DisCeRn:
Disease-Contrastive Representations from Multi-Modal Medical Data

CSC2541 Machine Learning for Healthcare: Project Report

Tom Ginsberg[*]
Department of Computer Science
University of Toronto
tomginsberg@cs.toronto.edu

Korbinian Koch[*]
Department of Computer Science
University of Toronto
korbinian.koch@mail.utoronto.ca

Vahid Balazadeh[*]
Department of Computer Science
University of Toronto
vahid@cs.toronto.edu

Vasudev Sharma[*]
Department of Computer Science
University of Toronto
vasu@cs.toronto.edu

Abstract

Self-supervised contrastive learning is a proven method to automatically learn representations of images and other modalities that encode meaningful low-dimensional information about complex features [1, 2]. Recent work has shown that contrastive learning is useful in the medical domain — a result of particular significance as medical datasets are highly complex and expensive to annotate. While contrastive learning methods designed for natural images have been extended to medical images, domain-specific insights have only been applied for positive sample selection [3] and seldom for the contrastive loss itself. In this work, we propose Disease-Contrastive Representations (DisCerRn), a method for modifying the contrastive loss by weighting negative pairs of samples differently based on only marginally related observed pathologies. In addition, we extend this method to the multi-modal regime where we consider how numerical measurements (e.g. blood pressure, heart rate, or O_2 saturation) could improve the utility of contrastive representations. We evaluate the downstream performance of DisCeRn on the task of Pneumonia prediction, by finetuning a linear layer on top of the encoder as well as using our self-supervised models as initialization for end-to-end training. DisCeRn achieves results that are superior or comparable to existing work (89% AUC on the CheXpert benchmark), although we suffer from an unexpected performance drop in the multi-modal setting. While our method is still preliminary, it emphasizes the fact that auxiliary labels found in many medical imagining datasets can be beneficial for improving representation learning techniques, even if these labels are not directly related to downstream objectives.

[*] Equal overall contribution. Contributions of specific team memebers can be found in Appendix A. The GitHub code can be found here.

1 Introduction

1.1 Background

In recent years we have seen rapidly growing progress in the application of machine learning (ML) to the healthcare domain [4, 3, 5]. On one hand, this is due to developments in building more powerful and robust models themselves, but on the other, it is due to improved access to large corpora of annotated medical data [6, 7, 8]. However, collecting large and high-quality labeled datasets is extremely expensive and hence difficult to scale to every predictive task a practitioner may be interested in. A proposed solution to this problem is self-supervised learning (SSL), a paradigm where an ML algorithm can teach itself inherent features about a dataset drastically reducing the amount of unlabeled data required to learn how to make specific predictions [1, 2]. The most common family of methods is self-supervised representation learning (SSRL), where a model learns a low-dimensional representation of data points that obey a set of rational assumptions (e.g. images with similar features map to representations that are nearby in Euclidean space). In our work, we extend existing strategies in self-supervised representation learning with two specific insights from the medical domain:

1. Many medical datasets contain annotations other than those of interest for a predictive task, but that may still be relevant for representation learning;

2. Most hospital patients have a baseline set of features recorded about them (e.g. age, gender, lab results) which may not be available during SSL[1], but would likely be available during a real-world inference scenario.

To expand on the first point we may consider a large chest X-ray dataset as an example: a likely assumption is that this dataset does not contain labels for all known respiratory diseases, however, it may contain sparse labels for a common disease like Pneumonia. Now suppose our end goal is to build a predictor for COVID-19 but we only have a small labelled dataset. A naive approach would apply SSRL to the large dataset and finetune on the small COVID-19 dataset, not taking advantage of the Pneumonia labels. On the other hand, transfer learning could be applied starting with a Pneumonia classifier trained on the large dataset, finetuned after for COVID-19 prediction. Our proposed solution is to take the best from both, we apply SSRL but modify the learning objective to encode information related to the available labels directly into the representations.

The second point is far less subtle. When a patient is sent for an X-ray there are likely many other baseline features recorded about them. Even if these features are not a part of whatever dataset we use for SSL (as they often are not) they should be available at inference for any model deployed in a hospital. The specific question we aim to answer is: if additional features are available for finetuning and inference, are they helpful for improving classification with self supervised image representations.

We will explore these ideas for the task of Pneumonia prediction in chest X-ray imagery — a problem that is notably difficult even for expert radiologists [9, 10]. We will consider both the scenarios outlined previously, e.g. where auxiliary labels are available during the SSL phase that may or may not be related to Pneumonia, and where common baseline covariates are available for finetuning and inference.

In summary, our main contributions are:

- A benchmark of existing SSRL algorithms on the downstream performance of Pneumonia prediction from chest X-rays;

- A novel self-supervised learning framework able to take advantage of partial labels for arbitrary classification tasks not related to the downstream learning objective;

- An investigation into combining data from non-imaging modalities with self-supervised representations learned on images.

[1] Most accessible medical datasets will not contain every value that may have been recorded for a particular patient while in the hospital.

1.2 Related Work

Self-Supervised Learning (SSL) methods are a form of unsupervised pre-training that use pretext tasks on unlabeled data to learn representation that can be leveraged for subsequent fine-tuning on labeled data [1].

Contrastive Learning is a form of SSL that tries to maximize the agreement between positive image pairs like different views of the same image, relative to negative or different sample images using contrastive loss. This paradigm enforces the representation to not distinguish different transformations of an image, and thus ignores nuisance factors in the data [11].

Momentum Contrast (MoCo) [2] proposes a contrastive loss mechanism, which views positive pair agreement as a dictionary-lookup and uses one encoder for queries, and one momentum encoder for keys. It then minimizes the InfoNCE loss [12] to match the encoded image to its positive pair in the dictionary of encoded keys. Sowrirajan et al. [13] show that contrastive learning can improve the performance in the context of medical imaging by applying MoCo to chest X-ray interpretation task. Though they demonstrate promising results, they still use augmented views of the *same* image as positive pairs. However, one can leverage the available non-image data to select more meaningful positive pairs. For example, **ConVIRT [14]** uses the existing textual reports for chest radiograph images as positive pairs, and pre-trains image encoders via a bidirectional constrastive objective between image and text modalities.

MedAug [3] leverages patient metadata to improve representation for chest X-ray interpretation task. Instead of using the same image for positive pairs, they choose different images from the same patient, imaging study or laterality as positive pairs. Also, they explore strategies for selecting hard negative pairs for contrastive learning.

2 Datasets

CheXpert [8] contains 224,316 labeled chest x-rays of 65,240 patients at the Stanford Hospital between October 2002 and July 2017.

MIMIC-CXR-JPG [7] contains 377,110 labeled chest x-rays of 65,379 patients at the Beth Israel Deaconess Medical Center Emergency Department between 2011 and 2016.

MIMIC-IV [6] contains data for over 60,000 patients admitted to the ICU or emergency department of the Beth Israel Deaconess Medical Center between 2008 and 2019. The patients included in MIMIC-CXR-JPG are a subset of MIMIC-IV.

Both CheXpert and MIMIC-CXR-JPG were labeled with the CheXpert labeler, which derives labels for up to 14 observations from free-text radiology reports using pre-defined rules. Labels are either positive (disease diagnosed), negative (disease ruled out) or uncertain (unable to rule out disease). If an observation was not mentioned in the report at all, its label is missing.

3 Method

3.1 Problem Definition

Given a chest X-ray image dataset X, a set of augmentations $\mathcal{T}$, and an encoder g, the goal of contrastive learning is to find a representation that minimizes the following InfoNCE loss:

$$\mathcal{L}(X) = \mathbb{E}_{x \in X}\left[-\log \frac{\exp(g(\hat{x}_1)\cdot g(\hat{x}_2))}{\exp(g(\hat{x}_1)\cdot g(\hat{x}_2)) + \sum_{i=1}^{K}\exp(g(\hat{x}_1)\cdot g(\hat{z}_i))}\right], \tag{1}$$

where the positive pairs $(\hat{x}_1 = t(x), \hat{x}_2 = t'(x))$ with $t, t' \in \mathcal{T}$ are the augmented views of the same input image x, and the negative pairs $(\hat{x}_1, \hat{z}_i)$, $1 \leq i \leq K$ are pairs of augmentations of different images, i.e., $\hat{z}_i = t''(z_i)$ with $t'' \in \mathcal{T}$ for $z_i \neq x$. We use MoCo architecture to encode negative pairs, where K is the size of queue. We also restrict $\mathcal{T}$ to flipping, cropping, and Gaussian blur.

We first aim to increase the number of positive pairs by using multiple images instead of one image. Similarly to MedAug [3], we use the available patient metadata such as the patient number to select positive pairs.

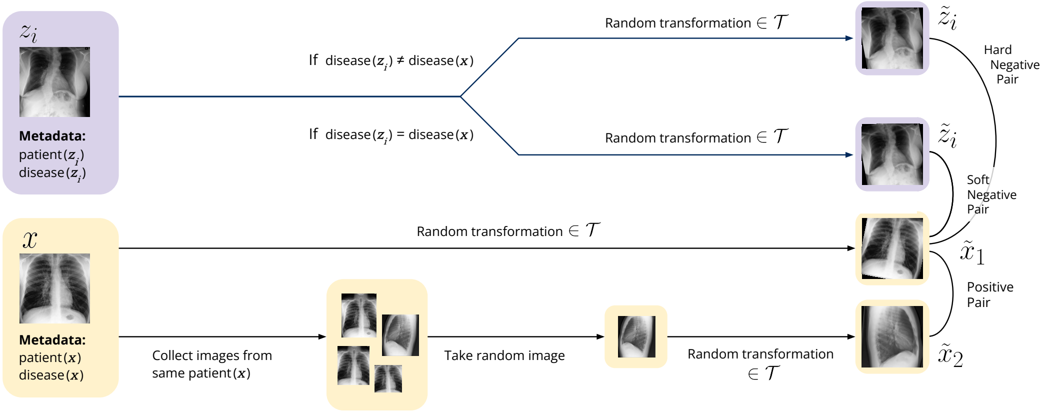
z_i
Metadata:
patient(z_i)
disease(z_i)
If disease(z_i) ≠ disease(x)
If disease(z_i) = disease(x)
Random transformation ∈ $\mathcal{T}$
Random transformation ∈ $\mathcal{T}$
$\tilde{z}_i$
$\tilde{z}_i$
Hard Negative Pair
Soft Negative Pair
x
Metadata:
patient(x)
disease(x)
Random transformation ∈ $\mathcal{T}$
$\tilde{x}_1$
Positive Pair
Collect images from same patient(x)
Take random image
Random transformation ∈ $\mathcal{T}$
$\tilde{x}_2$

Figure 1: Process of selecting transformed positive, soft negative and hard negative pairs from original images using disease metadata.

More formally, the new positive pairs are $(\tilde{x}_1, \tilde{x}_2)$, where

$$\begin{aligned} \tilde{x}_1 &= t(x), && \text{with } t \in \mathcal{T} \\ \tilde{x}_2 &= t'(x'), && \text{with } t' \in \mathcal{T}, \text{ patient}(x') = \text{patient}(x), \end{aligned} \tag{2}$$

such that images of the same patients are supposed to have representations close to each other in representation space.

3.2 Disease-Weighted Contrastive Loss

Our next aim is to leverage additional disease labels to select weighted negative pairs. In contrast to MedAug, which treats *all* different patients equally, we hypothesize that maximizing the distance to some negative samples with a higher weight than others would help getting better performance on downstream tasks. More specifically, we use disease labels available during training stage to define hard and soft negative pairs. Thus, the soft and hard negative samples for image x are defined as:

$$S(x) := \{t(z);\, z \in X,\, \text{disease}(x) = \text{disease}(z),\, t \in \mathcal{T}\} \tag{3}$$

$$H(x) := \{t(z);\, z \in X,\, \text{disease}(x) \neq \text{disease}(z),\, t \in \mathcal{T}\} \tag{4}$$

where image x and z are from different patients. Figure 1 demonstrates the process of positive and negative sample selection. Note that we call this approach a *weakly* supervised method, as we are injecting the labels into our pretraining phase. However, to avoid information leakage for our downstream supervised task, we exclude diseases related to the target disease (see Figure 2 in the appendix).

We now introduce our disease-weighted contrastive loss for DisCeRn:

$$\mathcal{L}_{\text{DisCeRn}}(X) = \mathbb{E}_{x \in X}\left[-\log \frac{\exp(g(\tilde{x}_1) \cdot g(\tilde{x}_2))}{\exp(g(\tilde{x}_1) \cdot g(\tilde{x}_2)) + G(\tilde{x}, z)}\right], \tag{5}$$

where

$$G(\tilde{x}, z) = \sum_{z_i \in S(x)} w_{\text{s}} \exp(g(\tilde{x}_1) \cdot g(z_i)) + \sum_{z_j \in H(x)} w_{\text{h}} \exp(g(\tilde{x}_1) \cdot g(z_i)) \tag{6}$$

is the weighted sum of negative pairs. In our experiments, we choose $w_s = \frac{0.1K}{|S|}$ and $w_h = \frac{0.9K}{|H|}$ to put more weight on hard (different disease) negative samples. Note that the loss defined in Equation 5 uses the newly defined positive pairs in Equation 2 derived from MedAug.

3.3 Augmenting Image Embeddings with Clinical Features

In the previous section, we used disease labels available during the training stage to get better representations. Now, we describe how we can leverage the available clinical features during the finetuning and deployment stage to get better performance on the downstream task. More specifically, for each image in the MIMIC-CXR-JPG dataset, we also access the patient's clinical features in MIMIC-IV, and MIMIC-IV-ED [6, 15]. We choose six clinical features from MIMIC-IV-ED, including temperature, heart rate, and oxygen saturation, available for almost more than 99% of the patients. Also, we select ten laboratory measurements with the lowest missing data, such as creatinine, hemoglobin, and red blood cell count. The list of all these 16 clinical features can be found in appendix Table 4. We take the average of each feature's values within two days before/after taking the corresponding chest X-ray image. We extend image embeddings during the finetuning stage by appending these extra features to it. Then, we train a linear classifier on top of it.

4 Experiments

Our experimental pipeline is divided into two major sections. Section 4.1 describes our pretraining strategies using baseline SSL methods as well as our proposed approach. Section 4.2 describes how we finetuned our models on the task of Pneumonia prediction. Numerical results are shown in Table 1 and a brief discussion is given in Section 4.3.

4.1 Pretraining

The first stage our experiment pipeline involves pretraining using existing self-supervised methods, as well as our proposed weakly supervised method using the disease weighted metric (DisCeRn) described in subsection 3.2. All models are pretraining using the MIMIC-CXR-JPG dataset [7] described in section 2. As a baseline, we begin by training a vanilla Momentum Contrast (MoCo) [2] encoder with a 121-layer Densenet [16], an architecture previously chosen for self-supervised learning and classification for chest X-ray images [4]. We proceed to train a second MoCo encoder using the MedAug strategy for sampling positive image pairs as any two images captured of the same patient [3]. Lastly, we train a third MoCo encoder still using the MedAug sampling strategy, but with our additional proposed addition of soft and hard negatives based on auxiliary disease labels. For the sake of consistency all experiments are run with the same hyperparameters and base data augmentations for contrastive learning (flipping, cropping, blurring). A table of all hyperparameters can be found in the appendix Table 2. The training time is consistent across each method and is approximately 1 week running on 2 Nvidia RTX A600 48GB graphics cards. The best model in each experiment was selected using the lowest average validation loss; in every case, this corresponded to approximately the 200th epoch.

4.2 Finetuning

Each self-supervised encoder was finetuned on the task of binary Pneumonia classification on the MIMIC-CXR-JPG [7] and CheXpert [8] datasets. Preprocessing only involved removing samples without a definite label for Pneumonia, the number of available training and test samples that remained is shown in Table 1. We explore two finetuning methods to build Pneumonia classifiers out of our self-supervised encoders. The first finetuning method involved fitting a *linear* classifier on top of the embedding space learned by each self-supervised encoder. The intention of training a linear classifier is not to produce the most accurate test results but to perform an unbiased test of the representation power of our embeddings for making clinical predictions. The second finetuning method also uses a linear classifier on top of the embedding layer but allows the entire encoder to train in an *end-to-end* fashion — as there are many more trainable parameters, this method is expected to produce better results than simple linear classification. Each encoder is finetuned again on the MIMIC-CXR-JPG dataset but our 16 selected numerical measurements from MIMIC-IV are concatenated to the representation of each image before the linear classification layer. It should be noted that the 16 additional features are specific to the patient and have been recorded at the time of image acquisition (see subsection 3.3 for additional details on this method). Lastly, a Densenet model pretrained on the ImageNet dataset [17] is finetuned in an *end-to-end* fashion on both MIMIC-CXR-JPG [7] and CheXpert [8] datasets to serve as a baseline. All models were trained for a total of 50 epochs and the best model was chosen based on the validation AUC score. A full table of hyperparameters can be found in the appendix Table 3. Table 1 summarizes the results of the finetuning experiments and the following section (subsection 4.3) goes into a discussion and interpretation.

Table 1: **AUC Scores for Binary Pneumonia Classification.** Three pretrained self-supervised models – MoCo, MedAug and DisCeRn (our approach) – are finetuned for Pneumonia classification on the CheXpert and MIMIC-CXR-JPG datasets. Two finetuning strategies are employed: *linear* finetuning trains only a linear classifier on top of the pretrained model's embedding space, while *End-to-End* finetuning uses the pretrained model as an initialization which can be fully updated during training. Additionally, each model is finetuned again for Pneumonia classification on MIMIC-CXR-JPG using additional input features from MIMIC-IV associated with the patient at the time of image acquisition. Lastly, a baseline model trained using standard supervised learning on ImageNet is used as a baseline. Best results are shown in bold, second-best are underlined. Discussion of these results can be found in subsection 4.3.

Method	CheXpert	MIMIC-CXR-JPG	MIMIC-CXR-JPG + MIMIC IV
	Train: 8838	Train: 68057	Train: 68057
	Test: 234	Test: 495	Test: 495
ImageNet *End-toEnd*	87.57%	76.91%	68.48%
MoCo *Linear*	84.98%	79.14%	71.82%
MoCo *End-to-End*	88.42%	79.63%	73.36%
MedAug *Linear*	88.11%	78.74%	73.99%
MedAug *End-to-End*	85.10%	**80.88%**	**78.00%**
DisCeRn *Linear*	86.25%	78.64%	72.76%
DisCeRn *End-to-End*	**89.03%**	80.51%	75.59%

4.3 Discussion of Results

We discuss the observed results from Table 1 in three settings: linear models, end-to-end models, and models using MIMIC-IV features. First, we notice MedAug linear is superior by 2% on CheXpert, but on MIMIC-CXR-JPG all linear models are nearly identical. When MIMIC-IV is used MedAug again dominates but only by a 1% margin. Additionally, we notice that linear models perform nearly as well as their end-to-end counterparts, with MedAug linear actually performing better than its counterpart on CheXpert — hinting that linear classifiers are often suitable for classifying self-supervised X-ray embeddings.

For end-to-end models, we see that DisCeRn is superior by a small margin on CheXpert, similar to MedAug on MIMIC-CXR-JPG and suboptimal on MIMIC-CXR-JPG+IV. However, the small percentage margin between results is likely within the realm of statistical uncertainty making it difficult to claim any single method as superior (see section 6 for further discussion).

Lastly, we observe a significantly lower AUC score for models trained using additional features from MIMIC-IV ($5.8\% \pm 1.8\%$ worse). This was an unexpected result that we do not fully understand. Our hypothesis is that combining image representations with numerical features creates too complex of a representation for the final layer of a deep network, even in end-to-end learning. There is little reason why image representations should contain linear correlations between arbitrary numerical measurements, forcing our network to search for patterns that do not generalize to unseen data. A potential solution is to require an additional deep network to classify these joint image/measurement representations as opposed to a single learn layer.

5 Conclusions

We propose Disease-Contrastive Representations (DisCeRn), a self-supervised learning approach to leverage domain knowledge, such as patient features (heart rate, O2, lab results or blood test), in addition to medical imaging. It shows promise for learning meaningful representations through self-supervised learning (SSL) for downstream tasks in a way such that it takes advantage of partial labels for uncorrelated classification tasks. A particular use case where our methodology would be useful is when a medical imaging dataset is accompanied with labels for diseases that are marginally or even unrelated to a target disease. For example, a label for Pleural effusion might be useful while pre-training to learn Pneumonia – as validated by our experimental results.

6 Limitations and Future directions

In this section, we highlight some of the limitations and explore possible future directions of our research. One of the limitations of our work is that simple aggregation methods like concatenation do not work well as evidenced by the experimental results, as shown when MIMIC-IV [6] patient features are concatenated with pretrained representations of images on MIMIC-CXR-JPG [7]. During the test time, this method leads to a degradation in the performance. One of the reasons we believe is that aggregation methods like concatenation are simple enough for the model to actually discriminate the classes. Thus, complex aggregation methods for combining learned representations and features should be explored. Also, it could be possible that since DisCeRn does not see the patient features during the pretraining phase, the performance degrades. A future direction we would like to explore is how to couple the images with additional features during the pretraining phase to learn better representations. As far as the aggregation of the features is concerned, instead of selecting the positive and negative pairs only on similar patients, the metadata, and the features would even be beneficial to select a “good” and “bad” pair. For instance, features such as lab values could be useful; the lab values which are close enough would suggest that the patients are similar enough as opposed to the lab values which are far apart.

Here are some of the limitations of the experimental results Table 1. First, our experiments did not evaluate the fine-tuning performance on small dataset sizes, e.g. 1%, 10% of the training dataset to analyze and compare the performance as training size decreases. One other limitation is that due to time constraints, minimal hyper-parameter tuning was performed, such as exploration of different learning rate schedulers, batch sizes, optimizers, or even tweaking the architecture of DisCeRn. Hence, it would be worth exploring these hyper-parameters to achieve optimal performance and compare the results objectively. Furthermore, we did not comprehensively evaluate our methodology on different disease pathology labels, such as Edema or Consolidation, etc. A natural future direction would be to test the validity of our results on other labels besides Pneumonia. To reduce the variance of results, we wish to explore cross-validations and statistical tests, such as K-fold cross-validation or permutation tests, to validate the conclusions of our proposed methodology.

References

[1] Ting Chen et al. "A Simple Framework for Contrastive Learning of Visual Representations". In: *CoRR* abs/2002.05709 (2020). arXiv: `2002.05709`. URL: `https://arxiv.org/abs/2002.05709`.

[2] Kaiming He et al. "Momentum Contrast for Unsupervised Visual Representation Learning". In: *CoRR* abs/1911.05722 (2019). arXiv: `1911.05722`. URL: `http://arxiv.org/abs/1911.05722`.

[3] Yen Nhi Truong Vu et al. "MedAug: Contrastive learning leveraging patient metadata improves representations for chest X-ray interpretation". In: *CoRR* abs/2102.10663 (2021). arXiv: `2102.10663`. URL: `https://arxiv.org/abs/2102.10663`.

[4] Anuroop Sriram et al. "COVID-19 Prognosis via Self-Supervised Representation Learning and Multi-Image Prediction". In: *CoRR* abs/2101.04909 (2021). arXiv: `2101.04909`. URL: `https://arxiv.org/abs/2101.04909`.

[5] Betty van Aken et al. "Clinical Outcome Prediction from Admission Notes using Self-Supervised Knowledge Integration". In: *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. Online: Association for Computational Linguistics, Apr. 2021, pp. 881–893. DOI: `10.18653/v1/2021.eacl-main.75`. URL: `https://aclanthology.org/2021.eacl-main.75`.

[6] Alistair E. W. Johnson et al. "MIMIC-IV: (version 1.0)". In: *PhysioNet* (2021). DOI: `https://doi.org/10.13026/s6n6-xd98`.

[7] Alistair E. W. Johnson et al. "MIMIC-CXR-JPG: A large publicly available database of labeled chest radiographs". In: *CoRR* abs/1901.07042 (2019). arXiv: `1901.07042`. URL: `http://arxiv.org/abs/1901.07042`.

[8] Jeremy Irvin et al. "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*. AAAI Press, 2019, pp. 590–597. DOI: `10.1609/aaai.v33i01.3301590`. URL: `https://doi.org/10.1609/aaai.v33i01.3301590`.

[9] Mark I Neuman et al. "Variability in the interpretation of chest radiographs for the diagnosis of pneumonia in children". In: *Journal of hospital medicine* 7.4 (2012), pp. 294–298.

[10] RM Hopstaken et al. "Inter-observer variation in the interpretation of chest radiographs for pneumonia in community-acquired lower respiratory tract infections". In: *Clinical radiology* 59.8 (2004), pp. 743–752.

[11] Ting Chen et al. "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*. Vol. 119. Proceedings of Machine Learning Research. PMLR, 2020, pp. 1597–1607. URL: `http://proceedings.mlr.press/v119/chen20j.html`.

[12] Aäron van den Oord, Yazhe Li, and Oriol Vinyals. "Representation Learning with Contrastive Predictive Coding". In: *CoRR* abs/1807.03748 (2018). arXiv: `1807.03748`. URL: `http://arxiv.org/abs/1807.03748`.

[13] Hari Sowrirajan et al. "MoCo Pretraining Improves Representation and Transferability of Chest X-ray Models". In: *Medical Imaging with Deep Learning, 7-9 July 2021, Lübeck, Germany*. Ed. by Mattias P. Heinrich et al. Vol. 143. Proceedings of Machine Learning Research. PMLR, 2021, pp. 728–744. URL: `https://proceedings.mlr.press/v143/sowrirajan21a.html`.

[14] Yuhao Zhang et al. "Contrastive Learning of Medical Visual Representations from Paired Images and Text". In: *CoRR* abs/2010.00747 (2020). arXiv: `2010.00747`. URL: `https://arxiv.org/abs/2010.00747`.

[15] Alistair Johnson et al. *MIMIC-IV-ED*. 2021. DOI: `10.13026/77Z6-9W59`. URL: `https://physionet.org/content/mimic-iv-ed/1.0/`.

[16] Gao Huang, Zhuang Liu, and Kilian Q. Weinberger. "Densely Connected Convolutional Networks". In: *CoRR* abs/1608.06993 (2016). arXiv: `1608.06993`. URL: `http://arxiv.org/abs/1608.06993`.

[17] Olga Russakovsky et al. "ImageNet Large Scale Visual Recognition Challenge". In: *CoRR* abs/1409.0575 (2014). arXiv: `1409.0575`. URL: `http://arxiv.org/abs/1409.0575`.

[18] Diederik P. Kingma and Jimmy Ba. “Adam: A Method for Stochastic Optimization”. In: *arXiv e-prints*, arXiv:1412.6980 (Dec. 2014), arXiv:1412.6980. arXiv: `1412.6980 [cs.LG]`.

A Team Contributions

In the subsections below a brief description of the main contributions of each team member is provided. Overall the contribution level was balanced among members and there are no complaints to be had.

A.1 Tom Ginsberg

- set up the repository, downloaded datasets
- contributed to code for dataloaders, training, Moco and Medaug
- ran baselines for Moco and Medaug
- contributed to code for finetuning
- ran finetuning experiments
- wrote the abstract, background and experiments section

A.2 Vahid Balazadeh

- contributed to the implementation of our modified loss function
- ran DisCeRn pretraining and some of finetuning experiments
- wrote the related work and large parts of the methods section
- investigated extra clinical features from MIMIC-IV and MIMIC-IV-ED and contributed to the implementation of finetuning with extra features

A.3 Korbinian Koch

- came up with the idea for disease-dependent hard and soft negatives
- contributed to the implementation of our modified loss function
- wrote the dataset section and parts of the methods section
- created all illustrations
- created large parts of the presentation

A.4 Vasudev Sharma

- proposed the problem for the project and some related works
- for report, wrote conclusion, limitations and future work
- contributed to evaluating models on downstream tasks

B Hyperparameters

Table 2: List of all hyperparamers used in pretraining experiments. Parameters are largely similar to previous efforts for self supervised learning and classification on chest X-ray images [4].

Hyperparameter	Value
Architecture	Densenet 121 [16]
Embedding Dimension	256
Image Input Dimension	224×224
Optimizer	ADAM [18]
Base Learning Rate	0.001
Train Epochs	200
Batch Size	512
Learning Rate Schedule	Cosine Annealing
Schedule Period	200 epochs
MoCo Momentum	0.9
Moco Queue Size	2^{16}
Random Crop Area	0.2-1
Random Flip Rate	0.5
Random Gaussian Blur Rate	0.5
Random Gaussian Blur Intensity	$\sigma \sim \mathcal{U}(0.2, 1)$
Model Selection Criteria	Validation Loss

Table 3: List of all hyperparamers used in finetuning experiments

Hyperparameter	Value
Architecture	Densenet 121 [16]
Embedding Dimension	256
Image Input Dimension	224×224
Optimizer	ADAM [18]
Base Learning Rate	0.001
Train Epochs	50
Batch Size	64
Learning Rate Schedule	Step $\gamma = 0.2$
Schedule Period	10 epochs
Random Crop Area	0.2-1
Random Flip Rate	0.5
Random Gaussian Blur Rate	0.5
Random Gaussian Blur Intensity	$\sigma \sim \mathcal{U}(0.2, 1)$
Model Selection Criteria	Validation AUC Score

C MIMIC-CXR Labels

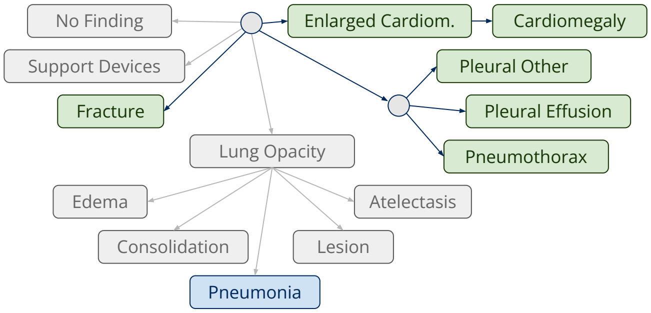

Figure 2: Labels used for pretraining (🟢) and evaluating the models (🔵).

D Code

The code can be found in this repository - GitHub

E Visualization of Embedding Space

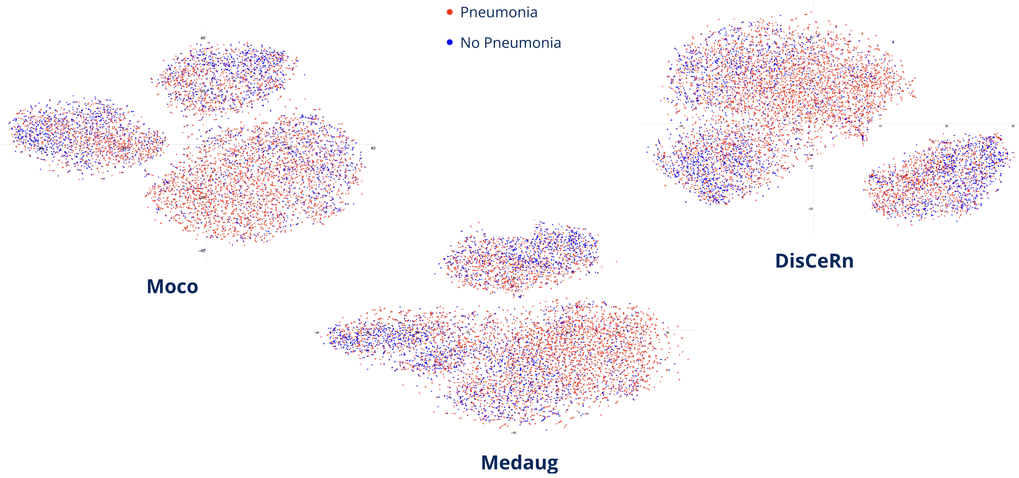

Figure 3: t-SNE visualization of pre-trained embeddings for MoCo, MedAug, and DisCeRn on MIMIC-CXR dataset. Blue and red points denote patients with and without Pneumonia, respectively.

F List of Clinical Features from MIMIC-IV and MIMIC-IV-ED

Table 4: List of all clinical features from MIMIC-IV and MIMIC-IV-ED used in finetuning experiments

Feature	Description
temperature	Patient Temperate
heartrate	Heart Rate
resprate	Respiratory Rate
o2sat	Oxygen Saturation
sbp	Systolic Blood Pressure
dbp	Diastolic Blood Pressure
chl	Chloride
crt	Creatinine
htc	Hematocrit
hbg	Hemoglobin
mch	Mean Corpuscular Hemoglobin
mchc	Mean Corpuscular Hemoglobin Concentration
mcv	Mean Corpuscular Volume
rdw	Red Cell Distribution Width
rbc	Red Blood Cells
bun	Blood Urea Nitrogen