

Sample-Specific Debiasing for Better Image-Text Models

Peiqi Wang¹

WPQ@MIT.EDU

Yingcheng Liu¹

LIUYC@MIT.EDU

Ching-Yun Ko¹

CYKO@MIT.EDU

William M. Wells¹

SW@BWH.HARVARD.EDU

Seth Berkowitz²

SBERKOWI@BIDMC.HARVARD.EDU

Steven Horng²

SHORNG@BIDMC.HARVARD.EDU

Polina Golland¹

POLINA@CSAIL.MIT.EDU

¹ *Massachusetts Institute of Technology, Cambridge, MA, USA*

² *Beth Israel Deaconess Medical Center, Harvard Medical School, Boston, MA, USA*

Abstract

Self-supervised representation learning on image-text data facilitates crucial medical applications, such as image classification, visual grounding, and cross-modal retrieval. One common approach involves contrasting semantically similar (positive) and dissimilar (negative) pairs of data points. Drawing negative samples uniformly from the training data set introduces false negatives, i.e., samples that are treated as dissimilar but belong to the same class. In healthcare data, the underlying class distribution is nonuniform, implying that false negatives occur at a highly variable rate. To improve the quality of learned representations, we develop a novel approach that corrects for false negatives. Our method can be viewed as a variant of debiased contrastive learning that uses estimated sample-specific class probabilities. We provide theoretical analysis of the objective function and demonstrate the proposed approach on both image and paired image-text data sets. Our experiments illustrate empirical advantages of sample-specific debiasing.

1. Introduction

In this paper, we propose and demonstrate a novel approach for contrastive learning of image-text representations. Specifically, we propose to estimate sample-specific class probabilities, i.e., the latent class probability of each data point, to appropriately compensate for the effect of false negative samples in the learning procedure. Our method achieves state-of-the-art performance on a wide range of downstream tasks.

Self-supervised representation learning on paired image-text data uses text as “labels”, requiring no further annotations beyond what has been routinely documented (Radford et al., 2021). By leveraging natural language to reference visual concepts and vice versa, the resulting image-text models trained using self-supervised objectives can perform a diverse set of vision-language tasks (Radford et al., 2021; Zhang et al., 2022b).

When applied to the medical domain, image-text models can (i) retroactively label images to select relevant patients for a clinical trial, (ii) help physicians verify the accuracy

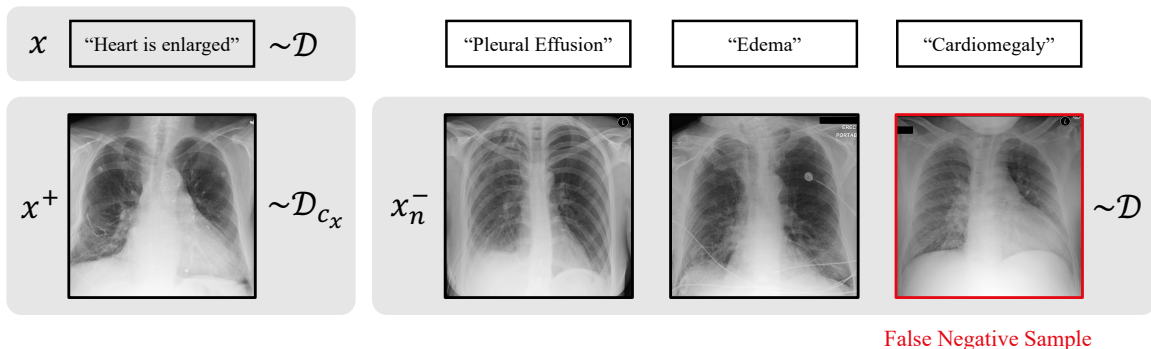


Figure 1: False negatives in paired image-text data. Drawing negative image samples x_n^- from the data distribution \mathcal{D} may result in samples that are semantically similar to the text x (“Heart is enlarged” and “Cardiomegaly” imply the same pathology). False negative samples occur at an uneven rate (e.g., depending on the pathology type) and degrade the performance of image-text models on downstream tasks.

of a report by noting whether the referred location (i.e., visual grounding of the text) is consistent with their impression of the image, and (iii) enable informed interpretation of medical images by retrieving similar patients from a database. Moreover, the abundance of paired image-text data (e.g., radiographs and radiology reports, histology slides and pathology reports) suggests the broad applicability of self-supervision using image-text data to improve healthcare.

In this paper, we focus on contrastive learning, a self-supervised approach that encourages the representations of semantically similar or *positive* pairs of data to be close, and those of dissimilar or *negative* pairs to be distant. Contrastive learning has been applied in the medical domain, demonstrating impressive transfer capabilities on a diverse set of downstream tasks (Chauhan et al., 2020; Huang et al., 2021; Liao et al., 2021; Müller et al., 2022; Zhang et al., 2022b; Boecking et al., 2022; Wang et al., 2023; Bannur et al., 2023). The biggest improvements come from addressing challenges unique to this domain. Examples include using cross-attention to localize areas of interest to handle the lack of effective pathology detectors (Huang et al., 2021), fine-tuning language models on medical corpora to address linguistic challenges in clinical notes (Boecking et al., 2022). In a similar spirit, our work aims to address the nonuniform class distribution typical of healthcare data.

Using text as “labels” induces a nonuniform class distribution with large support. Specifically, natural language descriptions of medical images can represent a vast number of possible classes. Most descriptions belong to a few common classes (e.g., cardiomegaly) while the remaining descriptions are spread across many rare classes (e.g., left apical pneumothorax). In our problem of chest X-ray representation learning, each CheXpert label (Irvin et al., 2019) representing common classes in chest radiographs is mentioned in 10-25% of the radiology reports in MIMIC-CXR, a large collection of chest X-ray images and radiology reports (Johnson et al., 2019). At the same time, the rare classes may be more descriptive and associated with high risk, which makes their accurate identification important in

clinical applications. Our goal is to handle the highly nonuniform class distribution that presents a challenge for many existing contrastive learning approaches.

When training an image-text model using contrastive learning, each text is positively paired with the associated image from the same imaging event and negatively paired with a batch of images uniformly drawn from the training data set (Li et al., 2019; Lu et al., 2019; Chen et al., 2020; Radford et al., 2021; Liao et al., 2021; Zhang et al., 2022b; Wang et al., 2023). If a negatively paired image is semantically similar to the text, it is considered as a *false negative* (Saunshi et al., 2019), as illustrated in Figure 1. It has been shown previously that false negatives cause a substantial decline in downstream classification performance when using image representations trained with contrastive learning (Chuang et al., 2020).

One approach to alleviating the problem of false negatives is to identify false negative pairs explicitly and reduce their effect. In some applications, ground truth class labels are available and can be used to ensure that no false negative pairs are generated (Khosla et al., 2020; Dwivedi et al., 2021). Unfortunately, deriving categorical labels from text is challenging because it can be difficult to determine the appropriate level of class granularity and ensure sufficient class coverage when the class distribution is nonuniform with large support. Label imputation with nearest neighbor methods (Zheng et al., 2021) or clustering (Chen et al., 2022) requires non-trivial implementation and additional computational cost. In the absence of class labels, a common approach is to treat negative samples whose embeddings are close to positive samples as false negatives and to eliminate them (Huynh et al., 2022; Zhang et al., 2022a; Zhou et al., 2022; Wang et al., 2022). This approach is easy to implement but it overlooks valuable information captured by text if only image embedding is used. Moreover, rejecting negative samples whose embeddings are close to positive samples runs risk of removing valuable “hard negatives”, i.e., visually similar true negative samples.

Alternatively, debiased contrastive learning uses (possibly) extra positive samples to offset the influence of false negatives without explicitly identifying them (Chuang et al., 2020). It assumes a uniform class distribution and applies a constant correction to each sample. This strategy can be suboptimal for healthcare data where the class distribution is nonuniform. We observe that naively applying debiased contrastive learning introduces a performance trade-off between coarse-grained tasks (e.g., classifying pneumonia) and fine-grained tasks (e.g., cross-modal retrieval of rare classes), as seen in Table 3.

Our method aims to reduce the effect of false negatives in contrastive representation learning while making no assumption on the underlying class distribution. Specifically, we estimate the class probability (or level of correction) for each data point based on the likelihood of the text provided by a language model. Our method (i) does not require or attempt to infer class labels, (ii) requires a few extra lines of code to implement when compared with contrastive objectives typically used in practice, (iii) incurs minimal computational overhead, and (iv) takes advantage of the class information implicitly represented by the text “labels” to adaptively mitigate the problem of false negatives in image-text contrastive learning. We study the advantages of using sample-specific class probability estimates on a small-scale image data set constructed with a nonuniform class distribution. We evaluate our approach on a large set of chest X-ray images and associated radiology reports (Johnson et al., 2019), demonstrating superior performance on image classification, visual grounding, and cross-modal retrieval tasks.

Generalizable Insights about Machine Learning in the Context of Healthcare

Our work (i) provides direct value to those interested in using self-supervised representation learning for healthcare applications, (ii) highlights the importance of considering the distinct characteristics of healthcare data that can pose challenges for methods designed for simpler scenarios (in our case, false negatives degrade the performance of image-text models trained on clinical data), and (iii) shows that a language model provides a useful prior that can be employed in other image-text modeling problems (e.g., brain tumor CT scans with imaging reports) or inspire future research to solve related problems in other data modalities.

2. Methods

In this section, we introduce notation and provide a brief overview of debiased contrastive learning (Chuang et al., 2020), followed by the derivation of our method to compensate for potential false negative samples and analysis of the relationship between our approach and the original formulation.

2.1. Notation and Problem Setup

Let \mathcal{X} be the set of all possible data points. In our application, this includes both images and text. Contrastive learning assumes access to similar (positive) data pairs (x, x^+) and N i.i.d. negative samples $\{x_n^-\}$ that are presumably unrelated to x . We use set \mathcal{C} of discrete latent classes to formalize the notion of semantic similarity, e.g., similar data (x, x^+) have the same latent class. Let ρ be the distribution over the latent class set \mathcal{C} . We use \mathcal{D}_c to denote the probability distribution over \mathcal{X} that captures the likelihood of a data point x belonging to a class $c \in \mathcal{C}$. The data distribution \mathcal{D} is simply the marginal distribution, i.e., $\mathcal{D}(x) \triangleq \sum_{c \in \mathcal{C}} \rho(c) \mathcal{D}_c(x)$. For convenience, we use c_x to denote the latent class of $x \in \mathcal{X}$. In practice, the class label c_x is unknown.

The positive data pair $(x, x^+) \sim \mathcal{D}_{\text{sim}}$ belongs to the same class by construction. This is achieved for example by (i) applying class-preserving data augmentation to generate an image x^+ from another image x (\mathcal{D}_{sim} is a distribution of image and its data augmentations) or (ii) treating x^+ as the image associated with text x (\mathcal{D}_{sim} is a distribution of image-text pairs). We define $\mathcal{D}_{\text{sim}}(x, x^+) \triangleq \mathcal{D}(x) \mathcal{D}_{c_x}(x^+)$ to capture the constraint that x^+ is generated from the same latent class c_x as x .

Negative samples $\{x_n^-\}$ should be drawn from semantically dissimilar (w.r.t. x) data

$$\mathcal{E}_{c_x}(x') \triangleq p(x' \mid c \neq c_x) = \sum_{c \neq c_x} \frac{\rho(c)}{1 - \rho(c_x)} \mathcal{D}_c(x'), \quad (1)$$

which is infeasible since we do not have access to class labels c_x . In practice, negative samples $\{x_n^-\}$ are typically drawn from the marginal \mathcal{D} . Doing so introduces false negative samples. Specifically, we observe that sampling from \mathcal{D} is equivalent to sampling from a mixture of semantically similar data \mathcal{D}_{c_x} and semantically dissimilar data \mathcal{E}_{c_x} , i.e.,

$$\mathcal{D}(x') = \sum_{c \in \mathcal{C}} \rho(c) \mathcal{D}_c(x') = \rho(c_x) \mathcal{D}_{c_x}(x') + (1 - \rho(c_x)) \mathcal{E}_{c_x}(x'). \quad (2)$$

With probability $\rho(c_x)$, $x_n^- \sim \mathcal{D}$ is a false negative that is drawn from the component \mathcal{D}_{c_x} .

In many computer vision data sets, the number of classes is large and the false negative rate $\rho(c_x)$ is likely small for any data point $x \in \mathcal{X}$. Thus, using \mathcal{D} instead of \mathcal{E}_{c_x} to sample negative examples is reasonable. Moreover, it is natural to assume that all classes are equally (un)likely. Unfortunately, neither assumption is true in the healthcare setting. In clinical image data sets, the probability of encountering common pathologies in a randomly chosen image is not negligible. Furthermore, common pathologies appear much more frequently than rare ones. Debiased contrastive learning (Chuang et al., 2020) addresses the former problem by modifying the optimization function to explicitly account for the chance of false negatives uniformly across data points.

2.2. Debiased Contrastive Learning

Contrastive learning aims to find a good representation function $f : \mathcal{X} \rightarrow \mathcal{S}^{d-1}(\gamma)$ that encodes data in \mathcal{X} to a hypersphere of radius γ where $\mathcal{S}^{d-1}(\gamma) \triangleq \{x \in \mathbb{R}^D \mid \|x\|_2 = \gamma\}$. Specifically, the contrastive learning objective forces representations of positive pairs to be closer than those of negative pairs (van den Oord et al., 2018) by minimizing

$$\mathbb{E}_{(x,x^+) \sim \mathcal{D}_{\text{sim}}, \{x_n^-\} \sim \mathcal{D}^N} \left[-\log \frac{e^{s(x,x^+)}}{e^{s(x,x^+)} + \sum_{n=1}^N e^{s(x,x_n^-)}} \right], \quad (3)$$

where the bounded function $s : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ measures the similarity (e.g., a dot-product) between learned representations captured by the encoder function f . Negative samples are uniformly drawn from the data \mathcal{D} , which poses a risk of sampling false negatives.

Re-arranging Equation 2, true negative sample distribution \mathcal{E}_{c_x} can be expressed in terms of \mathcal{D} and \mathcal{D}_{c_x} from which we can readily sample:

$$\mathcal{E}_{c_x}(x') = \frac{1}{1 - \rho(c_x)} \mathcal{D}(x') - \frac{\rho(c_x)}{1 - \rho(c_x)} \mathcal{D}_{c_x}(x'). \quad (4)$$

The asymptotic debiased contrastive learning objective (Chuang et al., 2020) considers the case where the number N of negative samples drawn from \mathcal{E}_{c_x} goes to infinity:

$$\mathbb{E}_{(x,x^+) \sim \mathcal{D}_{\text{sim}}} \left[-\log \frac{e^{s(x,x^+)}}{e^{s(x,x^+)} + N \mathbb{E}_{x^- \sim \mathcal{E}_{c_x}} [e^{s(x,x^-)}]} \right]. \quad (5)$$

Given N samples $\{u_n\}$ from \mathcal{D} and M samples $\{v_m\}$ from \mathcal{D}_{c_x} , the expected value in the denominator of Equation 5 can be estimated with

$$g(x, \{u_n\}, \{v_m\}; \eta) \triangleq \frac{1}{1 - \eta} \left[\frac{1}{N} \sum_{n=1}^N e^{s(x,u_n)} \right] - \frac{\eta}{1 - \eta} \left[\frac{1}{M} \sum_{m=1}^M e^{s(x,v_m)} \right]. \quad (6)$$

When the class distribution is uniform and $\rho(c_x) = \eta$ for all $x \in \mathcal{X}$, the estimator is consistent, i.e., $g(x, \{u_n\}, \{v_m\}; \eta) \xrightarrow{P} \mathbb{E}_{x^- \sim \mathcal{E}_{c_x}} [e^{s(x,x^-)}]$ as $N, M \rightarrow \infty$. In debiased contrastive learning (Chuang et al., 2020), η is treated as a hyperparameter and is assumed to be constant. To avoid numerical issues, the estimator $g(x, \{u_n\}, \{v_m\}; \eta)$ is lower bounded by its

theoretical minimum $e^{-\gamma^2}$ (Chuang et al., 2020). The resulting debiased contrastive loss

$$\mathcal{L} \triangleq \mathbb{E}_{(x,x^+) \sim \mathcal{D}_{\text{sim}}, \{u_n\} \sim \mathcal{D}^N, \{v_m\} \sim \mathcal{D}_{c_x}^M} \left[-\log \frac{e^{s(x,x^+)}}{e^{s(x,x^+)} + N g(x, \{u_n\}, \{v_m\}; \eta)} \right] \quad (7)$$

reweights the positive and negative terms in the denominator.

2.3. Sample-specific Class Probability Function $\eta(\cdot)$

The formulation above uses a single hyperparameter η to correct for false negatives uniformly for every data point. We propose to employ an estimate of the class probability $\rho(c_x)$ with a sample-specific class probability function $\eta(\cdot)$, by replacing η with $\eta(\cdot)$ in Equation 6:

$$g(x, \{u_n\}, \{v_m\}; \eta) \triangleq \frac{1}{1 - \eta(x)} \left[\frac{1}{N} \sum_{n=1}^N e^{s(x, u_n)} \right] - \frac{\eta(x)}{1 - \eta(x)} \left[\frac{1}{M} \sum_{m=1}^M e^{(x, v_m)} \right] \quad (8)$$

and use the objective function defined in Equation 7 with $g(\cdot)$ in Equation 8. In practice, $\eta(\cdot)$ may be an imperfect estimator for $\rho(c_x)$. The following proposition informs us how the quality of the estimator $g(\cdot)$ affects the approximation error.

Proposition 1 *Let f and η be arbitrary functions, N and M be finite. Then,*

$$|\mathcal{L} - \bar{\mathcal{L}}| \leq \frac{3e^2 \sqrt{\pi/2}}{\sqrt{N}} \mathbb{E}_{x \sim \mathcal{D}} \left[\frac{1}{1 - \rho(c_x)} \right] + \frac{3e^2 \sqrt{\pi/2}}{\sqrt{M}} \mathbb{E}_{x \sim \mathcal{D}} \left[\frac{\rho(c_x)}{1 - \rho(c_x)} \right] \quad (9)$$

$$+ 3e^2 \mathbb{E}_{x \sim \mathcal{D}} \left[\left| \frac{1}{1 - \eta(x)} - \frac{1}{1 - \rho(c_x)} \right| \right], \quad (10)$$

$$\text{where } \bar{\mathcal{L}} \triangleq \mathbb{E}_{(x,x^+) \sim \mathcal{D}_{\text{sim}}} \left[-\log \frac{e^{s(x,x^+)}}{e^{s(x,x^+)} + N \mathbb{E}_{x^- \sim \mathcal{L}_{c_x}} [e^{s(x,x^-)}]} \right]. \quad (11)$$

Proof is provided in Appendix A.1.

Proposition 1 bounds the approximation error due to (i) finite sample approximation in Equation 6 (first two terms $\mathcal{O}(\frac{1}{\sqrt{N}} + \frac{1}{\sqrt{M}})$) and (ii) misspecification of the sample-specific class probability η (last term). When ρ is uniform and $\eta(x) = \rho(c_x)$ is a constant, Proposition 1 reduces to the result in Chuang et al. (2020), up to constant factors. Assuming access to c_x and using the correct class distribution ρ , i.e., $\eta(x) = \rho(c_x)$, yields a tighter error bound since the last term in the approximation error bound (Equation 10) vanishes. Therefore, we aim to use sample-specific class probability function $\eta(x)$ that closely matches the true sample-specific class probability $\rho(c_x)$.

In Appendix A.2, we show classification generalization bounds on representations trained using debiased contrastive loss with the sample-specific probability function $\eta(\cdot)$ in Equation 8 for arbitrary class distribution.

2.4. Language Model Estimate of Class Probability

We employ the likelihood of the text $p_{\text{LM}}(x)$ provided by a language model (LM) for text x to construct the estimate $\eta(x)$ of the class probability $\rho(c_x)$. Language models naturally

provide estimates of token sequence probabilities. The estimates get better if the language model is fine-tuned on the data \mathcal{D} . We assume a log-linear relationship between text and class probabilities, i.e., $\eta_{LM}(x) = a \cdot p_{LM}(x)^k$ where a and k are hyperparameters. Figure 5 provides pseudocode for our proposed method, which requires a slight modification of the contrastive learning algorithm and incurs minimal computational overhead.

3. Experiments

We evaluate the advantages of estimating sample-specific class probability for contrastive learning in two different experiments.

3.1. CIFAR10

In this experiment, we evaluate image-only representations on data sets with a controlled class distribution. We do not attempt to estimate the class probabilities since no text information is available and would not be helpful since we control the frequency with which each class is included in the data set. Instead, we use the true class distribution during training and examine the effect of class distribution of the data set on the resulting representations.

Data For each $r \in \{0.05, 0.1, 0.25, 0.5, 0.75, 0.9\}$, we generate a CIFAR10- r subset of the CIFAR10 data set (Krizhevsky, 2009) as follows. The original data set includes 6,000 images for each class. For each of 5 selected classes (dog, frog, horse, ship, truck), we randomly draw and include r fraction of the images. We keep all images for each of the remaining 5 classes. The CIFAR10- r data set has a class probability of $\eta_{Low} = 0.2r/(1+r)$ for each of the 5 selected classes and $\eta_{High} = 0.2/(1+r)$ for the remaining classes. Larger values of r lead to a more uniform class distribution.

Representation Learning Similar to Chuang et al. (2020), we use SimCLR (Chen et al., 2022) for contrastive learning of image representations. We employ ResNet-18 (He et al., 2016) as the image encoder f and dot-product as the similarity function s . The encoder is followed by a 2-layer perceptron to create a 128-dimensional embedding. For a reference image x , we employ data augmentation to generate the positive sample x^+ . We set $v_1 = x^+$ ($M = 1$) and draw negative samples $\{u_n\}$ and $\{x_n^-\}$ randomly from the data set ($N = 254$). We set $\gamma = \sqrt{2}$. We use the Adam optimizer (Kingma and Ba, 2014) with a learning rate of 10^{-3} and weight decay of 10^{-6} . Each model is trained for 300 epochs.

Contrastive Loss Variants We learn image representations using four different types of contrastive objectives: (i) baseline contrastive learning without correction for false negatives (CL), (ii) debiased contrastive learning that uses a constant η_{Low} providing the true class probabilities for the 5 subsampled classes (DCL- η_{Low}) and misspecified class probabilities for the remaining classes (iii) debiased contrastive learning that uses a constant η_{High} providing misspecified class probabilities for the subsampled classes and the true class probability for the remaining 5 classes (DCL- η_{High}), and (iv) debiased contrastive learning with the true sample-specific class probability function for all samples (DCL- η_{True}).

Image Classification We evaluate the quality of image representations with linear classification. We train a linear classifier with the cross-entropy loss and a fixed pretrained image encoder. We use the Adam optimizer with a learning rate of 10^{-3} and weight decay

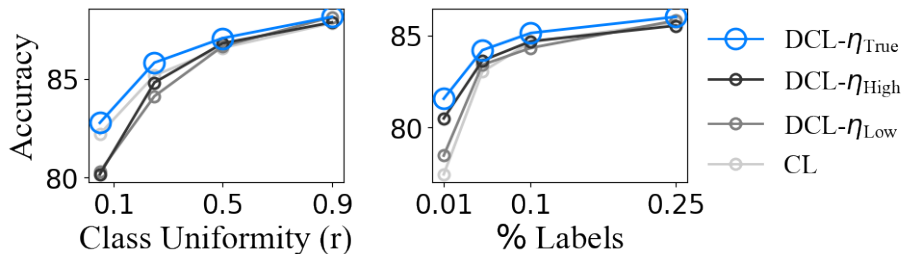


Figure 2: Evaluation using downstream classification task for CIFAR10 experiments. Left plot reports classification accuracy as a function of class distribution uniformity r . Right plot shows classification accuracy as a function of the fraction of images used for training the classifier on the CIFAR10-0.5 data set. Debiased contrastive learning with true class probabilities (DCL- η_{True}) consistently outperforms baseline contrastive learning (CL) and debiased contrastive learning with misspecified class probabilities (DCL- η_{Low} and DCL- η_{High}). The effect is more pronounced when fewer labels are used for training the classifier.

of 10^{-6} . Each model is trained for 100 epochs. We report classification accuracy with varying number of annotated examples available for training the classifier.

Results Figure 2 reports the effect of the choice of sample-specific class probability function η on classification accuracy. When the class distribution is nonuniform (i.e., smaller r), we observe that DCL- η_{True} consistently outperforms CL, DCL- η_{Low} , and DCL- η_{High} . When the class distribution is close to uniform (i.e., r is close to 1), all contrastive loss variants result in similar accuracy. Moreover, the gain from using the true sample-specific class probability is more pronounced when the classifier is trained with fewer labels. Figure 4 provides t-SNE (van der Maaten and Hinton, 2008) visualizations of the representations learned by contrastive and debiased contrastive objectives on CIFAR10-0.1. Using the true sample-specific class probability function η_{True} leads to better class separation, especially for the subsampled classes.

3.2. MIMIC-CXR

We learn image and text encoders for frontal chest X-ray images and associated radiology reports respectively and evaluate the resulting representations in a set of downstream tasks.

Data We use a subset of 234,073 frontal chest X-ray images and reports from MIMIC-CXR (Johnson et al., 2019). We normalize the images and resize them to 512x512 resolution. We apply random image augmentations, i.e., 480x480 random crops, brightness and contrast variations. We use PySBD (Sadvilkar and Neumann, 2020) for sentence tokenization. In all experiments, the data used for representation learning and downstream tasks are disjoint.

Representation Learning We employ ResNet-18 (He et al., 2016) as the image encoder and CXR-BERT (Boecking et al., 2022) as the sentence encoder. Each encoder is followed by a linear mapping to a 128-dimension embedding space. We use LSE+NL (Wang et al.,

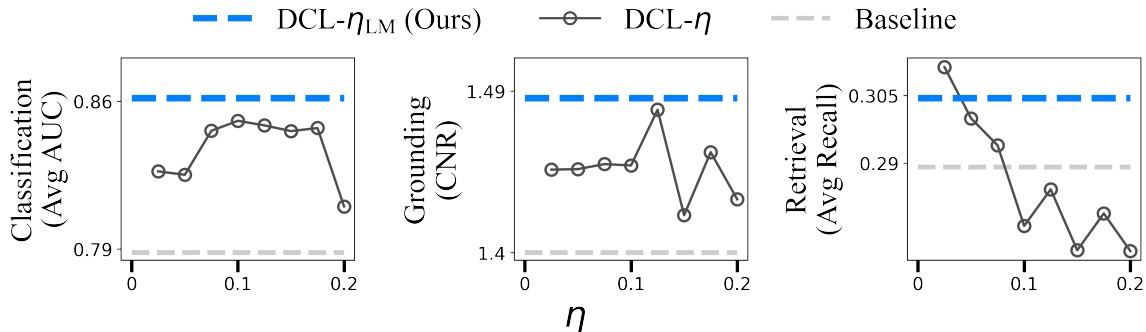


Figure 3: Our method (DCL- η_{LM}) outperforms debiased contrastive learning that applies a fixed amount of correction η to all samples (DCL- η) and LSE+NL that does not correct for the false negatives (Baseline). For each value of η , an image-text model is trained and subsequently evaluated in all three downstream tasks. Our method achieves consistently better performance than alternative approaches.

2023) as the similarity function s . Given a reference text x , we assign x^+ to be the associated image. We set $v_1 = x^+$ to avoid needing additional samples ($M = 1$) and treat all unpaired images in a batch as $\{x_n^-\}$ and $\{u_n\}$. After a grid search, we set $a = 0.2$ and $k = 0.35$. We precompute $p_{LM}(x)$ for all sentences x in the data set using CXR-BERT. Masked language models (e.g., CXR-BERT) cannot estimate sentence probabilities via the chain rule. Instead, we use pseudo-log-likelihood (Salazar et al., 2020) that scores a sentence x by adding the predicted log probabilities of every masked token as $\log p_{LM}(x)$. We use the AdamW optimizer (Loshchilov and Hutter, 2019) and decay the initial learning rate of 10^{-5} using a cosine schedule with 2k warmup steps. We initialize γ to $\sqrt{14}$ and optimize this hyperparameter alongside the encoder parameters. We employ a batch size of 64.

Baseline Methods We compare our method (DCL- η_{LM}) with strong baselines BioViL (Boecking et al., 2022) and LSE+NL (Wang et al., 2023) developed specifically for medical vision-language tasks. BioViL is an image-text model trained using symmetric contrastive learning and masked language modeling objective. LSE+NL is an image-text model that uses log-sum-exp and non-local aggregators for the similarity function s . Neither model corrects for the false negatives explicitly.

We also include in the evaluation several methods that explicitly identify false negatives based on the similarity measure between the positive sample x^+ and negative samples x_n^- , such as the intersection of their CheX5 labels (CheX5 Labels) or similarity of their text embeddings (Text Sim.). Negative samples that are too similar (i.e., above some threshold) are removed (Khosla et al., 2020; Huynh et al., 2022; Zhang et al., 2022a; Zhou et al., 2022). Alternatively, we can reweight the negative samples or even reduce the set of negative samples based on their similarity score (Robinson et al., 2021; Wang et al., 2022). We perform grid searches to select hyperparameters (e.g., the similarity threshold or the resampling size) and select the best model for each setup.

| Method | | Classification | | Grounding | | Retrieval | |
|------------------------------|-----------------|----------------|----------------|----------------|-----------------|-------------------|-------------------|
| | | AUC \uparrow | ACC \uparrow | CNR \uparrow | mIoU \uparrow | Recall \uparrow | MedR \downarrow |
| No Correction | BioViL | 0.78 | 0.62 | 1.14 | 0.17 | 0.25 | 148 |
| No Correction | LSE+NL | 0.79 | 0.65 | 1.40 | 0.19 | 0.29 | 115 |
| Remove | by CheX5 Labels | 0.86 | 0.71 | 1.37 | 0.19 | 0.29 | 113 |
| Resample | by Text Sim. | 0.82 | 0.71 | 1.37 | 0.19 | 0.30 | 111 |
| Remove | by Text Sim. | 0.84 | 0.72 | 1.39 | 0.19 | 0.30 | 112 |
| Reweight | by Text Sim. | 0.84 | 0.69 | 1.40 | 0.19 | 0.29 | 113 |
| DCL- η w/ $\eta = 0.05$ | | 0.80 | 0.72 | 1.46 | 0.19 | 0.30 | 104 |
| DCL- η w/ $\eta = 0.1$ | | 0.85 | 0.72 | 1.45 | 0.19 | 0.29 | 111 |
| DCL- η_{LM} (Ours) | | 0.86 | 0.72 | 1.49 | 0.20 | 0.30 | 104 |

Table 1: Zero-shot performance of the learned representations on downstream image classification, visual grounding, and cross-modal retrieval tasks. Debaised contrastive learning with sample-specific class probability η_{LM} (DCL- η_{LM}) outperforms state-of-the-art baseline methods BioViL and LSE+NL (no false negative correction) and alternative approaches to false negative correction. While methods that correct for false negatives by removing, resampling, or reweighting are effective in improving image classification results for commonly occurring classes, they do not yield comparable improvements in visual grounding or retrieval results.

In addition, we include the original debaised contrastive learning objective that uses a constant η (Chuang et al., 2020).

Downstream Tasks We assess the zero-shot image classification performance of 5 CheXpert labels (Cardiomegaly, Edema, Pleural Effusion, Pneumonia, Pneumothorax) on the MIMIC-CXR data set (Johnson et al., 2019) that we refer to as CheX5. There is roughly 1k images for each binary classification task. We first tokenize and encode class-specific text prompts (e.g., “No signs of pneumonia.” or “Findings suggesting pneumonia.”). Table 2 provides the prompts for each category. For every image, we assign a binary label that corresponds to the prompt with the higher image-sentence score. We report classification accuracy (ACC) and area under the curve (AUC).

We evaluate visual grounding performance using the MS-CXR region-sentence annotations (Boecking et al., 2022). This data set consists of 1,448 bounding boxes over 1,162 images, where each bounding box is associated with a sentence that describes its dominant radiological feature. We compute region-sentence scores to quantify how well the sentence is localized in the image. We report a measure of discrepancy between region-sentence scores inside and outside the bounding box, i.e., contrast-to-noise ratio (CNR) (Boecking et al., 2022), and how well the thresholded region-sentence scores overlap with the bounding box on average, i.e., mean intersection over union (mIoU). We use thresholds that span $[-1, 1]$ in 0.05 increments to compute the mIoU.

We evaluate cross-modal retrieval performance using the MS-CXR data set. To evaluate retrieval, we compute the bounding box features from the region features with RoIAlign

(He et al., 2017). We compute box-sentence scores and sort them to retrieve items in one modality given a query from the other modality. The correctly retrieved item is the one that is paired with the query item. We compute the fraction of times the correct item was found in the top K results (R@K), the median rank of the correct item in the ranked list (MedR), and the average recall over $K = 10, 50, 100$ and over both the image-to-text and the text-to-image retrieval direction (Recall).

Results Figure 3 illustrates performance trade-off for debiased contrastive learning that uses a constant class probability function η . Increasing the value of η improves image classification but harms cross-modal retrieval performance. Conversely, decreasing the value of η enhances cross-modal retrieval performance but harms image classification performance. DCL- η_{LM} provides an overall superior solution on all three tasks.

Table 1 reports the performance of DCL- η_{LM} and competing image-text models. Training LSE+NL with DCL- η_{LM} significantly improves its performance on these tasks compared to BioViL and LSE+NL. Using a sample-specific class probability function η_{LM} is more effective than using a constant function $\eta = 0.05$ and 0.1 . The methods that identify and remove, resample, or reweight the false negatives improve image classification performance but do not offer any gain for visual grounding and minimal improvement for cross-modal retrieval. Tables 3, 4, and 5 in Appendix C provide additional statistics for each task.

4. Discussion

We introduced a novel sample-specific approach that corrects the effect of false negative samples on the contrastive objective. Consistent with prior work (Chuang et al., 2020), this approach offers empirical advantages when the learned representation is used in downstream tasks. In addition, the sample-specific approach to correcting false negatives improves the performance over the original variant of debiased contrastive learning that applied the same correction for all data points.

Our experiments also demonstrate that reducing false negatives for tasks with varying levels of granularity are nuanced. In particular, the methods that attempt to remove, resample, or reweight likely false negative examples can improve image classification performance but offer minimal or no improvement for visual grounding and cross-modality retrieval tasks. We hypothesize that this is because the fine-grain classes that must be also handled for the latter two tasks make identifying false negative samples more error-prone. Similarly, the performance of representations learned via debiased contrastive learning with a fixed correction factor is sensitive to the value of the assumed class probability and varies substantially across the range of possible values. Moreover, the optimal choice of the correction factor varies across downstream tasks, making it challenging to train universally useful representations. To the best of our knowledge, our work is the first to identify the differences in performance of vision-language tasks due to choices in correcting for the effect of false negatives during representation learning. The proposed sample-specific approach applies adaptive correction and produces representations that achieve superior performance.

Using all CheX5 labels to remove false negatives consistently improves the performance of classifying the CheX5 classes. However, doing so has minimal impact on visual grounding and cross-modal retrieval tasks that require the ability to distinguish rare classes beyond those specified in CheX5. Methods that require a clear definition of the latent classes

(Khosla et al., 2020; Dwibedi et al., 2021) are not easily adaptable to situations where there are many classes or when the classes are difficult to define concretely. In contrast, methods that implicitly define the latent classes via clustering (Chen et al., 2022), or do not assume access to the latent classes (Chuang et al. (2020) or our approach) show promise in improving model performance on vision-language tasks.

In our experiments, visual grounding and cross-modal retrieval tasks use the MS-CXR data set that contains infrequently occurring sentences. Thus, one would assume both tasks would benefit more from using smaller values of the assumed uniform class probability η than image classification of commonly occurring pathologies. However, the observed performance improvement for grounding is less noticeable than for retrieval. One reason is that the overlap-based metrics (e.g., IoU) used to evaluate grounding may be less sensitive to the variations in text embeddings representing different but related classes. In contrast, to achieve a high value of ranking-based metrics (e.g., recall) requires the model to rank the correct sentence higher than closely related alternatives. This unexpected result suggests that in addition to the class distribution, target performance metrics that capture how the model will be used in the clinical application are also important when developing representation learning approaches.

Our work emphasizes the need to consider the unique characteristics of medical data that pose challenges for methods designed for simpler scenarios. When working with paired image-text data, we observe that false negatives occur at an uneven rate, making methods designed to address false negatives assuming a uniform class distribution less effective. Furthermore, our work shows the potential of using language models to develop more effective algorithms. Language models are versatile, i.e., capable of processing noisy language inputs, and can serve as useful priors.

Limitations

While the theory presented in this paper is informative, it has limitations. For example, the generalization bounds provided in Appendix A.2 assume a specific similarity function (e.g., dot product) that may not represent the various similarity functions used in practice. Moreover, the generalization bounds only apply to downstream classification tasks while we also evaluate on visual grounding and cross-modal retrieval tasks. While these tasks can be interpreted as some form of nearest neighbor classification, the theoretical results do not trivially extend to these scenarios and more analysis is needed to provide similar generalization guarantees for these tasks.

We use a language model to score text as a proxy for the sample-specific class probability. However, it is unclear how to apply this strategy to data that does not include associated text. While we can estimate data density with certain types of models (e.g., flow-based or autoregressive models), it is not well understood whether these estimates correlate well with the underlying class probabilities. Natural language data is unique in that it is created by humans, capturing important variations in data that align well with the types of problems that users typically wish to solve.

Moreover, it is uncertain how well our assumption that class probability is log-linear with respect to text sequence probability holds in practice. Verifying this assumption requires defining a concrete set of latent classes and annotating text with the defined classes.

However, the latent classes are difficult to define in practice, and we do not have access to the mapping from a data point to its latent class in most clinically important problems as having access to such mapping would eliminate the need for self-supervised learning.

5. Conclusion

We present a debiased contrastive learning framework that accommodates arbitrary class distributions and mitigates the impact of false negatives. We offer theoretical and empirical evidence for using accurate sample-specific class probabilities. To this end, we introduce a specific debiased contrastive objective that employs a language model to estimate the sentence likelihood as a proxy for the class probability. When applied to paired image-text data, our method outperforms strong image-text models on image classification, visual grounding, and cross-modal retrieval tasks.

Acknowledgments

This work was supported by NIH NIBIB NAC P41EB015902, Philips, Wistron, and MIT Lincoln Laboratory. We thank Neel Dey and Nalini Singh for proofreading and providing feedbacks on paper drafts.

References

- Shruthi Bannur, Stephanie Hyland, Qianchu Liu, Fernando Perez-Garcia, Maximilian Ilse, Daniel C. Castro, Benedikt Boecking, Harshita Sharma, Kenza Bouzid, Anja Thieme, Anton Schwaighofer, Maria Wetscherek, Matthew P. Lungren, Aditya Nori, Javier Alvarez-Valle, and Ozan Oktay. Learning to Exploit Temporal Structure for Biomedical Vision-Language Processing. In *CVPR*, January 2023.
- Benedikt Boecking, Naoto Usuyama, Shruthi Bannur, Daniel C. Castro, Anton Schwaighofer, Stephanie Hyland, Maria Wetscherek, Tristan Naumann, Aditya Nori, Javier Alvarez-Valle, Hoifung Poon, and Ozan Oktay. Making the Most of Text Semantics to Improve Biomedical Vision–Language Processing. In *ECCV*, October 2022.
- Geeticka Chauhan, Ruizhi Liao, William Wells, Jacob Andreas, Xin Wang, Seth Berkowitz, Steven Horng, Peter Szolovits, and Polina Golland. Joint Modeling of Chest Radiographs and Radiology Reports for Pulmonary Edema Assessment. In *MICCAI*, October 2020.
- Tsai-Shien Chen, Wei-Chih Hung, Hung-Yu Tseng, Shao-Yi Chien, and Ming-Hsuan Yang. Incremental False Negative Detection for Contrastive Learning. In *ICLR*, April 2022.
- Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. UNITER: UNiversal Image-TEXT Representation Learning. In *ECCV*. August 2020.
- Ching-Yao Chuang, Joshua Robinson, Yen-Chen Lin, Antonio Torralba, and Stefanie Jegelka. Debiased Contrastive Learning. In *NeuIPS*, December 2020.
- Debidatta Dwibedi, Yusuf Aytar, Jonathan Tompson, Pierre Sermanet, and Andrew Zisserman. With a Little Help from My Friends: Nearest-Neighbor Contrastive Learning of Visual Representations. In *ICCV*, October 2021.
- Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *CVPR*, June 2016.
- Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask R-CNN. In *ICCV*, October 2017.
- Shih-Cheng Huang, Liyue Shen, Matthew P. Lungren, and Serena Yeung. GLoRIA: A Multimodal Global-Local Representation Learning Framework for Label-Efficient Medical Image Recognition. In *ICCV*, October 2021.
- Tri Huynh, Simon Kornblith, Matthew R. Walter, Michael Maire, and Maryam Khademi. Boosting Contrastive Self-Supervised Learning with False Negative Cancellation. In *WACV*, January 2022.
- Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silvana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, Jayne Seekins, David A. Mong, Safwan S. Halabi, Jesse K. Sandberg, Ricky Jones, David B. Larson, Curtis P. Langlotz, Bhavik N. Patel, Matthew P. Lungren, and Andrew Y. Ng. CheXpert:

- A Large Chest Radiograph Dataset with Uncertainty Labels and Expert Comparison. In *AAAI*, January 2019.
- 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. MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. *Sci Data*, December 2019.
- Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschiot, Ce Liu, and Dilip Krishnan. Supervised Contrastive Learning. In *NeurIPS*, December 2020.
- Diederik Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. In *ICLR*, December 2014.
- Alex Krizhevsky. Learning Multiple Layers of Features from Tiny Images. 2009.
- Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. VisualBERT: A Simple and Performant Baseline for Vision and Language. *arXiv:1908.03557*, August 2019.
- Ruizhi Liao, Daniel Moyer, Miriam Cha, Keegan Quigley, Seth Berkowitz, Steven Horng, Polina Golland, and William M Wells. Multimodal Representation Learning via Maximization of Local Mutual Information. In *MICCAI*, September 2021.
- Ilya Loshchilov and Frank Hutter. Decoupled Weight Decay Regularization. In *ICLR*, May 2019.
- Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks. In *NeurIPS*, December 2019.
- Philip Müller, Georgios Kaissis, Congyu Zou, and Daniel Rueckert. Joint Learning of Localized Representations from Medical Images and Reports. In *ECCV*, October 2022.
- Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning Transferable Visual Models From Natural Language Supervision. In *ICML*, July 2021.
- Joshua Robinson, Ching-Yao Chuang, Suvrit Sra, and Stefanie Jegelka. Contrastive Learning with Hard Negative Samples. In *ICLR*, January 2021.
- Nipun Sadvilkar and Mark Neumann. PySBD: Pragmatic Sentence Boundary Disambiguation. In *NLP-OSS*, November 2020.
- Julian Salazar, Davis Liang, Toan Q. Nguyen, and Katrin Kirchhoff. Masked Language Model Scoring. In *ACL*, July 2020.

- Nikunj Saunshi, Orestis Plevrakis, Sanjeev Arora, Mikhail Khodak, and Hrishikesh Khandeparkar. A Theoretical Analysis of Contrastive Unsupervised Representation Learning. In *ICML*, May 2019.
- Aäron van den Oord, Yazhe Li, and Oriol Vinyals. Representation Learning with Contrastive Predictive Coding. *arXiv:/1807.03748*, July 2018.
- Laurens van der Maaten and Geoffrey Hinton. Visualizing Data using t-SNE. *Journal of Machine Learning Research*, November 2008.
- Peiqi Wang, William M. Wells, Seth Berkowitz, Steven Horng, and Polina Golland. Using Multiple Instance Learning to Build Multimodal Representations. In *IPMI*, March 2023.
- Yuyang Wang, Rishikesh Magar, Chen Liang, and Amir Barati Farimani. Improving Molecular Contrastive Learning via Faulty Negative Mitigation and Decomposed Fragment Contrast. *Journal of Chemical Information and Modeling*, June 2022.
- Shiwei Zhang, Jichao Sun, Yu Huang, Xueqi Ding, and Yefeng Zheng. Medical Symptom Detection in Intelligent Pre-Consultation Using Bi-directional Hard-Negative Noise Contrastive Estimation. In *SIGKDD*, August 2022a.
- Yuhao Zhang, Hang Jiang, Yasuhide Miura, Christopher D. Manning, and Curtis P. Langlotz. Contrastive Learning of Medical Visual Representations from Paired Images and Text. In *MLHC*, August 2022b.
- Mingkai Zheng, Fei Wang, Shan You, Chen Qian, Changshui Zhang, Xiaogang Wang, and Chang Xu. Weakly Supervised Contrastive Learning. In *ICCV*, October 2021.
- Kun Zhou, Beichen Zhang, Xin Zhao, and Ji-Rong Wen. Debiased Contrastive Learning of Unsupervised Sentence Representations. In *ACL*, May 2022.

Appendix A. Theoretical Results

In this section. We use g_0 to denote the estimator in Equation 8

$$g_0(x, \{u_n\}, \{v_m\}) \triangleq \frac{1}{1 - \eta(x)} \left(\frac{1}{N} \sum_{n=1}^N e^{s(x, u_n)} - \eta(x) \frac{1}{M} \sum_{m=1}^M e^{s(x, v_m)} \right) \quad (12)$$

and $g(x, \{u_n\}, \{v_m\}) = \max(g_0(x, \{u_n\}, \{v_m\}), e^{-\gamma^2})$ as the version of estimator lower bounded by its theoretical minimum.

A.1. Proof for Proposition 1

Proof

The goal is to show how well debiased contrastive loss \mathcal{L} approximates the asymptotic debiased contrastive loss $\bar{\mathcal{L}}$. Without loss of generality, we assume $\gamma = 1$. The proof holds as long as s is bounded. For simplicity, we assume $s(x, x') = f(x)^T f(x')$, implying $s(x) \in [-1, 1]$.

Let $\varepsilon > 0$ and $x, x^+ \in \mathcal{X}$ be arbitrary. We are interested in the tail probability of the difference between the integrands of \mathcal{L} and $\bar{\mathcal{L}}$, i.e., $\mathbb{P}(\Delta \geq \varepsilon)$ where

$$\Delta = \left| -\log \frac{e^{s(x, x^+)}}{e^{s(x, x^+)} + Ng(x, \{u_n\}, \{v_m\})} + \log \frac{e^{s(x, x^+)}}{e^{s(x, x^+)} + N\mathbb{E}_{x^- \sim \mathcal{E}_{c_x}} [e^{s(x, x^-)}]} \right|.$$

Note, Δ implicitly depends on x, x^+ . $\{u_n\}, \{v_m\}$ are random samples. Now simplify:

$$\begin{aligned} \mathbb{P}(\Delta \geq \varepsilon) &= \mathbb{P} \left(\left| \log \left(e^{s(x, x^+)} + Ng(x, \{u_n\}, \{v_m\}) \right) - \log \left(e^{s(x, x^+)} + N\mathbb{E}_{x^- \sim \mathcal{E}_{c_x}} [e^{s(x, x^-)}] \right) \right| \geq \varepsilon \right) \\ &= \mathbb{P} \left(\log \left(e^{s(x, x^+)} + Ng(x, \{u_n\}, \{v_m\}) \right) - \log \left(e^{s(x, x^+)} + N\mathbb{E}_{x^- \sim \mathcal{D}_{c_x}} [e^{s(x, x^-)}] \right) \geq \varepsilon \right) \\ &\quad + \mathbb{P} \left(-\log \left(e^{s(x, x^+)} + Ng(x, \{u_n\}, \{v_m\}) \right) + \log \left(e^{s(x, x^+)} + N\mathbb{E}_{x^- \sim \mathcal{D}_{c_x}} [e^{s(x, x^-)}] \right) \geq \varepsilon \right). \end{aligned}$$

The first term can be bounded, i.e.,

$$\begin{aligned} &\mathbb{P} \left(\log \left(e^{s(x, x^+)} + Ng(x, \{u_n\}, \{v_m\}) \right) - \log \left(e^{s(x, x^+)} + N\mathbb{E}_{x^- \sim \mathcal{E}_{c_x}} [e^{s(x, x^-)}] \right) \geq \varepsilon \right) \\ &= \mathbb{P} \left(\log \frac{e^{s(x, x^+)} + Ng(x, \{u_n\}, \{v_m\})}{e^{s(x, x^+)} + N\mathbb{E}_{x^- \sim \mathcal{E}_{c_x}} [e^{s(x, x^-)}]} \geq \varepsilon \right) \\ &\leq \mathbb{P} \left(\frac{Ng(x, \{u_n\}, \{v_m\}) - N\mathbb{E}_{x^- \sim \mathcal{E}_{c_x}} [e^{s(x, x^-)}]}{e^{s(x, x^+)} + N\mathbb{E}_{x^- \sim \mathcal{E}_{c_x}} [e^{s(x, x^-)}]} \geq \varepsilon \right) \quad (\log x \leq x - 1 \text{ for } x > 0) \\ &= \mathbb{P} \left(g(x, \{u_n\}, \{v_m\}) - \mathbb{E}_{x^- \sim \mathcal{E}_{c_x}} [e^{s(x, x^-)}] \geq \varepsilon \left(\frac{1}{N} e^{s(x, x^+)} + \mathbb{E}_{x^- \sim \mathcal{E}_{c_x}} [e^{s(x, x^-)}] \right) \right) \\ &\leq \mathbb{P} \left(g(x, \{u_n\}, \{v_m\}) - \mathbb{E}_{x^- \sim \mathcal{E}_{c_x}} [e^{s(x, x^-)}] \geq \varepsilon e^{-1} \right) \\ &\quad \left(\frac{1}{N} e^{s(x, x^+)} + \mathbb{E}_{x^- \sim \mathcal{E}_{c_x}} [e^{s(x, x^-)}] \geq \frac{1}{N} e^{-1} + e^{-1} \geq e^{-1} \right) \end{aligned}$$

Similarly, the second term is bounded in a similar manner. Since both term can be bounded,

$$\begin{aligned} \mathbb{P}(\Delta \geq \varepsilon) &\leq \mathbb{P}\left(\left|g(x, \{u_n\}, \{v_m\}) - \mathbb{E}_{x^- \sim \mathcal{E}_{c_x}} \left[e^{s(x, x^-)}\right]\right| \geq \varepsilon e^{-1}\right) \\ &\leq \mathbb{P}\left(\left|g_0(x, \{u_n\}, \{v_m\}) - \mathbb{E}_{x^- \sim \mathcal{E}_{c_x}} \left[e^{s(x, x^-)}\right]\right| \geq \varepsilon e^{-1}\right). \\ &\quad (\text{for any } b \geq e^{-1}, |\max(a, e^{-1}) - b| \leq |a - b|) \end{aligned}$$

We decompose the term inside the absolute value into 3 terms:

$$\begin{aligned} &g_0(x, \{u_n\}, \{v_m\}) - \mathbb{E}_{x^- \sim \mathcal{E}_{c_x}} \left[e^{s(x, x^-)}\right] \\ &= \frac{1}{1 - \eta(x)} \left(\frac{1}{N} \sum_{n=1}^N e^{s(x, u_n)} - \eta(x) \frac{1}{M} \sum_{m=1}^M e^{s(x, v_m)} \right) \\ &\quad - \frac{1}{1 - \rho(c_x)} \left(\mathbb{E}_{x^- \sim \mathcal{D}} \left[e^{s(x, x^-)}\right] - \rho(c_x) \mathbb{E}_{x^- \sim \mathcal{D}_{c_x}} \left[e^{s(x, x^-)}\right] \right) \\ &= \frac{1}{1 - \rho(c_x)} \left(\frac{1}{N} \sum_{n=1}^N e^{s(x, u_n)} - \mathbb{E}_{x^- \sim \mathcal{D}} \left[e^{s(x, x^-)}\right] \right) + \frac{\rho(c_x)}{1 - \rho(c_x)} \left(\frac{1}{M} \sum_{m=1}^M e^{s(x, v_m)} - \mathbb{E}_{x^- \sim \mathcal{D}_{c_x}} \left[e^{s(x, x^-)}\right] \right) \\ &\quad + \left(\frac{1}{1 - \eta(x)} - \frac{1}{1 - \rho(c_x)} \right) \left(\frac{1}{N} \sum_{n=1}^N e^{s(x, u_n)} - \frac{1}{M} \sum_{m=1}^M e^{s(x, v_m)} \right). \end{aligned}$$

Continue where we left off,

$$\begin{aligned} \mathbb{P}(\Delta \geq \varepsilon) &\leq \mathbb{P}\left(\left| \frac{1}{1 - \rho(c_x)} \left(\frac{1}{N} \sum_{n=1}^N e^{s(x, u_n)} - \mathbb{E}_{x^- \sim \mathcal{D}} \left[e^{s(x, x^-)}\right] \right) \right| \geq \frac{\varepsilon e^{-1}}{3}\right) \\ &\quad + \mathbb{P}\left(\left| \frac{\rho(c_x)}{1 - \rho(c_x)} \left(\frac{1}{M} \sum_{m=1}^M e^{s(x, v_m)} - \mathbb{E}_{x^- \sim \mathcal{D}_{c_x}} \left[e^{s(x, x^-)}\right] \right) \right| \geq \frac{\varepsilon e^{-1}}{3}\right) \\ &\quad + \mathbb{P}\left(\left| \left(\frac{1}{1 - \eta(x)} - \frac{1}{1 - \rho(c_x)} \right) \left(\frac{1}{N} \sum_{n=1}^N e^{s(x, u_n)} - \frac{1}{M} \sum_{m=1}^M e^{s(x, v_m)} \right) \right| \geq \frac{\varepsilon e^{-1}}{3}\right). \end{aligned}$$

Hoeffding's Inequality states that given independent bounded random variable Z_1, \dots, Z_n where $Z_i \in [a, b]$ for all i , then $\mathbb{P}(|\frac{1}{n} \sum_{i=1}^n Z_i - \mathbb{E}[Z_i]| \geq t) \leq 2 \exp(-2nt^2/(b-a)^2)$. Since $e^{-1} \leq e^{s(x, x')} \leq e$ for all $x' \in \mathcal{X}$, the first two terms can be bounded by Hoeffding's Inequality

$$\begin{aligned} &\mathbb{P}\left(\left| \frac{1}{1 - \rho(c_x)} \left(\frac{1}{N} \sum_{n=1}^N e^{s(x, u_n)} - \mathbb{E}_{x^- \sim \mathcal{D}} \left[e^{s(x, x^-)}\right] \right) \right| \geq \frac{\varepsilon e^{-1}}{3}\right) \leq 2 \exp\left(-\frac{2N\varepsilon^2}{9e^4} (1 - \rho(c_x))^2\right), \\ &\mathbb{P}\left(\left| \frac{\rho(c_x)}{1 - \rho(c_x)} \left(\frac{1}{M} \sum_{m=1}^M e^{s(x, v_m)} - \mathbb{E}_{x^- \sim \mathcal{D}_{c_x}} \left[e^{s(x, x^-)}\right] \right) \right| \geq \frac{\varepsilon e^{-1}}{3}\right) \leq 2 \exp\left(-\frac{2M\varepsilon^2}{9e^4} \left(\frac{1 - \rho(c_x)}{\rho(c_x)}\right)^2\right). \end{aligned}$$

The last term can be upper bounded with an indicator

$$\begin{aligned}
 & \mathbb{P} \left(\left| \left(\frac{1}{1-\eta(x)} - \frac{1}{1-\rho(c_x)} \right) \left(\frac{1}{N} \sum_{n=1}^N e^{s(x,u_n)} - \frac{1}{M} \sum_{m=1}^M e^{s(x,v_m)} \right) \right| \geq \frac{\varepsilon e^{-1}}{3} \right) \\
 & \leq \mathbb{P} \left(\left| \left(\frac{1}{1-\eta(x)} - \frac{1}{1-\rho(c_x)} \right) (e - e^{-1}) \right| \geq \frac{\varepsilon e^{-1}}{3} \right) \\
 & \leq \mathbb{1} \left\{ \varepsilon \leq 3e^2 \left| \frac{1}{1-\eta(x)} - \frac{1}{1-\rho(c_x)} \right| \right\}.
 \end{aligned}$$

Combine these 3 upper bounds together we have

$$\begin{aligned}
 \mathbb{P}(\Delta \geq \varepsilon) & \leq 2 \exp \left(-\frac{2N\varepsilon^2}{9e^4} (1 - \rho(c_x))^2 \right) + 2 \exp \left(-\frac{2M\varepsilon^2}{9e^4} \left(\frac{1 - \rho(c_x)}{\rho(c_x)} \right)^2 \right) \\
 & \quad + \mathbb{1} \left\{ \varepsilon \leq 3e^2 \left| \frac{1}{1-\eta(x)} - \frac{1}{1-\rho(c_x)} \right| \right\}.
 \end{aligned}$$

Now we bound the approximation error using the tail probability we just derived.

$$\begin{aligned}
 |\mathcal{L} - \bar{\mathcal{L}}| & \leq \mathbb{E}_{(x,x^+) \sim \mathcal{D}_{\text{sim}}} [\Delta] && \text{(Jensen's Inequality)} \\
 & = \mathbb{E}_{(x,x^+) \sim \mathcal{D}_{\text{sim}}} [\mathbb{E}[\Delta \mid x, x^+]] \\
 & = \mathbb{E}_{(x,x^+) \sim \mathcal{D}_{\text{sim}}} \left[\int_0^\infty \mathbb{P}(\Delta \geq \varepsilon \mid x, x^+) d\varepsilon \right] && \text{(Write } \mathbb{E} \text{ using CDF)} \\
 & \leq \mathbb{E}_{x \sim \mathcal{D}} \left[\int_0^\infty 2 \exp \left(-\frac{2N\varepsilon^2}{9e^4} (1 - \rho(c_x))^2 \right) d\varepsilon \right] \\
 & \quad + \mathbb{E}_{x \sim \mathcal{D}} \left[\int_0^\infty 2 \exp \left(-\frac{2M\varepsilon^2}{9e^4} \left(\frac{1 - \rho(c_x)}{\rho(c_x)} \right)^2 \right) d\varepsilon \right] \\
 & \quad + \mathbb{E}_{x \sim \mathcal{D}} \left[\int_0^\infty \mathbb{1} \left\{ \varepsilon \leq 3e^2 \left| \frac{1}{1-\eta(x)} - \frac{1}{1-\rho(c_x)} \right| \right\} d\varepsilon \right] \\
 & = 3e^2 \sqrt{\frac{\pi}{2N}} \mathbb{E}_{x \sim \mathcal{D}} \left[\frac{1}{1 - \rho(c_x)} \right] + 3e^2 \sqrt{\frac{\pi}{2M}} \mathbb{E}_{x \sim \mathcal{D}} \left[\frac{\rho(c_x)}{1 - \rho(c_x)} \right] \\
 & \quad + 3e^2 \mathbb{E}_{x \sim \mathcal{D}} \left[\left| \frac{1}{1-\eta(x)} - \frac{1}{1-\rho(c_x)} \right| \right]. \quad \text{(Use identity } \int_0^\infty e^{-cz^2} dz = \frac{1}{2} \sqrt{\frac{\pi}{c}} \text{)}
 \end{aligned}$$

■

A.2. Classification Generalization Bounds

To show relationship between the different loss functions, we re-define the losses to show their dependence on the encoder f and some additional parameters

$$\bar{\mathcal{L}}(f, Q) \triangleq \mathbb{E}_{(x, x^+) \sim \mathcal{D}_{\text{sim}}} \left[-\log \frac{e^{f(x)^T f(x^+)}}{e^{f(x)^T f(x^+)} + Q \mathbb{E}_{x^- \sim \mathcal{E}_{c_x}} [e^{f(x)^T f(x^-)}]} \right], \quad (13)$$

$$\mathcal{L}(f, N, M) \triangleq \mathbb{E}_{(x, x^+) \sim \mathcal{D}_{\text{sim}}, \{u_n\} \sim \mathcal{D}^N, \{v_m\} \sim \mathcal{D}_{c_x}^M} \left[-\log \frac{e^{f(x)^T f(x^+)}}{e^{f(x)^T f(x^+)} + N g(x, \{u_n\}, \{v_m\})} \right] \quad (14)$$

where the similarity function is dot product, i.e., $s(x, x') = f(x)^T f(x')$.

Following [Saunshi et al. \(2019\)](#); [Chuang et al. \(2020\)](#), we characterize a representation f on K -way classification task \mathcal{T} consisting of K classes $\{c_1, \dots, c_K\} \subset \mathcal{C}$. The supervised dataset is generated from $\mathcal{D}_{\mathcal{T}}(x, c) \propto \mathcal{D}_c(x) \rho(c)$. Specifically, we fix the representations and train a linear classifier $h(x) \triangleq W f(x)$ on task \mathcal{T} with softmax cross entropy loss $\mathcal{L}_{\text{softmax}}(\mathcal{T}, h)$. The supervised loss for classifier h on task \mathcal{T} is

$$\mathcal{L}_{\text{sup}}(\mathcal{T}, h) \triangleq \min_{W \in \mathbb{R}^{K \times D}} \mathcal{L}_{\text{softmax}}(\mathcal{T}, h). \quad (15)$$

We focus on the supervised loss of the mean classifier, where rows of W are the means of the representations of inputs with label c , i.e., $W_\mu \triangleq [\mu_1, \dots, \mu_K]^T$ where $\mu_c = \mathbb{E}_{x \in \mathcal{D}_c} [f(x)]$. The supervised loss for the mean classifier is

$$\mathcal{L}_{\text{sup}-\mu}(\mathcal{T}, f) \triangleq \mathcal{L}_{\text{sup}}(\mathcal{T}, W_\mu f) = \mathbb{E}_{(x, c) \sim \mathcal{D}_{\mathcal{T}}} \left[-\log \frac{e^{f(x)^T \mu_c}}{\sum_{c^- \in \mathcal{T}} e^{f(x)^T \mu_{c^-}}} \right]. \quad (16)$$

We consider the average classification performance over the distribution of K -way multi-class classification tasks $p_{\mathcal{T}}(\{c_i\}_{i=1}^K) \propto \prod_{k=1}^K \rho(c_k) \cdot \mathbb{1}[c_i \neq c_j \forall i \neq j]$. The average supervised loss over tasks is

$$\mathcal{L}_{\text{sup}}(f) \triangleq \mathbb{E}_{\mathcal{T} \sim p_{\mathcal{T}}} \mathcal{L}_{\text{sup}}(\mathcal{T}, f). \quad (17)$$

The following lemma bounds the supervised loss \mathcal{L}_{sup} with the asymptotic contrastive loss \mathcal{L} .

Lemma 2 *For any encoder f , whenever $N \geq \frac{1-\rho_{\min}}{\rho_{\min}}$ where $\rho_{\min} \triangleq \min_{c \in \mathcal{C}} \rho(c)$, we have*

$$\mathcal{L}_{\text{sup}}(f) \leq \mathcal{L}_{\text{sup}-\mu}(f) \leq \bar{\mathcal{L}}(f, N) \quad (18)$$

Proof We first show that $N = \frac{1-\rho_{\min}}{\rho_{\min}}$ gives the smallest loss:

$$\bar{\mathcal{L}}(f, N) = \mathbb{E}_{(x, x^+) \sim \mathcal{D}_{\text{sim}}} \left[-\log \frac{e^{f(x)^T f(x^+)}}{e^{f(x)^T f(x^+)} + N \mathbb{E}_{x^- \sim \mathcal{E}_{c_x}} [e^{f(x)^T f(x^-)}]} \right] \quad (19)$$

$$\geq \mathbb{E}_{(x, x^+) \sim \mathcal{D}_{\text{sim}}} \left[-\log \frac{e^{f(x)^T f(x^+)}}{e^{f(x)^T f(x^+)} + \frac{1-\rho_{\min}}{\rho_{\min}} \mathbb{E}_{x^- \sim \mathcal{E}_{c_x}} [e^{f(x)^T f(x^-)}]} \right] \quad (20)$$

$$= \bar{\mathcal{L}} \left(f, \frac{1-\rho_{\min}}{\rho_{\min}} \right). \quad (21)$$

The task-specific class distribution is $\rho_{\mathcal{T}}(\{c_k\}) \propto \prod_{k=1}^K \rho(c_k)$. Now we show the asymptotic debiased loss upper bounds the supervised classification loss:

$$\bar{\mathcal{L}}\left(f, \frac{1 - \rho_{\min}}{\rho_{\min}}\right) \quad (22)$$

$$= \mathbb{E}_{(x, x^+) \sim \mathcal{D}_{\text{sim}}} \left[-\log \frac{e^{f(x)^T f(x^+)}}{e^{f(x)^T f(x^+)} + \frac{1 - \rho_{\min}}{\rho_{\min}} \mathbb{E}_{x^- \sim \mathcal{E}_{c_x}} [e^{f(x)^T f(x^-)}]} \right] \quad (23)$$

$$= \mathbb{E}_{\mathcal{T} \sim p_{\mathcal{T}}, c^- \sim \rho_{\mathcal{T}}, x, x^+ \sim \mathcal{D}_c^2} \left[-\log \frac{e^{f(x)^T f(x^+)}}{e^{f(x)^T f(x^+)} + \frac{1 - \rho_{\min}}{\rho_{\min}} \mathbb{E}_{\mathcal{T} \sim p_{\mathcal{T}}, c^- \sim \rho_{\mathcal{T}}(\cdot | c^- \neq c), x^- \sim \mathcal{D}_{c^-}} [e^{f(x)^T f(x^-)}]} \right] \quad (24)$$

$$\geq \mathbb{E}_{\mathcal{T} \sim p_{\mathcal{T}}, (x, c) \sim \mathcal{D}_{\mathcal{T}}} \left[-\log \frac{e^{f(x)^T \mu_{c^+}}}{e^{f(x)^T \mu_{c^+}} + \frac{1 - \rho_{\min}}{\rho_{\min}} \mathbb{E}_{c^- \sim \rho_{\mathcal{T}}(\cdot | c^- \neq c)} [e^{f(x)^T \mu_{c^-}}]} \right] \quad (\text{Jensen's Inequality})$$

$$\geq \mathbb{E}_{\mathcal{T} \sim p_{\mathcal{T}}, (x, c) \sim \mathcal{D}_{\mathcal{T}}} \left[-\log \frac{e^{f(x)^T \mu_{c^+}}}{e^{f(x)^T \mu_{c^+}} + \sum_{c^- \in \mathcal{T}: c^- \neq c} e^{f(x)^T \mu_{c^-}}} \right] \quad \left(\frac{1 - \rho_{\min}}{\rho_{\min}} \geq \frac{\sum_{c_k: c_k \neq c} \rho(c_k)}{\rho(c^-)} \text{ for any } c^-, c_1, \dots, c_K \in \mathcal{C}^{K+1} \right)$$

$$= \mathcal{L}_{\text{sup-}\mu}(f). \quad (25)$$

We have showed the second inequality in the Lemma. The first inequality follows from the definition of \mathcal{L}_{sup} , i.e., \mathcal{L}_{sup} is the minimal loss over all weight matrices W including that of the mean classifier W_{μ} . \blacksquare

We wish to derive a data dependent bound. We follow the proof strategy detailed in [Saunshi et al. \(2019\)](#). We assume that the class-specific probability function depends on the representation, i.e., $\eta(x) = \kappa(f(x))$ for some $\kappa: \mathbb{R}^D \rightarrow [\eta_{\min}, \eta_{\max}]$. This assumption is introduced to use theoretical results from prior work; We leave generalization to arbitrary η to future work. First, we express the debiased contrastive loss in an alternative form:

$$\mathcal{L}(f, N, M) = \mathbb{E}_{(x, x^+ \sim \mathcal{D}_{\text{sim}}), \{u_n\} \sim \mathcal{D}^N, \{v_m\} \sim \mathcal{D}_{c_x}^M} \left[\ell(\phi(f(x), f(x^+), \{f(u_n)\}_n, \{f(v_m)\}_n)) \right]. \quad (26)$$

In Equation 26, $\phi: \mathbb{R}^{D(M+N+2)} \rightarrow \mathbb{R}^{M+N+2}$ computes some statistics of the representations

$$\phi(\bar{x}, \bar{x}^+, \{\bar{u}_n\}, \{\bar{v}_m\}) \triangleq (\{\bar{x}^T(\bar{u}_n - \bar{x}^+)\}_n, \{\bar{x}^T(\bar{v}_m - \bar{x}^+)\}_m, \bar{x}^T \bar{x}^+, \kappa(\bar{x})), \quad (27)$$

In Equation 26, $\ell: \mathbb{R}^{N+M+2} \rightarrow \mathbb{R}$ is the loss function

$$\ell(\{a_n\}, \{b_m\}, c, d) \triangleq \log \left[1 + N \max \left(\frac{1}{1-d} \left(\frac{1}{N} \sum_{n=1}^N e^{a_n} - d \frac{1}{M} \sum_{m=1}^M e^{b_m} \right), e^{-1-c} \right) \right]. \quad (28)$$

We can verify that Equation 26 is correct by noting that the integrand in Equation 14 can be expressed using a composition of ℓ and ϕ

$$-\log \left[\frac{e^{f(x)^T f(x^+)}}{e^{f(x)^T f(x^+)} + N g(x, \{u_n\}, \{v_m\})} \right] = \log \left[1 + N \frac{g(x, \{u_n\}, \{v_m\})}{e^{f(x)^T f(x^+)}} \right] \quad (29)$$

$$= \log \left[1 + N \max \left\{ \frac{1}{1 - \kappa(f(x))} \left(\frac{1}{N} \sum_{n=1}^N e^{f(x)^T (f(u_n) - f(x^+))} - \kappa(f(x)) \right) \frac{1}{M} \sum_{m=1}^M e^{f(x)^T (f(v_m) - f(x^+))} \right\} \right] \quad (30)$$

$$\left. , e^{-1 - f(x)^T f(x^+)} \right\} \quad (31)$$

$$= \ell(\phi(f(x), f(x^+), \{f(u_n)\}_n, \{f(v_m)\}_m)). \quad (32)$$

Given a data set $\mathcal{S} \triangleq (x_t, x_t^+, (u_{tn})_n, \{v_{tm}\}_m) \in \mathcal{D}_{\text{sim}} \times \mathcal{D}^N \times \mathcal{D}_{c_{x_t}}^M$ with T samples, the empirical estimate of the debiased contrastive loss in Equation 26 is

$$\hat{\mathcal{L}}(f, N, M) \triangleq \frac{1}{T} \sum_{t=1}^T \ell(\phi(f(x_t), f(x_t^+), \{f(u_{tn})\}_n, \{f(v_{tm})\}_m)). \quad (33)$$

The learning process finds a representation from a function class \mathcal{F} with bounded norm, i.e., $\mathcal{F} \triangleq \{f : \mathcal{X} \rightarrow \mathbb{R}^D \mid \|f(\cdot)\| \leq 1\}$. For example, \mathcal{F} can contain all functions that map to unit hypersphere. Specifically, the algorithm finds the encoder function \hat{f} via empirical risk minimization, i.e., $\hat{f} \in \arg \min_{f \in \mathcal{F}} \hat{\mathcal{L}}(f, N, M)$.

The following Lemma is a more specific version of Lemma A.2 in Saunshi et al. (2019), to account for the differences in our definition of ϕ, ℓ . The goal is to bound the debiased contrastive loss for the risk minimizer \hat{f} .

First we introduce some notations. σ are Rademacher random variables. The restriction of f onto \mathcal{S} is

$$f|_{\mathcal{S}} \triangleq (f_d(x_t), f_d(x_t^+), \{f_d(u_n)\}_n, \{f_d(v_m)\}_m)_{d \in [D], t \in [T]}. \quad (34)$$

$\mathcal{R}_{\mathcal{S}}(\mathcal{F})$ is the empirical Rademacher complexity of the function class \mathcal{F} with respect to the training data \mathcal{S} , i.e.,

$$\mathcal{R}_{\mathcal{S}}(\mathcal{F}) \triangleq \mathbb{E}_{\sigma \sim \{\pm 1\}^{(M+N+2)DT}} \left[\sup_{f \in \mathcal{F}} \langle \sigma, f|_{\mathcal{S}} \rangle \right]. \quad (35)$$

$h : \mathbb{R}^{(M+N+2)D} \rightarrow \mathbb{R}$ as $h \triangleq \ell \circ \phi$ is a function that computes the loss given the embeddings.

Lemma 3 *If h is L_h -Lipschitz and ℓ bounded by B . With probability $1 - \delta$, for all $f \in \mathcal{F}$*

$$\mathcal{L}(\hat{f}, N, M) \leq \mathcal{L}(f, N, M) + \mathcal{O} \left(\frac{L_h \mathcal{R}_{\mathcal{S}}(\mathcal{F})}{T} + B \sqrt{\frac{\log(\frac{1}{\delta})}{T}} \right). \quad (36)$$

Specifically,

$$B = \mathcal{O}(\log N) \quad (37)$$

and $L_h = L_\ell \cdot L_\phi$ where

$$L_\ell = \mathcal{O}\left(\sqrt{\frac{1}{(1-\eta_{\max})^2 N} + \frac{\eta_{\max}^2}{(1-\eta_{\max})^2 M} + \frac{1}{(1-\eta_{\max})^4}}\right), \quad (38)$$

$$L_\phi = \mathcal{O}\left(\sqrt{N + M + \|\nabla \kappa\|_2^2}\right). \quad (39)$$

Proof The first part of the Lemma is simply Lemma A.2 in [Saunshi et al. \(2019\)](#) without computing the Lipschitz constant for their definition of $\phi'(\bar{x}, \bar{x}^+, \{\bar{x}_n^-\}_n) = (\bar{x}^T(\bar{x}_n^- - \bar{x}^+))_n$. We refer reader to [Saunshi et al. \(2019\)](#) for more details.

Note ℓ in Equation 28 is lower bounded by $\log(1) = 0$. Note $a_n, b_m \in [-2, 2]$, $c \in [-1, 1]$, and $d \in [\eta_{\min}, \eta_{\max}]$. By plugging in $a_n = 2$, $b_m = 2$, $c = -1$, and $d = \eta_{\max}$, the upper bound of ℓ is

$$\ell(\{a_n\}, \{b_m\}, c, d) \leq \log\left(1 + N \max\left(\frac{1}{1-\eta_{\max}}(e^2 - \eta_{\max}e^{-2}), 1\right)\right) = \mathcal{O}(\log N). \quad (40)$$

Therefore, $B = \mathcal{O}(\log N)$.

To compute the Lipschitz constant for ℓ , we express ℓ as composition of two functions ψ, ω , i.e., $\ell = \omega \circ \psi$ where $\psi: \mathbb{R}^{N+M+2} \rightarrow \mathbb{R}$ and $\omega: \mathbb{R} \rightarrow \mathbb{R}$ are defined as

$$\omega(z) = \log(1 + Nz), \quad (41)$$

$$\psi(\{a_n\}, \{b_m\}, c, d) = \max\left(\frac{1}{1-d}\left(\frac{1}{N}\sum_{n=1}^N e^{a_n} - d\frac{1}{M}\sum_{m=1}^M e^{b_m}\right), e^{-1-c}\right). \quad (42)$$

Note $z \geq e^{-1-c} \geq e^{-2}$ due to max in Equation 42. Therefore, we can bound the Lipschitz constant for ω with

$$L_\omega \leq \left|\frac{d\omega}{dz}\right| \leq \frac{N}{1 + Ne^{-2}} \leq e^{-2}. \quad (43)$$

To bound the Lipschitz constant for ψ , we compute the Frobenius norm of the Jacobian J_ψ of ψ . We use facts $a_n, b_m \in [-2, 2]$, $c \in [-1, 1]$, and $d \in [\eta_{\min}, \eta_{\max}]$ to derive upper bounds to the absolute value of the partial derivatives:

$$\left|\frac{\partial \psi}{\partial a_n}\right| \leq \frac{e^{a_n}}{(1-d)N} \leq \frac{e^2}{(1-\eta_{\max})N} \quad n \in [N] \quad (44)$$

$$\left|\frac{\partial \psi}{\partial b_m}\right| \leq \frac{d \cdot e^{b_m}}{(1-d)M} \leq \frac{\eta_{\max}e^2}{(1-\eta_{\max})M} \quad m \in [M] \quad (45)$$

$$\left|\frac{\partial \psi}{\partial c}\right| = \frac{1}{(1-d)^2} \left(\frac{1}{N}\sum_{n=1}^N e^{a_n} - \frac{1}{M}\sum_{m=1}^M e^{b_m}\right) \leq \frac{e}{(1-\eta_{\max})^2} \quad (46)$$

$$\left|\frac{\partial \psi}{\partial d}\right| \leq e^{-1-c} \leq 1. \quad (47)$$

We use the Frobenius norm of the Jacobian to bound the Lipschitz constant

$$L_\psi \leq \|J_\psi\|_2 \leq \|J_\psi\|_F \leq \sqrt{\sum_{n=1}^N \frac{e^4}{(1-\eta_{\max})^2 N^2} + \sum_{m=1}^M \frac{\eta_{\max}^2}{(1-\eta_{\max})^2 M^2} + \frac{e^2}{(1-\eta_{\max})^4} + 1}. \quad (48)$$

Combine Lipschitz constants L_ω in Equation 43 and L_ψ in Equation 48, a Lipschitz constant $L_\ell = L_\omega \cdot L_\psi$ for ℓ in Equation 38 follows directly.

To compute a Lipschitz constant for ϕ , we use similar tactic as previous by computing the Jacobian J_ϕ of ϕ . We first expand out the inner products in Equation 27,

$$\phi(\bar{x}, \bar{x}^+, \{\bar{u}_n\}, \{\bar{v}_m\}) = \left(\left\{ \sum_d \bar{x}_d (\bar{u}_{dn} - \bar{x}_d^+) \right\}_n, \left\{ \sum_d \bar{x}_d (\bar{v}_{dm} - \bar{x}_d^+) \right\}_m, \sum_d \bar{x}_d \bar{x}_d^+, \kappa(\bar{x}) \right). \quad (49)$$

The partial derivatives of this vector-valued function ϕ , given $d \in [D]$, are

$$\frac{\partial \phi_i}{\partial \bar{x}_d} = \begin{cases} (\bar{u}_{di} - \bar{x}_d^+) & i \in [N] \\ (\bar{v}_{d(i-N)} - \bar{x}_d^+) & i = N+1, \dots, N+M \\ \bar{x}_d^+ & i = N+M+1 \\ \frac{\partial \kappa}{\partial \bar{x}_d} & i = N+M+2 \end{cases} \quad (50)$$

$$\frac{\partial \phi_i}{\partial \bar{x}_d^+} = \begin{cases} -\bar{x}_d & i \in [N+M] \\ \bar{x}_d & i = N+M+1 \\ 0 & \text{o.w.} \end{cases} \quad (51)$$

$$\frac{\partial \phi_i}{\partial \bar{u}_{dj}} = \begin{cases} \bar{x}_d & i = j, j \in [N] \\ 0 & \text{o.w.} \end{cases} \quad (52)$$

$$\frac{\partial \phi_i}{\partial \bar{v}_{dj}} = \begin{cases} \bar{x}_d & i = j + N, j \in [M] \\ 0 & \text{o.w.} \end{cases}. \quad (53)$$

We use the Frobenius norm of the Jacobian to bound the Lipschitz constant

$$L_\phi \leq \|J_\phi\|_2 \leq \|J_\phi\|_F \quad (54)$$

$$\leq \sqrt{\sum_{i=1}^N \sum_d (\bar{u}_{di} - \bar{x}_d^+)^2 + \sum_{i=N+1}^{N+M} \sum_d (\bar{v}_{d(i-N)} - \bar{x}_d^+)^2 + \sum_d (\bar{x}_d^+)^2 + \sum_d \left(\frac{\partial \kappa}{\partial \bar{x}_d} \right)^2 + (N+M) \sum_d (\bar{x}_d)^2 + \sum_d (\bar{x}_d)^2 + N \sum_d (\bar{x}_d)^2 + M \sum_d (\bar{x}_d)^2} \quad (55)$$

$$\leq \sqrt{4N + 4M + 1 + \|\nabla \kappa\|_2^2 + (N+M) + 1 + N + M} \quad (56)$$

$$= \mathcal{O} \left(\sqrt{N + M + \|\nabla \kappa\|_2^2} \right) \quad (57)$$

Here, norm of representations is bounded, i.e., $\|\bar{x}\|_2, \|\bar{x}^+\|_2, \|\bar{u}_n\|_2, \|\bar{v}_m\|_2 \leq 1$.

We can compute a Lipschitz constant for h by multiplying that for ℓ, ϕ , i.e., $L_h = L_\ell \cdot L_\phi$.

■

The following theorem provides the generalization bounds on classification tasks.

Theorem 4 *With probability at least $1 - \delta$, for all $f \in \mathcal{F}$, and $N \geq \frac{1-\rho_{\min}}{\rho_{\min}}$,*

$$\mathcal{L}_{\text{sup}}(\hat{f}) \leq \mathcal{L}(f, N, M) + \mathcal{O} \left(\sqrt{\frac{1}{M}} + \sqrt{\frac{1}{N}} + \mathbb{E}_{x \sim \mathcal{D}} \left[\left| \frac{1}{1 - \eta(x)} - \frac{1}{1 - \rho(c_x)} \right| \right] + \frac{L_h \mathcal{R}_S(\mathcal{F})}{T} + B \sqrt{\frac{\log(\frac{1}{\delta})}{T}} \right) \quad (58)$$

Proof Simply combine the results that we proved so far gives us the result

$$\begin{aligned} \mathcal{L}_{\text{sup}}(\hat{f}) &\leq \bar{\mathcal{L}}(\hat{f}, N) && \text{(Use Lemma 2)} \\ &\leq \mathcal{L}(\hat{f}, N, M) + \mathcal{O} \left(\sqrt{\frac{1}{M}} + \sqrt{\frac{1}{N}} + \mathbb{E}_{x \sim \mathcal{D}} \left[\left| \frac{1}{1 - \eta(x)} - \frac{1}{1 - \rho(c_x)} \right| \right] \right) && \text{(Use Proposition 1)} \\ &\leq \mathcal{L}(f, N, M) + \mathcal{O} \left(\sqrt{\frac{1}{M}} + \sqrt{\frac{1}{N}} + \mathbb{E}_{x \sim \mathcal{D}} \left[\left| \frac{1}{1 - \eta(x)} - \frac{1}{1 - \rho(c_x)} \right| \right] + \frac{L_h \mathcal{R}_S(\mathcal{F})}{T} + B \sqrt{\frac{\log(\frac{1}{\delta})}{T}} \right). && \text{(Use Lemma 3)} \end{aligned}$$

■

This bound states that if (i) the function class \mathcal{F} is sufficiently rich (i.e., contains some encoder function for which $\mathcal{L}(f, N, M)$ is small), (ii) trained over large data (i.e., T is large), (iii) the number of negatives sampled is large (i.e., large N, M), and (iv) η provides a good estimate of sample-specific class probabilities, then the encoder \hat{f} will perform well on the downstream classification tasks.

Appendix B. Additional Details for Experiments

| Label | Negative Prompt | Positive Prompt |
|--------------|-------------------------|---------------------------------|
| pneumonia | “No signs of pneumonia” | “Findings suggesting pneumonia” |
| cardiomegaly | “Heart size normal” | “Cardiomegaly” |
| edema | “No pulmonary edema” | “Pulmonary edema” |
| effusion | “No pleural effusions” | “Pleural effusions” |
| pneumothorax | “No pneumothorax” | “Pneumothorax” |

Table 2: Zero-shot image classification prompts.

Appendix C. Additional Results

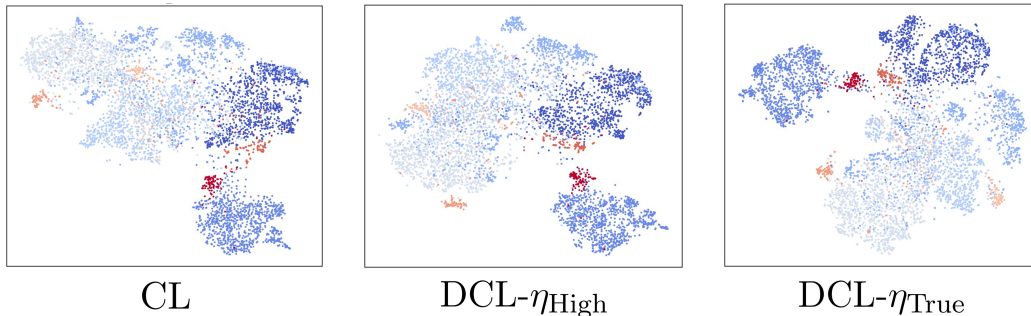


Figure 4: t-SNE visualization of learned representations on CIFAR10-0.1. Subsampled classes are in shades of red and the remaining classes are in shades of blue. Using the true sample-specific class probability function ($DCL-\eta_{True}$) leads to better class separation than both contrastive learning (CL) and using the incorrect sample-specific class probability function ($DCL-\eta_{High}$), especially for the subsampled classes.

| Method | Cardiomegaly | Edema | Effusion | Pneumonia | Pneumothorax | Avg |
|--------------------|--------------|--------------|--------------|--------------|--------------|--------------|
| BioViL | 0.752 | 0.887 | 0.914 | 0.768 | 0.594 | 0.783 |
| LSE+NL | 0.888 | 0.876 | 0.798 | 0.853 | 0.526 | 0.788 |
| w/ $DCL-\eta_{LM}$ | 0.884 | 0.889 | 0.913 | 0.861 | 0.760 | 0.862 |

Table 3: Zero-shot image classification performance. We report the AUC for 5 common radiology findings and their average (Avg) on CheX5 data set. Image-text model trained with our method outperforms state-of-the-art baselines.

| Method | CNR \uparrow | mIoU \uparrow |
|--------------------|----------------|-----------------|
| BioViL | 1.142 | 0.174 |
| LSE+NL | 1.400 | 0.190 |
| w/ $DCL-\eta_{LM}$ | 1.486 | 0.195 |

Table 4: Visual grounding performance. We report a measure of the discrepancy between region-sentence scores inside and outside the ground truth bounding box (contrast-to-noise ratio or CNR) and the mean IoU of a thresholded region-sentence map and the ground truth bounding box over a set of threshold (mIoU). Image-text model trained with our method outperforms BioViL and LSE+NL on both measures.

| Method | Image \rightarrow Text | | | | Text \rightarrow Image | | | |
|---------------------|--------------------------|-----------------|------------------|-------------------|--------------------------|-----------------|------------------|-------------------|
| | R@10 \uparrow | R@50 \uparrow | R@100 \uparrow | MedR \downarrow | R@10 \uparrow | R@50 \uparrow | R@100 \uparrow | MedR \downarrow |
| BioViL | 0.07 | 0.26 | 0.40 | 151 | 0.08 | 0.26 | 0.40 | 146 |
| LSE+NL | 0.09 | 0.29 | 0.44 | 123 | 0.10 | 0.32 | 0.49 | 107 |
| w/ DCL- η_{LM} | 0.10 | 0.31 | 0.48 | 106 | 0.10 | 0.33 | 0.50 | 102 |

Table 5: Cross-modal retrieval performance. We report recall for the top 10, 50 and 100 answers returned by the method (R@k), as well as the median rank of the ground truth element for sentence retrieval based on region queries and for region retrieval based on sentence queries (MedR). Our method improves the baselines.

```

# pos: exponential for positive example
# neg: sum of exponentials for negative examples
# N : number of negative examples
# eta : fixed class probability
# p: likelihood of the text scored by a language model
# a, k: hyperparameter that maps p to sample-specific class probability

standard_loss = -log(pos / (pos + neg))
dcl_loss = -log(pos / (pos + (neg - N * eta * pos) / (1 - eta)))
eta_LM = a * p**k
dcl_LM_loss = -log(pos / (pos + (neg - N * eta_LM * pos) / (1 - eta_LM)))

```

Figure 5: Pseudocode for our proposed sample-specific class probability estimate for de-biased contrastive learning with $M = 1$. The likelihood of the text scored by a language model can be precomputed and yield negligible computational overhead. Implementing our proposed changes requires minimal modification to code.