



CXR-CLIP: Toward Large Scale Chest X-ray Language-Image Pre-training

Kihyun You¹, Jawook Gu¹, Jiyeon Ham, Beomhee Park, Jiho Kim,
Eun K. Hong¹, Woonhyuk Baek, and Byungseok Roh[✉]

Kakaobrain, Seongnam, Republic of Korea
{ukihyun,jawook.gu,jiyeon.ham,brook.park,tyler.md,amy.hong,
wbaek,peter.roh}@kakaobrain.com

Abstract. A large-scale image-text pair dataset has greatly contributed to the development of vision-language pre-training (VLP) models, which enable zero-shot or few-shot classification without costly annotation. However, in the medical domain, the scarcity of data remains a significant challenge for developing a powerful VLP model. In this paper, we tackle the lack of image-text data in chest X-ray by expanding image-label pair as image-text pair via general prompt and utilizing multiple images and multiple sections in a radiologic report. We also design two contrastive losses, named ICL and TCL, for learning study-level characteristics of medical images and reports, respectively. Our model outperforms the state-of-the-art models trained under the same conditions. Also, enlarged dataset improve the discriminative power of our pre-trained model for classification, while sacrificing marginal retrieval performance. Code is available at <https://github.com/kakaobrain/cxr-clip>.

Keywords: Chest X-ray · Vision-Language Pre-training · Contrastive Learning

1 Introduction

Chest X-ray (CXR) plays a vital role in screening and diagnosis of thoracic diseases [19]. The effectiveness of deep-learning based computer-aided diagnosis has been demonstrated in disease detection [21]. However, one of the major challenges in training deep learning models for medical purposes is the need for extensive, high-quality clinical annotation, which is time-consuming and costly.

Recently, CLIP [22] and ALIGN [10] have shown the ability to perform vision tasks without any supervision. However, vision-language pre-training (VLP) in the CXR domain still lacks sufficient image-text datasets because many public datasets consist of image-label pairs with different class compositions. Med-CLIP [26] attempted to a rule-based labler to use both image-text data and image-label data. However, it relies on the performance of the rule-based labeler and is not scalable to other diseases that the labeler cannot address.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-3-031-43895-0_10.

In this paper, we propose a training method, *CXR-CLIP*, that integrates image-text data with image-label data using class-specific prompts made by radiologists. Our method does not depend on a rule-based labeler and can be applied to any image-label data. Also, inspired by DeCLIP [13], we used Multi-View Supervision (MVS) utilizing multiple images and texts in a CXR study to make more image-text pairs for efficient learning. In addition, we introduce two contrastive loss functions, named image contrastive loss (ICL) and text contrastive loss (TCL), to learn study-level characteristics of the CXR images and reports respectively.

The main contributions of this paper are summarized as follows. 1) We tackle the lack of data for VLP in CXR by generating image-text pairs from image-label datasets using prompt templates designed by radiologists and utilizing multiple images and texts in a study. 2) Two additional contrastive losses are introduced to learn discriminate features of image and text, improving image-text retrieval performances. 3) Performance of our model is validated on diverse datasets with zero-shot and few-shot settings.

2 Related Work

Data Efficient VLP. Recent studies [13, 17] have proposed data-efficient VLP via joint learning with self-supervision. DeCLIP [13] suggested MVS that utilizes image and text augmentation to leverage positive pairs along with other self-supervisions. In CXR domain, GloRIA [7] aligned words in reports and sub-regions in an image for label efficiency, and BioVIL [2] combined self-supervision for label efficiency. We modify MVS as two distinct images and texts from a study and present self-supervised loss functions, ICL and TCL for efficient learning.

Self-supervision Within CXR Study. A CXR study could include several images in different views and two report sections: 'findings' and 'impression'. The impression section includes the differential diagnosis inferred from the findings section. BioVIL [2] enhanced the text encoder by matching two sections during language pre-training. MedAug [24] shows that self-supervised learning by matching images in a study is better than differently augmented images. We utilize both of multiple images and texts from a single study in VLP in an end-to-end fashion.

Leveraging Image-Label Data in VLP. MedCLIP [26] integrated unpaired images, texts, and labels using rule-based labeler [8], which is less capable of retrieving the exact report for a given image due to the effect of decoupling image-text pairs. UniCL [28] suggested using prompts to leverage image-label dataset [4], considering the samples from the same label to be a positive pair. To our knowledge, this is the first work to utilize prompting for training in CXR domain.

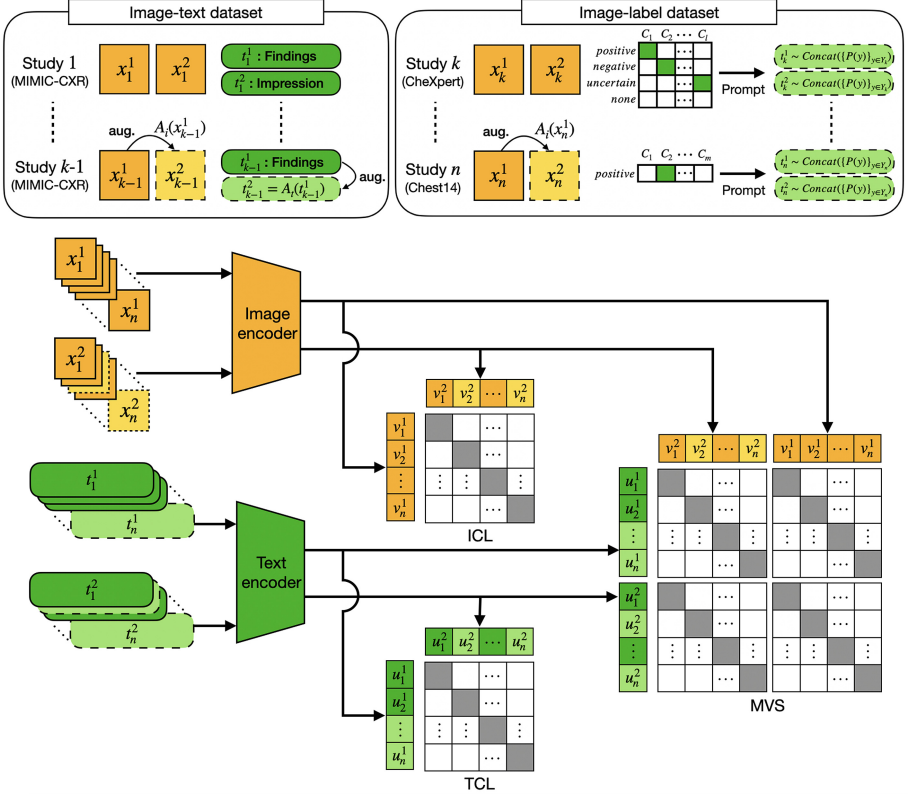


Fig. 1. Overview of the proposed method with a training batch sampling n studies, where each study has a pair of images (x^1 , x^2) and a pair of text (t^1 , t^2). If a study has one image or one text, data augmentation is conducted to make second examples. For the image-label data, two different prompts are generated from class labels as (t^1 , t^2). Using sampled pairs, the encoders are trained with three kinds of contrastive losses (MVS, ICL, and TCL).

3 Method

CXR-CLIP samples image-text pairs from not only image-text data but also image-label data, and learns study-level characteristics with two images and two texts per study. The overview of the proposed method is illustrated in Fig. 1.

3.1 Data Sampling

We define a CXR study as $s = \{X, T\}$, where X is a set of images, and T is a set of “findings” and “impression” sections. The study of image-label dataset has a set of image labels Y instead of T . For the image-label dataset, we make prompt-based texts $T = \text{Concat}(\{p \sim P(y)\}_{y \in Y})$, where p is a sampled prompt sentence, $P(y)$ is a set of prompts given the class name and value y , and $\text{Concat}(\cdot)$ means

concatenating texts. The set of prompts is used to generate sentences such as actual clinical reports, taking into account class labels and their values (positive, negative, etc.), unlike the previous prompt [7] for evaluation which randomly combines a level of severity, location, and sub-type of disease. Our prompts are available in Appendix.

We sample two images (x^1, x^2) in X if there are multiple images. Otherwise, we use augmented image $A_i(x^1)$ as x^2 , where A_i is image augmentation. To leverage various information from different views in CXR (AP, PA, or lateral), we sample images from two distinct views as possible. Similarly, we sample two texts (t^1, t^2) in T if there are both “findings” and “impression”. Otherwise, we use augmented text $A_t(t^1)$ as t^2 , where A_t is text augmentation. For the image-label data, we sample two prompt sentences as t^1 and t^2 from the constructed $T = \text{Concat}(\{p \sim P(y)\}_{y \in Y})$.

3.2 Model Architecture

We construct image encoder E^i and text encoder E^t to obtain global representations of image and text, and a projection layer f^i and f^t to match the size of final embedding vectors.

Image Encoder. We have tested two different image encoders; ResNet-50 [6] and Swin-Tiny [14] as follow [7, 26]. We extract global visual features from the global average pooled output of the image encoder. A linear layer is adopted to project the embeddings into the same size as text embeddings. The normalized visual embedding v is obtained by $v = f^i(E^i(x)) / \|f^i(E^i(x))\|$. We denote a batch of the visual embeddings as $V = \{v\}_{i=1}^n$, where n is a batch size.

Text Encoder. We use BioClinicalBERT [1] model, which is the same architecture as BERT [5] but pre-trained with medical texts [11] as follow [7, 26]. We use [EOS] token’s final output as the global textual representation. Also, a linear projection layer is adopted the same as the image encoder. The normalized text embedding u is denoted as $u = f^t(E^t(t)) / \|f^t(E^t(t))\|$. We denote a batch of the text embedding as $U = \{u\}_{i=1}^n$ and (v_i, u_i) are paired.

3.3 Loss Function

In this section, we first describe CLIP loss [22] and then describe our losses (MVS, ICL, TCL) in terms of CLIP loss. The goal of CLIP loss is to pull image embedding and corresponding text embedding closer and to push unpaired image and text farther in the embedding space. The InfoNCE loss is generally adopted as a type of contrastive loss, and CLIP uses the average of two InfoNCE losses; image-to-text and text-to-image. The formula for CLIP loss is given by

$$L_{CLIP}(U, V) = -\frac{1}{2n} \left(\sum_{u_i \in U} \log \frac{\exp(v_i^T u_i / \tau)}{\sum_{v_j \in V} \exp(u_i^T v_j / \tau)} + \sum_{v_i \in V} \log \frac{\exp(u_i^T v_i / \tau)}{\sum_{u_j \in U} \exp(v_i^T u_j / \tau)} \right), \quad (1)$$

where τ is a learnable temperature to scale logits.

In DeCLIP [13], MVS uses four L_{CLIP} loss with all possible pairs augmented views; (x, t) , $(x, A_t(t))$, $(A_i(x), t)$ and $(A_i(x), A_t(t))$. We modify DeCLIP’s MVS to fit the CXR domain by the composition of the second example. DeCLIP only utilizes an augmented view of the original sample, but we sample a pair of the second image and text as described in Sect. 3.1. We denote the first and the second sets of image embeddings as U^1, U^2 , and text embeddings as V^1, V^2 .

$$L_{MVS} = \frac{1}{4}(L_{CLIP}(U^1, V^1) + L_{CLIP}(U^2, V^1) + L_{CLIP}(U^1, V^2) + L_{CLIP}(U^2, V^2)) \quad (2)$$

The goal of ICL and TCL is to learn modality-specific characteristics in terms of image and text respectively. We design ICL and TCL as same as CLIP loss, but the input embeddings are different. ICL only uses image embeddings; $L_{ICL} = L_{CLIP}(V^1, V^2)$ and TCL only uses text embeddings; $L_{TCL} = L_{CLIP}(U^1, U^2)$. ICL pulls image embeddings from the same study and pushes image embeddings from the different studies, so that, the image encoder can learn study-level diversity. Similarly, TCL pulls embeddings of “findings” and “impression” in the same study or diverse expressions of prompts from the same label and pushes the other studies’ text embeddings, so that the text encoder can match diverse clinical expressions on the same diagnosis. Thereby, the final training objective consists of three contrastive losses balanced each component by λ_I and λ_T , formulated by $L = L_{MVS} + \lambda_I L_{ICL} + \lambda_T L_{TCL}$.

4 Experiment

4.1 Datasets

We used three pre-trained datasets and tested with various external datasets to test the generalizability of models. The statistics of the datasets used are summarized in Table 1.

MIMIC-CXR [12] consists of CXR studies, each with one or more images and free-form reports. We extracted “findings” and “impression” from the reports. We used the training split for pre-training and the test split for image-to-text retrieval.

CheXpert [8] is an image-label data with 14 classes, obtained from the impression section by its rule-based labeler, and each class is labeled as positive,

Table 1. The number of studies for each dataset and split in this paper

Data Split	Pre-training			Evaluation			
	MIMIC-CXR	CheXpert	ChestX-ray14	VinDR	RSNA	SIIM	Open-I
Train	222,628	216,478	89,696	12,000	18,678	8,422	
Valid	1,808	233	22,423	3,000	4,003	1,808	
Test	3,264	1,000		3,000	4,003	1,807	3,788

negative, uncertain, or none (not mentioned). We used the training split for pre-training with class-specific prompts. **CheXpert5x200** is a subset of CheXpert for 5-way classification, which has 200 exclusively positive images for each class. Note that only the reports of CheXpert5x200 are publicly available, but the reports of CheXpert are not. Following the previous works [7, 26], we excluded CheXpert5x200 from the training set and used it for test.

ChestX-ray14 [25] consists of frontal images with binary labels for 14 diseases. Prompts are generated by sampling 3 negative classes per study. We used 20% of the original training set for validation, and the remaining 80% for pre-training.

RSNA pneumonia [23] is binary-labeled data as pneumonia or normal. We split train/valid/test set 70%, 15%, 15% of the dataset following [7] for the external classification task.

SIIM Pneumothorax¹ is also binary labeled as pneumothorax or normal. We split the train/valid/test set same ratio as RSNA pneumonia following [7] and used it for the classification task.

VinDR-CXR [18] contains 22 local labels and 6 global labels of disease, which were obtained by experienced radiologists. We split the validation set from the original training set. Of 28 classes, “other diseases” and “other lesions” classes were excluded. Then, only 18 classes having 10 or more samples within the test set were evaluated for the binary classification of each class as follow [9].

Open-I [3] is an image-text dataset. From each study, one of the report sections and one frontal-view image were sampled and used for image-to-text retrieval.

4.2 Implementation Details

We used augmentations A_i and A_t to fit medical images and reports. For A_i , we resize and crop with scale [0.8, 1.1], randomly adapt CLAHE [20], and random color jittering; brightness, hue ratios from [0.9, 1.1] and contrast, saturation [0.8, 1.2]. For A_t , to preserve clinical meaning, sentence swap and back-translation² from Italian to English is used. The image size and final-embedding size are set to 224 and 512 respectively as in previous work [26]. We set λ_I and λ_T to 1.0, 0.5 for balancing total loss. Two encoders were trained for 15 epochs in a mixed-precision manner, early stopped by validation loss, and optimized by AdamW [16] with an initial learning rate 5e-5 and a weight decay 1e-4. We used cosine-annealing learning-rate scheduler [15] with warm-up for 1 epoch. A training batch consists of 128 studies with 256 image-text pairs. We implemented all experiments on PyTorch with 4 NVIDIA V100 GPUs.

4.3 Comparison with State-of-the-Arts

Zero-Shot and Few-Shot Classification. Table 2 shows performance on classification tasks of our models and state-of-the-art models. To evaluate zero-

¹ https://siim.org/page/pneumothorax_challenge.

² <https://huggingface.co/Helsinki-NLP>.

Table 2. Comparison with state-of-the-art for zero-shot(ZS) or few-shot(10%) classification tasks. M, C, and C14 mean MIMIC-CXR, CheXpert, and ChestX-ray14, respectively. C* means CheXpert with reports, which are not publicly available. ResNet50 (R50) and SwinTiny (SwinT) mean the image encoder used for each model.

Model Name	Pre-train Dataset	VinDR-CXR			RSNA			SIIM			C5x200
		ZS	10%	100%	ZS	10%	100%	ZS	10%	100%	ZS-ACC
GloRIA _{R50}	C*	78.0	73.0	73.1	80.6	88.2	88.5	84.0	91.5	91.9	62.4*
CXR-CLIP _{R50}	M	78.8	82.1	82.2	83.3	88.5	89.2	85.2	88.3	90.5	56.2
CXR-CLIP _{SwinT}	M	78.3	84.9	85.4	81.3	88.0	88.4	85.5	86.9	88.3	54.3
MedCLIP _{SwinT}	M,C	82.4	84.9	85.1	81.9	88.9	89.0	89.0	90.4	90.8	59.2
CXR-CLIP _{R50}	M,C	83.0	81.4	82.1	81.7	88.5	88.9	86.4	88.4	90.7	61.7
CXR-CLIP _{SwinT}	M,C	82.7	86.1	86.7	84.5	88.1	88.8	87.9	89.6	91.2	60.1
CXR-CLIP _{R50}	M,C,C14	78.1	80.2	81.0	81.8	88.7	89.3	85.2	91.5	92.8	60.3
CXR-CLIP _{SwinT}	M,C,C14	78.9	88.0	89.0	80.1	89.2	89.8	91.4	92.9	94.0	62.8

Table 3. Comparison with state-of-the-arts for image-to-text retrieval. The notations of datasets and models are same to Table 2.

Model Name	Pre-Train Dataset	CheXpert5x200			MIMIC-CXR			Open-I			Total RSUM
		R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	
GloRIA _{R50}	C*	17.8	38.8	49.9	7.2	20.6	30.3	1.5	4.4	6.5	177.0
CXR-CLIP _{R50}	M	9.4	23.0	32.6	21.4	46.0	59.2	3.8	8.2	12.3	216.9
CXR-CLIP _{SwinT}	M	8.4	21.5	30.2	21.6	48.9	60.2	3.6	8.3	11.5	214.2
MedCLIP _{SwinT}	M,C	2.6	3.0	3.6	1.1	1.4	5.5	0.1	0.4	0.7	18.4
CXR-CLIP _{R50}	M,C	5.5	19.2	27.4	20.2	45.9	58.2	3.5	8.2	12.0	200.1
CXR-CLIP _{SwinT}	M,C	8.5	23.0	31.6	19.6	44.2	57.1	3.1	8.3	11.6	207.0
CXR-CLIP _{R50}	M,C,C14	5.7	18.0	28.3	19.7	44.4	56.4	2.3	6.7	10.1	191.6
CXR-CLIP _{SwinT}	M,C,C14	7.0	20.1	29.7	20.9	46.2	58.8	2.4	6.6	9.4	201.1

shot classification fairly, we used evaluation prompts suggested from previous works [2, 7, 9]. The evaluation prompts are available in Appendix. We evaluate binary classification computed by Area Under ROC (AUC) and multi-class classification computed by accuracy (ACC). Our ResNet model trained with MIMIC-CXR outperforms GloRIA [7] except for CheXpert5x200, as GloRIA trained with image-text pair in CheXpert. Our SwinTiny model trained with MIMIC-CXR and CheXpert outperforms MedCLIP [26], which is the same architecture trained with the same datasets, in most of the metrics. Adding more pre-training datasets by prompting image-label datasets tends to improve performance for classifications, while the SwinTiny CXr-CLIP pre-trained with three datasets, performs the best for most of the metrics. More comparison with self-supervised models is available in Appendix.

Image-to-Text Retrieval. We evaluated image-to-text retrieval computed by $R@K$, the recall of the exact report in the top K retrieved reports for a given image. (Table 3) While GloRIA [7] uses image-text pairs in CheXpert(C*) which is not available in public, CXr-CLIP uses image-text in MIMIC-CXR. So we adapt an external image-text dataset Open-I [3] for a fair comparison. GloRIA

Table 4. Ablations and comparison with CLIP [22] and DeCLIP [13]. Our augmentations effectively preserves clinical meaning than EDA. Our full methodology (CXR-CLIP) outperforms DeCLIP.

Method	CheXpert 5x200				MIMIC-CXR			Total RSUM
	ACC	R@1	R@5	R@10	R@1	R@5	R@10	
Vanila CLIP	58.9	4.4	14.4	22.6	17.3	41.2	52.6	152.5
+ Study Level Sampling	58.7	4.6	15.1	23.2	17.8	42.5	54.2	157.4
+ Augmentations	60.6	5.7	17.0	24.9	16.1	40.2	51.5	155.4
+ MVS	61.2	5.4	17.1	24.7	16.3	40.6	53.3	157.4
+ ICL	61.6	6.8	20.3	28.6	17.5	41.6	53.2	168.0
+ TCL (CXR-CLIP)	61.7	6.2	18.2	29.1	19.6	44.8	56.6	174.5
MVS of DeCLIP (EDA)	59.5	3.2	15.5	22.9	15.8	39.1	51.5	148.0
MVS of DeCLIP (Our aug)	59.4	6.0	17.0	24.4	15.1	38.8	51.8	153.1
DeCLIP (Our aug)	59.4	5.7	16.1	24.6	18.1	44.0	55.3	163.8

has the best performance on CheXpert but our model trained with MIMIC-CXR, which has similar amounts of studies to CheXpert, outperforms on Open-I. MedCLIP almost lost the ability to retrieve image-text due to decoupling pairs of image and text during pre-training. In CXR-CLIP, adding more image-label datasets such as CheXpert and ChestX-ray14 degrades the image-text retrieval performance, possibly because the contribution of the text in original reports was diluted.

4.4 Ablations

For the ablation study, models with ResNet-50 [6] backbone were trained on MIMIC-CXR and CheXpert datasets and tested on zero-shot classification and image-to-text retrieval tasks with MIMIC-CXR and CheXpert5x200 datasets.

We conducted two ablations shown in Table 4. First, we analyzed the effect of each component of CXR-CLIP by adding the components to vanilla CLIP [22] one by one. To validate our data sampling closer, we divided the sampling method into three parts 1) study-level sampling 2) data augmentations 3) Multi-view and Multi-text sampling (MVS). Our study-level sampling strategy improves performance compared to vanilla CLIP, which uses a naive sampling method bringing an image and corresponding report. Additionally, the modified data augmentation to fit the CXR domain contributes to performance increment of classification, the similar performance on retrieval. MVS slightly improves performances in both classification and image-text retrieval. Adding more supervision (ICL and TCL) improves performance by utilizing better multi-views and multi-text inputs. However, TCL drops the performance of recalls in CheXpert5x200, TCL could be hard to optimize variation of the radiologic report and prompt not diverse as images.

In the second ablation study, CXR-CLIP was compared to DeCLIP [13] to confirm that our MVS using two image-text pairs per study is better than the MVS of DeCLIP which uses naively augmented images and texts. We show that our text augmentation outperforms DeCLIP’s text augmentation named EDA [27] in terms of image-to-text recall, which implies our text augmentation preserves clinical meaning. The superiority of our MVS over DeCLIP’s MVS confirms that using multiple images and texts from one study is better than using images and texts from augmented examples. Also, our full methodology (CXR-CLIP) outperforms DeCLIP, suggesting that our method efficiently learns in the CXR domain more than DeCLIP.

5 Conclusion

We presented a framework enlarging training image-text pair by using image-label datasets as image-text pair with prompts and utilizing multiple images and report sections in a study. Adding image-label datasets achieved performance gain in classification tasks including zero-shot and few-shot settings, on the other hand, lost the performance of retrieval tasks. We also proposed loss functions ICL and TCL to enhance the discriminating power within each modality, which effectively increases image-text retrieval performance. Our additional loss functions are designed to efficiently learn CXR domain knowledge along with image-text contrastive learning.

References

1. Alsentzer, E., et al.: Publicly available clinical BERT embeddings. CoRR abs/1904.03323 (2019). <http://arxiv.org/abs/1904.03323>
2. Boecking, B., et al.: Making the most of text semantics to improve biomedical vision-language processing. In: Avidan, S., Brostow, G., Cissé, M., Farinella, G.M., Hassner, T. (eds.) ECCV 2022. LNCS, vol. 13696, pp. 1–21. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-20059-5_1
3. Demner-Fushman, D., et al.: Preparing a collection of radiology examinations for distribution and retrieval. J. Am. Med. Inform. Assoc. **23**(2), 304–310 (2016)
4. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: a large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition, pp. 248–255. IEEE (2009)
5. Devlin, J., Chang, M., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. CoRR abs/1810.04805 (2018). <http://arxiv.org/abs/1810.04805>
6. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. CoRR abs/1512.03385 (2015). <http://arxiv.org/abs/1512.03385>
7. Huang, S.C., Shen, L., Lungren, M.P., Yeung, S.: 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 (ICCV), pp. 3942–3951 (2021)

8. Irvin, J., et al.: Chexpert: a large chest radiograph dataset with uncertainty labels and expert comparison. CoRR abs/1901.07031 (2019). <http://arxiv.org/abs/1901.07031>
9. Jang, J., Kyung, D., Kim, S.H., Lee, H., Bae, K., Choi, E.: Significantly improving zero-shot X-ray pathology classification via fine-tuning pre-trained image-text encoders (2022). <https://arxiv.org/abs/2212.07050>
10. Jia, C., et al.: Scaling up visual and vision-language representation learning with noisy text supervision. CoRR abs/2102.05918 (2021). <https://arxiv.org/abs/2102.05918>
11. Johnson, A., Pollard, T., Mark, R.: MIMIC-III clinical database (2020)
12. Johnson, A.E.W., Pollard, T., Mark, R., Berkowitz, S., Horng, S.: The MIMIC-CXR database (2019)
13. Li, Y., et al.: Supervision exists everywhere: a data efficient contrastive language-image pre-training paradigm. CoRR abs/2110.05208 (2021). <https://arxiv.org/abs/2110.05208>
14. Liu, Z., et al.: Swin transformer: hierarchical vision transformer using shifted windows. CoRR abs/2103.14030 (2021). <https://arxiv.org/abs/2103.14030>
15. Loshchilov, I., Hutter, F.: SGDR: stochastic gradient descent with restarts. CoRR abs/1608.03983 (2016). <http://arxiv.org/abs/1608.03983>
16. Loshchilov, I., Hutter, F.: Fixing weight decay regularization in adam. CoRR abs/1711.05101 (2017). <http://arxiv.org/abs/1711.05101>
17. Mu, N., Kirillov, A., Wagner, D.A., Xie, S.: SLIP: self-supervision meets language-image pre-training. CoRR abs/2112.12750 (2021). <https://arxiv.org/abs/2112.12750>
18. Nguyen, H.Q., et al.: VinDr-CXR: an open dataset of chest X-rays with radiologist's annotations. Sci. Data **9**(1), 429 (2022). <https://doi.org/10.1038/s41597-022-01498-w>
19. World Health Organization: Communicating radiation risks in paediatric imaging: information to support health care discussions about benefit and risk (2016)
20. Pisano, E.D., et al.: Contrast limited adaptive histogram equalization image processing to improve the detection of simulated spiculations in dense mammograms. J. Digit. Imaging **11**(4), 193 (1998). <https://doi.org/10.1007/BF03178082>
21. Qin, C., Yao, D., Shi, Y., Song, Z.: Computer-aided detection in chest radiography based on artificial intelligence: a survey. Biomed. Eng. Online **17**(1), 113 (2018). <https://doi.org/10.1186/s12938-018-0544-y>
22. Radford, A., et al.: Learning transferable visual models from natural language supervision. CoRR abs/2103.00020 (2021). <https://arxiv.org/abs/2103.00020>
23. Shih, G., et al.: Augmenting the national institutes of health chest radiograph dataset with expert annotations of possible pneumonia. Radiol. Artif. Intell. **1**, e180041 (2019). <https://doi.org/10.1148/ryai.2019180041>
24. Vu, Y.N.T., Wang, R., Balachandar, N., Liu, C., Ng, A.Y., Rajpurkar, P.: Medaug: contrastive learning leveraging patient metadata improves representations for chest x-ray interpretation. In: Jung, K., Yeung, S., Sendak, M., Sjoding, M., Ranganath, R. (eds.) Proceedings of the 6th Machine Learning for Healthcare Conference. Proceedings of Machine Learning Research, vol. 149, pp. 755–769. PMLR (2021). <https://proceedings.mlr.press/v149/vu21a.html>
25. Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M., Summers, R.M.: Chestx-ray8: hospital-scale chest X-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017)

26. Wang, Z., Wu, Z., Agarwal, D., Sun, J.: Medclip: contrastive learning from unpaired medical images and text (2022). <https://arxiv.org/abs/2210.10163>
27. Wei, J.W., Zou, K.: EDA: easy data augmentation techniques for boosting performance on text classification tasks. CoRR abs/1901.11196 (2019). <http://arxiv.org/abs/1901.11196>
28. Yang, J., et al.: Unified contrastive learning in image-text-label space (2022). <https://arxiv.org/abs/2204.03610>