Supervised Contrastive Learning with Hard Negative Samples

1st Ruijie Jiang
Dept. of ECE
Tufts University
Medford, USA
Ruijie.Jiang@tufts.edu

1st Thuan Nguyen
Dept. of CS
Tufts University
Medford, USA
Nguyen.Thuan@tufts.edu

2nd Prakash Ishwar
Dept. of ECE
Boston University
Boston, USA
pi@bu.edu

3rd Shuchin Aeron
Dept. of ECE
Tufts University
Medford, USA
shuchin@ece.tufts.edu

***Abstract*—Through minimization of an appropriate loss function such as the InfoNCE loss, contrastive learning (CL) learns a useful representation function by pulling positive samples close to each other while pushing negative samples far apart in the embedding space. The positive samples are typically created using "label-preserving" augmentations, i.e., domain-specific transformations of a given datum or anchor. In absence of class information, in unsupervised CL (UCL), the negative samples are typically chosen randomly and independently of the anchor from a preset negative sampling distribution over the entire dataset. This leads to class-collisions in UCL. Supervised CL (SCL), avoids this class collision by conditioning the negative sampling distribution to samples having labels different from that of the anchor. In hard-UCL (H-UCL), which has been shown to be an effective method to further enhance UCL, the negative sampling distribution is conditionally *tilted*, by means of a *hardening function*, towards samples that are closer to the anchor. Motivated by this, in this paper we propose hard-SCL (H-SCL) wherein the class conditional negative sampling distribution is tilted via a hardening function. Our simulation results confirm the utility of H-SCL over SCL with significant performance gains in downstream classification tasks. Analytically, we show that in the limit of infinite negative samples per anchor and a suitable assumption, the H-SCL loss is upper bounded by the H-UCL loss, thereby justifying the utility of H-UCL for controlling the H-SCL loss in the absence of label information. Through experiments on several datasets, we verify the assumption as well as the claimed inequality between H-UCL and H-SCL losses. We also provide a plausible scenario where H-SCL loss is lower bounded by UCL loss, indicating the limited utility of UCL in controlling the H-SCL loss.** [1]

***Index Terms*—contrastive representation learning, hard negative sampling**

I. Introduction

Contrastive representation learning (CL) has received considerable attention in the machine learning literature as a method to learn representations of data for use in downstream inference tasks, both in the absence of class information via unsupervised CL (UCL) [1], [2], as well as with known class labels via supervised CL (SCL) [3]. Contrastive learning has impacted a number of applications ranging from image classification [4]–[6], text classification [7], [8], and natural language processing [9], [10], to learning and inference with time-series data [11], [12]. Contrastive learning methods learn a representation map that pulls positive samples together while pushing the negative samples apart in the representation space by minimizing a suitable loss such as the widely-used InfoNCE loss [13]. Given an anchor datum, the positive samples are often constructed by applying domain-specific augmentations or transformations that are highly likely to preserve the latent label [5]. For example, crop, blur, rotation, and occlusion transformations for image data, and word masking for natural language processing (NLP) data. For a given augmentation mechanism, the performance of CL highly depends on the choice of the negative sampling mechanism that provides adequate *contrast* with the given anchor. In UCL, the negative samples are typically chosen randomly and independently of the anchor from a preset negative sampling distribution over the entire dataset. This leads to class-collisions in UCL. Supervised CL (SCL), avoids this class collision via conditioning the negative sampling distribution on the label of the anchor. In hard-UCL (H-UCL), which has been empirically shown to be an effective method for further enhancing the effectiveness of UCL on downstream inference tasks [14]–[16], the negative sampling distribution is conditionally *tilted*, by means of a *hardening function*, towards samples that are closer to the anchor [17]. Motivated by the success of H-UCL, in this paper we propose hard-negative sampling for SCL, where we tilt the class conditional negative sampling distribution via a hardening function similarly to H-UCL. To the best of our knowledge, this work is the first one that jointly combines the label information and the hard-negative sampling strategies to improve downstream performance. We make the following main contributions:

1) Via extensive numerical comparisons on standard datasets we show that downstream performance of H-SCL is significantly higher compared to SCL. A preview of the results is shown in Fig. 1 and more results are provided in Sec. V.
2) In Sec. IV, for a general class of hardening functions recently introduced in [17] for H-SCL and H-UCL, in the limit of negative samples going to infinity and under a suitable assumption, in Lemma 1 we show that the H-SCL loss is upper bounded by the H-UCL loss. Since in our experiments H-SCL outperforms SCL we posit that this result takes a step towards theoretically justifying

[1] Our code is publicly available at https://github.com/rjiang03/H-SCL.

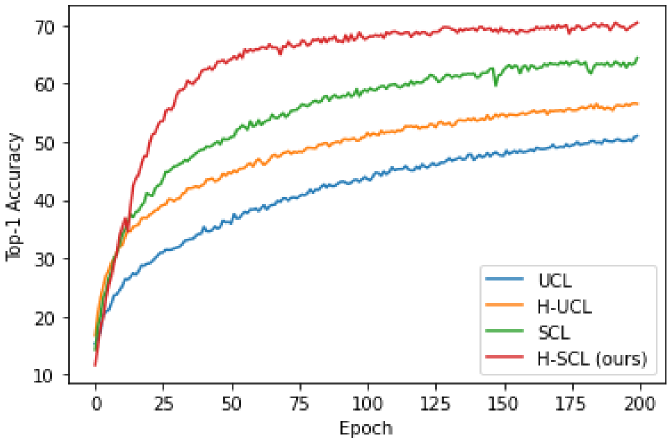

Fig. 1. Top-1 accuracy (in %) of UCL, H-UCL, SCL, and H-SCL on the CIFAR100 dataset.

the utility of H-UCL over UCL, addressing a question left open in [18].

3) In Sec. V, we conduct experiments to numerically verify to what extent the assumption needed for the main theoretical result is satisfied. We also numerically verify Lemma 1.

The rest of the paper is organized as follows. In Sec. II, we discuss related work within the context of hard negative sampling for contrastive learning. Section III outlines the mathematical set-up for the UCL, SCL, H-UCL, and H-SCL scenarios. In Sec. IV we state and prove the main theoretical result of this paper characterizing the relationship between the losses under the H-UCL and H-SCL settings along with the assumptions needed. Finally, we provide numerical results in Sec. V and conclude in Sec. VI.

II. Related work

While design of positive sampling is also important in contrastive learning [19], in this work we focus mainly on negative sampling. For negative sample selection, recent works focus on designing "hard" negative samples, *i.e.,* negative samples coming from a different classes than the anchor, but close to the anchor. Robinson *et al.* [14] derive a simple but practical hard-negative sampling strategy that improves the downstream task performance on image, graph, and text data. Tabassum *et al.* [15] introduce an algorithm called UnReMix which takes into account both the anchor similarity and the model uncertainty to select hard negative samples. Kalantidis *et al.* [20] propose a method called "hard negative mixing" which synthesizes hard negative samples directly in the embedding space to improve the downstream-task performances. Although many studies observed that H-UCL outperforms UCL, there is no theoretical justification for this observation. Specifically, Wu *et al.* [18] observe that compared to the UCL loss, the H-UCL loss is, indeed, a looser lower bound of mutual information between two random variables derived from the dataset and raise the question "why is a looser bound ever more useful" in practice? Our main theoretical result, Lemma 1 takes a first step towards answering this question. A recent work [17] analyzes and establishes that the H-UCL loss is lower bounded by the UCL loss and the H-SCL loss[2] is lower bounded by the SCL loss for general hard negative sampling strategies. In this paper we extend the results there and relate the H-UCL and H-SCL losses under a suitable technical assumption.

III. Problem Formulation

Notation and preliminaries: We denote the *sample space* (or *input space*) by $\mathcal{X}$, the *label space* by $\mathcal{Y}$, the unit-sphere in $\mathbb{R}^d$ by S^{d-1}, and the indicator function of an event $\mathcal{E}$ by $1(\mathcal{E})$. For integers i, j with $i < j$, we define $i : j := i, i+1, \ldots, j$ and $a_{i:j} := a_i, a_{i+1}, \ldots, a_j$. If $i > j$, $i : j$ and $a_{i:j}$ are "null". For any $f : \mathcal{X} \to S^{d-1}$ we define $g(x, x') := f^\top(x) f(x')/\gamma$, with $\gamma \in (0, \infty)$ a "temperature parameter". We measure the "similarity" of $u, v \in S^{d-1}$ by their inner product, with larger inner products corresponding to greater similarity. Since $||u - v||^2 = 2(1 - u^\top v)$ for all $u, v \in S^{d-1}$, it follows that u and v are more similar if, and only if, they are more closer. For future reference, we note the following useful result:

Proposition 1. *Let p be a probability distribution over $\mathcal{Z}$ and $\rho : \mathcal{Z} \to [0, \infty)$ a nonnegative function such that $\alpha := \mathbb{E}_{z \sim p}[\rho(z)] \in (0, \infty)$. Then,*

$$r(z) := \frac{\rho(z) p(z)}{\alpha}$$

is also a probability distribution and for any measurable function $s : \mathcal{Z} \to \mathbb{R}$ we have

$$\mathbb{E}_{z \sim r}[s(z)] = \frac{\mathbb{E}_{z \sim p}[\rho(z) s(z)]}{\alpha}$$

Proof. Firstly, r is nonnegative since ρ, p, and α are nonnegative. Since α is nonnegative, nonzero, and finite, we have

$$\int_{\mathcal{Z}} r(z) dz = \int_{\mathcal{Z}} \frac{\rho(z)}{\alpha} p(z) dz = \frac{\mathbb{E}_{z \sim p}[\rho(z)]}{\alpha} = 1.$$

Therefore,

$$\mathbb{E}_{z \sim r}[s(z)] = \int_{\mathcal{Z}} s(z) r(z) dz = \int_{\mathcal{Z}} s(z) \frac{\rho(z)}{\alpha} p(z) dz \\ = \frac{\mathbb{E}_{z \sim p}[\rho(z) s(z)]}{\alpha}.$$

□

A. Contrastive learning

Contrastive learning assumes access to pairs of *similar* samples $(x, x^+) \sim p_{\text{sim}}(x, x^+)$, with x referred to as the *anchor* and x^+ as the *positive sample*, and k *negative samples* $x^-_{1:k}$ that are conditionally independent and identically distributed (iid) given anchor x, with distribution $q_{\text{CL}}(x^- | x)$, and are presumably not similar (in a representation space) to the anchor x.

Let $y : \mathcal{X} \to \mathcal{Y}$ be a deterministic labeling function mapping the input space to a label space. For a sample $x \in \mathcal{X}$, $y(x)$ is the (groundtruth) label assigned to x. The labels are available

[2] Note that [17] cites an earlier draft of this paper for H-SCL.

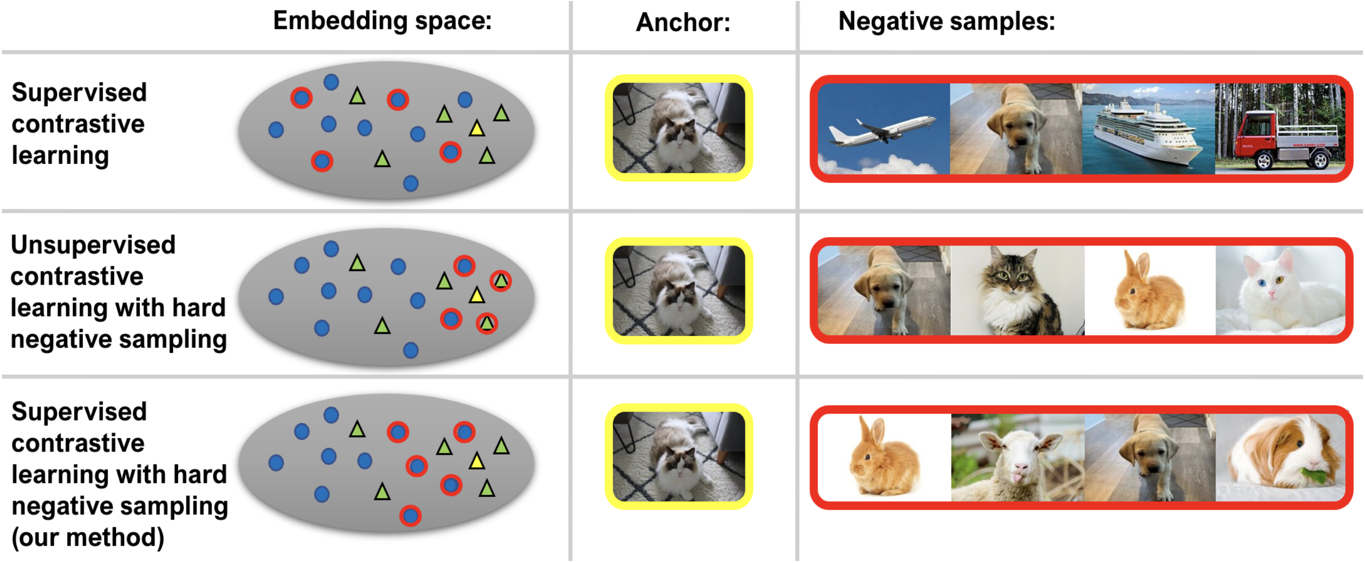
Embedding space:
Anchor:
Negative samples:
Supervised contrastive learning
Unsupervised contrastive learning with hard negative sampling
Supervised contrastive learning with hard negative sampling (our method)

Fig. 2. Schematic illustration of negative sampling strategies under H-UCL, SCL, and H-SCL settings in classifying species of cat. Top row (SCL): the negative samples (red rings) are randomly sampled from the set of circle samples which belongs to different classes of the anchor (yellow triangle). Middle row (H-UCL): the negative samples (red rings) are only selected from the neighbors of the anchor (yellow triangle). Since H-UCL prefers samples that are close to the anchor, it may select false negative samples (green triangles) which come from the same class as the anchor. Bottom row (H-SCL): the negative samples (red rings) are selected such that they are not only the "true negative" samples (circle samples) but also are close to the anchor (yellow triangle).

during training in the supervised settings (SCL, H-SCL), but unknown in the unsupervised settings (UCL, H-UCL).

The goal is to learn a representation function $f : \mathcal{X} \to S^{d-1}$, mapping the input space to the latent space of unit-norm vectors in $\mathbb{R}^d$, that minimizes a contrastive loss function

$$\mathcal{L}^{(k)}_{\text{CL}}(f) := \mathbb{E}_{(x,x^+)\sim p_{\text{sim}}}\left[\mathbb{E}_{x^-_{1:k}\sim \text{ iid } q_{\text{CL}}}\left[\psi_k(x,x^+,x^-_{1:k},f)\right]\right]$$

over some family of representation functions $\mathcal{F}$, e.g., all deep neural networks with a specified architecture. In this work we only consider the widely used InfoNCE contrastive loss function [13] given by

$$\psi_k(x,x^+,x^-_{1:k},f) = \log\left(1 + e^{-g(x,x^+)}\frac{1}{k}\sum_{j=1}^{k} e^{g(x,x^-_j)}\right).$$

We will assume that for all $f \in \mathcal{F}$ and all $x \in \mathcal{X}$, we have $\mathbb{E}_{x^-\sim q_{\text{CL}}(x^-|x)}[e^{g(x,x^-)}] < \infty$. Then, in the limit as $k \to \infty$, by the strong law of large numbers,

$$\psi_k(x,x^+,x^-_{1:k},f) \xrightarrow[k\to\infty]{a.s.} \psi_\infty(x,x^+,f,q_{\text{CL}})$$

where

$$\psi_\infty(x,x^+,f,q_{\text{CL}}) := \log\left(1 + e^{-g(x,x^+)}\mathbb{E}_{x^-\sim q_{\text{CL}}}\left[e^{g(x,x^-)}\right]\right).$$

Since all representation vectors have unit-norm, $|g(x,x^+)| = |f^\top(x)f(x^+)|/\gamma \le 1/\gamma < \infty$. This implies that both $\psi_k(x,x^+,x^-_{1:k},f)$ and $\psi_\infty(x,x^+,f,q_{\text{CL}})$ are globally bounded functions. From the dominated convergence theorem it follows that

$$\mathcal{L}^{(k)}_{\text{CL}}(f) \xrightarrow{k\to\infty} \mathcal{L}^{(\infty)}_{\text{CL}}(f)$$

where

$$\mathcal{L}^{(\infty)}_{\text{CL}}(f) := \mathbb{E}_{(x,x^+)\sim p_{\text{sim}}}\left[\psi_\infty(x,x^+,f,q_{\text{CL}})\right]$$

B. UCL and SCL settings

Let (x,x^+) be drawn from a joint distribution p_{sim}. For a given (v,v^+), the main difference between UCL, SCL, and H-UCL settings arises from the negative sampling distribution $q_{\text{CL}}(x^-|x)$ and the sampling strategy.

1) In the UCL setting, for all $x, x^+ \in \mathcal{X}$,

$$q_{\text{CL}}(x^-|x) = q_{\text{UCL}}(x^-),$$

for some probability distribution $q_{\text{UCL}}(x^-)$ over $\mathcal{X}$. Thus in the UCL setting, the negative samples are selected independently of the anchor and positive samples.

2) In the SCL setting, for the given labeling function $y(\cdot)$ and all $x, x^- \in \mathcal{X}$,

$$q_{\text{CL}}(x^-|x) = q_{\text{SCL}}(x^-|x)$$

where

$$q_{\text{SCL}}(x^-|x) := \frac{1(y(x^-) \neq y(x))\, q_{\text{UCL}}(x^-)}{\alpha_{\text{SCL}}(x)},$$
$$\alpha_{\text{SCL}}(x) := \mathbb{E}_{x^-\sim q_{\text{UCL}}}\left[1(y(x^-) \neq y(x))\right]$$

and we assume that $\alpha_{\text{SCL}} > 0$ (this would be true if for all classes y, we have $\mathbb{E}_{x^-\sim q_{\text{UCL}}}\left[1(y(x^-) = y\right] > 0$). To enable comparison of the UCL and SCL settings, we assume that the q_{UCL} distribution used in the SCL setting is identical to the negative sampling distribution used in the UCL setting. Thus, the distribution of negative samples in the SCL setting is q_{UCL} conditioned on the event that the negative samples have labels *different* from that of the anchor.

C. H-UCL and H-SCL settings

We consider the very general class of negative sample hardening mechanisms introduced in [17] which is based on a ***hardening function*** defined as follows.

Definition 1 (Hardening function)**.** *[17] $\eta : \mathbb{R} \to \mathbb{R}$ is a hardening function if it is non-negative and nondecreasing throughout $\mathbb{R}$.*

Examples of hardening functions include the exponential tilting hardening function $\eta_{\exp}(t) := e^{\beta t}$, $\beta > 0$, employed in [14], [16] and $\eta_{\text{thresh}}(t) := 1(e^t \geq \tau)$ for some threshold τ.

1) In the H-UCL setting, for all $x, x^- \in \mathcal{X}$, all $f \in \mathcal{F}$, and a given hardening function $\eta(\cdot)$,

$$q_{\text{CL}}(x^-|x) = q_{\text{H-UCL}}(x^-|x, f)$$

where,

$$q_{\text{H-UCL}}(x^-|x, f) := \frac{\eta(g(x, x^-))\, q_{\text{UCL}}(x^-)}{\alpha_{\text{H-UCL}}(x, f)},$$
$$\alpha_{\text{H-UCL}}(x, f) := \mathbb{E}_{x^- \sim q_{\text{UCL}}}\left[\eta(g(x, x^-))\right],$$

and we assume that $\alpha_{\text{H-UCL}}(x, f) \in (0, \infty)$ for all $x \in \mathcal{X}$ and all $f \in \mathcal{F}$. The hardening function is nondecreasing. Thus negative samples x^- that are more similar to the anchor x in the representation space, i.e., $g(x, x^-)$ is large, are more likely to be sampled under $q_{\text{H-UCL}}$ than q_{UCL}.

2) In the H-SCL setting we utilize both hard-negative sampling and label information. This is the first key contribution of this paper. This is motivated by the effectiveness of hard-negative sampling strategies in H-UCL and the usefulness of label information in SCL. The main difference between H-SCL and other contrastive learning methods (UCL, H-UCL, and SCL) comes from the way the negative samples are selected.

Formally, in H-SCL, the positive pair (x, x^+) is first sampled from p_{sim}, i.e., using the same sampling strategy as in UCL, SCL, and H-UCL. Next, for the given labeling function $y(\cdot)$, all $x, x^+ \in \mathcal{X}$, all $f \in \mathcal{F}$, and a given hardening function $\eta(\cdot)$,

$$q_{\text{CL}}(x^-|x) = q_{\text{H-SCL}}(x^-|x, f)$$

where,

$$q_{\text{H-SCL}}(x^-|x, f)$$
$$:= \frac{1(y(x^-) \neq y(x))\, \eta(g(x, x^-))\, q_{\text{UCL}}(x^-)}{\alpha_{\text{H-SCL}}(x, f)}$$
$$\alpha_{\text{H-SCL}}(x, f)$$
$$:= \mathbb{E}_{x^- \sim q_{\text{UCL}}}\left[1(y(x^-) \neq y(x))\, \eta(g(x, x^-))\right],$$

and we assume that $\alpha_{\text{H-SCL}}(x, f) \in (0, \infty)$ for all $x \in \mathcal{X}$ and all $f \in \mathcal{F}$.

In other words, in the H-SCL setting, we only select a negative sample x^- which simultaneously satisfies two conditions:

(i) x^-'s label is different from x, i.e., $y(x^-) \neq y(x)$, and

(ii) x^- is hard to discern from x.

The top, middle, and bottom rows of Fig. 2 illustrate the negative sampling strategies in SCL, H-UCL, and H-SCL, respectively. In Sec. V, we will numerically demonstrate the advantages of H-SCL compared to SCL, UCL, and H-UCL.

The contrastive losses in the UCL, SCL, H-UCL, and H-SCL settings are given by

$$\mathcal{L}^{(k)}_{\text{UCL}}(f) := \mathbb{E}_{(x,x^+) \sim p_{\text{sim}}}\left[\mathbb{E}_{x^-_{1:k} \sim \text{ iid } q_{\text{UCL}}}\left[\psi_k(x, x^+, x^-_{1:k}, f)\right]\right]$$
$$\mathcal{L}^{(k)}_{\text{SCL}}(f) := \mathbb{E}_{(x,x^+) \sim p_{\text{sim}}}\left[\mathbb{E}_{x^-_{1:k} \sim \text{ iid } q_{\text{SCL}}}\left[\psi_k(x, x^+, x^-_{1:k}, f)\right]\right]$$
$$\mathcal{L}^{(k)}_{\text{H-UCL}}(f) := \mathbb{E}_{(x,x^+) \sim p_{\text{sim}}}\left[\mathbb{E}_{x^-_{1:k} \sim \text{ iid } q_{\text{H-UCL}}}\left[\psi_k(x, x^+, x^-_{1:k}, f)\right]\right]$$
$$\mathcal{L}^{(k)}_{\text{H-SCL}}(f) := \mathbb{E}_{(x,x^+) \sim p_{\text{sim}}}\left[\mathbb{E}_{x^-_{1:k} \sim \text{ iid } q_{\text{H-SCL}}}\left[\psi_k(x, x^+, x^-_{1:k}, f)\right]\right]$$

respectively and their limits as $k \to \infty$ by $\mathcal{L}^{(\infty)}_{\text{UCL}}(f), \mathcal{L}^{(\infty)}_{\text{SCL}}(f), \mathcal{L}^{(\infty)}_{\text{H-UCL}}(f)$, and $\mathcal{L}^{(\infty)}_{\text{H-SCL}}(f)$ respectively.

IV. Connection between H-SCL and H-UCL losses

There is, in general, no known simple relationship between $\mathcal{L}_{\text{H-SCL}}$ and $\mathcal{L}_{\text{UCL}}$. In this section, however, we will show that under certain technical conditions $\mathcal{L}^{(\infty)}_{\text{H-SCL}} \leq \mathcal{L}^{(\infty)}_{\text{H-UCL}}$. This implies that when k is large, minimizing $\mathcal{L}_{\text{H-UCL}}$ can act as a proxy for minimizing $\mathcal{L}_{\text{H-SCL}}$ whereas minimizing $\mathcal{L}_{\text{UCL}}$ cannot in general. As we will demonstrate in Sec. V, H-SCL empirically outperforms other contrastive learning methods. Thus our theoretical results provide a plausible explanation for why H-UCL outperforms UCL in practice and partially answer an open question in [18].

For the given labeling function $y(\cdot)$, all $x, x^+ \in \mathcal{X}$, all $f \in \mathcal{F}$, and a given hardening function $\eta(\cdot)$, let

$$q_{\text{Hcol}}(x^-|x, f) := \frac{1(y(x^-) = y(x))\, \eta(g(x, x^-))\, q_{\text{UCL}}(x^-)}{\alpha_{\text{Hcol}}(x, f)}$$
$$\alpha_{\text{Hcol}}(x, f) := \mathbb{E}_{x^- \sim q_{\text{UCL}}}\left[1(y(x^-) = y(x))\, \eta(g(x, x^-))\right],$$

and assume that $\alpha_{\text{Hcol}}(x, f) \in (0, \infty)$ for all $x \in \mathcal{X}$ and all $f \in \mathcal{F}$. Negative samples generated using q_{Hcol} have the same label as that of the anchor $y(x)$, i.e., we have a label collision, and are also hard to distinguish from the anchor x in the representation space $\mathcal{F}$.

Proposition 2. *For all $x \in \mathcal{X}, f \in \mathcal{F}$ and any hardening function $\eta(\cdot)$ common to H-UCL, H-SCL, and Hcol we have*

$$\alpha_{\text{H-UCL}}(x, f) = \alpha_{\text{H-SCL}}(x, f) + \alpha_{\text{Hcol}}(x, f)$$

Proof. Adding

$$\alpha_{\text{H-SCL}}(x, f) = \mathbb{E}_{x^- \sim q_{\text{UCL}}}\left[1(y(x^-) \neq y(x))\, \eta(g(x, x^-))\right]$$
$$\alpha_{\text{Hcol}}(x, f) = \mathbb{E}_{x^- \sim q_{\text{UCL}}}\left[1(y(x^-) = y(x))\, \eta(g(x, x^-))\right] \Rightarrow$$
$$\alpha_{\text{H-SCL}} + \alpha_{\text{Hcol}} = \mathbb{E}_{x^- \sim q_{\text{UCL}}}\left[\eta(g(x, x^-))\right] = \alpha_{\text{H-UCL}}(x, f).$$

□

We make the following technical assumption:

Assumption 1. *For any given $f \in \mathcal{F}$ and hardening function $\eta(\cdot)$, for all $x \in \mathcal{X}$,*

$$\mathbb{E}_{x^- \sim q_{\text{Hcol}}}\left[e^{g(x,x^-)}\right] \geq \mathbb{E}_{x^- \sim q_{\text{H-SCL}}}\left[e^{g(x,x^-)}\right].$$

Assumption 1 asserts that in expectation, the exponentiated similarity (respectively, distance) between the anchor x and hard-to-distinguish samples sharing the anchor's label is greater (respectively, smaller) than the exponentiated similarity (respectively, distance) between the anchor x and hard-to-distinguish samples that also have a different label from the anchor.

In practice, Assumption 1 is reasonable if the representation function f is a "good" mapping, i.e., under the mapping f, samples having the same label are pulled closer to each other whereas samples having different labels are pushed far apart. In Sec. V, we will provide some empirical evidence for Assumption 1.

Lemma 1. *Under Assumption 1,*

$$\mathcal{L}^{(\infty)}_{\text{H-UCL}} \geq \mathcal{L}^{(\infty)}_{\text{H-SCL}}.$$

Proof. We will show that for any given $f \in \mathcal{F}$, hardening function $\eta(\cdot)$, and all $x \in \mathcal{X}$,

$$\mathbb{E}_{x^- \sim q_{\text{H-UCL}}}\left[e^{g(x,x^-)}\right] \geq \mathbb{E}_{x^- \sim q_{\text{H-SCL}}}\left[e^{g(x,x^-)}\right]$$

from which the desired inequality would follow since $\log(\cdot)$ is a strictly increasing function. Using Proposition 1,

$$\begin{aligned}
&\mathbb{E}_{x^- \sim q_{\text{H-UCL}}}\left[e^{g(x,x^-)}\right] \\
&= \frac{\mathbb{E}_{x^- \sim q_{\text{UCL}}}[\eta(g(x,x^-))e^{g(x,x^-)}]}{\alpha_{\text{H-UCL}}} \\
&= \frac{\mathbb{E}_{x^- \sim q_{\text{UCL}}}[1(y(x) = y(x^-)\eta(g(x,x^-))e^{g(x,x^-)}]}{\alpha_{\text{H-UCL}}} \\
&\quad + \frac{\mathbb{E}_{x^- \sim q_{\text{UCL}}}[1(y(x) \neq y(x^-)\eta(g(x,x^-))e^{g(x,x^-)}]}{\alpha_{\text{H-UCL}}} \\
&= \frac{\alpha_{\text{Hcol}}}{\alpha_{\text{H-UCL}}}\mathbb{E}_{x^- \sim q_{\text{Hcol}}}\left[e^{g(x,x^-)}\right] + \frac{\alpha_{\text{H-SCL}}}{\alpha_{\text{H-UCL}}}\mathbb{E}_{x^- \sim q_{\text{H-SCL}}}\left[e^{g(x,x^-)}\right] \\
&\geq \frac{\alpha_{\text{Hcol}}}{\alpha_{\text{H-UCL}}}\mathbb{E}_{x^- \sim q_{\text{H-SCL}}}\left[e^{g(x,x^-)}\right] + \frac{\alpha_{\text{H-SCL}}}{\alpha_{\text{H-UCL}}}\mathbb{E}_{x^- \sim q_{\text{H-SCL}}}\left[e^{g(x,x^-)}\right] \qquad (1) \\
&= \left(\frac{\alpha_{\text{Hcol}} + \alpha_{\text{H-SCL}}}{\alpha_{\text{H-UCL}}}\right)\mathbb{E}_{x^- \sim q_{\text{H-SCL}}}\left[e^{g(x,x^-)}\right] \\
&= \mathbb{E}_{x^- \sim q_{\text{H-SCL}}}\left[e^{g(x,x^-)}\right] \qquad (2)
\end{aligned}$$

where inequality (1) follows from Assumption 1 and equality (2) follows from Proposition 2. □

Lemma 1 shows that the loss function of H-UCL can be used as a proxy to optimize the loss function of H-SCL under certain conditions. Lemma 1 requires $k \to +\infty$ which suggests using a large value of k in practice. This is consistent with the numerical results in [18], [21]–[23] where large values of k lead to higher accuracies in downstream tasks.

We shall now loosely explain why the loss function of UCL ($\mathcal{L}_{\text{UCL}}$) cannot upper bound the loss function of H-SCL ($\mathcal{L}_{\text{H-SCL}}$) for all hardening functions. Suppose we have a hardening function such that for any given $f \in \mathcal{F}$ and any $x \in \mathcal{X}$ we have

$$\mathbb{E}_{x^- \sim q_{\text{H-SCL}}}\left[e^{g(x,x^-)}\right] \geq \mathbb{E}_{x^- \sim q_{\text{UCL}}}\left[e^{g(x,x^-)}\right].$$

Then, $\mathcal{L}_{\text{H-SCL}} \geq \mathcal{L}_{\text{UCL}}$ since $\log(\cdot)$ is a strictly increasing function. To design such a hardening function, let

$$\tau(x,f) := \mathbb{E}_{x^- \sim q_{\text{UCL}}}\left[e^{g(x,x^-)}\right]$$

and define

$$\eta(g(x,x^-)) = 1(e^{g(x,x^-)} \geq \tau).$$

Then, all the hard negative samples generated by $q_{\text{H-SCL}}$ will, by design, satisfy the inequality $e^{g(x,x^-)} \geq \tau$. If there is a nonzero probability of at least some of them having a label different from that of the anchor, then we would have $\mathbb{E}_{x^- \sim q_{\text{H-SCL}}}\left[e^{g(x,x^-)}\right] \geq \tau(x,f) = \mathbb{E}_{x^- \sim q_{\text{UCL}}}\left[e^{g(x,x^-)}\right]$.

A. Thresholded similarity hardening function

To gain better intuition for the theoretical concepts and results we have developed, we now consider a special hardening function based on thresholding the similarity between samples, specifically,

$$\eta_{\text{thresh}}(t) := 1(e^t \geq \tau)$$

where $\tau > 0$ is a threshold that controls the hardness of negative samples. A large value of τ makes it harder to distinguish between the anchor and negative samples in the representation space. For this hardening function,

$$\begin{aligned}
\alpha_{\text{SCL}} &= q_{\text{UCL}}(\mathcal{H}_{\text{SCL}}) \\
\alpha_{\text{H-UCL}} &= q_{\text{UCL}}(\mathcal{H}_{\text{H-UCL}}) \\
\alpha_{\text{H-SCL}} &= q_{\text{UCL}}(\mathcal{H}_{\text{H-SCL}}) \\
\alpha_{\text{Hcol}} &= q_{\text{UCL}}(\mathcal{H}_{\text{Hcol}})
\end{aligned}$$

where for any set $\mathcal{H}$,

$$q_{\text{UCL}}(\mathcal{H}) := \Pr_{x^- \sim q_{\text{UCL}}}(x^- \in \mathcal{H}),$$

and

$$\begin{aligned}
\mathcal{H}_{\text{SCL}}(x) &:= \left\{x^- \in \mathcal{X} \mid y(x^-) \neq y(x)\right\}. \\
\mathcal{H}_{\text{H-UCL}}(x,f,\tau) &:= \left\{x^- \in \mathcal{X} | e^{g(x,x^-)} \geq \tau\right\} \\
\mathcal{H}_{\text{H-SCL}}(x,f,\tau) &:= \mathcal{H}_{\text{SCL}}(x) \cap \mathcal{H}_{\text{H-UCL}}(x,f,\tau) \\
\mathcal{H}_{\text{Hcol}}(x,f,\tau) &:= \mathcal{H}^c_{\text{SCL}}(x) \cap \mathcal{H}_{\text{H-UCL}}(x,f,\tau)
\end{aligned}$$

and "^{c}" denotes set complement. The negative sampling distributions for SCL, H-UCL, H-SCL, and Hcol are given by

$$q_{\text{SCL}}(x^-|x) = \begin{cases} \frac{q_{\text{UCL}}(x^-)}{q_{\text{UCL}}(\mathcal{H}_{\text{SCL}})} & x^- \in \mathcal{H}_{\text{SCL}}(v) \\ 0 & \text{otherwise} \end{cases}$$

$$q_{\text{H-UCL}}(x^-|x,f,\tau) = \begin{cases} \frac{q_{\text{UCL}}(x^-)}{q_{\text{UCL}}(\mathcal{H}_{\text{H-UCL}})} & x^- \in \mathcal{H}_{\text{H-UCL}} \\ 0 & \text{otherwise} \end{cases}$$

$$q_{\text{H-SCL}}(x^-|x,f,\tau) = \begin{cases} \frac{q_{\text{UCL}}(x^-)}{q_{\text{UCL}}(\mathcal{H}_{\text{H-SCL}})} & \text{if } x^- \in \mathcal{H}_{\text{H-SCL}} \\ 0 & \text{otherwise} \end{cases}$$

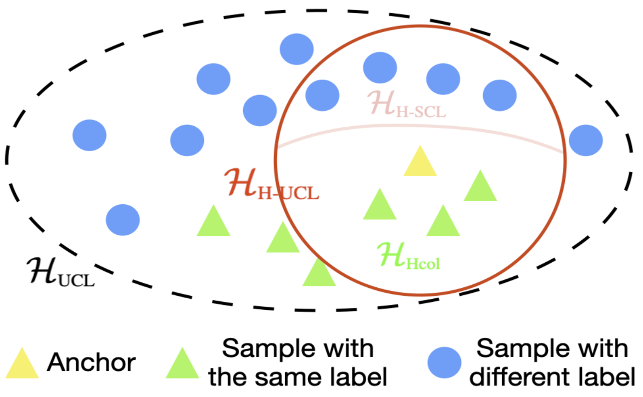
$\mathcal{H}_{\text{H-SCL}}$
$\mathcal{H}_{\text{H-UCL}}$
$\mathcal{H}_{\text{Hcol}}$
$\mathcal{H}_{\text{UCL}}$
Anchor
Sample with the same label
Sample with different label

Fig. 3. For a given representation function f, anchor x (yellow triangle) and a threshold τ, $\mathcal{H}_{\text{H-UCL}}(x, f, \tau)$ contains all the samples x^- that satisfy the constraint $e^{g(x,x^-)} \geq \tau$ (samples within the solid-line circle in the figure) that which are difficult to distinguish from the anchor in the representation space. $\mathcal{H}_{\text{H-SCL}}(x, f, \tau)$ is a subset of $\mathcal{H}_{\text{H-UCL}}(x, f, \tau)$ and only contains samples that are hard to distinguish from the anchor and have labels different from the anchor's (blue discs within the solid-line circle). $\mathcal{H}_{\text{Hcol}}(x, f, \tau)$ only contains samples that are hard to distinguish from the anchor and have the same label as the anchor (triangles within the solid-line circle). The set $\mathcal{H}_{\text{SCL}}(x)$ consists of all samples having labels different from the anchor's, irrespective of whether they are easy or hard to distinguish from the anchor (all blue discs).

$$q_{\text{Hcol}}(x^-|x, f, \tau) = \begin{cases} \frac{q_{\text{UCL}}(x^-)}{q_{\text{UCL}}(\mathcal{H}_{\text{Hcol}})} & x^- \in \mathcal{H}_{\text{Hcol}} \\ 0 & \text{otherwise} \end{cases}$$

respectively. We note that $\mathcal{H}_{\text{SCL}}(x)$ is the set of all samples having labels different from that of the anchor x. The set $\mathcal{H}_{\text{H-UCL}}(x, f, \tau)$ consists of all samples for which $e^{g(x,x^-)} \geq \tau$, i.e., samples whose similarity with the anchor x in the representation space is greater than or equal to $\gamma\tau$, where γ is the temperature parameter of g, and are therefore harder to distinguish from the anchor than other samples. The set $\mathcal{H}_{\text{H-SCL}}(x, f, \tau)$ consists of all samples having labels different from that of the anchor x and for which $e^{g(x,x^-)} \geq \tau$, i.e., they are also hard to distinguish from the anchor in the representation space. Finally, the set $\mathcal{H}_{\text{Hcol}}(x, f, \tau)$ consists of all samples having the *same* label as the anchor's, i.e., $y(x)$, but are hard to distinguish from the anchor in the representation space, specifically, $e^{g(x,x^-)} \geq \tau$.

Figure 3 illustrates the relationships between $\mathcal{H}_{\text{UCL}}, \mathcal{H}_{\text{SCL}}, \mathcal{H}_{\text{H-UCL}}, \mathcal{H}_{\text{H-SCL}}, \mathcal{H}_{\text{Hcol}}$.

V. Numerical Results

In this section, we first demonstrate the efficiency of H-SCL over other competing methods on four image datasets. In addition, we also empirically verify Assumption 1 which supports our claim that $\mathcal{L}^{(k)}_{\text{H-SCL}} \leq \mathcal{L}^{(k)}_{\text{H-UCL}}$. Since we are using stochastic gradient descent methods, in all the experiments for each batch, k is chosen to be the number of all negative samples in a given batch, which may vary across batches and with batch sizes for different datasets. Then we present additional experimental results for 5 graph datasets.

A. Image datasets

We evaluated UCL, H-UCL, SCL, and H-SCL on the STL10 [24], CIFAR10, and CIFAR100 [25] datasets for visual representation learning which contain images with 10, 10, and 100 classes, respectively.

Experiment setup: We adopt the simulation set-up and practical implementation from [3]. For UCL and H-UCL, the positive samples are generated using augmentations (crop, flip, color-jitter, and Gaussian noise) only, while for the SCL and H-SCL the positive sample are generated using both the augmentations and the label information. We employ two different methods for selecting hard negative samples via hardening functions. In the first method we use $\eta_{\text{thresh}}(t) = 1(e^t \geq \tau)$ for sampling hard-negatives and call it the H-SCL(τ) method and in the second method we use $\eta_{\exp}(t) = e^{\beta t}$ for sampling hard-negatives and call it the H-SCL(β) method. For both methods we use the large negative sample limit of the Info-NCE loss function:

$$\log\left(1 + M\ e^{-g(x,x^+)}\mathbb{E}_{x^-\sim q_{\text{CL}}}\left[e^{g(x,x^-)}\right]\right), \tag{3}$$

where we approximate the inner expectation by averaging over all the negative samples in a given batch. The additional parameter M is a positive scalar that is used in benchmark implementations in [3], [5], [14] and for a fair comparison we include it in our simulations.

We use the simCLR set-up [5] with the projection head dimension of 128 with ResNet-50 [26] architecture to parameterize the representation function. After fixing the representation function generated by the trained ResNet-50, we train a linear classifier using the available labeled data for each dataset and report the classification accuracies.

TABLE I
The best accuracies of tested methods on four datasets (in %).

Method	STL10	CIFAR10	CIFAR100	TinyImageNet
UCL [5]	64.36 ± 0.92	89.16	64.02	53.40
H-UCL [14]	67.82 ± 1.41	90.35	67.77	56.22
SCL	68.28 ± 0.92	93.46	71.68	62.06
H-SCL(β)	**72.52 ± 1.94**	**93.98**	**75.11**	**65.39**
H-SCL(τ)	71.02 ± 2.03	92.95	72.97	63.84

Training procedure: All models are trained for 200 epochs with a batch size of 512. We use the Adam optimizer with a learning rate of 0.001 and weight decay of 10^{-6}. We set γ to 0.5 following [14] for a fair comparison. For the H-SCL(τ) method, we set $\tau = e^{l(\text{start, end, epoch})/\gamma}$, where l(start, end, epoch) is a function defined as

$$l(\text{start, end, epoch}) = \text{start} + \frac{\text{epoch} - 1}{199}(\text{end} - \text{start}) \tag{4}$$

we searched start from a set of$\{-0.5, -0.3, -0.1\}$, the end from a set $\{-0.1, 0, 0.1\}$ testing all nine possible combinations. For the H-SCL(β) method, there is only one hyper-parameter that needs to be tuned namely β. Here, we perform a grid search of β over the set $\{0.1, 0.5, 1, 2, 5\}$ and set M (in Eq. (3)) equal to the batch size minus 2 following the standard implementations [3], [5], [14]. We used NVIDIA A100 32 GB GPU for computations and it takes about 10 hours to train one model (200 epochs) for each dataset. Since labeled STL10 is a small dataset, we repeated our experiment five times on

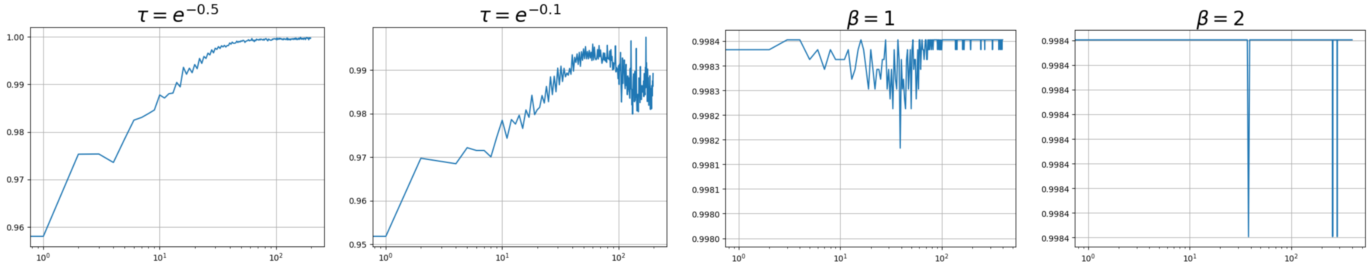

Fig. 4. Fraction of anchors satisfying Assumption 1 at the end of each epoch for $\tau = e^{-0.5}$ (first figure), $\tau = e^{-0.1}$ (second figure), $\beta = 1$ (third figure) and $\beta = 2$ (forth figure).

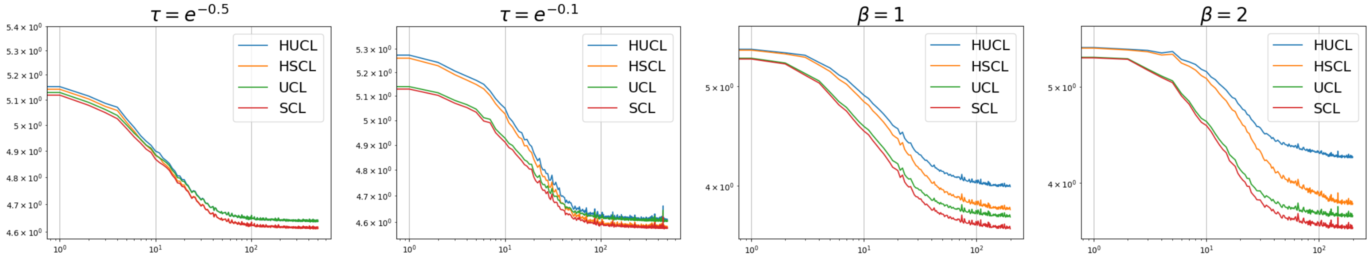

Fig. 5. Comparison of four different loss functions across epochs, with $\tau = e^{-0.5}$ (first figure), $\tau = e^{-0.1}$ (second figure), $\beta = 1$ (third figure) and $\beta = 2$ (forth figure)

this dataset and reported the average accuracy together with its standard deviation.

Results: Table I compares the best (after grid search for the best parameters) accuracies of UCL, H-UCL, SCL, H-SCL(β) and H-SCL(τ) attained on four image datasets. As seen, both H-SCL(β) and H-SCL(τ) methods are better than the baseline for most cases. H-SCL(β) consistently outperforms other methods with margins of at least 3% points on CIFAR100 and 4% points. However, H-SCL(β) is just slightly better than SCL on CIFAR10. The accuracies of the tested methods on CIFAR100 as a function of epochs is shown in Fig. 1. We can observe that H-SCL(β) only requires less than 50 epochs to achieve the same accuracy as SCL at 200 epochs.

B. Verification of Assumption 1 and Lemma 1

Note that Lemma 1 requires that the positive sampling distribution be the same for H-SCL and H-UCL. To ensure this we use both augmentation and label information to generate the positive samples for H-UCL and H-SCL. In order to verify Assumption 1, we compute the fraction of anchors that satisfy the Assumption 1 at the end of each epoch on the CIFAR100 dataset using both the H-SCL(τ) and H-SCL(β) methods and plot the fraction against epochs in Fig. 4. Our results indicate that, for both methods, this assumption is satisfied for over 95% of all anchors across all epochs.

Finally, Fig. 5 empirically confirm the correctness of Lemma 1. For both H-SCL(τ) and H-SCL(β) methods, $\mathcal{L}_{\text{H-SCL}}$ (orange curves) is always upper bounded by $\mathcal{L}_{\text{H-UCL}}$ (blue curves). However, the relationship between $\mathcal{L}_{\text{H-SCL}}$ and $\mathcal{L}_{\text{UCL}}$ is not consistent. Specifically, as shown in the first two subplots in Fig. 5, UCL loss (green curve) is less than H-SCL loss (orange curve) in the earlier epochs and is greater than the H-SCL loss in the later epochs.

C. Graph dataset

We also applied our method to learn graph representations on five graph datasets: MUTAG, ENZYMES, PTC, IMDB-BINARY, and IMDB-MULTI by [27]. We employ InfoGraph [28] as a baseline UCL method.

Training Procedure: As the H-SCL(β) method is consistently better than the H-SCL(τ) method, for the graph dataset, we only conduct the simulation using the H-SCL(β) method. We search for the best values of β over the set $\{1, 2, 10\}$, which is also used in [14]. We report the the accuracy for the best value of β for each dataset. All models are trained for 200 epochs and we use the Adam optimizer with a learning rate 0.01. We used the 3-layer GIN [29] for the representation function with a representation dimension equal to 32. Then we train an SVM classifier based on the learned graph-embedding. Each model is trained 10 times with 10-fold cross-validation.

Result: We report the performance accuracy of the different methods in Table II with boldface numbers indicating the best performance for each dataset. We observe that H-SCL(β) is consistently better than other methods across 5 datasets.

VI. Conclusion and Discussion

In this paper we introduced hard-negative supervised contrastive learning (H-SCL) which utilizes both label information and hard-negative sampling to improve downstream performance. On several real-world datasets we empirically demonstrated that H-SCL can substantially improve performance of downstream tasks compared to other contrastive learning approaches. We showed that in the asymptotic setting where

TABLE II
ACCURACY ON GRAPH DATASETS.

Method	MUTAG	ENZYMES	PTC	IMDB-B	IMDB-M
UCL [28]	86.8	50.4	55.3	72.2	49.6
H-UCL [14]	**87.2**	50.4	57.3	72.8	49.6
SCL	86.9	50.4	55.8	72.4	49.9
H-SCL (β)	**87.2**	**50.7**	**57.7**	**73.0**	**50.1**

the number of negative samples goes to infinity and a technical assumption, the hard unsupervised contrastive learning loss upper bounds the hard supervised contrastive learning loss. Our future work aims to weaken the technical assumption that is required for this relationship to hold true. We further aim to establish similar results in the non-asymptotic setting having a finite number of negative samples.

VII. Acknowledgements

This research is supported by NSF:DRL:1931978.

References

[1] A. van den Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," *arXiv preprint arXiv:1807.03748*, 2018.

[2] S. Arora, H. Khandeparkar, M. Khodak, O. Plevrakis, and N. Saunshi, "A theoretical analysis of contrastive unsupervised representation learning," *arXiv preprint arXiv:1902.09229*, 2019.

[3] P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, A. Maschinot, C. Liu, and D. Krishnan, "Supervised contrastive learning," *Advances in Neural Information Processing Systems*, vol. 33, pp. 18 661–18 673, 2020.

[4] Y. Tian, D. Krishnan, and P. Isola, "Contrastive multiview coding," in *European conference on computer vision*. Springer, 2020, pp. 776–794.

[5] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *International conference on machine learning*. PMLR, 2020, pp. 1597–1607.

[6] Z. Huang, R. Jiang, S. Aeron, and M. C. Hughes, "Accuracy versus time frontiers of semi-supervised and self-supervised learning on medical images," *arXiv preprint arXiv:2307.08919*, 2023.

[7] J. Chen, R. Zhang, Y. Mao, and J. Xu, "Contrastnet: A contrastive learning framework for few-shot text classification," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, 2022, pp. 10 492–10 500.

[8] J. Gouvea, R. Jiang, and S. Aeron, "Analyzing students' written arguments by combining qualitative and computational approaches," in *Proceedings of the 15th International Conference on Computer-Supported Collaborative Learning-CSCL 2022, pp. 163-170.* International Society of the Learning Sciences, 2022.

[9] R. Jiang, J. Gouvea, E. Miller, D. Hammer, and S. Aeron, "Interpretable contrastive word mover's embedding," *arXiv preprint arXiv:2111.01023*, 2021.

[10] N. Rethmeier and I. Augenstein, "A primer on contrastive pretraining in language processing: Methods, lessons learned, and perspectives," *ACM Computing Surveys*, vol. 55, no. 10, pp. 1–17, 2023.

[11] M. N. Mohsenvand, M. R. Izadi, and P. Maes, "Contrastive representation learning for electroencephalogram classification," in *Machine Learning for Health*. PMLR, 2020, pp. 238–253.

[12] M. T. Nonnenmacher, L. Oldenburg, I. Steinwart, and D. Reeb, "Utilizing expert features for contrastive learning of time-series representations," in *International Conference on Machine Learning*. PMLR, 2022, pp. 16 969–16 989.

[13] A. v. d. Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," *arXiv preprint arXiv:1807.03748*, 2018.

[14] J. D. Robinson, C.-Y. Chuang, S. Sra, and S. Jegelka, "Contrastive learning with hard negative samples," in *International Conference on Learning Representations*, 2021. [Online]. Available: https://openreview.net/forum?id=CR1XOQ0UTh-

[15] A. Tabassum, M. Wahed, H. Eldardiry, and I. Lourentzou, "Hard negative sampling strategies for contrastive representation learning," *arXiv preprint arXiv:2206.01197*, 2022.

[16] R. Jiang, P. Ishwar, and S. Aeron, "Hard negative sampling via regularized optimal transport for contrastive representation learning," in *2023 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2023, pp. 1–8.

[17] R. Jiang, T. Nguyen, S. Aeron, and P. Ishwar, "On neural and dimensional collapse in supervised and unsupervised contrastive learning with hard negative sampling," *arXiv preprint arXiv:2311.05139*, 2023.

[18] M. Wu, M. Mosse, C. Zhuang, D. Yamins, and N. Goodman, "Conditional negative sampling for contrastive learning of visual representations," in *International Conference on Learning Representations*, 2020.

[19] Y. Tian, C. Sun, B. Poole, D. Krishnan, C. Schmid, and P. Isola, "What makes for good views for contrastive learning?" *Advances in neural information processing systems*, vol. 33, pp. 6827–6839, 2020.

[20] Y. Kalantidis, M. B. Sariyildiz, N. Pion, P. Weinzaepfel, and D. Larlus, "Hard negative mixing for contrastive learning," *Advances in Neural Information Processing Systems*, vol. 33, pp. 21 798–21 809, 2020.

[21] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 9729–9738.

[22] C. Zhuang, A. L. Zhai, and D. Yamins, "Local aggregation for unsupervised learning of visual embeddings," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 6002–6012.

[23] M. Laskin, A. Srinivas, and P. Abbeel, "Curl: Contrastive unsupervised representations for reinforcement learning," in *International conference on machine learning*. PMLR, 2020, pp. 5639–5650.

[24] A. Coates, A. Ng, and H. Lee, "An analysis of single-layer networks in unsupervised feature learning," in *Proceedings of the fourteenth international conference on artificial intelligence and statistics*. JMLR Workshop and Conference Proceedings, 2011, pp. 215–223.

[25] A. Krizhevsky, G. Hinton *et al.*, "Learning multiple layers of features from tiny images," 2009.

[26] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *arXiv preprint arXiv:1512.03385*, 2015.

[27] C. Morris, N. M. Kriege, F. Bause, K. Kersting, P. Mutzel, and M. Neumann, "Tudataset: A collection of benchmark datasets for learning with graphs," *arXiv preprint arXiv:2007.08663*, 2020.

[28] F.-Y. Sun, J. Hoffman, V. Verma, and J. Tang, "Infograph: Unsupervised and semi-supervised graph-level representation learning via mutual information maximization," in *International Conference on Learning Representations*, 2019.

[29] K. Xu, W. Hu, J. Leskovec, and S. Jegelka, "How powerful are graph neural networks?" in *International Conference on Learning Representations*, 2019. [Online]. Available: https://openreview.net/forum?id=ryGs6iA5Km