Rethinking Prototypical Contrastive Learning through Alignment, Uniformity and Correlation

Shentong Mo[1]
shentonm@andrew.cmu.edu
Zhun Sun†[2]
zhunsun@gmail.com
Chao Li[3]
chao.li@riken.jp

[1] Carnegie Mellon University
Pittsburgh, PA 15213, United States

[2] Tohoku University
Sendai, Miyagi, Japan

[3] Center for Advanced Intelligence
Project (AIP), RIKEN
Tokyo, Japan

Abstract

Contrastive self-supervised learning (CSL) with a prototypical regularization has been introduced in learning meaningful representations for downstream tasks that require strong semantic information. However, optimizing CSL with a loss that performs the prototypical regularization aggressively, *e.g.*, the ProtoNCE loss, might cause the "coagulation" of examples in the embedding space. That is, the intra-prototype diversity of samples collapses to trivial solutions for their prototype being well-separated from others. Motivated by previous works, we propose to mitigate this phenomenon by learning **P**rototypical representation through **A**lignment, **U**niformity and **C**orrelation (**`PAUC`**). Specifically, the ordinary ProtoNCE loss is revised with: (1) an alignment loss that pulls embeddings from positive prototypes together; (2) a uniformity loss that distributes the prototypical level features uniformly; (3) a correlation loss that increases the diversity and discriminability between prototypical level features. We conduct extensive experiments on various benchmarks where the results demonstrate the effectiveness of our method on improving the quality of prototypical contrastive representations. Particularly, in the classification down-stream tasks with linear probes, our proposed method outperforms the state-of-the-art instance-wise and prototypical contrastive learning methods on the ImageNet-100 dataset by 2.96% and the ImageNet-1K dataset by 2.46% under the same settings of batch size and epochs.

1 Introduction

Contrastive learning in a prototypical style [5, 22, 24, 34] has achieved remarkable progress in terms of learning meaningful representations under the self-supervised setting. The major goal of introducing prototypes is to encourage representations to be closer to clusters of samples that contain certain semantic meaning (*i.e.* positive prototypes), while being away from the clusters that do not (*i.e.* negative prototypes).

†Corresponding author

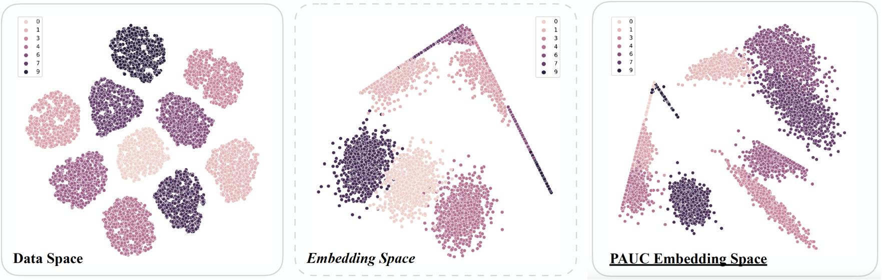

Figure 1: Illustration of the "coagulation" phenomenon of prototypes in 2D-toy examples: (Top Left) Original data space; (Top Right) unnormalized CSL embedding space; (Bottom) unnormalized PAUC embedding space. The prototypes in the proposed PAUC embedding space are more proportioned then the CSL one.

In contrast to the instance-wise contrastive learning, where the InfoNCE loss [28] is computed between instance-level features, a typical prototypical contrastive self-supervised learning (CSL) framework calculates the contrastive loss between features at the cluster level. However, optimizing the prototype-level contrastive loss tends to lead to trivial solutions, termed *prototype collapse*. Instead, it would sacrifice the variety and diversity of features in the embedding space for the samples with similar semantic information. Meanwhile, once a sample is assigned to an unrelated prototype during the optimization progress, the quality of learned representations further deteriorates since features are less informative except its prototype.

To estimate the severity of the collapse, we treat the prototypes as independent probability distributions which describe the uncertainty of data being a certain prototype, and evaluate the quality of pre-trained representations at *cluster-level*, on the **N**ormalized hyper-sphere by **E**arth **M**over's **D**istance (NEMD). Intuitively, prototypes containing different semantic information should uniformly distribute on a normalized hyper-sphere, with moderate overlapping (due to the ambiguity of boundary samples). If they are collapsing, then the distance between them should increase, causing "void" on the hyper-sphere. This fact will be empirically analyzed in this work by experiments on the 2D embedding space, under the PCL [22] framework. Subsequently, we tackle this collapsing issue by revising the vanilla PCL loss with three additional terms, dubbed, **P**rototypical representation through **A**lignment, **U**niformity and **C**orrelation (**PAUC**). Specifically, we impose the prototypical alignment loss between positive prototypes to pull embeddings from positive prototypes together; to achieve a uniform distribution of the prototypes on the normalized hyper-sphere, we then consider the prototypical uniformity loss between all prototypes; to increase the diversity and discriminability of prototypical level features, we further take the prototypical correlation loss motivated by mutual information into account.

We conduct extensive experiments and ablation studies on ImageNet-100 and ImageNet-1K datasets, and compare with state-of-the-art instance-wise and prototypical frameworks. The contributions in this work can be summarized as follows:

- We demonstrate the *prototype* collapse issue of pre-trained features in a typical PCL framework and analyze it with the proposed NEMD score defined on the normalized embedding space.
- We improve the prototypical learning of representations through alignment, uniformity, and correlation, mitigating the prototype collapse issue.

- Extensive experiments on various benchmarks demonstrate the effectiveness of our method on enriching the diversity and alleviating the collapsing issue in meaningful prototypical representations.

2 Related Work

Taxonomy of Self-supervised Learning. Self-supervised learning [7, 30, 31, 35] has achieved remarkable progress in recent years, in representation learning without the need for class label supervision. The core part of self-supervised algorithms [14, 26, 39] is to build a specific task for networks to learn. Typically, most algorithms utilize data augmentation to generate different views of an anchor sample. Then the networks are optimized to maximize the mutual information between different views. The maximization can be achieved in both contrastive [7, 17] and non-contrastive approach [9, 38]. Likewise, the mutual information in contrastive framework can be also estimated in both instance-wise approach [7, 15, 17] and prototypical approach [5, 22, 24]. Below, we provide related backgrounds on these two approaches.

Instance-wise Contrastive Learning (ICL). The goal of ICL [7, 8, 9, 10, 15, 17] is to bring the embedding of different views from the same instance closer, and push embeddings of views from different instances far apart using instance-level contrastive loss. This is commonly achieved by a large batch size that can accumulate positive and negative pairs in the same batch [7, 8], or a momentum encoder to update negative instances from a large and consistent dictionary on the fly [10, 17]. Other related works [6, 33] generalize the instance-wise contrastive loss to the alignment of representations from positive pairs and uniformity of the induced distribution of the normalized embeddings on the hyper-sphere. We are motivated by these works yet explore those properties under the prototypical contrastive learning framework.

Prototypical Contrastive Learning. Compared to the extensive ICL frameworks, there are fewer works focusing on the prototypical contrastive learning task. SwAV [5] proposes to simply predict the code of a view from the representation of the augmented view, where the code is obtained by multiplying the cluster assignments of the data with the same linear transformation matrix. PCL [22] replaces the InfoNCE loss with ProtoNCE loss for contrastive learning to encourage representations closer to their assigned prototypes and far from negative prototypes. CLD [34] introduces the cross-level discrimination between instances and local instance groups to increase the positive/negative instance ratio of contrastive learning for better invariant mapping. SPCL [24] employs an offline prototype spawn approach with k-means clustering and regularizes samples to their corresponding prototypes explicitly. Although the prototypical framework demonstrates its superiority in classification downstream tasks, *e.g.,* the ImageNet-1K, the inferior quality of features caused by the prototype collapse has been barely studied.

Table 1: Comparison with existing works.

Method	Publication	Loss	Focus
DIM [1]	NeurIPS'2019	mutual information	instance
MoCo+A/U [33]	ICML'2020	align/uniform	instance
VICReg [2]	ICLR'2022	covariance	instance
PAUC (ours)	–	align/uniform/correlation	prototype

Collapse in Contrastive Learning. A recent work [19] discovers the feature collapse problem in self-supervised settings. The authors compartmentalize the collapse into two categories: *complete collapse* when the network produces constant; and *dimensional collapse* when some dimensions of the embedded feature representations are non-informative. Decorrelated batch normalization is then utilized to prevent the later one. Yet another concurrent work [2] in Table 1 also aims to prevent collapsed solutions by explicitly applying regularization termed Invariance and Covariance. However, in this study, we focus on the *prototype collapse*, a different type of collapse happens when the samples inside a prototype lose their diversity.

3 Preliminary

Under the self-supervised setting, there are two main branches of contrastive learning frameworks, including instance-wise contrastive learning (ICL) [4, 7, 8, 10, 17] and prototypical contrastive learning (PCL) [5, 22, 34]. The ICL framework is optimized using an ordinary NT-Xent (the normalized temperature-scaled cross-entropy loss) as the contrastive loss, called InfoNCE, to maximize the similarity between positive samples and minimize the similarity between negative samples. The PCL idea mainly focuses on applying a similar InfoNCE loss in terms of the prototypical level to distinguish positive and negative prototypes, where prototypes are defined under the clustering mechanism. To explain them in a unified manner, we define notations as follows.

Notations Given a set of training examples $\mathcal{X} = \{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n\}$, a mapping function $f(\cdot)$ is applied to generate representations $\mathcal{V} = \{\mathbf{v}_1, \mathbf{v}_2, \cdots, \mathbf{v}_n\}$, *i.e.,* $\mathbf{v}_i = f(\mathbf{x}_i), \forall i$. In the ICL setting, we denote the number of negative instances as r. In the PCL framework, we denote the times of clustering as M, and the number of prototypes as $k_m, m \in \{1, 2, \cdots, M\}$. Therefore, we have a set of different number of prototypes $K = \{k_1, k_2, \cdots, k_M\}$. The prototypes of the samples using k_m clusters are marked as $\mathcal{C}^m = \{\mathbf{c}_1, \mathbf{c}_2, \cdots, \mathbf{c}_{k_m}\}$.

The InfoNCE [25, 28, 35] objective of the instance-wise contrastive learning is defined by

$$\mathcal{L}_{\text{InfoNCE}} = \sum_{i=1}^{n} -\log \frac{\exp(\mathbf{v}_i \cdot \mathbf{v}_i'/\tau)}{\sum_{j=1}^{r} \exp(\mathbf{v}_i \cdot \mathbf{v}_j'/\tau)} \tag{1}$$

where $\mathbf{v}_i, \mathbf{v}_i', \mathbf{v}_j$ represent the anchor, positive, and negative embedding, respectively for each training sample i, and τ is a temperature hyper-parameter. The operation $\cdot$ denotes the trivial inner product of two vectors.

In prototypical contrastive learning, we replace r negative instances with r negative prototypes to calculate the normalization term. For robust estimation, we average the prototypical probability by sampling M steps with a set of different numbers of clusters K. Thus, the prototypical InfoNCE objective with $\mathcal{L}_{\text{InfoNCE}}$ is defined as:

$$\mathcal{L}_{\text{ProtoNCE}} = \mathcal{L}_{\text{InfoNCE}} + \sum_{i=1}^{n} -\frac{1}{M} \sum_{m=1}^{M} \log \frac{\exp(\mathbf{v}_i \cdot \mathbf{c}_s^m / \phi_s^m)}{\sum_{j=1}^{r} \exp(\mathbf{v}_i \cdot \mathbf{c}_j^m / \phi_j^m)} \tag{2}$$

where $\mathbf{v}_i$ is the anchor embedding for each training sample i, and $\mathbf{c}_s^m, \mathbf{c}_j^m$ are the positive prototype s that the sample i belongs to and the negative prototype j at m step, and ϕ_s^m, ϕ_j^m are concentration estimation indicators for the distribution of embeddings around the prototype s, j at m step. This term is used for pulling similar embeddings together around the target prototype.

Normalized Earth Moving Distance (NEMD) We present the *Normalized Earth Mover's Distance* to quantify the level of prototype collapse issue existing in current PCL methods. We start with the formulation of distance among prototypes on "earth", to calculate how far two prototypes are separated. To achieve this, we assume the prototypes as empirical distributions supported on a normalized hyper-sphere. NEMD is then calculated between distributions associated with prototypes to measure the distance of separate prototypes generated in PCL, which is formally defined by

$$\mathcal{L}_{\text{NEMD}}(p,q) = \inf_{\gamma \in \prod(\mathbf{c}_p^m, \mathbf{c}_q^m)} \mathbb{E}_{(x,y)\sim\gamma} \|x - y\| \tag{3}$$

where $\prod(\mathbf{c}_p^m, \mathbf{c}_q^m)$ denotes the set of all joint distributions $\gamma(x,y)$ whose marginals are $\mathbf{c}_p^m$ and $\mathbf{c}_q^m$ associated with prototypes p and q, respectively. Here $\gamma(x,y)$ intuitively indicates how many "clusters of samples" should be transported from the embedding x to the embedding y, in order to transform the prototype $\mathbf{c}_p^m$ into the prototype $\mathbf{c}_q^m$. In the experiments, we adopt the Sinkhorn [13] algorithm to calculate this score for fast computation of optimal transport between prototypes. It is also worth mentioning that, in theory there exists the case when all the embedded features collapse to one point regardless of the prototypes, and NEMD thus equals 0 in this case. However, in practice we barely fall into such trivial solutions due to the regularization of the prototype loss. Therefore, in this study we assume that NEMD is computed on well-optimized networks. To evaluate the prototype collapsing problem, we conduct extensive analytical experiments on the synthetic 3D isotropic Gaussian data, embedding them into a 2D space. Specifically, we generate isotropic Gaussian blobs with various classes (10, 20, 30, 40, 50), and each class has 1000 samples. Note that the original dimension of data is 3 in this tiny experiment. On the synthetic data we then carry out the vanilla PCL framework and its variant called PAUC with alignment, uniformity, and correlation (proposed in the next section), respectively. To further evaluate the quality of prototypical solutions, we calculate the NEMD score between pre-trained prototypical embeddings and also visualize pre-trained embeddings by projecting them into a unit circle feature space.

As shown in Fig. 2, we can observe that there exist collapsing solutions in representations generated by PCL (in the second row), leading to a high NEMD score. Our PAUC achieves a lower NEMD score compared to the general PCL framework under all class settings. This implies that the pre-trained representations generated by PAUC are more uniformly distributed on the normalized hyper-sphere. Results in Fig. 2 also validate the uniformity of our prototypical representations on the normalized hyper-sphere.

4 PAUC: Prototypical Alignment, Uniformity and Correlation

In this section, we elaborate on the three proposed properties of prototypes, *i.e.,* alignment, uniformity, and correlation, for learning prototypical representations, based on the vanilla PCL framework.

Alignment To pull embeddings from positive prototypes together and push negative prototypes away, the prototypical alignment loss is defined with the expected distance between positive prototypes:

$$\mathcal{L}_{\text{p-align}}(\mathbf{c}^m; s) = \mathbb{E}_{(p,q)\sim p_{\text{pos}}}[\|\mathbf{c}_p^m - \mathbf{c}_q^m\|_2^s] \tag{4}$$

where $\mathbf{c}_p^m$, $\mathbf{c}_q^m$ are the *positive* prototypes p,q that the sample i belongs to at m step, p_{pos} denotes the distribution of positive prototypes in the hyper-sphere, and s is a positive factor to define the distance metric between $\mathbf{c}_p^m$ and $\mathbf{c}_q^m$.

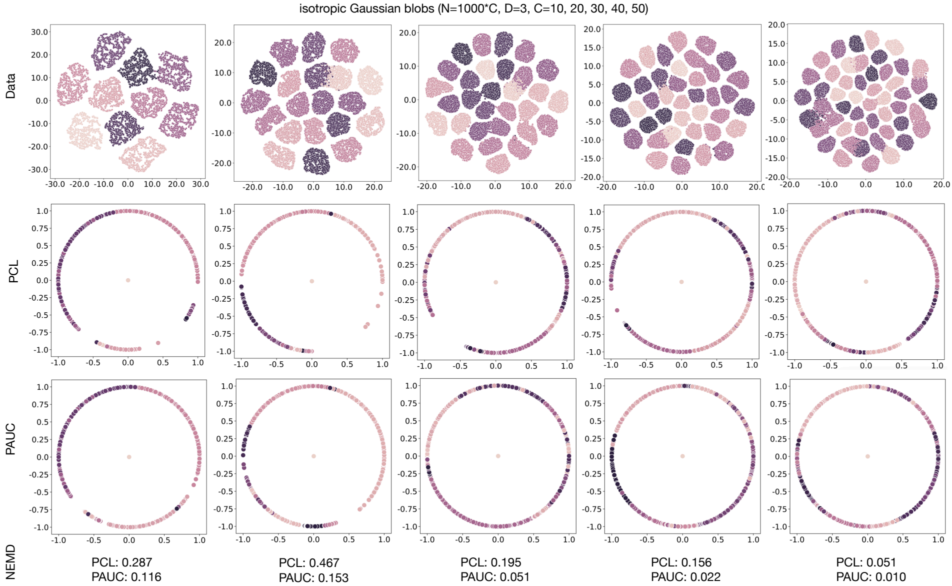

Figure 2: Comparison results of tiny experiments between prototypical contrastive learning (PCL) and our PAUC in terms of **qualitative** (the projection of pre-trained embedding on a unit circle feature space) and **quantative** ($\mathcal{L}_{\text{NEMD}}$ score).

It is worth mentioning that, in the ordinary protoNCE loss as Eq. (2), the alignment is achieved by matching other samples to the anchor. From the perspective of numerical optimization, when reducing over different number of prototypes, all the samples from positive prototypes are pulled together to their anchor sample, which might result in unwanted "coagulation" effect in a long-term training progress.

Uniformity To alleviate the inter-prototype collapsing issues, we adopt a uniformity loss to learn a uniform distribution of the prototypical features on the hyper-sphere. Motivated by the uniformity property proposed in ICL, we consider the Gaussian potential kernel [3] and define the prototypical uniformity loss by

$$\mathcal{L}_{\text{p-uniform}}(\mathbf{c}^m;t) = \mathbb{E}_{(p,q)\sim p_{\text{proto}}}[e^{-t\|\mathbf{c}_p^m - \mathbf{c}_q^m\|_2^2}] \tag{5}$$

where $\mathbf{c}_p^m$, $\mathbf{c}_q^m$ are the prototypes p,q that the sample i belongs to at m step, p_{proto} denotes the distribution of all prototypes in the hyper-sphere, and t is a positive factor to define the weight of the ℓ_2 distance between $\mathbf{c}_p^m$ and $\mathbf{c}_q^m$.

Correlation In order to distinguish the difference between each prototype further to avoid inter-prototype collapsing, we borrow the idea of mutual information and define the correlation loss between positive prototypes and negative prototypes by

$$\mathcal{L}_{\text{p-corr}} = \mathbb{E}_{(p,q)\sim p_{\text{proto}}}[\mathbf{c}_p^m \log(\mathbf{c}_q^m \odot (\mathbf{c}_p^m)^{-1})] \tag{6}$$

where $\mathbf{c}_p^m$, $\mathbf{c}_q^m$ are prototypes p,q that the sample i belongs to at m step, and $\odot$ is for element-wise product. Intuitively, this loss term drives the samples from different prototypes being unrelated among all the dimensions of the embedding space.

Overall Loss The overall loss $\mathcal{L}$ is therefore defined with the weighted summation of the

Table 2: Linear classification on ImageNet-100. Bold and underline denote the first and second place.

Method	Arch.	Param.(M)	Batch	Epochs	Top-1(%)	Top-5(%)
CMC[29]	ResNet-50	24	256	200	66.20	87.00
MoCo[17]	ResNet-50	24	256	200	72.80	91.64
Biased CMC [12]	ResNet-50	24	256	200	73.58	92.06
Debiased CMC [12]	ResNet-50	24	256	200	74.60	92.08
MoCo+align/uniform[33]	ResNet-50	24	256	200	74.60	92.74
NPID [35]	ResNet-50	24	256	200	75.30	–
BYOL [15]	ResNet-50	24	4096	1000	75.80	–
PCL-v1 [22]	ResNet-50	24	256	200	76.17	93.52
PCL-v2 [22]	ResNet-50+MLP	28	256	200	78.35	94.25
SwAV [6]	ResNet-50	24	4096	200	80.20	95.02
LooC [36]	ResNet-50	24	256	200	81.10	95.30
CLD [34]	ResNet-50	24	256	200	<u>81.50</u>	<u>95.48</u>
PAUC (ours)	ResNet-50	24	256	200	**84.46**	**97.15**

three losses by

$$\mathcal{L} = \mathcal{L}_{\text{InfoNCE}} - \sum_{i=1}^{n} \frac{1}{M} \sum_{m=1}^{M} \left(\log \frac{\exp(\mathbf{v}_i \cdot \mathbf{c}_s^m / \phi_s^m)}{\sum_{j=1}^{r} \exp(\mathbf{v}_i \cdot \mathbf{c}_j^m / \phi_j^m)} + \alpha \mathcal{L}_{\text{p-align}} + \beta \mathcal{L}_{\text{p-uniform}} + \gamma \mathcal{L}_{\text{p-corr}} \right) \tag{7}$$

where α, β, and γ denote the weights of the alignment, uniformity and correlation loss, respectively. We conduct extensive ablation studies in the supplementary material to explore the effects of each loss on the quality of prototypical representations generated by our PAUC.

5 Experiments

5.1 Dataset and Configurations

Following previous methods [6, 7, 17, 22, 29], we employ two benchmark datasets (ImageNet-100 and ImageNet-1K) for evaluating classification performance. For a fair comparison with state-of-the-arts, we adopt a ResNet-50 [16] as the encoder, where the last fully connected layer outputs a 128-D embedding with ℓ_2-normalization. Same as PCL [22], we apply data augmentation methods with the random crop, random color jittering, random horizontal flip, and random grayscale conversion.

For ImageNet-100 pre-training, we set number of clusters $K = 2500, 5000, 10000$, $r = 1024$. We apply SGD as our optimizer, with a weight decay of 0.0001, a momentum of 0.9, and a batch size of 256. We train for 200 epochs and use the first 20 epochs as a warm-up step using only the InfoNCE loss. We set the initial learning rate as 0.03, and multiply it by 0.1 at 120 and 160 epochs.

For ImageNet-1K, we set $K = 25000, 50000, 100000$, $r = 16000$. For other hyper-parameters, we follow the same setting as ImageNet-100 pre-training. For an efficient k-means clustering, we adopt the faiss library [20] during the pre-training. The whole training time for ImageNet-1K is 132 hours using 8 Tesla V100 GPUs, and 15 hours for ImageNet-100.

5.2 Comparison with State-of-the-arts

ImageNet-100. We evaluate the linear classification for the ImageNet-100 dataset, where linear models are trained on frozen features from different self-supervised methods shown in Table 2. Our PAUC substantially outperforms existing methods in terms of both instance-wise and prototypical contrastive learning. We achieve new state-of-the-art performance for linear

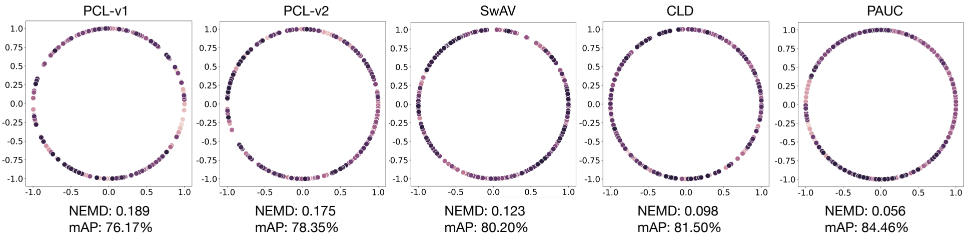

Figure 3: Comparison results on ImageNet-100 dataset between current prototypical contrastive learning methods (PCL-v1, PCL-v2, SwAV, CLD) and our PAUC in terms of **qualitative** (pre-trained embeddings on a unit circle feature space) and **quantative** ($\mathcal{L}_{\text{NEMD}}$ and top-1 accuracy).

classification on the ImageNet-100 dataset. This demonstrates the effectiveness of our PAUC pre-trained representations in transfer learning on image classification.

Table 3: Top-1 accuracy for linear classification on ImageNet-1K, where models are trained on frozen features from different methods. Bold and underline numbers denote the first and second place.

Method	**Arch.**	**Param.(M)**	**Batch**	**Epochs**	**Top-1(%)**
Instance-wise Contrastive:					
CPC [25]	ResNet-101	28	512	200	48.70
MoCo [17]	ResNet-50	24	256	200	60.60
PIRL [23]	ResNet-50	24	1024	800	63.60
CMC [29]	ResNet-50+MLP$\{L, ab\}$	47	256	200	64.00
CPCv2 [18]	ResNet-170	303	512	200	65.90
MoCo+align/uniform [15]	ResNet-50	24	256	200	67.69
AMDIM [1]	Custom-ResNet	192	1008	150	68.10
LoCo [37]	ResNet-50	24	4096	800	69.50
SimCLR [7]	ResNet-50+MLP	28	4096	400	70.00
InfoMin [30]	ResNet-50	24	256	200	70.10
MoCHi [21]	ResNet-50+MLP	28	512	200	70.60
PIC [4]	ResNet-50	24	512	1600	70.80
SWD [6]	ResNet-50+MLP	28	2048	800	70.90
MoCov2 [10]	ResNet-50+MLP	28	256	200	71.10
SimSiam [9]	ResNet-50+MLP	28	256	800	71.30
SimCLRv2 [8]	ResNet-50+MLP	28	4096	800	71.70
MoCov3 [11]	ResNet-50+MLP	28	4096	300	72.80
VICReg [2]	ResNet-50+MLP	28	2048	1000	73.20
AdCo [27]	ResNet-50	24	256	200	73.20
Barlow Twins [38]	ResNet-50	24	2048	1000	73.20
BYOL [15]	ResNet-50+MLP	35	4096	400	73.20
BYOL [15]	ResNet-50+MLP	35	4096	1000	74.30
MoCov3 [11]	ResNet-50+MLP	28	4096	1000	<u>74.60</u>
Prototypical Contrastive:					
PCL [22]	ResNet-50	24	256	200	61.50
PCLv2 [22]	ResNet-50+MLP	28	256	200	67.60
CLD [34]	ResNet-50	24	256	200	71.50
SwAV [5]	ResNet-50+MLP	28	256	200	72.70
PAUC (ours)	ResNet-50	24	256	200	**75.16**

Additionally, we report the comparison results with current prototypical contrastive

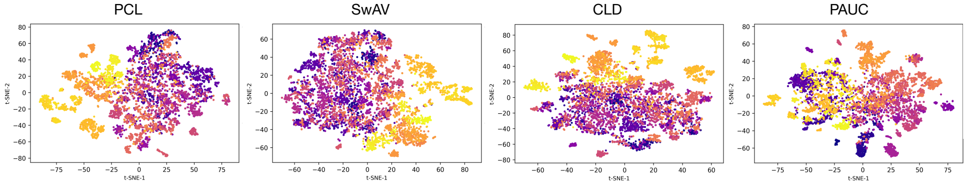

Figure 4: Visualization of pre-trained representations generated by PCL, SwAV, CLD, and PAUC (ours) from random 100 classes in the ImageNet validation set.

learning methods (PCL-v1 [22], PCL-v2 [22], SwAV [5], CLD [34]) in terms of qualitative and quantitative aspects. Notably, we project pre-trained embeddings generated by different methods on a normalized unit circle feature space and compare the $\mathcal{L}_{\text{NEMD}}$ score and top-1 accuracy in Fig. 3. Our PAUC achieves the lowest $\mathcal{L}_{\text{NEMD}}$ score while performing the best linear classification on ImageNet-100. This validates the rationality of our proposed $\mathcal{L}_{\text{NEMD}}$ in evaluating the quality of prototypical representations. As can be seen in Fig. 3, our PAUC pre-trained representations achieve better uniformity on the normalized hyper-sphere compared to previous methods, which validates the effectiveness of our PAUC in mitigating collapsing issues for the general PCL framework.

ImageNet-1K. Following previous self-supervised methods [5, 7, 17, 22, 29, 34], we compare the top-1 accuracy for linear classification on ImageNet-1K dataset as shown in Table 3. Our PAUC achieves the best results compared to existing prototypical contrastive learning methods. Under the same setting, our PAUC outperforms SwAV [5] by a large margin, *i.e.*, by 2.46%. This validates the advantage of our PAUC pre-trained representations on transfer learning for image classification again. Furthermore, it also demonstrates that, addressing the collapse is beneficial for improving the quality of pre-trained representations for downstream tasks. We also evaluate our PAUC pre-trained representations on object detection and report the detailed comparison in the supplementary material. Our PAUC achieves better or comparable performance compared to existing approaches.

When compared to ICL learning methods, PAUC pre-trained with small batch size and epochs also outperforms BYOL [15] with large batch size and pre-training epochs. This further implies that our PAUC has great potential in saving much pre-training time and GPU memories for contrastive self-supervised learning. This also validates the advantage of prototypical contrastive learning frameworks over the instance-wise contrastive learning methods.

Visualization of Pre-trained Representations. To better evaluate the quality of pre-trained representations, we visualize the self-supervised learned embeddings generated by different approaches (PCL [22], SwAV [5], CLD [34]) in Fig. 4. Typically, we project the 2048-dimensional embeddings pre-trained from ResNet-50 onto 2 dimension using tSNE [32] tools, and we randomly select 100 classes in ImageNet-1K validation set for visualization. Compared to the pre-trained representations generated by previous PCL frameworks, our PAUC pre-trained representations form more separated clusters, which are distributed more uniformly on the space in terms of all classes. This further validates the effectiveness of our PAUC in improving the quality of prototypical representations.

6 Conclusion

In this work, we reveal the prototype collapse problems existing in the general prototypical contrastive learning settings. We propose a Normalized Earth Mover's Distance (NEMD) between prototypes to measure the quality of pre-trained representations at the prototypical level. To improve the quality of prototypical representations further, our model is trained to learn the alignment of prototypes, the uniformity of prototypes on the normalized hypersphere, and the correlation between prototypes. Extensive experiments on tiny experiments and state-of-the-art benchmarks validate the effectiveness of our method on pre-training better prototypical representations.

Acknowledgement

This work was partially supported by JSPS KAKENHI (Grant No. 20H04249, 20H04208) and the National Natural Science Foundation of China (Grant No. 62006045).

References

[1] Philip Bachman, R Devon Hjelm, and William Buchwalter. Learning representations by maximizing mutual information across views. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

[2] Adrien Bardes, Jean Ponce, and Yann LeCun. Vicreg: Variance-invariance-covariance regularization for self-supervised learning. *Proceedings of International Conference on Learning Representations (ICLR)*, 2022.

[3] Albert P. Bartók and Gábor Csányi. Gaussian Approximation Potentials: a brief tutorial introduction. *arXiv preprint arXiv:1502.01366*, 2015.

[4] Yue Cao, Zhenda Xie, Bin Liu, Yutong Lin, Zheng Zhang, and Han Hu. Parametric instance classification for unsupervised visual feature learning. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, pages 15614–15624, 2020.

[5] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[6] Ting Chen and Lala Li. Intriguing properties of contrastive losses. *arXiv preprint arXiv:2011.02803*, 2020.

[7] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *Proceedings of International Conference on Machine Learning (ICML)*, 2020.

[8] Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey Hinton. Big self-supervised models are strong semi-supervised learners. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[9] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

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

[11] Xinlei Chen*, Saining Xie*, and Kaiming He. An empirical study of training self-supervised vision transformers. In *Proceedings of the International Conference on Computer Vision (ICCV)*, pages 9640–9649, 2021.

[12] Ching-Yao Chuang, Joshua Robinson, Yen-Chen Lin, Antonio Torralba, and Stefanie Jegelka. Debiased contrastive learning. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[13] Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 2013.

[14] Christoph Feichtenhofer, Haoqi Fan, Bo Xiong, Ross Girshick, and Kaiming He. A large-scale study on unsupervised spatiotemporal representation learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[15] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, Bilal Piot, koray kavukcuoglu, Remi Munos, and Michal Valko. Bootstrap your own latent - a new approach to self-supervised learning. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[17] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9729–9738, 2020.

[18] Olivier J. Hénaff, Aravind Srinivas, Jeffrey De Fauw, Ali Razavi, Carl Doersch, S. M. Ali Eslami, and Aäron van den Oord. Data-efficient image recognition with contrastive predictive coding. *CoRR*, 2019.

[19] Tianyu Hua, Wenxiao Wang, Zihui Xue, Sucheng Ren, Yue Wang, and Hang Zhao. On feature decorrelation in self-supervised learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9598–9608, 2021.

[20] Jeff Johnson, Matthijs Douze, and Hervé Jégou. Billion-scale similarity search with gpus. *arXiv preprint arXiv:1702.08734*, 2017.

[21] Yannis Kalantidis, Mert Bulent Sariyildiz, Noe Pion, Philippe Weinzaepfel, and Diane Larlus. Hard negative mixing for contrastive learning. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[22] Junnan Li, Pan Zhou, Caiming Xiong, and Steven Hoi. Prototypical contrastive learning of unsupervised representations. In *Proceedings of International Conference on Learning Representations (ICLR)*, 2021.

[23] Ishan Misra and Laurens van der Maaten. Self-supervised learning of pretext-invariant representations. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[24] Shentong Mo, Zhun Sun, and Chao Li. Siamese prototypical contrastive learning. In *Proceedings of the British Machine Vision Conference (BMVC)*, 2021.

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

[26] Senthil Purushwalkam and Abhinav Gupta. Demystifying contrastive self-supervised learning: Invariances, augmentations and dataset biases. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[27] Hu Qianjiang, Wang Xiao, Hu Wei, and Qi Guo-Jun. AdCo: adversarial contrast for efficient learning of unsupervised representations from self-trained negative adversaries. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[28] Kihyuk Sohn. Improved deep metric learning with multi-class n-pair loss objective. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016.

[29] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. In *Proceedings of European Conference on Computer Vision (ECCV)*, 2020.

[30] Yonglong Tian, Chen Sun, Ben Poole, Dilip Krishnan, Cordelia Schmid, and Phillip Isola. What makes for good views for contrastive learning? In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, pages 6827–6839, 2020.

[31] Yuandong Tian, Xinlei Chen, and Surya Ganguli. Understanding self-supervised learning dynamics without contrastive pairs. In *Proceedings of International Conference on Machine Learning (ICML)*, 2021.

[32] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9:2579–2605, 2008.

[33] Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *Proceedings of International Conference on Machine Learning (ICML)*, 2020.

[34] Xudong Wang, Ziwei Liu, and Stella X Yu. CLD: unsupervised feature learning by cross-level instance-group discrimination. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[35] 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 (CVPR)*, 2018.

[36] Tete Xiao, Xiaolong Wang, Alexei A Efros, and Trevor Darrell. What should not be contrastive in contrastive learning. In *Proceedings of International Conference on Learning Representations (ICLR)*, 2021.

[37] Yuwen Xiong, Mengye Ren, and Raquel Urtasun. LoCo: local contrastive representation learning. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, pages 11142–11153, 2020.

[38] Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction. *Proceedings of International Conference on Machine Learning (ICML)*, 2021.

[39] Nanxuan Zhao, Zhirong Wu, Rynson W.H. Lau, and Stephen Lin. What makes instance discrimination good for transfer learning. In *Proceedings of International Conference on Learning Representations (ICLR)*, 2021.