Language over Labels: Contrastive Language Supervision Exceeds Purely Label-Supervised Classification Performance on Chest X-Rays

Anton Orell Wiehe[1,2], **Florian Schneider**[1], **Sebastian Blank**[2], **Xintong Wang**[1]
Hans-Peter Zorn[2], **Chris Biemann**[1]
[1]Universität Hamburg, 20146 Hamburg, Germany
[2]inovex GmbH, 75179 Pforzheim, Germany
antonwiehe@gmail.com
{florian.schneider,xintong.wang,christian.biemann}@uni-hamburg.de
{sebastian.blank,hzorn}@inovex.de

Abstract

The multi-modal foundation model CLIP computes representations from texts and images that achieved unprecedented performance on tasks such as zero-shot image classification. However, CLIP was pretrained on public internet data. Thus it lacks highly domain-specific knowledge. We investigate the adaptation of CLIP-based models to the chest radiography domain using the MIMIC-CXR dataset. We show that the features of the pretrained CLIP models do not transfer to this domain. We adapt CLIP to the chest radiography domain using contrastive language supervision and show that this approach yields a model that outperforms supervised learning on labels on the MIMIC-CXR dataset while also generalizing to the CheXpert and RSNA Pneumonia datasets. Furthermore, we do a detailed ablation study of the batch and dataset size. Finally, we show that language supervision allows for better explainability by using the multi-modal model to generate images from texts such that experts can inspect what the model has learned.

1 Introduction

Multi-modal models that understand text and images, as well as the relations between them, surged in performance due to the pioneering work of CLIP (Radford et al., 2021). Through a contrastive loss based on language supervision, the model embeds matching text-image pairs closely in latent space. This enables various applications, such as image classification (Radford et al., 2021), object detection (Alex Shonenkov, 2021), semantic segmentation, (Zhou et al., 2021; Rao et al., 2021), and text-to-image generation (Crowson et al., 2022).

As the CLIP models were trained on data scraped from the internet, they work remarkably well for data of the general domain and excel at tasks such as food (Bossard et al., 2014), car brand (Krause et al., 2013), or animal classification (Parkhi et al., 2012). However, for more specialized tasks such as satellite image (Helber et al., 2019, 2018) and cancer cell classification (Veeling et al., 2018), they do not perform much better than a random guess (Radford et al., 2021). To make these models work for these tasks, they require adaptation to the specific domain.

In this paper, we study the adaptation of CLIP models to the domain of chest x-ray images of the MIMIC-CXR (Johnson et al., 2019b; Goldberger et al., 2000; Johnson et al., 2019a,c) dataset. We show that the CLIP model pretrained on data scraped from the internet (Radford et al., 2021) does not transfer well to MIMIC-CXR. Furthermore, two approaches to adapting the model are compared: contrastive language supervision (CLS) and supervised fine-tuning (FT) on labels. We show that CLS combined with linear probing performs better than only using FT on labels. Furthermore, we show that the same language-supervised model can be used to achieve good performance with only a linear probe on other chest radiograph datasets without retraining.

Our first ablation study investigates the batch size, as the massive batch size of 32,768 used for the original CLIP training would impose an obstacle for any CLS fine-tuning. We show that a small batch size is sufficient to achieve good CLS performance. We also find that a batch size that is too large hurts performance, contrary to the findings of prior work (Chen et al., 2020; Grill et al., 2020; Radford et al., 2021).

Next to the large batch size, CLIP also used a large dataset of over 400 million image-text pairs. In a second ablation study, we investigate whether CLS needs a large dataset size to outperform supervised learning. We show that CLS can be superior to FT even with only 20,000 image-text pairs (10% of the MIMIC-CXR dataset).

In the last experiment, we display how to get

more interpretable neural network classifiers. The language-supervised model can compare the similarity of the features of a text and an image. Through the gradient of the similarity towards the image, an image can be generated purely from a text. This generation allows clinicians and machine learning scientists to visualize model representations. This approach, inspired by CLIP-based text-to-image approaches such as VQGAN-CLIP (Crowson et al., 2022) resembles the work of DeepDream (Mordvintsev et al., 2015). Instead of visualizing classes or neuron activations, it visualizes texts.

2 Related Work

In a closely related work named ConVirt, (Zhang et al., 2020) train a model using CLS on the image-text pairs of the MIMIC-CXR dataset and compare it to supervised learning on the labels. Their pioneering work partially inspired the creation of CLIP (Radford et al., 2021). Our work is complementary to their work by using the widely adapted architecture and simplified loss function of CLIP, evaluating the performance of the OpenAI-pretrained CLIP model on MIMIC-CXR, running ablation studies on the batch size and dataset size, and introducing the text-to-image visualization of diagnoses.

In CLIP-art (Conde and Turgutlu, 2021), CLIP was fine-tuned using CLS on a large dataset of museum artworks with descriptions. The features of the fine-tuned CLIP model do not lead to a significantly better classification performance than the features of the base CLIP model. More related to the approach of this paper is PubMedCLIP (Eslami et al., 2021). The authors fine-tune CLIP using the CLS objective on image-text pairs from medical papers. They show that the pretrained CLIP features improve visual question-answering performance over the current state-of-the-art baseline. The continued pretraining using CLS only slightly improves the performance over the base CLIP model.

A current preprint follows a similar approach as our paper. (Seibold et al., 2022) compare the zero-shot performance of a model trained using CLS-like loss functions to the performance of supervision on labels. Their work confirms the benefits of training using language supervision over labels. However, their work focuses on selecting training data and loss functions. In contrast, our work analyzes batch sizes, dataset sizes, and an explainability approach.

3 Background

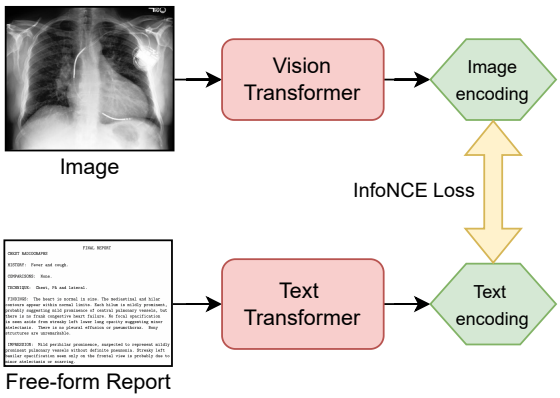
This work investigates the effect of pretraining using CLS on text-image pairs before FT on labels. We are given a set of images S, corresponding texts T, and labels Y. For both the CLS and FT stages, a network (in this case, a Transformer (Vaswani et al., 2017) network from Radford et al., 2021) first transforms its input into an encoding, leading to the encodings e_{text} and e_{image}. In the CLS stage these encodings are improved by training the weights of both transformers using a contrastive loss, whereas the FT stage only uses the vision transformer and its encoding, followed by a linear layer. An overview of the two stages is given in Figure 1.

For FT, a prediction $\hat{y}_{ni}$ of the target label $n \in Y$ for image $s_i \in S$ is made by a network f: $f(s_i) = y_{ni}$. The binary cross entropy loss L_{BCE} is calculated per label y_n and then averaged over all N labels to get the supervised loss L_{SL}, as was done in previous work in multi-label settings (Liu et al., 2021; Nam et al., 2014).

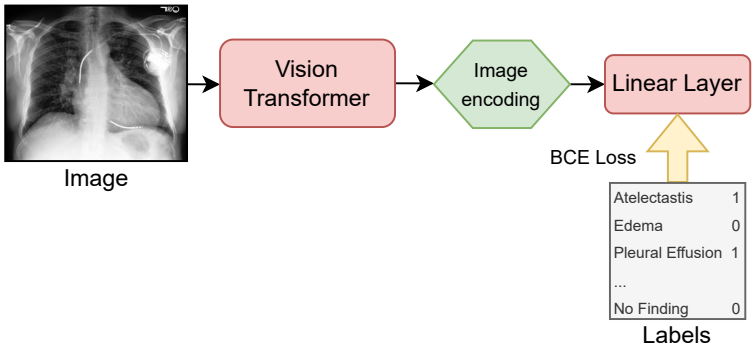
The BCE loss L_{BCE} of label n is calculated using the ground truth y_{ni} and the prediction y_{ni} for sample i. It assigns a loss that is high initially and drops off logarithmically as the prediction approaches the ground truth:

$$L_{BCE_n}(y_{ni}, \hat{y}_{ni}) = \\ -1(y_{ni} \log(\hat{y}_{ni}) + (1 - y_{ni}) \log(1 - \hat{y}_{ni})) \quad (1)$$

The pretraining stage of CLS utilizes text and image representations e_{text} and e_{image} computed by the text and image encoders of CLIP, respectively. During the later FT stage, the linear probe is trained based on the image encoding e_{image}, and the full fine-tuning also tunes the weights of the image encoder. During pretraining, a batch of size K image-text pairs is sampled and encoded. The loss is calculated by using every encoding of both modalities once as the anchor sample x_i. The matching positive sample x_i^+ is the paired encoded sample of the other modality and all other encodings from the other modality of the sampled batch are the negative samples X^-. For each anchor sample, the InfoNCE loss (Oord et al., 2018) is calculated with



(a) Contrastive Language Supervision Stage (CLS)



(b) Supervised Fine-tuning Stage (FT)

Figure 1: Loss calculation flowcharts for the different training stages. The Contrastive Language Supervision (CLS) stage is always followed by the Supervised Fine-tuning Stage (FT) to adapt the model to predict the labels. The FT stage either trains only a new linear layer head (linear probe via a logistic regression) or it also trains the weights of the vision transformer. Red rectangles are networks with trainable weights, green shadings indicate encodings, and yellow arrows indicate the gradient flow.

a similarity function $\text{sim}(x, y)$:

$$L_{\text{InfoNCE}}(x_i, x_i^+, X^-) = -\log \frac{\exp(\text{sim}(x_i, x_i^+))}{\sum_{j=0}^{K} \exp(\text{sim}(x_i, x_j^-))} \quad (2)$$

The total InfoNCE loss is the average of all individual losses of the samples from the batch. We use the cosine similarity as a similarity function, as in the original CLIP paper.

4 Methods

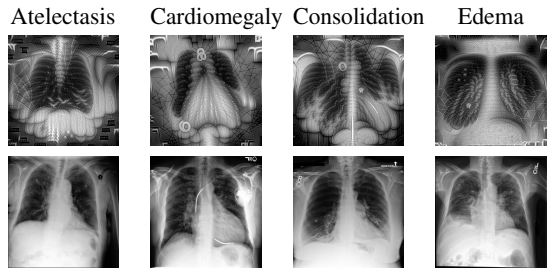
The code for training and evaluating is available online [1]. All models were evaluated using the macro average of the area under the receiver-operator curve (ROC-AUC or AUC) (Bradley, 1997) averaged over all labels of the dataset. This metric was used to enable a comparison with prior work. For a clinical evaluation, the sensitivity and specificity should be studied in more detail.

4.1 MIMIC-CXR dataset

The MIMIC-CXR dataset contains 227,827 studies of chest radiographs with a written report by expert radiologists. There are one or multiple radiography images present for each study, leading to 377,095 total image-text pairs. The labels were extracted by the automatic labeler from the CheXpert dataset (Irvin et al., 2019). For each report, 14 diagnoses can appear individually and in conjunction. The official validation and test splits were used. No images were excluded. Examples for images and extracts of reports can be seen in Figure 2.

We marked all labels which are either not contained in a report or contained with an uncertainty quantifier as negative. All others were marked as positive. The report text was cleaned for language

[1] https://github.com/NotNANtoN/master_thesis



Mild atelectasis at the left base

Heart size is enlarged

Bibasal consolidations have slightly progressed

Improved pulmonary edema

Figure 2: Text-to-image generations for a subset of four diagnoses: atelectasis, cardiomegaly, consolidation, and edema. The first row shows the generated images. The second row shows real radiography images of MIMIC-CXR with the sentence of the report relevant to the labeling of the diagnosis.

supervision by filtering repetitive headers, censored personal information, newlines, and other unnecessary characters. The images were resized such that the smaller side has a length of 256 pixels.

4.2 External Test Datasets

The RSNA Pneumonia (Wang et al., 2017; Shih et al., 2019) and cheXpert (Irvin et al., 2019) datasets were used to evaluate if the model pretrained on MIMIC-CXR generalizes to data from other hospitals with other labels. Linear probes were trained on the features of the pretrained models to predict the labels of the external datasets. The cheXpert dataset contains 223,648 images labeled with the same diagnoses as MIMIC-CXR. The official validation split was used as our test set. The RSNA dataset contains 30,227 images of which 9,555 are annotated with the pneumonia diagnosis, forming a single-class, single-label classification task. A random subset of 10% of the data was used as a test set, the rest was used for training.

4.3 Model training

In preliminary FT experiments, the CLIP models RN50, RN50x4, ViT-B/32, ViT-B/16, and ViT-L/14 were investigated. The ViT-B/32 model was chosen as it is the fastest model and as the performances of all models were nearly equal. The aim of this paper is not the best performance but rather a comparison of the training procedure.

The training setup follows the setup of the original CLIP paper (Radford et al., 2021), using Adam (Kingma and Ba, 2015) with a weight decay (Loshchilov and Hutter, 2019) of 0.2, β_1 value of 0.9, and β_2 value of 0.98 for training all models. A learning rate schedule with a linear warmup from zero to the maximum learning rate was used during the first 5% of training and a cosine decay schedule for the rest of the training. During training, the images were augmented by rotating them randomly by up to 45 degrees, shifting them randomly in the x and y-axis by up to 15% of the image length, and zooming into and out of the image by up to 10% of the image size.

All model runs used the pretrained weights from Radford et al., 2021. The FT models were trained for 10 epochs with a batch size of 256. The learning rates $\{1e{-}6, 3e{-}5, 1e{-}5, 3e{-}5, 1e{-}4\}$ were evaluated, of which $1e{-}5$ performed best on the validation set. The CLS model was trained for 10 epochs with a batch size of 196. The sentences of each report text were randomly shuffled during training to avoid always truncating the final part of the report if it is longer than 75 tokens (tokenized with the pretrained CLIP tokenizer). The learning rate for the CLS stage was tuned with the same set of learning rates as above. The best learning rate for a linear probe on the validation set was again $1e{-}5$. After the CLS stage, the model was continued to be trained on the labels with either a linear probe using logistic regression or with the FT setup from above.

4.4 Ablation Studies

The first ablation study varied the batch size while keeping other parameters constant. It measures the impact of the number of negative samples in CLS, which is dependent on the batch size. We varied the batch size from 6 to 1,536. The maximum batch size for a single GPU with 12 GB of VRAM is 192. Training runs with batch sizes below 192 accumulate the gradients for as many steps to match the number of update steps done with a batch size of 192. To accommodate the reduction in update steps due to the increased batch size, we tested scaling the learning rate linearly proportional to the batch size and compared it to keeping the learning rate constant.

The second ablation study varied the dataset size to a minimum of 1% to understand whether CLS is performant on smaller datasets. We trained once for 10 and 50 epochs for each dataset size. Training for more epochs increases the training duration. There-

fore it balances the effect of having fewer batches in an epoch for smaller dataset sizes. The learning rate and other hyperparameters stayed unchanged.

4.5 Text-to-Image Generation

In the text-to-image generation approach, a language-supervised model was used after only 3 epochs of training to avoid any overfitting. To generate an image from a text, first, the text of a diagnosis is encoded into a text feature vector. The image is randomly initialized as a single-channel tensor of size 224x224, randomly sampled from a normal distribution with a mean of 0.5 and a standard deviation of 0.25. The gradient of the cosine similarity between the image's features and the diagnosis's features towards the image is applied repeatedly to the image to iteratively increase the similarity to the text. Optimization was done with Adam (Kingma and Ba, 2015) with a learning rate of 0.03 and a weight decay (Loshchilov and Hutter, 2019) of 0.1.

Directly optimizing the pixels without any regularization creates adversarial examples (Crowson et al., 2022). The generated image is augmented before encoding it with CLIP to avoid this . We use the augmentation pipeline proposed by (Crowson et al., 2022).

Multiple images of different resolutions overlaying each other are optimized simultaneously to increase image quality. Images of pixel sizes [224, 112, 61, 30, 15] are randomly initialized and optimised. During the iterative generation process, the images are resized to 224x224 pixels and then averaged. The images of smaller resolutions learn general shapes, and the higher resolution ones focus on the details. The average of all resized images forms the generated image. The augmentations are applied to this image. The loss to be optimized is the cosine similarity between the features of this image and the features of the target text.

5 Results

5.1 Language Supervision Compared to Supervised Learning

The results of the comparison between FT and CLS are shown in Table 1. The CLIP ViT-B/32 model performs worst when using randomly initialized weights with a linear probe. The improvement when using pretrained weights is only marginal, showing that the features of the general CLIP model do not transfer to the chest radiographs of MIMIC-CXR. Training the model using FT increases the AUC significantly from 0.66 to 0.77. CLS beats this score by a slight margin. CLS with only a linear probe is competitive with and slightly superior to pure FT.

Table 1: Table comparing the results of CLS and SL, set into context with prior work. CLS stands for contrastive language supervision, FT for supervised fine-tuning, ZS for zero-shot, and LP for linear probe. *Rand.* indicates that the weights of the network were randomly initialized - in all other cases the pretrained weights from Radford et al., 2021 are used. *AUC* stands for the macro ROC-AUC, averaged over all labels and multiplied by 100 for legibility. *Ours* stands for the CLIP ViT-B/32 model.

(a) MIMIC-CXR (Johnson et al., 2019c)

Model	Type	AUC
Nunes et al., 2019	FT	65.6
Seibold et al., 2022	ZS	79.4
Ours	Rand. + LP	66.5
Ours	LP	66.7
Ours	FT	77.2
Ours	CLS + LP	77.8
Ours	CLS + FT	77.3

(b) CheXpert (Irvin et al., 2019)

Model	Type	AUC
Seibold et al., 2022	ZS	78.9
Zhang et al., 2020	CLS + LP	87.3
Zhang et al., 2020	CLS + FT	88.1
Azizi et al., 2021	FT	77.0
Ours	CLS + LP	87.2

(c) RSNA (Wang et al., 2017; Shih et al., 2019)

Model	Type	AUC
Zhang et al., 2020	CLS + LP	92.1
Zhang et al., 2020	CLS + LP	92.7
Han et al., 2021	FT	92.3
Ours	CLS + LP	90.7

The comparison with the results of prior work shows that similar performance has been reached for all datasets. For both external datasets, CLS with a linear probe reaches competitive performance, which displays the generality of the learned features.

5.2 Batch Size Ablation

The results of the batch size ablation experiment can be seen in Figure 3. For smaller batch sizes, the performance drops but stays above 0.75. Notably, the best batch size is 576. The AUC drops for larger

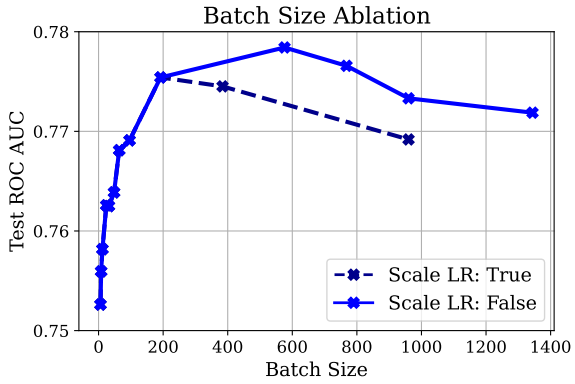


Figure 3: The batch size plotted against the test ROC AUC score. The batch size is varied to investigate whether a larger pool of negative samples is necessary for CLS. The optimal batch size peaks at 576. *Scale LR* indicates whether the learning rate is scaled linearly with the batch size for batch sizes beyond 196.

batch sizes independent of the learning rate scaling method. This drop demonstrates an upper limit of the optimal batch size for our model and dataset.

5.3 Dataset Size Ablation

The results of the dataset size ablation study in Figure 4 show that the main results hold at varying dataset sizes. Pretraining using CLS on the whole dataset, followed by fine-tuning on a fraction of the labels consistently performs best. CLS with a linear probe outperforms FT for all dataset sizes greater or equal than 10% (around 20,000 image-text pairs) when trained for 50 epochs. With only 10% of the dataset, CLS nearly matches the performance of applying it to the full dataset. The difference between the performance of the 10 and 50 epoch runs is large for the CLS runs that use at least 10% of the dataset size and small lower dataset sizes. This discrepancy could indicate that a critical dataset size of around 10% of the total dataset size exists that CLS requires to learn good representations.

5.4 Explainability via Text-to-image Generation

The interpretability results are shown in the top row of Figure 2. Qualitatively, one can observe that the generated images display a lung and a heart. They also greatly differ depending on which text they are conditioned on. We consulted two radiologists from a local clinic who both were able to assign 2 out of 4 diagnoses correctly to the generated images. These qualitative analyses open the door for further empirical studies.

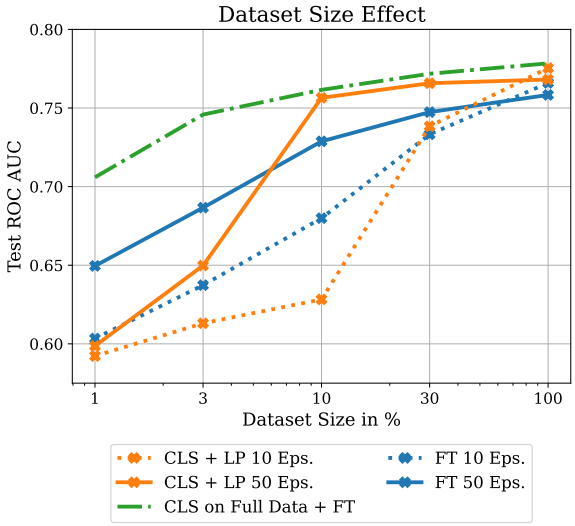


Figure 4: The dataset size plotted against the test ROC AUC score for FT and CLS. *FT* stands for supervised fine-tuning, *CLS* for contrastive language supervision. *CLS + FT* is the two-stage approach of first applying CLS to text-image pairs, followed by full fine-tuning via labels. *CLS + LP* is CLS followed by a linear probe. The *CLS on Full Data + FT* approach uses all data for CLS and a reduced dataset size for FT.

6 Conclusion

We show that CLS with a simple linear probe outperforms FT on the MIMIC-CXR dataset, even when using small batch sizes on a single GPU. Models trained using CLS generalize to datasets of the same domain. CLS outperforms FT for all dataset sizes down to 20,000 image-text pairs.

The optimal batch size in our experiments was 576. Furthermore, CLS stopped being performant when using fewer than 20,000 training pairs. Future work could investigate how the optimum batch size changes depending on the dataset size and if this critical dataset size is replicable for other datasets.

Acknowledgments

This research was partially funded by the German Research Foundation – "DFG Transregio SFB 169: Crossmodal Learning" and by inovex GmbH.

References

Denis Karachev Alex Shonenkov, Sergey Shtekhin. 2021. CLIP ODS: CLIP object detection & segmentation. https://github.com/shonenkov/CLIP-ODS.

Shekoofeh Azizi, Basil Mustafa, Fiona Ryan, Zachary Beaver, Jan Freyberg, Jonathan Deaton, Aaron Loh, Alan Karthikesalingam, Simon Kornblith, Ting Chen, Vivek Natarajan, and Mohammad Norouzi. 2021. Big self-supervised models advance medical image classification. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 3458–3468. IEEE.

Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. 2014. Food-101 - mining discriminative components with random forests. In *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part VI*, volume 8694 of *Lecture Notes in Computer Science*, pages 446–461. Springer.

Andrew P. Bradley. 1997. The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognit.*, 30(7):1145–1159.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey E. Hinton. 2020. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 1597–1607. PMLR.

Marcos V. Conde and Kerem Turgutlu. 2021. Clip-art: Contrastive pre-training for fine-grained art classification. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2021, virtual, June 19-25, 2021*, pages 3956–3960. Computer Vision Foundation / IEEE.

Katherine Crowson, Stella Biderman, Daniel Kornis, Dashiell Stander, Eric Hallahan, Louis Castricato, and Edward Raff. 2022. VQGAN-CLIP: Open domain image generation and editing with natural language guidance. *arXiv preprint arXiv:2204.08583*.

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

Ary L. Goldberger, Luis A. Nunes Amaral, L Glass, Jeffrey M. Hausdorff, Plamen Ch. Ivanov, Roger G. Mark, Joseph E. Mietus, George B. Moody, Chung-Kang Peng, and Harry Eugene Stanley. 2000. PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. *Circulation*, 101 23:E215–20.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H. Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Ávila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, Bilal Piot, Koray Kavukcuoglu, Rémi Munos, and Michal Valko. 2020. Bootstrap your own latent - A new approach to self-supervised learning. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Yan Han, Chongyan Chen, Ahmed H. Tewfik, Ying Ding, and Yifan Peng. 2021. Pneumonia detection on chest x-ray using radiomic features and contrastive learning. In *18th IEEE International Symposium on Biomedical Imaging, ISBI 2021, Nice, France, April 13-16, 2021*, pages 247–251. IEEE.

Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. 2018. Introducing eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. In *2018 IEEE International Geoscience and Remote Sensing Symposium, IGARSS 2018, Valencia, Spain, July 22-27, 2018*, pages 204–207. IEEE.

Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. 2019. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE J. Sel. Top. Appl. Earth Obs. Remote. Sens.*, 12(7):2217–2226.

Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn L. Ball, Katie S. 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. 2019. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pages 590–597. AAAI Press.

Alistair E. W. Johnson, Tom J. Pollard, Seth J. Berkowitz, Nathaniel R. Greenbaum, Matthew P. Lungren, Chih-ying Deng, Roger G. Mark, and Steven Horng. 2019a. MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific Data*, 6(1):317.

Alistair E. W. Johnson, Tom J. Pollard, Nathaniel R. Greenbaum, Matthew P. Lungren, Chih-ying Deng, Yifan Peng, Zhiyong Lu, Roger G. Mark, Seth J. Berkowitz, and Steven Horng. 2019b. MIMIC-CXR-JPG, a large publicly available database of labeled chest radiographs. *arXiv preprint arXiv:1901.07042*.

Alistair E. W. Johnson, Tom J. Pollard, Roger G. Mark, Seth J. Berkowitz, and Steven Horng. 2019c. MIMIC-CXR database (version 2.0.0). *PhysioNet*.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 2013. 3D object representations for fine-grained

categorization. In *2013 IEEE International Conference on Computer Vision Workshops, ICCV Workshops 2013, Sydney, Australia, December 1-8, 2013*, pages 554–561. IEEE Computer Society.

Weiwei Liu, Haobo Wang, Xiaobo Shen, and Ivor Tsang. 2021. The emerging trends of multi-label learning. *IEEE transactions on pattern analysis and machine intelligence*.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

A. Mordvintsev, Christopher Olah, and Mike Tyka. 2015. Inceptionism: Going deeper into neural networks. *Blog Article*.

Jinseok Nam, Jungi Kim, Eneldo Loza Mencía, Iryna Gurevych, and Johannes Fürnkranz. 2014. Large-scale multi-label text classification — revisiting neural networks. In *Machine Learning and Knowledge Discovery in Databases*, pages 437–452. Springer Berlin Heidelberg.

Nelson Nunes, Bruno Martins, Nuno André da Silva, Francisca Pais Leite, and Mário J. Silva. 2019. A multi-modal deep learning method for classifying chest radiology exams. In *Progress in Artificial Intelligence - 19th EPIA Conference on Artificial Intelligence, EPIA 2019, Vila Real, Portugal, September 3-6, 2019, Proceedings, Part I*, volume 11804 of *Lecture Notes in Computer Science*, pages 323–335. Springer.

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

Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and C. V. Jawahar. 2012. Cats and dogs. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3498–3505.

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. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.

Yongming Rao, Wenliang Zhao, Guangyi Chen, Yansong Tang, Zheng Zhu, Guan Huang, Jie Zhou, and Jiwen Lu. 2021. DenseCLIP: Language-guided dense prediction with context-aware prompting. *arXiv preprint arXiv:2112.01518*.

Constantin Seibold, Simon Reiß, M. Saquib Sarfraz, Rainer Stiefelhagen, and Jens Kleesiek. 2022. Breaking with fixed set pathology recognition through report-guided contrastive training. *arXiv preprint arXiv:2205.07139*.

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, Ritu R. Gill, Myrna C.B. Godoy, Stephen Hobbs, Jean Jeudy, Archana Laroia, Palmi N. Shah, Dharshan Vummidi, Kavitha Yaddanapudi, and Anouk Stein. 2019. Augmenting the national institutes of health chest radiograph dataset with expert annotations of possible pneumonia. *Radiology: Artificial Intelligence*, 1(1):e180041. PMID: 33937785.

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

Bastiaan S. Veeling, Jasper Linmans, Jim Winkens, Taco Cohen, and Max Welling. 2018. Rotation equivariant cnns for digital pathology. In *Medical Image Computing and Computer Assisted Intervention - MICCAI 2018 - 21st International Conference, Granada, Spain, September 16-20, 2018, Proceedings, Part II*, volume 11071 of *Lecture Notes in Computer Science*, pages 210–218. Springer.

Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M. Summers. 2017. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 3462–3471. IEEE Computer Society.

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

Chong Zhou, Chen Change Loy, and Bo Dai. 2021. DenseCLIP: Extract free dense labels from CLIP. *arXiv preprint arXiv:2112.01071*.