CONDITIONAL NEGATIVE SAMPLING FOR CONTRASTIVE LEARNING OF VISUAL REPRESENTATIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Recent methods for learning unsupervised visual representations, dubbed contrastive learning, optimize a bound on mutual information between two views of an image, where "negative samples" are used to approximate a partition function. In this paper, we show that choosing difficult negatives, or those more similar to the current instance, can build stronger representations in the constrastive setting. To do this, we introduce a family of new estimators of mutual information that sample more difficult negatives from conditional distributions. Theoretically, we show these estimators remain a lower bound, and have lower variance than NCE. Experimentally, we find our approach, applied on top of IR, CMC, and MoCo, achieves 2-5% points higher in each case, measured by linear evaluation on four standard image datasets. Moreover, we find continued benefits when transferring features (1) to a variety of new image distributions from the Meta-Dataset collection and (2) to a variety of downstream tasks such as object detection, instance segmentation, and keypoint detection. While much of the attention in contrastive learning has focused on augmentations and architecture, we showcase negative sampling as a less explored but important direction for future research.

1 INTRODUCTION

Supervised learning algorithms have given rise to human-level performance in several visual tasks (Russakovsky et al., 2015; Redmon et al., 2016; He et al., 2017), relying heavily on large image datasets paired with semantic annotations. These annotations vary in difficulty and cost, spanning from simple class labels (Deng et al., 2009) to more granular descriptions like bounding boxes (Everingham et al., 2010; Lin et al., 2014) and key points (Lin et al., 2014). As it is impractical to scale high quality annotations to the size that modern deep learning demands, this reliance on supervision poses a barrier to widespread adoption. In response, we have seen the growth of *un*-supervised approaches to learning representations, or embeddings, that are general and can be re-used for many tasks at once. In natural language processing, this approach has been highly successful, resulting in the popular GPT (Radford et al., 2018; 2019; Brown et al., 2020) and BERT (Devlin et al., 2018; Liu et al., 2019) models. While supervised pretraining is still dominant in computer vision, recent approaches using "contrastive" objectives, have sparked great interest from the research community (Wu et al., 2018; Oord et al., 2018; Hjelm et al., 2018; Zhuang et al., 2019; Hénaff et al., 2019; Misra & Maaten, 2020; He et al., 2019; Chen et al., 2020a;b; Grill et al., 2020). In the last two years, contrastive methods have already achieved remarkable results, quickly closing the gap to supervised methods (He et al., 2016; 2019; Chen et al., 2020a; Grill et al., 2020).

Recent contrastive algorithms were developed as estimators of mutual information (Oord et al., 2018; Hjelm et al., 2018; Bachman et al., 2019), building on the intuition that a good low-dimensional "representation" would be one that linearizes the useful information embedded within a high-dimensional data point. In the visual domain, these estimators optimize an encoder by maximizing the similarity of encodings for two augmentations (i.e. transformations) of the same image. Doing so is trivial unless this similarity function is normalized. This is done by using "negative examples", contrasting an image (e.g. of a cat) with a set of possible other images (e.g. of dogs, tables, cars, etc.). We hypothesize that the manner in which we choose these negatives greatly effects the quality of the representations. For instance, differentiating a cat from other breeds of cats is visually more difficult than differentiating a cat from other classes. The encoder may thus have to focus on more granular, semantic information (e.g. fur patterns) that may be useful for downstream visual

tasks (e.g. object classification). While research in contrastive learning has explored architectures, augmentations, and pretext tasks, there has been little attention given to how one chooses negative samples beyond the common tactic of uniformly sampling from the training dataset.

While choosing more difficult negatives seems promising, there are several unanswered theoretical and practical questions. Naively choosing difficult negatives may yield an objective that no longer bounds mutual information. Since such bounds are the basis for many contrastive objectives, and have been used for choosing good augmentations (Tian et al., 2020) and other innovations, it is desirable to use harder negatives without losing this property. Moreover, even if choosing difficult negatives is theoretically justified, we do not know if it will yield representations better for downstream tasks. In this paper, we present a new family of estimators that supports sampling negatives from a particular class of conditional distributions. We then prove that this family remains a lower bound of mutual information. Moreover, we show that while they are a looser bound than the well-known noise contrastive estimator, estimators in this family have lower variance. We propose a particular method, the Ring model, within this family for choosing negatives that are close, but not too close, to the positive example. We then apply Ring to representation learning, where it is straightforward to adjust state-of-the-art contrastive objectives (e.g. MoCo, CMC) to sample harder negatives. We find that Ring negatives improve transfer performance across datasets and across underlying objectives, making this an easy and useful addition to contrastive learning methods.

2 BACKGROUND

Recent contrastive learning has focused heavily on exemplar-based objectives, where examples, or instances, are compared to one another to learn a representation. Many of these exemplar-based losses (Hjelm et al., 2018; Wu et al., 2018; Bachman et al., 2019; Zhuang et al., 2019; Tian et al., 2019; Chen et al., 2020a) are equivalent to noise contrastive estimation, or NCE (Gutmann & Hyvärinen, 2010; Oord et al., 2018; Poole et al., 2019), which is a popular lower bound on the mutual information, denoted by $\mathcal{I}$, between two random variables. This connection is well-known and stated in several works (Chen et al., 2020a; Tschannen et al., 2019; Tian et al., 2020), as well as explicitly motivating several algorithms (e.g. Deep InfoMax (Hjelm et al., 2018; Bachman et al., 2019)), and choices of image views (Tian et al., 2020). To review, recall the NCE objective:

$$\mathcal{I}(U;V) \geq \mathcal{L}_{\text{NCE}}(u_i, v_i) = \mathbf{E}_{u_i, v_i \sim p(u,v)} \mathbf{E}_{v_{1:k} \sim p(v)} \left[\log \frac{e^{f_\theta(u_i, v_i)}}{\frac{1}{k+1} \sum_{j \in \{i, 1:k\}} e^{f_\theta(u_i, v_j)}} \right] \tag{1}$$

where u, v are realizations of two random variables of interest, U and V. We call $v_{1:k} = \{v_1, \dots v_k\}$ "negative examples" that normalize the numerator with respect to other possible realizations of V. A proof of the inequality in Eq. 1 can be found in Poole et al. (2019).

Now, suppose U and V are derived from the same random variable X, and we are given a dataset $\mathcal{D} = \{x_i\}_{i=1}^n$ of n values that X can take, sampled from a distribution $p(x)$. Define $\mathcal{T}$ as a family of functions where each member $t : X \to X$ maps one realization of X to another. We call a transformed input $t(x)$ a "view" of x. In vision, $t \in \mathcal{T}$ is user-specified to be a composition of cropping, adding color jitter, gaussian blurring, among many others (Wu et al., 2018; Bachman et al., 2019; Chen et al., 2020a). The choice of view family is a primary determinant of how successful a contrastive algorithm is (Tian et al., 2020; 2019; Chen et al., 2020a). Finally, let $p(t)$ denote a distribution over $\mathcal{T}$ from which we can sample, a common choice being uniform over $\mathcal{T}$.

Next, introduce an encoder $g_\theta : X \to \mathbf{R}^d$ that maps an instance to a representation. Then, a general contrastive objective for the i-th example in $\mathcal{D}$ is:

$$\mathcal{L}(x_i) = \mathbf{E}_{t, t', t_{1:k} \sim p(t)} \mathbf{E}_{x_{1:k} \sim p(x)} \left[\log \frac{e^{g_\theta(t(x_i))^T g_\theta(t'(x_i))/\tau}}{\frac{1}{k+1} \sum_{j \in \{i, 1:k\}} e^{g_\theta(t(x_i))^T g_\theta(t_j(x_j))/\tau}} \right] \tag{2}$$

where τ is a temperature used to scale the dot product. The equivalence to NCE is immediate given $f_\theta(u, v) = g_\theta(u)^T g_\theta(v)/\tau$. We can interpret maximizing Eq. 2 as choosing an embedding that pulls two views of the same instance together while pushing two views of distinct instances apart. As a result, the learned representation is invariant to the transformations in $\mathcal{T}$. The output of g_θ is L_2 normalized to prevent trivial solutions. That is, it is optimal to uniformly disperse representations across the surface of the hypersphere. A drawback to NCE, and consequently to this

class of contrastive objectives, is that the number of negative examples k must be large to faithfully approximate the true partition function. However, the size of k in Eq. 2 is limited by compute and memory when optimizing. Thus, recent innovations have focused on tackling this challenge.

Instance Discrimination (Wu et al., 2018), or IR, introduces a memory bank M to cache embeddings of each $x_i \in \mathcal{D}$. Since every epoch we observe each instance once, the memory bank will save the embedding of the view of x_i observed last epoch in its i-th entry. Then, the objective is:

$$\mathcal{L}_{\text{IR}}(x_i; M) = \mathbf{E}_{t\sim p(t)}\mathbf{E}_{j_{1:k}\sim U(1,n)}\left[\log \frac{e^{g_\theta(t(x_i))^T M[i]/\tau}}{\frac{1}{k+1}\sum_{j\in\{i,j_{1:k}\}} e^{g_\theta(t(x_i))^T M[j]/\tau}}\right] \tag{3}$$

where $M[i]$ represents fetching the i-th entry in M, and $j_{1:k} \sim U(1, n)$ indicates uniformly sampling k integers from 1 to n, or equivalently entries from M. Observe that sampling uniformly k times from M is equivalent to $x_{1:k} \sim p(x)$. Representations stored in the memory bank are removed from the automatic differentiation tape, but in return, we can choose a large k. Several later approaches (Zhuang et al., 2019; Tian et al., 2019; Iscen et al., 2019; Srinivas et al., 2020) build on the IR framework. In particular, Contrastive Multiview Coding (Tian et al., 2019), or CMC, decomposes an image into luminance and AB-color channels. Then, CMC is the sum of two IR losses where the memory banks for each "modality" are swapped, encouraging the representation of the luminance of an image to be "close" to the representation of the AB-color of that image, and vice versa.

Momentum Contrast (He et al., 2019; Chen et al., 2020b), or MoCo, observed that the representations stored in the memory bank grow stale, since possibly thousands of optimization steps pass before updating an entry twice. This is problematic as stale entries could bias gradients. So, MoCo makes two important changes to the IR framework. First, it replaces the memory bank with a first-in first-out (FIFO) queue Q of size k. During each minibatch, representations are cached into the queue while the most stale ones are removed. Since elements in a minibatch are chosen i.i.d. from $p(x)$, using the queue as negatives is equivalent to drawing $x_{1:k} \sim p(x)$ i.i.d. Second, MoCo introduces a second (momentum) encoder $g'_{\theta'}$. Now, the primary encoder g_θ is used to embed one view of instance x_i whereas the momentum encoder is used to embed the other. Again, gradients are not propagated to $g'_{\theta'}$. Instead, its parameters are deterministically set by $\theta' = m\theta' + (1 - m)\theta$ where m is a "momentum" coefficient. In summary, the MoCo objective is

$$\mathcal{L}_{\text{MoCo}}(x_i; Q) = \mathbf{E}_{t\sim p(t)}\mathbf{E}_{j\sim U(1,n)}\left[\log \frac{e^{g_\theta(t(x_i))^T g'_{\theta'}(t'(x_i))/\tau}}{\frac{1}{k+1}\sum_{j\in\{i,1:k\}} e^{g_\theta(t(x_i))^T Q[j]/\tau}}\right], \tag{4}$$

again equivalent to NCE under a slightly different implementation.

3 Conditional Noise Contrastive Estimation

In NCE, the negative examples are sampled i.i.d. from the marginal distribution. Indeed, the existing proof that NCE lower bounds mutual information (Poole et al., 2019) assumes this to be true. However, choosing negatives in this manner may not be the best choice for learning a good representation. For instance, prior work in metric learning has shown the effectiveness of hard negative mining in optimizing triplet losses (Wu et al., 2017; Yuan et al., 2017; Schroff et al., 2015). In this work, we similarly wish to exploit choosing negatives *conditional on the current instance* but to do so in a manner that preserves the relationship of contrastive algorithms to mutual information.

We consider the general case of two random variables U and V according to a distribution $p(u, v)$, although the application to the contrastive setting is straightforward. To start, suppose we define a new distribution $q(v|v^*)$ conditional on a specific realization v^* of V. Ideally, we would like for $q(v|v^*)$ to belong to any distribution family but not all choices of q preserve a bound. We provide a counterexample in the Appendix. This does not, however, imply that we can only sample negatives from the marginal $p(v)$ (Poole et al., 2019; Oord et al., 2018). One of our theoretical contributions is to formally define a family of conditional distributions $\mathcal{Q}$ such that for any $q \in \mathcal{Q}$, we can draw negative examples from it, instead of p, in the NCE estimator while maintaining a lower bound on $\mathcal{I}(U; V)$. We call our new lower bound on mutual information the Conditional Noise Contrastive Estimator, or CNCE. The next Theorem shows CNCE to bound $\mathcal{I}(U; V)$.

Theorem 3.1. *Define U and V_1 by $p(u, v_1)$ and let $V_1, ..., V_k$ be i.i.d. Fix any $f : (U, V_j) \to \mathbf{R}$ and put $c = \mathbf{E}_{p(v_1)}[e^{f(u,v_1)}]$. Pick $B \subset \mathbf{R}$ strictly lower-bounded by c. Assume $p(S_B) > 0$ for*

(a) IR, CMC, MoCo (b) Ring (c) Annealed Ring

Figure 1: Black: view of instance x_i; gray: second view of x_1 i.e. the positive example; red: negative samples; gray area: negative distribution $q(x|t(x_i))$. In subfigure (c), the negative samples are annealed to be closer to $t(x_i)$ through training. In other words, the support of q shrinks.

$S_B = \{v|e^{f(u,v)} \in B\}$. *For Borel* $A = A_2 \times .... \times A_k$, *put* $q(V_{2:k} \in A) = \prod_{j=2}^{k} p(A_j|S_B)$. *Let* $\mathcal{L}_{CNCE}(U;V_1) = \mathbf{E}_{u,v_1\sim p(u,v_1)}\mathbf{E}_{v_{2:k}\sim q}\left[\log \frac{e^{f(u,v_1)}}{\frac{1}{k}\sum_{j=1}^{k} e^{f(u,v_j)}}\right]$. *Then* $\mathcal{L}_{CNCE} \leq \mathcal{L}_{NCE} \leq \mathcal{I}(U;V_1)$.

Proof. To show $\mathcal{L}_{\text{CNCE}} \leq \mathcal{L}_{\text{NCE}}$, we show $\mathbf{E}_p[\log(\sum_{j=1}^{k} e^{f(u,v_j)})] < \mathbf{E}_q[\log(\sum_{j=1}^{k} e^{f(u,v_j)})]$. To see this, apply Jensen's to the left-hand side of $\log \mathbf{E}_p[\sum_{j=1}^{k} e^{f(u,v_j}] < \log \sum_{j=1}^{k} e^{f(u,v_j)}$, which holds if $v_j \in S_B$ for $j \in [k]$, and then take the expectation $\mathbf{E}_q$ of both sides. The last inequality holds by monoticity $\log$, linearity of expectation, and the fact that $\mathbf{E}_p[e^{f(u,v_j)}] \leq e^{f(u,v_j)}$. □

Theorem Intuition. As a high level summary, although using arbitrary negative distributions in NCE does not bound mutual information, we have found a restricted class of distributions $\mathcal{Q}$ that "subsets the support of the marginal $p(v)$". In other words, given some fixed v^*, we have defined $q(v|v^*) \in \mathcal{Q}$ to constrain the support of p to a set S_B whose members are "close" to v^* as measured by the "similarity function" f. The remaining probability mass assigned by p to elements outside S_B is renormalized to sum to one (i.e. $p(A_j|S_B) = \frac{p(A_j \cap S_B)}{p(S_B)}$) for q to be well-defined. Intuitively, q cannot change p too much: it must redistribute mass proportionally. The primary distinction then, is the smaller support of q, which forces the negatives we sample to be harder for f to distinguish from v^*. Thm. 3.1 shows that substituting such a distribution into NCE still bounds mutual information.

Interestingly, we also find that CNCE is a looser bound than NCE, which raises the question: *when is a looser bound ever more useful?* In reply, we show that while CNCE is a more biased estimator than NCE, in return it has lower variance. This is natural to expect: because q is the result of shifting p around a smaller support, samples from q have less opportunity to deviate, hence lower variance.

Theorem 3.2. *Define* U *and* V_1 *by* $u, v_1 \sim p(u, v_1)$. *Fix* q *as stated in Theorem 3.1. Define* $Z(v_{2:k}) = \log \frac{e^{f(u,v_1)}}{\frac{1}{k}\sum_{j=1}^{k} e^{f(u,v_j)}}$. *By Theorem 3.1,* $\mathbf{E}_{p(v_{2:k})}[Z]$ *and* $\mathbf{E}_{q(v_{2:k})}[Z]$ *are estimators for the mutual information between* U *and* V_1. *Suppose that* S_B *is chosen to ensure* $Var_{q(v_{2:k})}[Z] \leq Var_{\tilde{q}(v_{2:k})}[Z]$, *where* $\tilde{q}(A) = p(A|S_B^c)$. *Then* $\text{Bias}_p(Z) \leq \text{Bias}_q(Z)$ *and* $\text{Var}_p(Z) \geq \text{Var}_q(Z)$. *That is, sampling* $v_{2:k} \sim q$ *instead of* p *trades higher bias for lower variance.*

The proof is in Sec. A.2. Given a good similarity function f, the elements inside S_B contain values of v truly similar to the fixed point u as measured by f. Thus, the elements outside of S_B occupy a larger range of f, and thereby are more varied, satisfying the assumption. Thm. 3.2 provides one answer to our question of looseness. In stochastic optimization, a lower variance objective may lead to better local optima. For representation learning, using CNCE to sample more difficult negatives may (1) encourage the representation to distinguish fine-grained features useful in transfer tasks, and (2) provide less noisy gradients. Thm. 3.2 also raises a warning: for a bad similarity function f, such as a randomly initialized neural network, we may not get the benefits of lower variance. We will explore the consequences of this for representation learning in the next section.

4 Ring Discrimination

We have shown that the CNCE objective provides a lower variance bound on the mutual information between two random variables. Now, we wish to apply CNCE to contrastive learning where the two random variables are derived from two views of a complex random variable X. To do so, we must

specify a concrete proposal for the support set S_B. Suppose we take the i-th example $x_i \in \mathcal{D}$, and choose a percentile $w_\ell \in [0, 100]$. Given the dataset $\mathcal{D}$, we consider each x as a negative example if and only if the normalized distance, $g_\theta(t(x_i))^T g_\theta(t'(x))$, is above the w_ℓ-th percentile of all $x \in \mathcal{D}$ for fixed transforms $t, t' \sim p(t)$. That is, we construct $q(x|t(x_i))$ such that we remove examples from the dataset whose representation dot producted with the representation of $t(x_i)$ is "too low". (Note that $w_\ell = 0$ recovers Eq. 2.) Under this formulation, w_ℓ uniquely specifies a set of distances B (recall Thm. 3.1) no lower than a threshold. For a smaller enough choice of w_ℓ, this threshold will be greater than expected distance with respect to p. In effect, the pre-image set S_B contains all $x \in \mathcal{D}$ whose distance to $t(x_i)$ in representation space is above the w_ℓ-th percentile.

Algorithm 1: MoCoRing

```
# g_q, g_k: encoder networks
# m: momentum; t: temperature
# omega_u: ring upper threshold
# omega_\ell: ring lower threshold
tx1=aug(x)  # random augmentation
tx2=aug(x)
emb1=norm(g_q(tx1))
emb2=norm(g_k(tx2)).detach()
dps=sum(tx1*tx2)/t  # dot product
# sort from farthest to closest neg
all_dps=sort(emb1@queue.T/t)
# find indices of thresholds
ix_\ell=omega_\ell*len(queue)
ix_u=omega_u*len(queue)
ring_dps=all_dps[:,ix_\ell:ix_u]
# nonparametric softmax
loss=-dps+logsumexp(ring_dps)
loss.backward()
step(g_q.params)
# moco updates
g_k.params = m*g_k.params+\
             (1-m)*g_q.params
enqueue(queue,emb2); dequeue(queue)
# threshold updates
anneal(omega_\ell); anneal(omega_u)
```

However, picking the closest examples to $t(x_i)$ as its negative examples may be inappropriate, as these examples might be better suited as positive views rather than negatives (Zhuang et al., 2019; Xie et al., 2020). As an extreme case, if the same image is included in the dataset twice, we would not like to select it as a negative example for itself. Furthermore, choosing negatives "too close" to the current instance may result in representations that pick up on fine-grain details only, ignoring larger semantic concepts. For instance, we may find representations that can distinguish two cats based on fur but are unable to classify animals from cars. This suggests removing a set from S_B of instances we consider "too close" to the current example. In practice, this translates to picking two percentiles $w_\ell < w_u \in [0, 100]$. Now, we consider each example x as a negative example for x_i if and only if $g_\theta(t(x_i))^T g_\theta(t'(x))$ is within the w_ℓ-th to w_u-th percentiles of all $x \in \mathcal{D}$. We are free to define the support set S_B in this manner as Thm. 3.1 does not require S_B to contain *all* elements with high similarity to $t(x_i)$. Intuitively, we construct a conditional distribution for negative examples that are (1) not too easy since their representations are fairly similar to that of $t(x_i)$ and (2) not too hard since we remove the "closest" instances to x_i from S_B. We call this algorithm *Ring Discrimination*, or Ring, inspired by the shape of negative set (see Fig. 1).

Ring can be easily added to popular contrastive algorithms. For IR and CMC, this amounts to simply sampling entries in the memory bank that fall within the w_ℓ-th to w_u-th percentile of all distances to the current instance view (in representation space). Similarly, for MoCo, we sample from a subset of the queue (chosen to be in the w_ℓ-th to w_u-th percentile), preserving the FIFO ordering. In our experiments, we refer to these as IRing, CMCRing, MoCoRing, respectively. Alg. 1 shows PyTorch-like pseudocode for MoCoRing. One of the strengths of this approach is the simplicity: the algorithm requires only a few lines of code on top of existing implementations.

Annealing Policy. Naively using Ring can collapse to a poor representation, as hinted by Thm. 3.2. Early in training, when the representations are still disorganized, choosing negatives that are close in representation may detrimentally exclude those examples that are "actually" close. This could lock in poor local minima. To avoid this possibility we propose to use Ring with an annealing policy that reduces the size of S_B throughout training. To do this, early in training we choose w_ℓ to be small. Over many epochs, we slowly anneal w_ℓ to approach w_u thereby selecting more difficult negatives. We explored several annealing policies and found a linear schedule to be well-performing and simple (see Appendix). In our experiments, we found annealing thresholds to be crucial: being too aggressive with negatives early in training resulted in convergence to poor optima.

5 Experiments

We explore our method applied to IR, CMC, and MoCo in four commonly used visual datasets. As in prior work (Wu et al., 2018; Zhuang et al., 2019; He et al., 2019; Misra & Maaten, 2020; Hénaff et al., 2019; Kolesnikov et al., 2019; Donahue & Simonyan, 2019; Bachman et al., 2019; Tian et al., 2019; Chen et al., 2020a), we evaluate each method by linear classification on frozen embeddings. That is, we optimize a contrastive objective on a pretraining dataset to learn a representation; then,

using a transfer dataset, we fit logistic regression on representations only. A better representation would contain more "object-centric" information, thereby achieving a higher classification score.

Training Details. We resize input images to be 256 by 256 pixels, and normalize them using dataset mean and standard deviation. The temperature τ is set to 0.07. We use a composition of a 224 by 224-pixel random crop, random color jittering, random horizontal flip, and random grayscale conversion as our augmentation family $\mathcal{T}$. We use a ResNet-18 encoder with a output dimension of 128. For CMC, we use two ResNet-18 encoders, doubling the number of parameters. For linear classification, we treat the pre-pool output (size $512 \times 7 \times 7$) after the last convolutional layer as the input to the logistic regression. Note that this setup is equivalent to using a linear projection head (Chen et al., 2020a;b). In pretraining, we use SGD with learning rate 0.03, momentum 0.9 and weight decay 1e-4 for 300 epochs and batch size 256 (128 for CMC). We drop the learning rate twice by a factor of 10 on epochs 200 and 250. In transfer, we use SGD with learning rate 0.01, momentum 0.9, and no weight decay for 100 epochs without dropping learning rate. Future work can explore orthogonal factors such as choice of architecture or pretext task.

Model	Transfer Acc.
IR	81.2
IRing	83.9 (+2.7)
CMC*	85.6
CMCRing*	**87.6** (+2.0)
MoCo	83.1
MoCoRing	86.1 (+3.0)
LA	83.9

(a) CIFAR10

Model	Transfer Acc.
IR	60.4
IRing	**62.3** (+1.9)
CMC*	56.0
CMCRing*	56.0 (+0.0)
MoCo	59.1
MoCoRing	61.5 (+2.4)
LA	61.4

(b) CIFAR100

Model	Transfer Acc.
IR	61.4
IRing	64.3 (+2.9)
CMC*	63.8
CMCRing*	**66.4** (+2.6)
MoCo	63.8
MoCoRing	65.2 (+1.4)
LA	63.0

(c) STL10

Model	Transfer Acc.
IR	43.2
IRing	48.4 (+5.2)
CMC*	48.2
CMCRing*	**50.4** (+2.2)
MoCo	52.8
MoCoRing	54.6 (+1.8)
LA	48.0

(d) ImageNet

Table 1: Comparison of contrastive algorithms on three image domains. Superscript (*) indicates models that use twice as many parameters as others e.g. CMC has "L" and "ab" encoders.

The results for CIFAR10, CIFAR100, STL10, and ImageNet are in Table 1. Overall, IR, CMC, and MoCo all benefit from using more difficult negatives as shown by 2-5% absolute points of improvement across the four datasets. While we find different contrastive objectives to perform best in each dataset, the improvements from Ring are consistent: the Ring variant outperforms the base for every model and every dataset. We also include as a baseline Local Aggregation, or LA (Zhuang et al., 2019), a popular contrastive algorithm (see Sec. F) that implicitly uses hard negatives without annealing. We find our methods to outperform LA by up to 4% absolute.

Model	Acc.
IR	81.2
IRing	83.9
IRing (No Anneal)	81.4
IRing ($w_u = 100$)	82.1

(a) CIFAR10

Model	Acc.
IR	43.2
IRing	48.4
IRing (No Anneal)	41.3
IRing ($w_u = 100$)	47.3

(b) ImageNet

Table 2: Lesioning the effects of annealing and choice of w_u.

Lesions: Annealing and Upper Boundary. Having found good performance with Ring Discrimination, we want to assess the importance of the individual components that comprise Ring. We focus on the annealing policy and the exclusion of very close negatives from S_B. Concretely, we measure the transfer accuracy of (1) IRing without annealing and (2) IRing with an upper percentile w_u set to 100, thereby excluding no close negatives. That is, S_B contains *all* examples in the dataset with representation similarity greater than the w_ℓ-th percentile (a "ball" instead of a "ring"). Table 2 compares these lesions to IR and full IRing on CIFAR10 and ImageNet classification transfer. We observe that both lesions result in worse transfer accuracy, with proper annealing being especially important, confirming the suspicions raised by Thm. 3.2.

Transferring Features. Thus far we have only evaluated the learned representations on unseen examples from the training distribution. As the goal of unsupervised learning is to capture *general* representations, we are also interested in their performance on new, unseen distributions. To gauge this, we use the same linear classification paradigm on a suite of image datasets from the "Meta Dataset" collection (Triantafillou et al., 2019) that have been used before in contrastive literature (Chen et al., 2020a). All representations were trained on CIFAR10. For each transfer dataset, we compute mean and variance from a training split to normalize input images, which we found important for generalization to new visual domains.

Model	Aircraft	CUBirds	DTD	Fungi	MNIST	FashionMNIST	TrafficSign	VGGFlower	MSCOCO
IR	40.9	17.9	39.2	2.7	96.9	91.7	97.1	68.1	52.4
IRing	40.6 (-0.3)	17.9 (+0.0)	39.5 (+0.3)	3.4 (+0.7)	97.8 (+0.9)	91.6 (+0.1)	98.8 (+1.7)	68.5 (+0.4)	52.5 (+0.1)
MoCo	41.5	18.0	39.7	3.1	96.9	90.9	97.3	64.5	52.0
MoCoRing	41.6(+0.1)	18.6 (+0.6)	39.5 (-0.2)	3.6 (+0.5)	97.9 (+1.0)	91.3 (+0.4)	99.3 (+2.0)	69.1 (+4.6)	52.6 (+0.6)
CMC	40.1	15.8	38.3	4.3	97.5	91.5	94.6	67.1	51.4
CMCRing	40.8 (+0.7)	16.8 (+1.0)	40.6 (+2.3)	4.2 (-0.1)	97.9 (+0.4)	92.1 (+0.6)	97.1 (+2.5)	69.1 (+2.0)	52.1 (+0.7)
LA	41.3	17.8	39.0	2.3	97.2	92.3	98.2	66.9	52.3

Table 3: Transferring CIFAR10 embeddings to various image distributions.

We find in Table 3 that the Ring models are competitive with the non-Ring analogues, with increases in transfer accuracies of 0.5 to 2% absolute. Most notable are the TrafficSign and VGGFlower datasets in which Ring models surpass others by a larger margin. We also observe that IRing largely outperforms LA. This suggests the features learned with more difficult negatives are not only useful for the training distribution but may also be transferrable to many visual datasets.

More Downstream Tasks. Object classification is a popular transfer task, but we want our learned representations to capture holistic knowledge about the contents of an image. We must thus evaluate performance on transfer tasks such as detection and segmentation that require different kinds of visual information. We study four additional downstream tasks: object detection on COCO (Lin et al., 2014) and Pascal VOC'07 (Everingham et al., 2010), instance segmentation on COCO, and keypoint detection on COCO. In all cases, we employ embeddings trained on ImageNet with a ResNet-18 encoder. We base these experiments after those found in He et al. (2019) with the same hyperparameters. However, we use a smaller backbone (ResNet-18 versus ResNet-50) and we freeze its parameters instead of finetuning them. We adapt code from Detectron2 (Wu et al., 2019).

	COCO: Object Detection			COCO: Inst. Segmentation			COCO: Keypoint Detection			VOC: Object Detection		
Arch.	Mask R-CNN, R_{18}-FPN, 1x schedule						R-CNN, R_{18}-FPN			Faster R-CNN, R_{18}-C4		
Model	AP^{bb}	AP^{bb}_{50}	AP^{bb}_{75}	AP^{mk}	AP^{mk}_{50}	AP^{mk}_{75}	AP^{kp}	AP^{kp}_{50}	AP^{kp}_{75}	AP^{bb}	AP^{bb}_{50}	AP^{bb}_{75}
IR	8.6	19.0	6.6	8.5	17.4	7.4	34.6	63.0	32.9	5.5	14.5	3.3
IRing	**10.9**	**22.9**	**8.7**	11.0	20.9	9.6	37.2	66.1	35.7	7.6	20.3	4.4
MoCo	6.0	14.3	4.0	10.8	21.4	9.7	37.6	66.5	36.9	7.3	17.9	4.1
MoCoRing	9.4	20.3	7.6	**12.0**	**22.9**	**10.8**	**38.7**	**67.7**	**37.9**	**8.0**	**22.1**	**4.8**
LA	10.2	22.0	8.1	10.0	20.3	9.0	36.3	65.3	35.1	7.6	20.0	4.3

Table 4: Evaluation of ImageNet representations using four visual transfer tasks.

We find IRing outperforms IR by around 2.3 points in COCO object detection, 2.5 points in COCO Instance Segmentation, 2.6 points in COCO keypoint detection, and 2.1 points in VOC object detection. Similarly, MoCoRing finds consistent improvements of 1-3 points over MoCo on the four tasks. Future work can investigate orthogonal directions of using larger encoders (e.g. ResNet-50) and finetuning ResNet parameters for these individual tasks.

6 Related Work

Several of the ideas in Ring Discrimination relate to existing work. Below, we explore these connections, and at the same time, place our work in a fast-paced and growing field.

Hard negative mining. While it has not been deeply explored in modern contrastive learning, negative mining has a rich line of research in the metric learning community. Deep metric learning utilizes triplet objectives of the form $\mathcal{L}_{\text{triplet}} = d(g_\theta(x_i), g_\theta(x_+)) - d(g_\theta(x_i), g_\theta(x_-) + \alpha)$ where d is a distance function (e.g. L_2 distance), x_+ and x_- are a positive and negative example, respectively, relative to x_i, the current instance, and $\alpha \in \mathbf{R}^+$ is a margin. In this context, several approaches pick semi-hard negatives: Schroff et al. (2015) treats the furthest (in L_2 distance) example in the same minibatch as x_i as its negative, whereas Oh Song et al. (2016) weight each example in the minibatch by its distance to $g_\theta(x_i)$, thereby being a continuous version of Schroff et al. (2015). More sophisticated negative sampling strategies developed over time. In Wu et al. (2017), the authors pick negatives from a fixed normal distribution that is shown to approximate L_2 normalized embeddings in high dimensions. The authors show that weighting by this distribution samples more diverse negatives. Similarly, HDC (Yuan et al., 2017) simulataneously optimizes a triplet loss using many levels

of "hardness" in negatives, again improving the diversity. Although triplet objectives paved the way for modern NCE-based objectives, the focus on negative mining has largely been overlooked. Ring Discrimination, being inspired by the deep metric learning literature, reminds that negative sampling is still an effective way of learning stronger representations in the new NCE framework. As such, an important contribution was to do so while retaining the theoretical properties of NCE, namely in relation to mutual information. This, to the best of our knowledge, is novel as negative mining in metric learning literature was not characterized in terms of information theory.

That being said, there are some cases of negative mining in contrastive literature. In CPC (Oord et al., 2018), the authors explore using negatives from the same speaker versus from mixed speakers in audio applications, the former of which can be interpreted as being more difficult. A recent paper, InterCLR (Xie et al., 2020), also finds that using "semi-hard negatives" is beneficial to contrastive learning whereas negatives that are too difficult or too easy produce worse representations. Where InterCLR uses a margin-based approach to sample negatives, we explore a wider family of negative distributions and show analysis that annealing offers a simple and easy solution to choosing between easy and hard negatives. Further, as InterCLR's negative sampling procedure is a special case of CNCE, we provide theory grounding these approaches in information theory. Finally, a separate line of work in contrastive learning explores using neighboring examples (in embedding space) as "positive" views of the instance (Zhuang et al., 2019; Xie et al., 2020; Asano et al., 2019; Caron et al., 2020; Li et al., 2020). That is, finding a set $\{x_j\}$ such that we consider $x_j = t(x_i)$ for the current instance x_i. While this does not deal with negatives explicitly, it shares similarities to our approach by employing other examples in the contrastive objective to learn better representations. In the Appendix, we discuss how one of these algorithms, LA (Zhuang et al., 2019), implicitly uses hard negatives and expand the Ring family with ideas inspired by it.

Contrastive learning. We focused primarily on comparing Ring Discrimination to three recent and highly performing contrastive algorithms, but the field contains much more. The basic idea of learning representations to be invariant under a family of transformations is an old one, having been explored with self-organizing maps (Becker & Hinton, 1992) and dimensionality reduction (Hadsell et al., 2006). Before IR, the idea of instance discrimination was studied (Dosovitskiy et al., 2014; Wang & Gupta, 2015) among many pretext objectives such as position prediction (Doersch et al., 2015), color prediction (Zhang et al., 2016), multi-task objectives (Doersch & Zisserman, 2017), rotation prediction (Gidaris et al., 2018; Chen et al., 2019), and many other "pretext" objectives (Pathak et al., 2017). As we have mentioned, one of the primary challenges to instance discrimination is making such a large softmax objective tractable. Moving from a parametric (Dosovitskiy et al., 2014) to a nonparametric softmax reduced issues with vanishing gradients, shifting the challenge to efficient negative sampling. The memory bank approach (Wu et al., 2018) is a simple and memory-efficient solution, quickly being adopted by the research community (Zhuang et al., 2019; Tian et al., 2019; He et al., 2019; Chen et al., 2020b; Misra & Maaten, 2020). With enough computational resources, it is now also possible to reuse examples in a large minibatch and negatives of one another (Ye et al., 2019; Ji et al., 2019; Chen et al., 2020a). In our work, we focus on hard negative mining in the context of a memory bank or queue due to its computational efficiency. However, the same principles should be applicable to batch-based methods (e.g. SimCLR): assuming a large enough batch size, for each example, we only use a subset of the minibatch as negatives as in Ring. Finally, more recent work (Grill et al., 2020) removes negatives altogether, which is speculated to implicitly use negative samples via batch normalization (Ioffe & Szegedy, 2015). We leave a more thorough understanding of negatives in BYOL to future work.

7 Conclusion

In this work, we presented a family of mutual information estimators that approximate the partition function using samples from a class of conditional distributions. We proved several theoretical statements about this family, showing a bound on mutual information and a tradeoff between bias and variance. Then, we applied these estimators as objectives in contrastive representation learning. In doing so, we found that our representations outperform existing approaches consistently across a spectrum of contrastive objectives, data distributions, and transfer tasks. Overall, we hope our work to encourage more exploration of negative sampling in the recent growth of research in contrastive learning. Future work can investigate better annealing protocols to ensure diversity.

REFERENCES

Yuki Markus Asano, Christian Rupprecht, and Andrea Vedaldi. Self-labelling via simultaneous clustering and representation learning. *arXiv preprint arXiv:1911.05371*, 2019.

Philip Bachman, R Devon Hjelm, and William Buchwalter. Learning representations by maximizing mutual information across views. In *Advances in Neural Information Processing Systems*, pp. 15535–15545, 2019.

Suzanna Becker and Geoffrey E Hinton. Self-organizing neural network that discovers surfaces in random-dot stereograms. *Nature*, 355(6356):161–163, 1992.

Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020.

Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *arXiv preprint arXiv:2006.09882*, 2020.

Ting Chen, Xiaohua Zhai, Marvin Ritter, Mario Lucic, and Neil Houlsby. Self-supervised gans via auxiliary rotation loss. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 12154–12163, 2019.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. *arXiv preprint arXiv:2002.05709*, 2020a.

Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020b.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Carl Doersch and Andrew Zisserman. Multi-task self-supervised visual learning. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2051–2060, 2017.

Carl Doersch, Abhinav Gupta, and Alexei A Efros. Unsupervised visual representation learning by context prediction. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1422–1430, 2015.

Jeff Donahue and Karen Simonyan. Large scale adversarial representation learning. In *Advances in Neural Information Processing Systems*, pp. 10542–10552, 2019.

Alexey Dosovitskiy, Jost Tobias Springenberg, Martin Riedmiller, and Thomas Brox. Discriminative unsupervised feature learning with convolutional neural networks. In *Advances in neural information processing systems*, pp. 766–774, 2014.

Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2): 303–338, 2010.

Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations. *arXiv preprint arXiv:1803.07728*, 2018.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent: A new approach to self-supervised learning. *arXiv preprint arXiv:2006.07733*, 2020.

Michael Gutmann and Aapo Hyvärinen. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pp. 297–304, 2010.

Raia Hadsell, Sumit Chopra, and Yann LeCun. Dimensionality reduction by learning an invariant mapping. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, pp. 1735–1742. IEEE, 2006.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pp. 2961–2969, 2017.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. *arXiv preprint arXiv:1911.05722*, 2019.

Olivier J Hénaff, Aravind Srinivas, Jeffrey De Fauw, Ali Razavi, Carl Doersch, SM Eslami, and Aaron van den Oord. Data-efficient image recognition with contrastive predictive coding. *arXiv preprint arXiv:1905.09272*, 2019.

R Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Phil Bachman, Adam Trischler, and Yoshua Bengio. Learning deep representations by mutual information estimation and maximization. *arXiv preprint arXiv:1808.06670*, 2018.

Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015.

Ahmet Iscen, Giorgos Tolias, Yannis Avrithis, and Ondrej Chum. Label propagation for deep semi-supervised learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5070–5079, 2019.

Xu Ji, João F Henriques, and Andrea Vedaldi. Invariant information clustering for unsupervised image classification and segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 9865–9874, 2019.

Alexander Kolesnikov, Xiaohua Zhai, and Lucas Beyer. Revisiting self-supervised visual representation learning. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pp. 1920–1929, 2019.

Junnan Li, Pan Zhou, Caiming Xiong, Richard Socher, and Steven CH Hoi. Prototypical contrastive learning of unsupervised representations. *arXiv preprint arXiv:2005.04966*, 2020.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pp. 740–755. Springer, 2014.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

Ishan Misra and Laurens van der Maaten. Self-supervised learning of pretext-invariant representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6707–6717, 2020.

Hyun Oh Song, Yu Xiang, Stefanie Jegelka, and Silvio Savarese. Deep metric learning via lifted structured feature embedding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4004–4012, 2016.

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

Deepak Pathak, Ross Girshick, Piotr Dollár, Trevor Darrell, and Bharath Hariharan. Learning features by watching objects move. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2701–2710, 2017.

Ben Poole, Sherjil Ozair, Aaron van den Oord, Alexander A Alemi, and George Tucker. On variational bounds of mutual information. *arXiv preprint arXiv:1905.06922*, 2019.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training, 2018.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8):9, 2019.

Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 779–788, 2016.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015.

Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 815–823, 2015.

Aravind Srinivas, Michael Laskin, and Pieter Abbeel. Curl: Contrastive unsupervised representations for reinforcement learning. *arXiv preprint arXiv:2004.04136*, 2020.

Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. *arXiv preprint arXiv:1906.05849*, 2019.

Yonglong Tian, Chen Sun, Ben Poole, Dilip Krishnan, Cordelia Schmid, and Phillip Isola. What makes for good views for contrastive learning. *arXiv preprint arXiv:2005.10243*, 2020.

Eleni Triantafillou, Tyler Zhu, Vincent Dumoulin, Pascal Lamblin, Utku Evci, Kelvin Xu, Ross Goroshin, Carles Gelada, Kevin Swersky, Pierre-Antoine Manzagol, et al. Meta-dataset: A dataset of datasets for learning to learn from few examples. *arXiv preprint arXiv:1903.03096*, 2019.

Michael Tschannen, Josip Djolonga, Paul K Rubenstein, Sylvain Gelly, and Mario Lucic. On mutual information maximization for representation learning. *arXiv preprint arXiv:1907.13625*, 2019.

Xiaolong Wang and Abhinav Gupta. Unsupervised learning of visual representations using videos. In *Proceedings of the IEEE international conference on computer vision*, pp. 2794–2802, 2015.

Chao-Yuan Wu, R Manmatha, Alexander J Smola, and Philipp Krahenbuhl. Sampling matters in deep embedding learning. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2840–2848, 2017.

Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. Detectron2. https://github.com/facebookresearch/detectron2, 2019.

Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3733–3742, 2018.

Jiahao Xie, Xiaohang Zhan, Ziwei Liu, Yew Soon Ong, and Chen Change Loy. Delving into inter-image invariance for unsupervised visual representations. *arXiv preprint arXiv:2008.11702*, 2020.

Mang Ye, Xu Zhang, Pong C Yuen, and Shih-Fu Chang. Unsupervised embedding learning via invariant and spreading instance feature. In *Proceedings of the IEEE Conference on computer vision and pattern recognition*, pp. 6210–6219, 2019.

Yuhui Yuan, Kuiyuan Yang, and Chao Zhang. Hard-aware deeply cascaded embedding. In *Proceedings of the IEEE international conference on computer vision*, pp. 814–823, 2017.

Richard Zhang, Phillip Isola, and Alexei A Efros. Colorful image colorization. In *European conference on computer vision*, pp. 649–666. Springer, 2016.

Chengxu Zhuang, Alex Lin Zhai, and Daniel Yamins. Local aggregation for unsupervised learning of visual embeddings. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 6002–6012, 2019.