen M Mastorides. Lung and colon cancer histopathological image dataset (lc25000). *arXiv preprint arXiv:1912.12142*, 2019.

Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3558–3568, 2021.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning. In *European conference on computer vision*, pp. 104–120. Springer, 2020.

Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive language-image learning. *arXiv preprint arXiv:2212.07143*, 2022.

Kenneth Clark, Bruce Vendt, Kirk Smith, John Freymann, Justin Kirby, Paul Koppel, Stephen Moore, Stanley Phillips, David Maffitt, Michael Pringle, et al. The cancer imaging archive (tcia): maintaining and operating a public information repository. *Journal of digital imaging*, 26(6): 1045–1057, 2013.

Quan Cui, Boyan Zhou, Yu Guo, Weidong Yin, Hao Wu, Osamu Yoshie, and Yubo Chen. Contrastive vision-language pre-training with limited resources. In *European Conference on Computer Vision*, pp. 236–253. Springer, 2022.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Karan Desai and Justin Johnson. Virtex: Learning visual representations from textual annotations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11162–11173, 2021.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Zi-Yi Dou, Yichong Xu, Zhe Gan, Jianfeng Wang, Shuohang Wang, Lijuan Wang, Chenguang Zhu, Pengchuan Zhang, Lu Yuan, Nanyun Peng, Zicheng Liu, and Michael Zeng. An empirical study of training end-to-end vision-and-language transformers. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. URL `https://arxiv.org/abs/2111.02387`.

Sedigheh Eslami, Gerard de Melo, and Christoph Meinel. Does clip benefit visual question answering in the medical domain as much as it does in the general domain? *arXiv preprint arXiv:2112.13906*, 2021.

Fartash Faghri, David J Fleet, Jamie Ryan Kiros, and Sanja Fidler. Vse++: Improving visual-semantic embeddings with hard negatives. *arXiv preprint arXiv:1707.05612*, 2017.

Andrea Frome, Greg S Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Marc'Aurelio Ranzato, and Tomas Mikolov. Devise: A deep visual-semantic embedding model. *Advances in neural information processing systems*, 26, 2013.

Jevgenij Gamper and Nasir Rajpoot. Multiple instance captioning: Learning representations from histopathology textbooks and articles. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16549–16559, 2021.

Alba García Seco de Herrera, Henning Müller, and Stefano Bromuri. Overview of the ImageCLEF 2015 medical classification task. In *Working Notes of CLEF 2015 (Cross Language Evaluation Forum)*, September 2015.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)*, 3(1): 1–23, 2021.

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*, pp. 770–778, 2016.

Lars Heiliger, Anjany Sekuboyina, Bjoern Menze, Jan Egger, and Jens Kleesiek. Beyond medical imaging-a review of multimodal deep learning in radiology. 2022.

Kexin Huang, Jaan Altosaar, and Rajesh Ranganath. Clinicalbert: Modeling clinical notes and predicting hospital readmission. *arXiv:1904.05342*, 2019.

Shih-Cheng Huang, Anuj Pareek, Saeed Seyyedi, Imon Banerjee, and Matthew P Lungren. Fusion of medical imaging and electronic health records using deep learning: a systematic review and implementation guidelines. *NPJ digital medicine*, 3(1):1–9, 2020.

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*, pp. 3942–3951, 2021.

Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. Openclip, July 2021. URL `https://doi.org/10.5281/zenodo.5143773`. If you use this software, please cite it as below.

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

Niveditha S. Iyer, Aditya Gulati, Oishi Banerjee, Cécile Logé, Maha Farhat, Agustina D. Saenz, and Pranav Rajpurkar. Self-supervised pretraining enables high-performance chest x-ray interpretation across clinical distributions. *medRxiv*, 2022. doi: 10.1101/2022.11.19.22282519. URL `https://www.medrxiv.org/content/early/2022/11/25/2022.11.19.22282519`.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning*, pp. 4904–4916. PMLR, 2021.

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

Armand Joulin, Laurens van der Maaten, Allan Jabri, and Nicolas Vasilache. Learning visual features from large weakly supervised data. In *European Conference on Computer Vision*, pp. 67–84. Springer, 2016.

Andrej Karpathy, Armand Joulin, and Li F Fei-Fei. Deep fragment embeddings for bidirectional image sentence mapping. *Advances in neural information processing systems*, 27, 2014.

Jamie Kiros, William Chan, and Geoffrey Hinton. Illustrative language understanding: Large-scale visual grounding with image search. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 922–933, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1085. URL `https://aclanthology.org/P18-1085`.

Taku Kudo and John Richardson. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 66–71, Brussels, Belgium, November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-2012. URL `https://aclanthology.org/D18-2012`.

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

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240, 2020.

Ang Li, Allan Jabri, Armand Joulin, and Laurens Van Der Maaten. Learning visual n-grams from web data. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 4183–4192, 2017.

Chunyuan Li, Haotian Liu, Liunian Harold Li, Pengchuan Zhang, Jyoti Aneja, Jianwei Yang, Ping Jin, Yong Jae Lee, Houdong Hu, Zicheng Liu, and Jianfeng Gao. Elevater: A benchmark and toolkit for evaluating language-augmented visual models. *Neural Information Processing Systems*, 2022a.

Kunpeng Li, Yulun Zhang, Kai Li, Yuanyuan Li, and Yun Fu. Visual semantic reasoning for image-text matching. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4654–4662, 2019.

Shen Li, Yanli Zhao, Rohan Varma, Omkar Salpekar, Pieter Noordhuis, Teng Li, Adam Paszke, Jeff Smith, Brian Vaughan, Pritam Damania, et al. Pytorch distributed: Experiences on accelerating data parallel training. *arXiv preprint arXiv:2006.15704*, 2020.

Yanghao Li, Haoqi Fan, Ronghang Hu, Christoph Feichtenhofer, and Kaiming He. Scaling language-image pre-training via masking. *arXiv preprint arXiv:2212.00794*, 2022b.

Ruizhi Liao, Daniel Moyer, Miriam Cha, Keegan Quigley, Seth Berkowitz, Steven Horng, Polina Golland, and William M Wells. Multimodal representation learning via maximization of local mutual information. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 273–283. Springer, 2021.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pp. 740–755. Springer, 2014.

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

Fenglin Liu, Yuanxin Liu, Xuancheng Ren, Xiaodong He, and Xu Sun. Aligning visual regions and textual concepts for semantic-grounded image representations. *Advances in Neural Information Processing Systems*, 32, 2019.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021b.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32, 2019.

Renqian Luo, Liai Sun, Yingce Xia, Tao Qin, Sheng Zhang, Hoifung Poon, and Tie-Yan Liu. Biogpt: generative pre-trained transformer for biomedical text generation and mining. *Briefings in Bioinformatics*, 23(6), 2022.

Michael McCloskey and Neal J Cohen. Catastrophic interference in connectionist networks: The sequential learning problem. In *Psychology of learning and motivation*, volume 24, pp. 109–165. Elsevier, 1989.

Philip Müller, Georgios Kaissis, Congyu Zou, and Daniel Rueckert. Joint learning of localized representations from medical images and reports. In *European Conference on Computer Vision*, pp. 685–701. Springer, 2022.

Hyeonseob Nam, Jung-Woo Ha, and Jeonghee Kim. Dual attention networks for multimodal reasoning and matching. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 299–307, 2017.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

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

Obioma Pelka, Sven Koitka, Johannes Rückert, Felix Nensa, and Christoph M Friedrich. 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*, pp. 180–189. Springer, 2018.

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

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

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, pp. 8821–8831. PMLR, 2021.

Richard J Roberts. Pubmed central: The genbank of the published literature, 2001.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10684–10695, 2022.

J Saltz, R Gupta, L Hou, et al. Tumor-infiltrating lymphocytes maps from tcga h&e whole slide pathology images. *Cancer Imaging Arch*, 2018a.

Joel Saltz, Rajarsi Gupta, Le Hou, Tahsin Kurc, Pankaj Singh, Vu Nguyen, Dimitris Samaras, Kenneth R Shroyer, Tianhao Zhao, Rebecca Batiste, et al. Spatial organization and molecular correlation of tumor-infiltrating lymphocytes using deep learning on pathology images. *Cell reports*, 23(1):181–193, 2018b.

Mert Bulent Sariyildiz, Julien Perez, and Diane Larlus. Learning visual representations with caption annotations. In *European Conference on Computer Vision*, pp. 153–170. Springer, 2020.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *arXiv preprint arXiv:2210.08402*, 2022.

Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1715–1725, Berlin, Germany, August 2016. Association for Computational Linguistics. doi: 10.18653/v1/P16-1162. URL `https://aclanthology.org/P16-1162`.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 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)*, pp. 2556–2565, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1238. URL `https://aclanthology.org/P18-1238`.

George Shih, Carol C Wu, Safwan S Halabi, Marc D Kohli, Luciano M Prevedello, Tessa S Cook, Arjun Sharma, Judith K Amorosa, Veronica Arteaga, Maya Galperin-Aizenberg, et al. Augmenting the national institutes of health chest radiograph dataset with expert annotations of possible pneumonia. *Radiology. Artificial intelligence*, 1(1), 2019.

Richard Socher, Andrej Karpathy, Quoc V. Le, Christopher D. Manning, and Andrew Y. Ng. Grounded compositional semantics for finding and describing images with sentences. *Transactions of the Association for Computational Linguistics*, 2:207–218, 2014. doi: 10.1162/tacl_a_00177. URL `https://aclanthology.org/Q14-1017`.

Krishna Srinivasan, Karthik Raman, Jiecao Chen, Michael Bendersky, and Marc Najork. Wit: Wikipedia-based image text dataset for multimodal multilingual machine learning. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 2443–2449, 2021.

Hao Tan and Mohit Bansal. LXMERT: Learning cross-modality encoder representations from transformers. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 5100–5111, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1514. URL `https://aclanthology.org/D19-1514`.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Bastiaan S Veeling, Jasper Linmans, Jim Winkens, Taco Cohen, and Max Welling. Rotation equivariant CNNs for digital pathology. June 2018.

Yen Nhi Truong Vu, Richard Wang, Niranjan Balachandar, Can Liu, Andrew Y. Ng, and Pranav Rajpurkar. Medaug: Contrastive learning leveraging patient metadata improves representations for chest x-ray interpretation. In Ken Jung, Serena Yeung, Mark Sendak, Michael Sjoding, and Rajesh Ranganath (eds.), *Proceedings of the 6th Machine Learning for Healthcare Conference*, volume 149 of *Proceedings of Machine Learning Research*, pp. 755–769. PMLR, 06–07 Aug 2021. URL `https://proceedings.mlr.press/v149/vu21a.html`.

Xiaosong Wang, Ziyue Xu, Leo Tam, Dong Yang, and Daguang Xu. Self-supervised image-text pre-training with mixed data in chest x-rays. *arXiv preprint arXiv:2103.16022*, 2021.

Zifeng Wang, Zhenbang Wu, Dinesh Agarwal, and Jimeng Sun. Medclip: Contrastive learning from unpaired medical images and text. *arXiv preprint arXiv:2210.10163*, 2022.

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

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

A Hyperparameters

Hyperparameters	**Value**
optimizer	AdamW (Loshchilov & Hutter, 2017)
peak learning rate	5.0e-4
weight decay	0.2
optimizer momentum	β_1, β_2 = 0.9, 0.98
eps	1.0e-6
learning rate schedule	cosine decay
epochs	40
warmup (in steps)	2000
random seed	0
image mean	(0.48145466, 0.4578275, 0.40821073)
image std	(0.26862954, 0.26130258, 0.27577711)
augmentation	RandomResizedCrop
validation frequency	every epoch

Table 12: Pretraining settings.

B More results for image classification

B.1 Prompts used in BiomedCLIP for zero-shot classification

	BiomedCLIP	
dataset	classes	templates

PCam	`normal lymph node`
	`lymph node metastasis`
		`this is an image of {}`
		`{} presented in image`

LC25000 (Lung)	`lung adenocarcinomas`
	`normal lung tissue`
	`lung squamous cell carcinomas`
		`this is an image of {}`
		`{} presented in image`

LC25000 (Colon)	`colon adenocarcinomas`
	`normal colonic tissue`
		`a photo of {}`
		`{} presented in image`

TCGA-TIL	`none`
	`tumor infiltrating lymphocytes`
		`{} presented in image`

RSNA	`normal lung`
	`pneumonia`
		`a photo of {}`
		`{} presented in image`

Table 13: Prompts used for zero-shot image classification.

B.2 Supervised results on PCam

model	pretraining data	zero-shot	1%-shot	10%-shot	100%-shot
CLIP	WIT-400M	54.02	81.01	82.98	83.16
PubMedCLIP	ROCO	**63.58**	83.04	84.63	84.66
BiomedCLIP	PMC-15M	62.16	**83.80**	**84.51**	**84.91**

Table 14: Zero-shot and fine-tuned (linear probing) accuracy on PCam image classification dataset.