Vision-Language Modelling For Radiological Imaging and Reports In The Low Data Regime

Author name(s) withheld EMAIL(S) WITHHELD
Address withheld

Editors: Under Review for MIDL 2023

Abstract

This paper explores training medical vision-language models (VLMs) – where the visual and language inputs are embedded into a common space – with a particular focus on scenarios where paired image-text training data is limited, as is often the case in medical imaging domains. Using subsets of the MIMIC-CXR dataset of chest X-rays and associated reports, we study the ability of several candidate methods to improve low-data performance at a range of different dataset sizes. These methods are: (i) initializing models with weights from image-only and text-only self-supervised pre-training. (ii) using a variety of contrastive loss-functions including (a) standard approaches (i.e. InfoNCE) (b) additional image-only and text-only supervision (e.g. FILIP, GLoRIA) (c) different model architectures, i.e. ResNet, ViT. As well as exploring the trained model's performance at the global image-text retrieval, we also measure this performance transfers to other downstream tasks. This report concludes with a set of recommendations for researchers aiming to train vision-language models on other medical imaging modalities where data may be more scarce than in the Chest X-ray case. To facilitate this further research, we will make our code and models publicly available.

Keywords: Vision-Language Modelling, Contrastive Learning, Chest X-Ray

1. Introduction

Recently, there has been significant progress in the field of vision-language modelling (VLM), for example (Radford et al., 2021; Alayrac et al., 2022; Jia et al., 2021). This has a wide range of implications, however it is particularly important to medical imaging for multiple reasons. Firstly, a major bottleneck in the advancement of deep learning in the medical domain is the difficulty in obtaining large amounts of labelled data; annotations must usually be performed by a human expert, which is both time-consuming and expensive. This ultimately limits the size of training datasets. with many possible diagnoses being rare. Therefore categorising images into a discrete set of classes to annotate is challenging and requires significant domain knowledge. A possible solution to this is using supervision directly from radiological reports, since these exist for almost every scan taken and can be retroactively extracted from hospital PACS and thus are scalable to very large datasets. Furthermore, since reports are free-text, they are more flexible and can much wider range of diseases and conditions than a strict, tabular annotation scheme.

Finally, many medical imaging tasks are intrinsically multi-modal, requiring both visual and textual information. For example, in the context of chest X-ray reports, the text often provides important information about the patient's medical history; an automated disease classification model, for example, would likely benefit from this information.

© 2023 Anonymous.

There are, however, important differences between medical vision-language tasks (applied to medical scans and associated reports) and general-domain vision-language problems (usually applied to image-caption pairs scraped from the internet). Firstly, general-domain VLM techniques use incredibly large datasets of paired images and text. For example, CLIP is trained on 400 million image-text pairs. Clearly, this is not feasible in the medical domain, where the size of datasets is limited by factors such as the availability of medical imaging equipment, concerns about patient privacy and the cost of obtaining images. For some medical domains, researchers can realistically expect no more than a few thousand image-text pairs. It should be also be noted that the nature of 'captions' provided by radiological reports is very different to that of image captions scraped from the internet. Reports will often comment on changes from previous scans, or explicitly mention the *absence of a particular feature*, rather than describing the image in a more general sense. Many will contain very similar information, for example 'no evidence of pneumonia' and 'healthy lungs'. Standard contrastive learning approaches used in VLM, e.g. InfoNCE (Oord et al., 2018), would treat these reports and their corresponding images as negative matches for each other, despite there being little semantic difference between them. This is a problem, since it means that the model will learn spuriously from such false-negative examples, resulting in a noisy training signal. In addition, reports are usually much longer than alt-text image captions and contain a large amount of specialized vocabulary, abbreviations and short-hand.

In this paper we conduct a study of possible methods to improve vision-language modelling in this unique setting. This done using data from MIMIC-CXR, a large dataset of approximately 200,000 paired chest X-rays and associated reports (Johnson et al., 2019; Bustos et al., 2020). Specifically, we explore: (i) the importance of pre-training the image and text encoders using uni-modal self-supervision (ii) the loss-function used to train the VLM; this includes *local* loss functions (e.g. FILIP, GLoRIA), where the aim is to maximise similarity between individual word embeddings and image patches and also methods which add additional text-only and image-only supervision, e.g. (Boecking et al., 2022; **?**). We also evaluate how these methods perform on downstream tasks, such as **[tbc]**.

By training on progressively smaller subsets of this dataset, we explore how VLM training methods vary in performance across a range of different dataset sizes. To investigate whether these results generalise across modalities, we also apply the proposed methods to a dataset of spinal MRIs and their associated reports, observing similar performance improvements.

This paper's layout is as follows: Section 1.1 outlines related work. In Section 2, we discuss the methods of training VLMs explored in this study. We then report results of experiments applying these methods to chest X-rays (section 3). Section 4 concludes the report, making recommendations for training low-data VLMs and suggesting possible areas of future work.

1.1. Related Work

[refine]

Recent years have seen significant advances in vision-language modelling, largely through applying large models to datasets of millions of image-text pairs (Radford et al., 2021; Jia et al., 2021; Ilharco et al., 2021). There have also been several works published related to

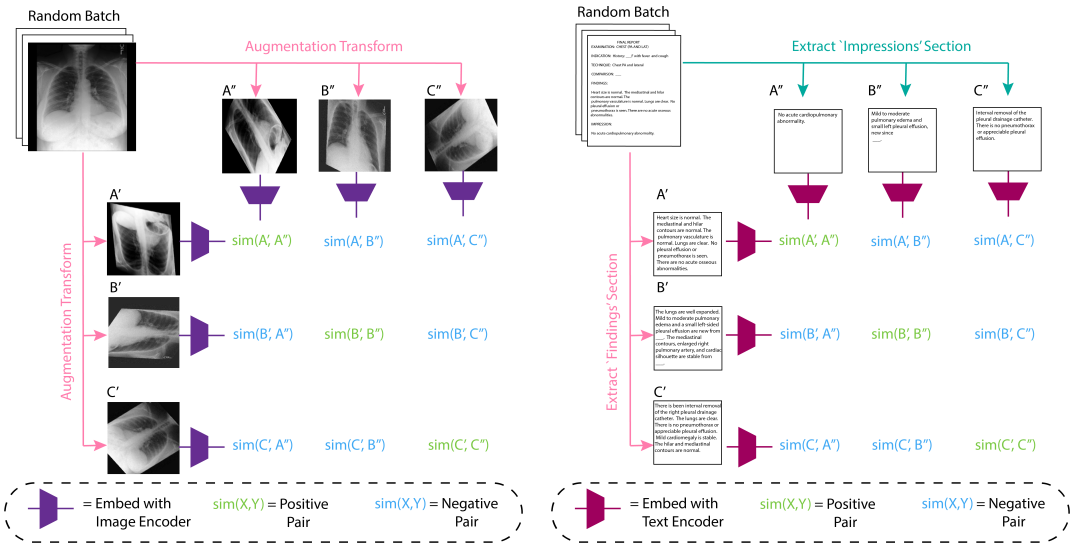
vision-language modelling in medical imaging, mostly in the domain of chest X-rays (CXRs) and their associated reports. This is likely due to the existence of several large-scale, publically available CXR datasets such as MIMIC-CXR (Johnson et al., 2019), CheXpert (Irvin et al., 2019) and PadChest (Bustos et al., 2020). The majority of these works aim to learn representations of images and their associated reports using a contrastive learning paradigm (i.e. matching associated image-text pairs together across a mini-batch) (Boecking et al., 2022; Huang et al., 2021; Zhang et al., 2022; Liao et al., 2021; Müller et al., 2022). Several works have demonstrated that strong performance at this pre-training task correlates with performance across a wide range of downstream tasks, such as classification, segmentation and natural language inference tasks. In a related line of research, several works attempt automated scan report generation (Yang et al.).

While most of these works explore the performance of VLM methods with limited data for downstream tasks, to our best knowledge none consider limited data *for contrastive pretraining*. We believe this is a particularly important problem for medical imaging since few modalities other than CXR have publicly-available datasets of 100,000s image-text pairs. For many researchers, getting datasets of even thousands image-text pairs for their modality of interest is a significant challenge. A few recent works (Segal et al., 2022; Li et al., 2022; Mu et al., 2021) explore the idea of reduced bimodal training data for generic image-text pairs and emphasise the importance of strong self-supervision in each modality both before and during VLM pre-training. However, both these works consider subsets of exceptionally large datasets (millions of image-text pairs), rather than 1000s as considered here.

2. Increasing Data Efficiency During Vision-Language Modelling

We are interested in reducing the amount of image-text pairs used to train a VLM for medical data. In this paper we explore possible methods to achieve this goal, which can broadly separated into two categories: (i) Initializing the image and text encoders with weights from models already trained to perform well on related tasks (ii) adding additional supervision terms during bimodal training or using a more finegrained 'local' contrastive loss function. These categories are described below:

Domain-adapting image and text encoders: Previous works on vision language modelling with small to moderate sized pretraining datasets have shown there can be significant performance gains by fine-tuning already strong image and text encoders rather than training a model from scratch (Segal et al., 2022). In addition, in medical imaging tasks it important to consider how *domain-adapted* the encoder models are; for example, a generic language model is unlikely to outperform a medical specific language-model on a medical NLP task, even if trained on significantly more data. Similarly, (Raghu et al., 2019) demonstrates that for a range of common architectures pre-training on ImageNet have little in terms of downstream performance on medical imaging tasks, compared to random initialization. Therefore we adopt the following paradigm; we start with strong, generic image and text encoders (e.g. an ImageNet-pretrained image model or language model pre-trained on a large, domain-relevant corpora). These encoders are then adapted to the VLM image and text domains by uni-modal training via self-supervision.



(*a*) SimCLR Image Pretraining (*b*) CXR-BERT Specialization

Figure 1: The unimodal pretraining methods used to domain adapt image and text encoders before VLM training.

Image Encoder: To pretrain the image encoder, uni-modal domain adaption is achieved using SimCLR (Chen et al., 2020). In this method, multiple views of each image in the dataset are created using a range of augmentations. The image encoder is then trained contrastively such that embedding vectors corresponding to different views of the same original image should have cosine similarity. This process is illustrated in Figure 1(*a*).

Text Encoder: To train a domain-adapted text encoder, we follow the suggestions of (Boecking et al., 2022). We begin with CXR-BERT, a BERT model and tokenizer trained on a large corpora of generic clinical text (PubMed Abstracts and MIMIC clinical notes), with a comparatively small amount of domain-specific data added in (MIMIC-CXR reports). CXR-BERT is then further domain adapted by an additional pre-training step, whereby the language model to trained to contrastively match a report's 'findings' section to the 'impressions' section (i.e. summary) from the same report. Conceptually, this is analogous to SimCLR, except for instead of different views of the same image, different sections from the same report are matched together. this is shown in Figure 1(*b*).

Local vs. Global Loss Functions: Standard VLM training methods such as CLIP and ALIGN represent each image and each caption as a single global embedding vectors and attempt to maximise similarity between vectors from the same pair, usually via the InfoNCE loss function (Oord et al., 2018):

$$\mathcal{L}_{InfoNCE} = -\frac{1}{2B}\sum_{i=0}^{B}\left[\log\frac{\exp(-\mathbf{v}_i.\mathbf{t}_i/\tau)}{\sum_j^B \exp(-\mathbf{v}_i.\mathbf{t}_j/\tau)} + \log\frac{\exp(-\mathbf{v}_i.\mathbf{t}_i/\tau)}{\sum_j^B \exp(-\mathbf{v}_j.\mathbf{t}_i/\tau)}\right]. \quad (1)$$

Here, $\mathbf{v}_i$ and $\mathbf{t}_i$ represent L2-normalized embedding vectors the image and text respectively for the i-th image-text pair in a batch of size B. τ represents the softmax temperature.

However, in medical imaging, reports often contain multiple unrelated statements, each referring to distinct regions within the image (e.g. 'left lung base atelectasis, mild cardiomegaly, right lung normal'). Since global embedings are usually produced by a pooling operation across the whole image, maintaining this spatial information in the embeddings is challenging. A possible solution to this problem are *local representations*, whereby image patches and text tokens are each represented by their own embedding vector. The model is trained such that the mean maximum patch-to-token and token-to-patch similarity is high for matching image-text pairs and low otherwise. In this paper, we consider GLoRIA, the local loss function proposed in (Huang et al., 2021). Concretely, for a image-text pair with W word embeddings $\mathbf{t} \in \mathbb{R}^{W \times E}$ and local embedding vectors describing M image regions, $\mathbf{v} \in \mathbb{R}^{M \times E}$ (both L2-normalized along the embedding dimension), a similarity matrix $S \in \mathbb{R}^{W \times M}$ is calculated, where s_{ij} represents the similarity between the i-th word and j-th image region. Using, this an attention score is calculated,

$$a_{ij} = \frac{\exp(s_{ij}/\tau_2)}{\sum_{k=1}^{M} \exp(s_{ik}/\tau_2)}. \tag{2}$$

This is then used to calculate an attention-weighted representation of the image for word i in the report, $\mathbf{c}_i = \sum_{j=0}^{M} a_{ij}\mathbf{v}_{ij} \in \mathbb{R}^E$. The similarity of the entire image and report is then given by the matching function

$$Z(\mathbf{v}, \mathbf{t}) = \log(\sum_{i=1}^{W} \exp(\mathbf{c}_i.\mathbf{t}_i/\tau_3)^{\tau_3}. \tag{3}$$

$Z(\mathbf{v}, \mathbf{t})$ then replaces $\mathbf{v}.\mathbf{t}$ in equation 1, leaving us with the following symmetric contrastive loss function for a batch:

$$\mathcal{L}_{GLoRIA} = -\frac{1}{2B} \sum_{k=1}^{B} \left[\log \frac{\exp(-Z(v,t)/\tau)}{\sum_l^B \exp(-Z(v_k, t_l)/\tau)} + \log \frac{\exp(-Z(v,t)/\tau)}{\sum_l^B \exp(-Z(v_l, t_k)/\tau)} \right]. \tag{4}$$

Similar to the original GLoRIA paper, we also explore linear combinations of global and local loss functions, i.e. $\mathcal{L} = \mathcal{L}_{InfoNCE} + \mathcal{L}_{GLoRIA}$.

Additional Supervision During VLM Training: As well as InfoNCE contrastive vision-language matching, we also explore additional text-only and image-only self-supervision during training. To do this we use the DeCLIP framework (Li et al., 2022). For images, this adds SimCLR-style self-supervision. For text, masked-language modelling is used. As well as this additional uni-modal supervision, additional positive text-image pairs are generated positive text-image pairs are generated. Starting with a matching image-text pair, $\{\mathbf{v}_i, \mathbf{t}_i\}$, SimCLR-augmentations are used to generate another view of the image, $\mathbf{v}_i'$. Similarly, another 'view' of the text, $\mathbf{t}_i'$, is created by EDA (Wei and Zou, 2019) text augmentation, involving random synonym replacement, word insertion and deletion as well as swapping word positions. Finally, a nearest-neighbour to the text, $\mathbf{t}_i^{NN}$ is sampled as follows: text embeddings are cached during training in a first-in, first-out (FIFO) queue. Once the queue is populated, the nearest neighbour for each text embedding is

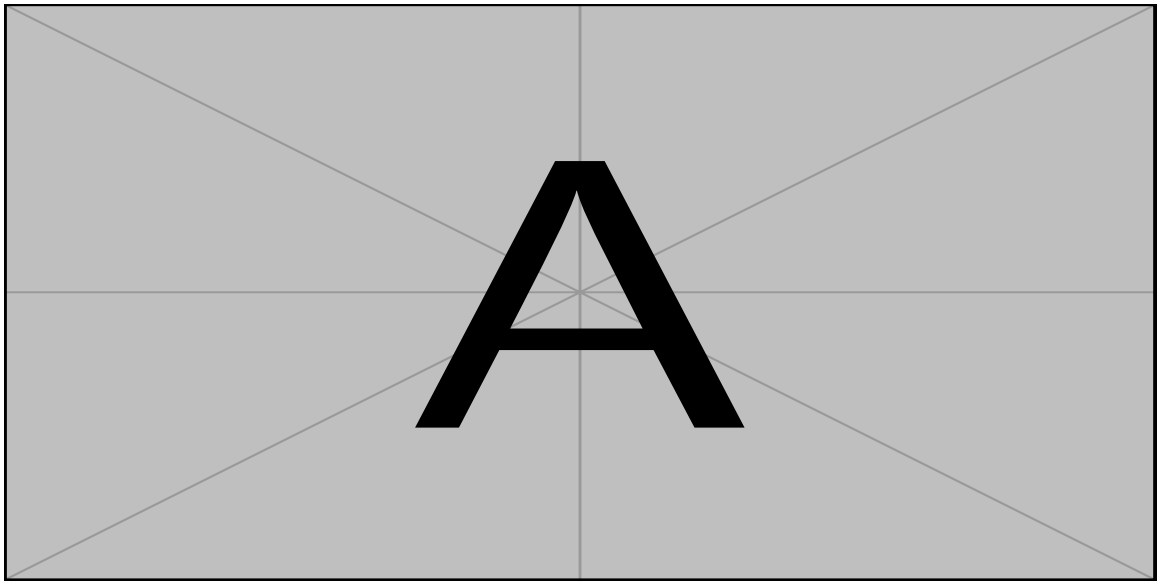


Figure 2: The DeCLIP framework used to add additional supervision during VLM training.

found by cosine similarity and used as a positive example for both the corresponding image. This method results in 6 positive pairs which can be used for contrastive training; $\{\mathbf{v}_i, \mathbf{t}_i\}, \{\mathbf{v}'_i, \mathbf{t}_i\}, \{\mathbf{v}_i, \mathbf{t}'_i\}, \{\mathbf{v}'_i, \mathbf{t}'_i\}, \left\{\mathbf{v}_i, \mathbf{t}_i^{NN}\right\}$ and $\left\{\mathbf{v}'_i, \mathbf{t}_i^{NN}\right\}$. This process is illustrated in Figure 2. For more information, consult the original DeCLIP paper.

Removing False Negatives from Training: Semantically, many reports will contain identical information said in different words, e.g. 'There are no cardiac abnormalities' and 'Heart appears normal'. In a manual check of 200 reports from MIMIC-CXR-JPG, around 50% of reports were ruling out conditions, rather than ruling them in. These examples could equally be treated as positive matches for each others scans, however in standard CLIP-style contrastive learning, they would be treated as negative pairs. To mitigate this issue, we remove all negative image-text pairs $\{\mathbf{v}_i, \mathbf{t}_j\}$ from the contrastive loss where the cosine similarity of the image's positive text and the randomly sampled negative text exceeds some threshold, τ, i.e. $\text{sim}(\mathbf{t}_i, \mathbf{t}_j) \geq \tau$.

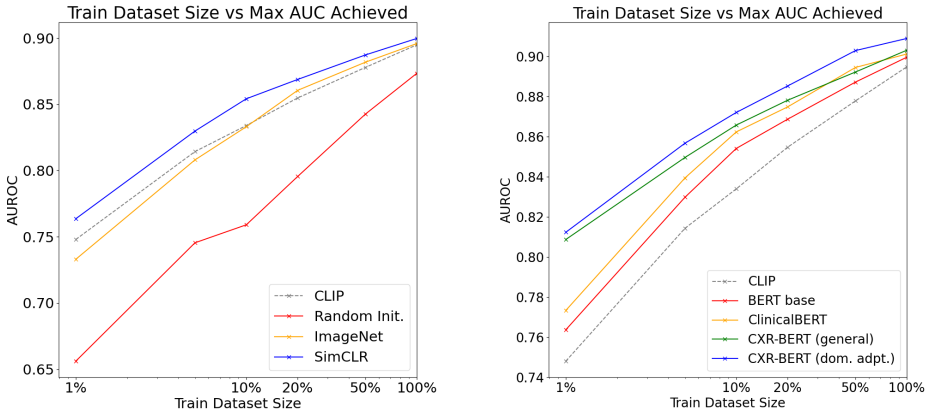
3. Experiments

Dataset: For the experiments in this paper, we use the MIMIC-CXR dataset. (Johnson et al., 2019). We include all reports with at least one frontal X-ray (AP or PA) and an 'Impression' section in the associated report, leaving 181,112 studies (144,781/18,679/17,652 in the training, validation and test sets respectively). Each image is resized to 512×512 pixels by bilinear interpolation. For the reports, we used only the 'Impression' section as input to the text encoder.

Architecture & Implementation Details: In all experiments, we use a ResNet50 image encoder and a BERT text encoder. All configurations are trained using a Adam opti-

mizer with a learning rate of 10^{-3} and $\beta = (0.9, 0.99)$ with a batch size of 20 and softmax temperature of 0.5. Models are trained for 500 iterations and then evaluated for 100 iterations on the same subset of the validation set. Training continues until the validation loss does not improve for 10 successive validation epochs. We evaluate models by measuring the global image-to-text recall on the test set after training has concluded. For all models, standard image augmentations are used; ±5px translation, 10° rotation, 20% brightness, 20% scaling and 10% gaussian noise. Our experiments report AUROC for models trained with 1,2,5,10,20,50 and 100% of the pairs from MIMIC-CXR-JPG training split.

Varying Image and Text Encoders: Figures 3(a) and 3(b) shows results of experiments varying the weight initialization for both the image and text encoders. For the experiments varying image encoder pretrained weights, a standard, uncased BERT base model is used to initialize the text encoder. We experiment with three schemes: random initialization, standard ImageNet pretrained weights and also ImageNet weights with further domain adaption via image-only simCLR pretraining. For the varying text encoder initialization, the simCLR domain adapted image encoder is used. Four BERT models are tested: uncased BERT-base, ClinicalBERT and CXR-BERT with and without further domain adaption. For both sets of experiments, we also plot results of initializing the text and image encoders with ResNet50-CLIP pretrained weights. These figures show a clear benefit of initializing the encoders using weights from models trained via uni-modal self-supervision at all dataset sizes, with the effects particularly pronounced at very small training dataset sizes.



(a) Varying Image Encoder Init. (b) Varying Text Encoder Init.

Figure 3: The results of varying the initialization schemes for both the image and text encoders. For varying image encoder, BERT-base-uncased is used as the inital text encoder. Likewise, for varying text encoder, the simCLR-domain adapted model is used as an image encoder.

Local vs. Global Loss Functions: Figure 4 shows experiments varying the form of contrastive loss function used. In total, three configurations are tested: (i) InfoNCE (equa-

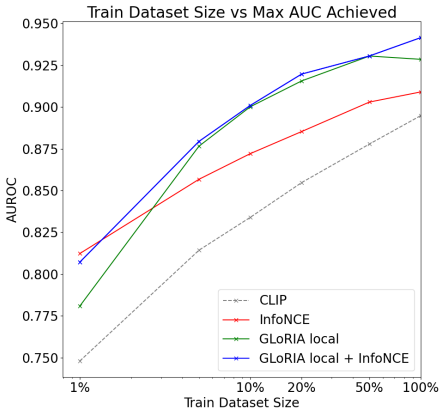


Figure 4: Performance of local loss functions such as GLoRIA compared to global loss functions, such as InfoNCE, as well as a combination of both types.

tion 1) (ii) the local loss function component of GLoRIA (equation 4) (iii) a combination of (i) and (ii). In these experiments we use the best-performing initialization schemes from the previous experiment (i.e. the SimCLR domain-adapted image encoder and domain-adapted CXR-BERT text encoder). Note that for local loss functions, calculating pairwise similarities is a more expensive operation than for InfoNCE and thus these local loss functions are less suitable for applications which require retrieval over a large dataset. Again, we plot the performance of CLIP for comparison purposes. The experiments suggest that local loss functions like GLoRIA do provide a benefit for large dataset sizes, however tend to overfit when there are limited data pairs. Fortunately, the combination of local and global contrastive objectives results in optimal performance in at all training split sizes.

4. Conclusion

Acknowledgments

Acknowledgments withheld.

References

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, Andrew Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karen Simonyan. Flamingo: a visual language model for few-shot learning. In *NeurIPS*, 2022.

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 *Proc. ECCV*, 2022.

Aurelia Bustos, Antonio Pertusa, Jose-Maria Salinas, and Maria de la Iglesia-Vayá. PadChest: A large chest x-ray image dataset with multi-label annotated reports. *Medical Image Analysis*, 66:101797, 2020.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *Proc. ICML*, 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 *Proc. ICCV*, 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, 2021.

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

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 *Proc. ICML*, 2021.

A.E.W. Johnson, T.J. Pollard, S.J. Berkowitz, et al. MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific Data*, 6(1):317, 2019.

Yangguang Li, Feng Liang, Lichen Zhao, Yufeng Cui, Wanli Ouyang, Jing Shao, Fengwei Yu, and Junjie Yan. Supervision exists everywhere: A data efficient contrastive language-image pre-training paradigm. In *Proc. ICLR*, 2022.

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 *Proc. MICCAI*, 2021.

Norman Mu, Alexander Kirillov, David Wagner, and Saining Xie. Slip: Self-supervision meets language-image pre-training. *arXiv preprint arXiv:2112.12750*, 2021.

Philip Müller, Georgios Kaissis, Congyu Zou, and Daniel Rueckert. Joint learning of localized representations from medical images and reports. In *Proc. ECCV*, 2022.

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

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proc. ICML*, 2021.

Maithra Raghu, Chiyuan Zhang, Jon Kleinberg, and Samy Bengio. Transfusion: Understanding transfer learning for medical imaging. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Elad Segal, Ben Bogin, and Jonathan Berant. Training vision-language models with less bimodal supervision. In *4th Conference on Automated Knowledge Base Construction*, 2022.

Jason Wei and Kai Zou. EDA: Easy data augmentation techniques for boosting performance on text classification tasks. 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)*, pages 6383–6389, Hong Kong, China, Nov 2019. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/D19-1670.

Shuxin Yang, Xian Wu, Shen Ge, S. Kevin Zhou, and Li Xiao. Knowledge matters: Chest radiology report generation with general and specific knowledge. *Medical Image Analysis*, 80:102510. ISSN 1361-8415.

Yuhao Zhang, Hang Jiang, Yasuhide Miura, Christopher D. Manning, and Curtis P. Langlotz. Contrastive learning of medical visual representations from paired images and text. In *Proc. MLHC*, 2022.

Appendix A. Appendix

[fillin]