Geometric Contrastive Learning

Yeskendir Koishekenov Sharvaree Vadgama* Riccardo Valperga* Erik J. Bekkers
University of Amsterdam
yeskendir.koishekenov@student.uva.nl

Abstract

Contrastive learning has been a long-standing research area due to its versatility and importance in learning representations. Recent works have shown improved results if the learned representations are constrained to be on a hypersphere. However, this prior geometric constraint is not fully utilized during training. In this work, we propose making use of geodesic distances on the hypersphere to learn contrasts between representations. Through empirical results, we show that this contrastive learning approach improves downstream tasks across different contrastive learning frameworks. We show that having geometric inductive priors perform even better in contrastive learning if used along with other correct geometric information.

1. Introduction

Learning good representations of data is a key challenge in deep learning. Until recent years, training models with ground truth labels, i.e. supervised learning, was the most popular process. However, the main problem with the supervised approach to learning features from labeled data is the high cost of annotating millions of data samples. Self-supervised learning techniques have been instrumental in significantly accelerating the development of universally applicable representations for several downstream tasks. It has gained popularity due to enabling systems to learn from data without explicit supervision, i.e. avoiding the cost of annotating large-scale datasets. The general pipeline of self-supervised learning algorithms consists of two phases: pretraining on large unlabeled data with pseudo labels and fine-tuning on labeled data for downstream tasks. Self-supervised algorithms have achieved promising results, and the performance gap with supervised algorithms in downstream tasks has significantly decreased. Asano et al. [6] showed that even on only a single image, self-supervised learning algorithms can surprisingly produce low-level characteristics that generalize well.

*equal contribution

Contrastive learning, a predominant technique in self-supervised learning, has notably bolstered performance in numerous downstream tasks such as classification, object detection, segmentation, and pose estimation, to name a few [35]. At its core, contrastive learning is about differentiating between similar and dissimilar samples. This technique leverages the principle of learning by comparison, enabling models to identify patterns and features in the data by contrasting positive (similar) and negative (dissimilar) examples. Contrastive learning objectives, such as the family of InfoNCE losses [16, 52], encourage similarity between representations of transformed versions of a data point while discouraging that between other data points. Already classic methods that match or even outperform supervised learning methods are SimCLR [16], MoCo v1 & v2 [32, 18], BYOL [25], SimSIAM [17], and many more.

Within this context, many recent empirical studies, which include various unsupervised contrastive representation learning approaches [53, 29, 67, 48, 51, 55, 34], have proposed to learn representations with a unit norm constraint. This constraint confines the output space to be on a unit hypersphere. Normalizing vectors has been shown to lead to more stable representation learning in various settings [70]. Although achieving state-of-the-art performances on a variety of downstream tasks, these methods seem to not fully exploit the geometric structure of the representation space. In this work, we propose *geometric contrastive loss*, a geometric interpretation to computing contrasts using geodesic distances. It improves on contrastive learning methods where geodesics lengths on the representation manifold can be computed. We argue that the geodesic distance is a natural choice to use as a similarity measure in the case of non-Euclidean representation manifolds with given metrics. In this work, we apply this idea to hyperspherical representation spaces. We focus on the latter, although the proposed method is not limited to hyperspheres.

We summarize the contribution of this work as follows :

1. We propose a *geometric* contrastive loss that utilizes geodesic distance on the hypersphere to measure contrasts between samples.

2. We empirically demonstrate the benefit of using geodesic distance in the objective of popular contrastive learning frameworks.

3. We validate our approach on both the curated balanced datasets as well as datasets following long-tail distributions.

4. We demonstrate that our proposed method works in the low-data domain and gives a comparable performance.

2. Related Work

2.1. Contrastive Learning

Self-supervised representation learning (SSL) from unlabeled visual data is a quickly evolving field. Recent methods are based on various forms of comparing embeddings between transformations of input images. This idea of making representations of an image agree with each other under small transformations, for example, the consecutive two-dimensional versions of a rotating three-dimensional object, dates back to Becker and Hinton [10]. Current methods in SSL can be divided into two categories: contrastive learning [32, 18, 52] and non-contrastive learning [7, 25, 14, 73, 17, 9, 24, 68, 33]. In this work, we focus our analysis on contrastive learning methods.

Contrastive learning methods employ instance discrimination to learn representations by forming positive pairs of images through augmentations and a loss formulation that maximizes their similarity while simultaneously minimizing the similarity to other samples, i.e. negative samples. The contrastive loss was first introduced by Bromley et al. [12] and then more formally defined in [19, 27]. Some contrastive learning methods have been motivated by the InfoMax principle [45] which maximizes the mutual information between two views of the same image, formed by some transformations such as cropping or color jittering [60, 8]. Tschannen et al. [62] shows that in practice, having a tighter lower bound on mutual information can lead to worse representations. Arora et al. [1] shows some theoretical insights on the representational capacity of contrastive loss frameworks with the number of negative pairs, although that is not consistent with the empirical results of these frameworks [60, 18, 32]. While the work developed so far aiming at understanding the behavior of SSL provides insights into its various aspects, they overlook prior geometric knowledge of representations lying on a well-known compact manifold, the hypersphere.

For ease of comparison, we focus on two well-known contrastive learning frameworks: SimCLR [16] and MoCo v2 [18] and show that our proposed method improves performances in these frameworks.

SimCLR SimCLR [16] learns representations by encouraging similarity between two augmented views of an image. Two views are formed by applying a series of transformations including random resizing, cropping, color jittering, and random blurring. After encoding each view, SimCLR uses a projector, often a multi-layer perception (MLP) followed by a ReLU activation, to map the initial embeddings into another space where the contrastive loss is applied to encourage similarity between the views. Given a minibatch of N example images, with augmentations, we have $2N$ data points. SimCLR does not sample negative examples: given a positive pair, the other $2(N-1)$ augmented examples within a minibatch are treated as negative examples. SimCLR demonstrated that simple end-to-end architectures with large batch sizes, a higher number of epochs, and a carefully chosen set of augmentations can perform well. The number of negative samples available in this approach is proportional to the batch size as it accumulates negative samples from the current batch. Since the batch size is limited by the GPU memory size, the scalability factor with these methods remains an issue.

MoCo v2 Momentum Contrast (MoCo) [32] learns visual representations by building a dynamic dictionary with a queue and a moving-averaged encoder. MoCo maintains the dictionary as a queue of data samples: the encoded representations of the current mini-batch are enqueued, and the oldest are removed. The queue decouples the dictionary size from the mini-batch size, allowing it to be large. Moreover, as the dictionary keys come from the preceding several mini-batches, a slowly progressing key encoder, is implemented as a momentum-based moving average of the query encoder. When SimCLR [16] introduced the use of a projector and stronger data augmentations, MoCo v2 [18] followed by implementing these design improvements to boost the performance of MoCo.

2.2. Geometry-aware representation learning

SSL is usually composed of the backbone encoder and the projector. The backbone encoder aims to encode data into a more compact, lower-dimensional representation. The manifold hypothesis states that in a high dimensional space, the data has a low dimensional nonlinear geometric structure. One way to compute distances that respects this structure is by using discrete shortest paths on neighborhood graphs [59]. Although, this strategy does not allow performing continuous analysis, as for example Riemannian statistics [54]. Therefore, methods based on latent variable models have been developed to enable computing continuous shortest paths.

Generative models provide a way to estimate the probability density of the given data lying in an ambient space. While most of the models utilize a latent space Z, the Vari-

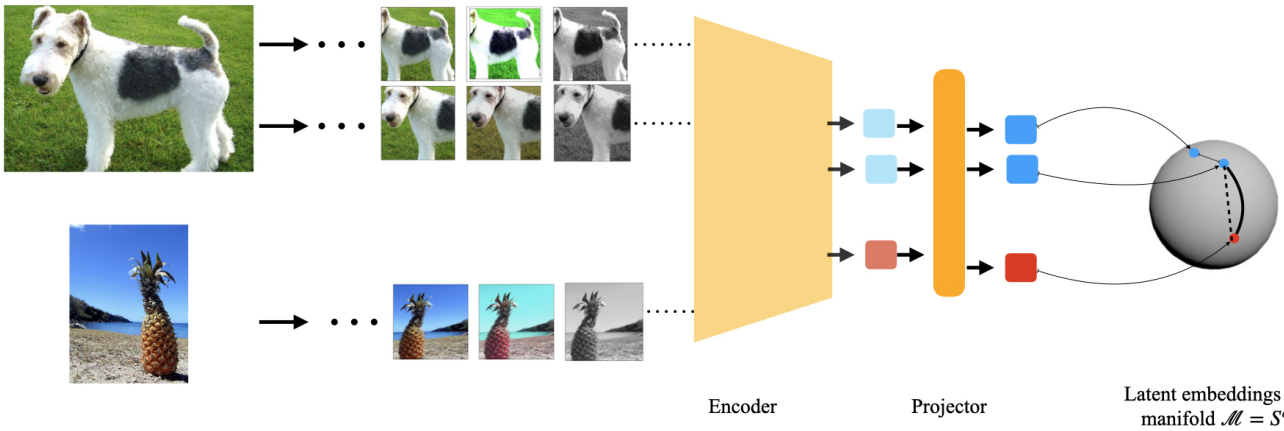

Figure 1: The pipeline for contrastive learning with latent embeddings that are constrained to lie on a $(d-1)$ dimension hypersphere (S^{d-1}). We show the embeddings of two different classes and their positive pairs generated by augmentations. The geodesic distance (depicted by a bold line) on a sphere is different from the straight line aka Euclidean distance (depicted by a dashed line) on a hypersphere.

ational Auto-Encoder (VAE) also learns a low dimensional representation of the data [40]. Unfortunately, using straight lines to compute distances in the latent space is misleading, and in addition, is not identifiable [2, 30].

Considering the latent space as a Riemannian manifold allows for encoding domain knowledge through the associated Riemannian metric, one solution is to compute shortest paths in $\mathcal{Z}$ using a Riemannian metric that is induced by the generator [61, 2]. This gives a natural and identifiable distance measure since it is actually computed directly on the data manifold in $\mathcal{X}$. However, we need to estimate meaningfully the generator's uncertainty $\mathcal{Z}$.

Arvanitidis et al. [3] proposed a fast, simple, and robust algorithm for computing shortest paths and distances on Riemannian manifolds learned from data. Arvanitidis et al. [4] considered the ambient space of generative models, in addition to latent space, as a Riemannian manifold. Arvanitidis et al. [5] captures the geometry of a data manifold in the latent space of a generative model using a Riemannian metric that is inversely proportional to a learnable prior.

2.3. Representation learning with Hyperspheres

Many representation learning approaches normalize their features such that they lie on a unit sphere [69, 16, 32, 8, 18, 60]. In latent variable models for representation learning like Autoencoders [46] and Variational Autoencoders [41], hyperspherical latent space have been shown to learn efficiently and outperform Euclidean latent space latent models [55, 63]. Intuitively, having the features live on the unit hypersphere leads to several favorable traits. Fixed-norm vectors improve training stability in modern machine learning where dot products are ubiquitous [70, 65]. Additionally, sufficiently well-clustered features of a class allow linear separability with the rest of the feature space, a common criterion used to evaluate representation quality. In contrastive learning, normalizing feature vectors is proven to optimize for alignment and uniformity in the latent representations [66] and these have led to improvement in downstream tasks as shown in [48, 51, 18, 32].

2.4. Imbalanced Self-Supervised Learning

Since natural data commonly follows long-tailed distributions [56, 67, 49] it is critical to address learning on imbalanced data instead of curated balanced datasets. Classical long-tail recognition approaches mainly attempt to amplify the impact of tail class samples, either by re-sampling the data distribution [15, 28, 58, 50] or re-weighting the loss for each class [20, 39, 13, 38]. However, in recent works, Kang et al. [37], Yang and Xu [71], Liu et al. [47], Zhong et al. [74], Gwilliam and Shrivastava [26] it is shown that self-supervised learning generally allows one to learn a more robust embedding space than a supervised counterpart.

3. Method

In this section, we discuss the preliminaries and introduce *geometric contrastive loss*, our proposed approach to learning contrasts in visual representation learning. We motivate the use of geodesic distances to learn contrasts on data manifolds. A template for the contrastive learning pipeline is shown in Fig. 1.

3.1. Contrastive Loss

The core idea with contrastive loss is to pull positive pairs closer while pushing negatives apart in the embedding space, thereby learning similarities and contrasts. The contrastive loss for training an encoder $f : \mathbb{R}^n \to \mathcal{S}^{d-1}$ maps data to l_2 normalized feature vectors of dimension d. For this, we use the Info-NCE loss which leads to learning useful representations from unlabelled data [52, 69, 32, 16].

Given a set of inputs $x_1, ..., x_N$, a similarity measure $s_{ij} = sim(z_i, z_j)$ between learned representations $z_i = f(\mathcal{A}(x_i))$ and $z_j = f(\mathcal{A}(x_j))$, the loss is defined by

$$\mathcal{L}_{contr.} = \sum_{i=1}^{N} -\log \frac{\exp(s_{ii}/\tau)}{\exp(s_{ii}/\tau) + \sum_{i \neq j}^{K} \exp(s_{ij}/\tau)} \quad (1)$$

where $\mathcal{A}(\cdot)$ is the set of random augmentations applied to its input, K is a fixed number of negative samples, and the $\tau > 0$ is the temperature of the Info-NCE loss and has been found to crucially impact the representation learning [66, 64, 57, 43]. We refer to z_i as an anchor, to z_j as a positive example if $i = j$ and as a negative example if $i \neq j$.

3.2. Geometric Contrastive Loss

In this work, we propose to model the latent space as a Riemannian manifold. By doing so, we are able to quantify the notion of contrast between any two points by a distance that is informed by the Riemannian metric. In fact, from the Riemannian metric structure stem geometric properties like curvature, symmetries, and, most importantly for our method, distances. Formally, let us denote a Riemannian manifold with $\mathcal{M}$, and the tangent space at a certain location $z \in \mathcal{M}$ in with $T_z(\mathcal{M})$. $\mathcal{M}$ is equipped with a metric tensor (inner product) $g_z : T_z(\mathcal{M}) \times T_z(\mathcal{M}) \to \mathbb{R}$ on the tangent spaces $T_z(\mathcal{M})$ that can be used to quantify the length of tangent vectors $v \in T_z(\mathcal{M})$ via $\|v\|_{g_z} := \sqrt{g_z(v, v)}$. The metric tensor allows us to quantify the lengths of curves via

$$L(\gamma) = \int_0^1 \|\dot{\gamma}(t)\|_{g_{\gamma(t)}} \mathrm{d}t \,,$$

i.e., as the integral of the tangent vector lengths along the curve. This introduces a notion of a distance $d(z_0, z_1)$ between any two points $z_0, z_1 \in \mathcal{M}$, as the length of the shortest possible path $\gamma : [0, 1] \to \mathcal{M}$ connecting these points, i.e.

$$d(z_0, z_1) = \inf_{\gamma} L(\gamma) \quad \text{s.t. } z_0 = \gamma(0), z_1 = \gamma(1). \quad (2)$$

The length-minimizing curves of (2) are called *geodesics*. Figure 2 illustrates the difference between geodesic and Euclidean distances on a manifold. For many known manifolds, the distances and geodesics have closed-form solutions and thus do not need to be found by solving (2). For example, the distance between any two points $z_i, z_j \in S^d$ on a hypersphere $M = S^3$ is given by

$$d(z_i, z_j) = \arccos(z_i^T z_j). \quad (3)$$

As a trivial example, the Euclidean space $\mathbb{R}^n$ is naturally equipped with the Euclidean metric $g_e = id_n$ which results in the geodesic distance between two points is simply the length of the straight line that connects them.

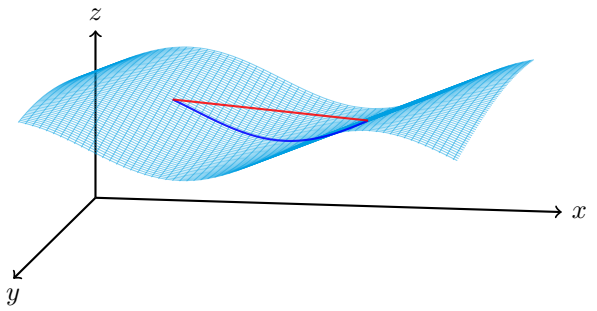

Figure 2: The geodesic (blue) distance vs Euclidean (red) distance on a toy curved manifold surface.

A vast number of recent unsupervised contrastive representation learning methods learn representations with a unit l_2 norm constraint [69, 8, 60, 16]. This results in features being on the *spherically symmetric* hypersphere $\mathcal{S}^{n-1} \subset \mathbb{R}^n$. Given two points $z_1, z_2 \in \mathcal{S}^{n-1}$, the spherically symmetric hypersphere is equipped with a metric that results in the geodesic distance $d(z_1, z_2) = \arccos(\langle z_1, z_2 \rangle)$ where $\langle \cdot, \cdot \rangle$ is the standard inner product in $\mathbb{R}^n$. For representations constrained on a hypersphere, it is more natural to consider the geodesic distance as a similarity measure. For two features z_i, z_j on the unit sphere, we, therefore, use the negative geodesic distance between them as the similarity measure.

$$sim(z_i, z_j) = 1 - \arccos(\langle z_i, z_j \rangle)/\pi \quad (4)$$

For a different non-Euclidean embedding, the respective geodesic distance can be considered. In general, though, a closed form of the geodesic distance is not available.

3.3. Distinction between $\arccos$ and cosine similarity

Cosine similarity between two unnormalized vectors $u, v \in \mathbb{R}^d$ is given by

$$sim(u, v) = \frac{\langle u, v \rangle}{||u|| \cdot ||v||} \quad (5)$$

It gives a similarity measure between two vectors in Euclidean space $\mathbb{R}^d$ and ranges between -1 and 1. This can be interpreted as the cosine of the angle between the two vectors in $\mathbb{R}^d$. In our method, we take the $\arccos$ between two vectors, the geodesic distance on the unit sphere. The main difference between these two measures is when two vectors

are close to each other. For small angles, the approximation of $cosine$ is $cos(\theta) = 1 - \frac{\theta^2}{2}$ in contrast to geodesic distance θ, which makes the model more difficult to learn the contrast between two similar vectors. We can observe it in our experiments in Sec. 5.3.

4. Experimental Setup

In this chapter, we provide details of our experimental setup. We first introduce datasets in Sec. 4.1. Sec. 4.2 describes the contrastive methods we use as baselines. Finally, in Sec. 4.3 and Sec. 4.4 we provide training and evaluation protocols.

4.1. Datasets

We run experiments on different datasets commonly used for contrastive learning frameworks. We divide these datasets into three categories: balanced, long-tail, and low-diversity datasets. Balanced datasets, as the name suggests, are multiclass datasets with an equal number of examples for each class. Long-tail datasets are the datasets that follow a long-tail distribution while the low-diversity datasets contain classes that are difficult to classify because they belong to the same parent class.

Balanced datasets For balanced datasets, we consider the following well-known datasets `CIFAR10`, `CIFAR100`, and `ImageNet100`. `CIFAR10` [42] is the dataset that consists of 60000 color images of size 32×32 in 10 classes with 6000 images per class. The train/test split is 50000/10000. `CIFAR100` [42] is like `CIFAR10` dataset, except it has 100 classes containing 600 images each. `ImageNet100` [60] is a subset of the original `ImageNet` [22] consisting of 100 classes for a total of 12210 images.

Long-tail datasets For long-tailed (LT) datasets we consider LT versions of the above-mentioned datasets for the experiments: `CIFAR10-LT`, `CIFAR100-LT`, and `ImageNet100-LT`. Long-tail versions of the datasets were introduced by Cui et al. [20] and consist of a subset of the original datasets with an exponential decay in the number of images per class. The imbalance ratio controls the uniformity of the dataset and is calculated as the ratio of the sizes of the biggest and the smallest classes. As a standard practice, we use an imbalance ratio of 100 if not stated otherwise [36].

***Low-diversity* datasets** As a *low-diversity* dataset, we use the `Imagewoof` dataset introduced by fast.ai. `Imagewoof` is a subset of 10 dog breed classes from `ImageNet`[22]. We consider this dataset as a difficult dataset compared to the other datasets, as all images belong to the same ancestor in the ImageNet hierarchy, which is the dog in our case.

https://github.com/fastai/imagenette

4.2. Baselines

We evaluate the effect of using negative geodesic distance as a similarity measure (or geodesic distance for learning contrasts) in two standard contrastive methods: SimCLR and MoCo v2. SimCLR and MoCo v2 are self-supervised learning frameworks to learn representations from unlabeled data. We set ResNet18 as an encoder backbone [31], the projection head in the pipeline is set up with an output size of 128, and temperature values in the loss function are set to $\tau = 0.05$ for all baselines and datasets.

4.3. Training

For all experiments, we build our setup off of the implementation of the baseline models from the Solo-Learn library [21]. Our geometric contrastive loss is agnostic to self-supervised learning frameworks and their related training components. Therefore we keep the same training settings when making comparisons. It is fair to assume that larger gains could be expected with further hyperparameter tuning in on our experiments, but for the current work, we just show the improvement for the pre-set hyperparameters.

We use the same experimental setup for pairs `CIFAR10`/`CIFAR100` and `ImageNet100`/`Imagewoof` datasets, so we will refer to them as `CIFAR` and `Imagenet` datasets. We train `CIFAR` and `Imagenet` for 1000 and 400 epochs. For the experiments with long-tail versions of the datasets, we reduce the number of epochs. We keep it to 250 epochs for `CIFAR10-LT`/`CIFAR100-LT` and 100 epochs for `ImageNet100-LT`. We train with a batch size of 512. As for the learning rate, we utilize linear warm-up for 10 epochs which is followed by a cosine annealing schedule.

SimCLR: we follow Chen et al. [16] to choose hyperparameters and use the LARS optimizer [72] for all SimCLR experiments with a weight decay of 1e-4, LARS coefficient of 0.4/0.3 for `CIFAR` and `Imagenet` datasets.

MoCo v2: we use an SGD optimizer for all MoCo v2 experiments with a weight decay of 1e-4. We use a dictionary of size 4096 for `CIFAR10`, `CIFAR100`, `Imagewoof` and 8192 for `ImageNet100`.

4.4. Evaluation

We use linear classification as well as the k nearest neighbors (kNN) to assess the features learned through the contrastive framework. For kNN, we compute $l2$-normalized distances between samples from the train set and the test set. For each test image, we assign it to the majority class among the top-k closest train images. We report accuracy for kNN with $k = 1$ (kNN@1) and $k = 10$

Table 1: **Effect of geometric contrastive loss (GCL) on balanced datasets.** Comparison of top-1 accuracy of SimCLR vs SimCLR + GCL and MoCo v2 vs MoCo v2 + GCL on `CIFAR10`, `CIFAR100`, and `ImageNet100` with kNN@1, kNN@10, and linear probe (LP).

Method	CIFAR-10			CIFAR-100			ImageNet-100		
	kNN@1	kNN@10	LP	kNN@1	kNN@10	LP	kNN@1	kNN@10	LP
SimCLR	80.03	83.55	86.57	47.14	45.04	58.76	58.70	64.40	74.05
SimCLR + GCL	**87.50**	**89.61**	**90.53**	**57.39**	**62.27**	**63.92**	**73.35**	**76.75**	**80.05**
MoCo v2	79.25	82.95	86.68	48.59	54.01	61.53	55.75	63.20	74.80
MoCo v2 + GCL	**89.70**	**91.67**	**92.55**	**61.47**	**65.70**	**68.61**	**75.25**	**78.90**	**83.15**

Table 2: **Effect of geometric contrastive loss (GCL) on long-tail datasets.** Comparison of top-1 accuracy of SimCLR vs SimCLR + GCL and MoCo v2 vs MoCo v2 + GCL on CIFAR10-LT, CIFAR100-LT, and ImageNet100-LT with kNN@1, kNN@10, and linear probe (LP).

Method	CIFAR-10-LT			CIFAR-100-LT			ImageNet-100-LT		
	kNN@1	kNN@10	LP	kNN@1	kNN@10	LP	kNN@1	kNN@10	LP
SimCLR	45.17	45.04	58.48	15.23	16.74	26.55	10.8	12.25	25.45
SimCLR + GCL	**52.76**	**51.28**	**59.67**	**19.52**	**19.46**	**28.57**	**15.1**	**18.35**	**33.05**
MoCo v2	46.64	45.69	56.53	17.6	19.11	27.46	12.20	13.90	26.00
MoCo v2 + GCL	**50.51**	**49.22**	**58.07**	**20.58**	**20.97**	**28.87**	**12.85**	**14.55**	**26.45**

(kNN@10) as well. Compared to linear probing, kNN directly evaluates the learned embedding since it relies on the learned metric and local structure of the space. We evaluate the linear separability and generalization of the space with linear probing. We train a linear classifier on top of the frozen pre-trained model. Linear evaluation is done simply by appending a linear layer at the end of the frozen backbone encoder. A linear classifier is trained for 100 epochs with an initial learning rate of 10.0 multiplied by 0.1 at the $60th$ and $80th$ epochs.

5. Results

5.1. Results on balanced data

In Table 1 we present the efficacy of geometric contrastive loss for SimCLR and MoCo v2. We find that both frameworks benefit from a negative geodesic distance as a similarity measure and we observe consistent improvements in all evaluation metrics for `CIFAR10`, `CIFAR100`, and `ImageNet100`, i.e. the local structure of the embedding space (kNN) and the global structure (linear probe) are both improved.

The improvement is more noticeable as the difficulty of the task increases. For example, SimCLR linear probe performance increases by 7.47%, 10.25%, and 14.65% for `CIFAR10`, `CIFAR100`, `ImageNet100`. Similarly, MoCo v2 improves by 10.45%, 12.88%, and 19.5% respectively.

5.2. Results on long-tail data

In Table 2 we present the effect of using geometric contrastive loss in contrastive learning frameworks on the datasets with a long-tail distribution. We can observe consistent improvements over the baselines, SimCLR and MoCo v2, for `CIFAR10-LT`, `CIFAR100-LT`, `ImageNet100-LT` datasets, and evaluation metrics (kNN@1, kNN@10, LP).

5.3. Qualitative analysis of similarity

In addition to the well-known evaluation metrics, we show how the geodesic distances compare with cosine similarities in learning contrasts for positive and negative pairs. To do so, for each example in the test set of `CIFAR100` dataset, we measure and save the distances between positive examples, i.e. its two augmentations, and between negative examples, i.e. the anchor and another random example. Then, we plot the histogram of measured distances between positive and negative examples. For both methods, we measure the distances they were trained on and normalize geodesic distance to have values between -1 and 1 to compare it with standard cosine similarity. In Figure 3 we show the results through a distance plot. We can observe that in both figures the distribution of positive (in blue) and negative examples (in red) are well separated. The main difference is that the method utilizing geodesic distance is more sensitive to differences between positive examples.

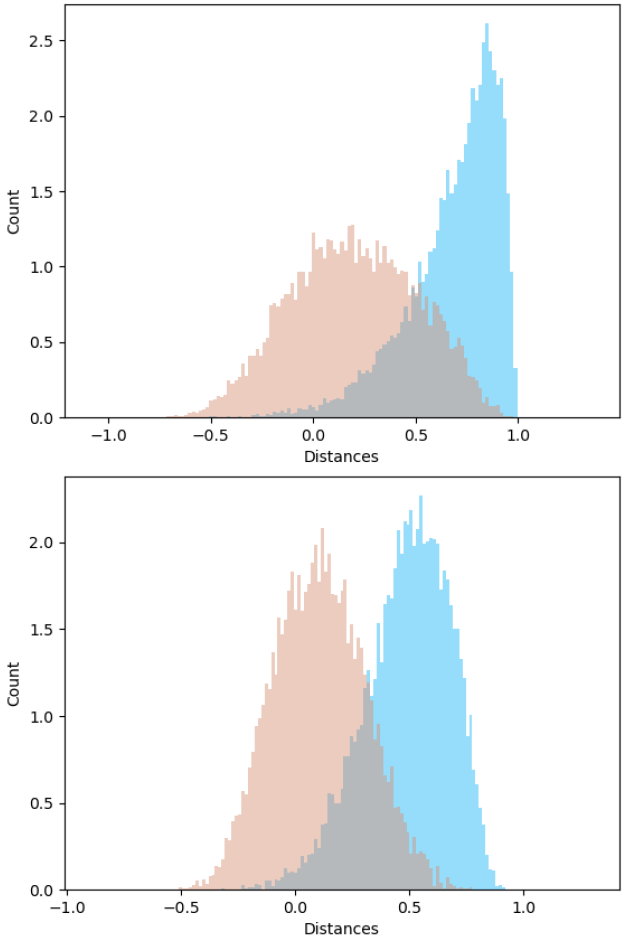

Figure 3: The histogram of distances of positive (in blue) and negative (in red) examples CIFAR100 dataset for simCLR (top) and simCLR + geodesic distance (bottom).

5.4. Results on low-diversity datasets

In Table 3 we present the performance of SimCLR and MoCo v2 with the geometric contrastive loss on the *low-diversity* dataset, the dataset of images with the same ancestor class in the ImageNet hierarchy. We see that utilizing negative geodesic distance boosts performance in this scenario too.

5.5. On the data efficiency of Geometric Contrastive Loss

Figure 4 shows the linear classifier accuracy of SimCLR and MoCo v2 trained on different fractions of the CIFAR100 training set, namely 40%, 60%, and 80%. We see that geometric contrastive loss outperforms the baseline in these scenarios. Additionally, we note that the difference is increasing with a larger training dataset size.

Table 3: Effect of geometric contrastive loss (GCL). Comparison of SimCLR with SimCLR + GCL and MoCo v2 with MoCo v2 + GCL on Imagewoof dataset with evaluation metric kNN@1, kNN@10, and linear probe (LP).

Method	Imagewoof kNN@1	Imagewoof kNN@10	Imagewoof LP
SimCLR	60.91	68.16	75.01
SimCLR+GCL	**70.81**	**74.55**	**78.26**
MoCo v2	60.45	67.27	75.52
MoCo v2+GCL	**63.76**	**71.16**	**76.38**

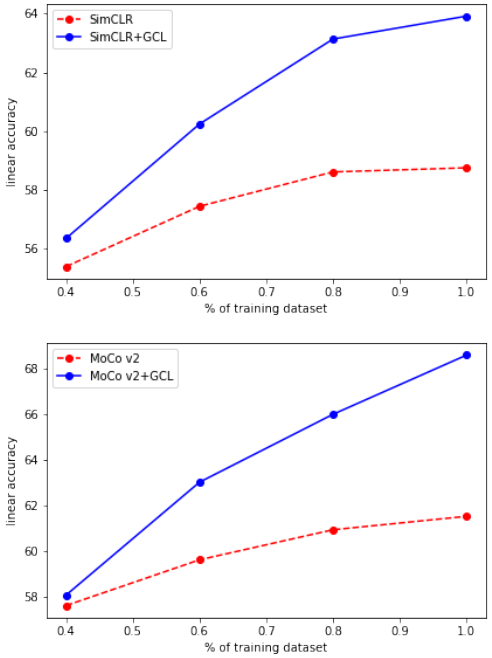

Figure 4: Efficacy of geometric contrastive loss (GCL) on SimCLR and MoCo v2 trained on different fractions of CIFAR100 dataset.

6. Conclusion

In this paper, we introduce a new perspective on the classical contrastive learning framework drawing inspiration from the recent success of hyperspherical embeddings and their applications in various representation learning domains. We propose a simple, yet effective geometric contrastive loss that utilizes a similarity measure based on the geodesic distance of the representation manifold. By evaluating our method on different datasets, we demonstrate its promising performance in a range of downstream tasks. We further demonstrate how this simple change in the similarity measure used in the contrastive objective distributes the dis-

tances between positive and negative pairs. The simplicity of the proposed method makes it model-agnostic and can be used in other contrastive learning frameworks beyond SimCLR and MoCo v2. Notably, our approach is applicable to any representation manifold as long as geodesic distances can be computed efficiently. As a potential future research direction, this method could be combined with Riemannian metric learning methods such as [44, 11, 23], to learn the metric and utilize geodesic distances computed using the learned metric.

Acknowledgements This work is part of the research program VENI with project "context-aware AI" with number 17290, which is (partly) financed by the Dutch Research Council (NWO). Sharvaree Vadgama is funded by the Hybrid Intelligence Center, a 10-year program funded by the Dutch Ministry of Education, Culture, and Science through the Netherlands Organisation for Scientific Research.

References

[1] S. Arora, H. Khandeparkar, M. Khodak, O. Plevrakis, and N. Saunshi. A theoretical analysis of contrastive unsupervised representation learning, 2019.

[2] G. Arvanitidis, L. K. Hansen, and S. Hauberg. Latent space oddity: on the curvature of deep generative models. *arXiv preprint arXiv:1710.11379*, 2017.

[3] G. Arvanitidis, S. Hauberg, P. Hennig, and M. Schober. Fast and robust shortest paths on manifolds learned from data. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1506–1515. PMLR, 2019.

[4] G. Arvanitidis, S. Hauberg, and B. Schölkopf. Geometrically enriched latent spaces, 2020.

[5] G. Arvanitidis, B. Georgiev, and B. Schölkopf. A prior-based approximate latent riemannian metric. *arXiv preprint arXiv:2103.05290*, 2021.

[6] Y. M. Asano, C. Rupprecht, and A. Vedaldi. A critical analysis of self-supervision, or what we can learn from a single image. *arXiv preprint arXiv:1904.13132*, 2019.

[7] Y. M. Asano, C. Rupprecht, and A. Vedaldi. Self-labelling via simultaneous clustering and representation learning. *arXiv preprint arXiv:1911.05371*, 2019.

[8] P. Bachman, R. D. Hjelm, and W. Buchwalter. Learning representations by maximizing mutual information across views, 2019.

[9] A. Bardes, J. Ponce, and Y. LeCun. Vicreg: Variance-invariance-covariance regularization for self-supervised learning. *arXiv preprint arXiv:2105.04906*, 2021.

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

[11] H. Beik-Mohammadi, S. Hauberg, G. Arvanitidis, G. Neumann, and L. Rozo. Learning riemannian manifolds for geodesic motion skills. *arXiv preprint arXiv:2106.04315*, 2021.

[12] J. Bromley, I. Guyon, Y. LeCun, E. Säckinger, and R. Shah. Signature verification using a" siamese" time delay neural network. *Advances in neural information processing systems*, 6, 1993.

[13] K. Cao, C. Wei, A. Gaidon, N. Arechiga, and T. Ma. Learning imbalanced datasets with label-distribution-aware margin loss. *Advances in neural information processing systems*, 32, 2019.

[14] M. Caron, I. Misra, J. Mairal, P. Goyal, P. Bojanowski, and A. Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *Advances in neural information processing systems*, 33:9912–9924, 2020.

[15] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002.

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

[17] X. Chen and K. He. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15750–15758, 2021.

[18] X. Chen, H. Fan, R. Girshick, and K. He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020.

[19] S. Chopra, R. Hadsell, and Y. LeCun. Learning a similarity metric discriminatively, with application to face verification. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 1, pages 539–546. IEEE, 2005.

[20] Y. Cui, M. Jia, T.-Y. Lin, Y. Song, and S. Belongie. Class-balanced loss based on effective number of samples. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9268–9277, 2019.

[21] V. G. T. Da Costa, E. Fini, M. Nabi, N. Sebe, and E. Ricci. solo-learn: A library of self-supervised methods for visual representation learning. *J. Mach. Learn. Res.*, 23(56):1–6, 2022.

[22] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[23] N. S. Detlefsen, S. Hauberg, and W. Boomsma. Learning meaningful representations of protein sequences. *Nature communications*, 13(1):1914, 2022.

[24] S. Gidaris, A. Bursuc, G. Puy, N. Komodakis, M. Cord, and P. Pérez. Obow: Online bag-of-visual-words generation for self-supervised learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6830–6840, 2021.

[25] J.-B. Grill, F. Strub, F. Altché, C. Tallec, P. Richemond, E. Buchatskaya, C. Doersch, B. Avila Pires, Z. Guo, M. Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33:21271–21284, 2020.

[26] M. Gwilliam and A. Shrivastava. Beyond supervised vs. unsupervised: Representative benchmarking and analysis of image representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9642–9652, 2022.

[27] R. Hadsell, S. Chopra, and Y. LeCun. Dimensionality reduction by learning an invariant mapping. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, pages 1735–1742. IEEE, 2006.

[28] H. Han, W.-Y. Wang, and B.-H. Mao. Borderline-smote: a new over-sampling method in imbalanced data sets learning. In *Advances in Intelligent Computing: International Conference on Intelligent Computing, ICIC 2005, Hefei, China, August 23-26, 2005, Proceedings, Part I 1*, pages 878–887. Springer, 2005.

[29] M. A. Hasnat, J. Bohné, J. Milgram, S. Gentric, and L. Chen. von mises-fisher mixture model-based deep learning: Application to face verification, 2017.

[30] S. Hauberg. Only bayes should learn a manifold (on the estimation of differential geometric structure from data). *arXiv preprint arXiv:1806.04994*, 2018.

[31] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[32] 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*, pages 9729–9738, 2020.

[33] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16000–16009, 2022.

[34] S. Hou, X. Pan, C. C. Loy, Z. Wang, and D. Lin. Learning a unified classifier incrementally via rebalancing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[35] A. Jaiswal, A. R. Babu, M. Z. Zadeh, D. Banerjee, and F. Makedon. A survey on contrastive self-supervised learning, 2021.

[36] Z. Jiang, T. Chen, B. J. Mortazavi, and Z. Wang. Self-damaging contrastive learning. In *International Conference on Machine Learning*, pages 4927–4939. PMLR, 2021.

[37] B. Kang, Y. Li, S. Xie, Z. Yuan, and J. Feng. Exploring balanced feature spaces for representation learning. In *International Conference on Learning Representations*, 2021.

[38] S. Khan, M. Hayat, S. W. Zamir, J. Shen, and L. Shao. Striking the right balance with uncertainty. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 103–112, 2019.

[39] S. H. Khan, M. Hayat, M. Bennamoun, F. A. Sohel, and R. Togneri. Cost-sensitive learning of deep feature representations from imbalanced data. *IEEE transactions on neural networks and learning systems*, 29(8):3573–3587, 2017.

[40] D. P. Kingma and M. Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[41] D. P. Kingma and M. Welling. Auto-encoding variational bayes, 2022.

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

[43] A. Kukleva, M. Böhle, B. Schiele, H. Kuehne, and C. Rupprecht. Temperature schedules for self-supervised contrastive methods on long-tail data. *arXiv preprint arXiv:2303.13664*, 2023.

[44] G. Lebanon. Learning riemannian metrics. *arXiv preprint arXiv:1212.2474*, 2012.

[45] R. Linsker. Self-organization in a perceptual network. *Computer*, 21(3):105–117, 1988. doi: 10.1109/2.36.

[46] C.-Y. Liou, W.-C. Cheng, J.-W. Liou, and D.-R. Liou. Autoencoder for words. *Neurocomputing*, 139:84–96, 2014.

[47] H. Liu, J. Z. HaoChen, A. Gaidon, and T. Ma. Self-supervised learning is more robust to dataset imbalance. *arXiv preprint arXiv:2110.05025*, 2021.

[48] W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj, and L. Song. Sphereface: Deep hypersphere embedding for face recognition, 2018.

[49] Z. Liu, Z. Miao, X. Zhan, J. Wang, B. Gong, and S. X. Yu. Large-scale long-tailed recognition in an open world. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2537–2546, 2019.

[50] D. Mahajan, R. Girshick, V. Ramanathan, K. He, M. Paluri, Y. Li, A. Bharambe, and L. Van Der Maaten. Exploring the limits of weakly supervised pretraining. In *Proceedings of the European conference on computer vision (ECCV)*, pages 181–196, 2018.

[51] P. Mettes, E. van der Pol, and C. G. M. Snoek. Hyperspherical prototype networks, 2019.

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

[53] O. M. Parkhi, A. Vedaldi, and A. Zisserman. Deep face recognition. In X. X. M. W. Jones and G. K. L. Tam, editors, *Proceedings of the British Machine Vision Conference (BMVC)*, pages 41.1–41.12. BMVA Press, September 2015. ISBN 1-901725-53-7. doi: 10.5244/C.29.41. URL https://dx.doi.org/10.5244/C.29.41.

[54] X. Pennec. Intrinsic statistics on riemannian manifolds: Basic tools for geometric measurements. *Journal of Mathematical Imaging and Vision*, 25:127–154, 2006.

[55] D. T. R, F. Luca, D. C. Nicola, K. Thomas, and T. J. M. Hyperspherical variational auto-encoders. *arXiv preprint arXiv:1804.00891*, 2018.

[56] W. J. Reed. The pareto, zipf and other power laws. *Economics letters*, 74(1):15–19, 2001.

[57] J. Robinson, L. Sun, K. Yu, K. Batmanghelich, S. Jegelka, and S. Sra. Can contrastive learning avoid shortcut solutions? *Advances in neural information processing systems*, 34:4974–4986, 2021.

[58] L. Shen, Z. Lin, and Q. Huang. Relay backpropagation for effective learning of deep convolutional neural networks. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part VII 14*, pages 467–482. Springer, 2016.

[59] J. B. Tenenbaum, V. d. Silva, and J. C. Langford. A global geometric framework for nonlinear dimensionality reduction. *science*, 290(5500):2319–2323, 2000.

[60] Y. Tian, D. Krishnan, and P. Isola. Contrastive multiview coding, 2020.

[61] A. Tosi, S. Hauberg, A. Vellido, and N. D. Lawrence. Metrics for probabilistic geometries. *arXiv preprint arXiv:1411.7432*, 2014.

[62] M. Tschannen, J. Djolonga, P. K. Rubenstein, S. Gelly, and M. Lucic. On mutual information maximization for representation learning, 2020.

[63] S. Vadgama, J. M. Tomczak, and E. J. Bekkers. Kendall shape-VAE : Learning shapes in a generative framework. In *NeurIPS 2022 Workshop on Symmetry and Geometry in Neural Representations*, 2022. URL https://openreview.net/forum?id=nzh4N6kdl2G.

[64] F. Wang and H. Liu. Understanding the behaviour of contrastive loss. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2495–2504, 2021.

[65] F. Wang, X. Xiang, J. Cheng, and A. L. Yuille. Normface: L2 hypersphere embedding for face verification. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 1041–1049, 2017.

[66] T. Wang and P. Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *International Conference on Machine Learning*, pages 9929–9939. PMLR, 2020.

[67] Y.-X. Wang, D. Ramanan, and M. Hebert. Learning to model the tail. *Advances in neural information processing systems*, 30, 2017.

[68] C. Wei, H. Fan, S. Xie, C.-Y. Wu, A. Yuille, and C. Feichtenhofer. Masked feature prediction for self-supervised visual pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14668–14678, 2022.

[69] Z. Wu, Y. Xiong, S. X. Yu, and D. Lin. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3733–3742, 2018.

[70] J. Xu and G. Durrett. Spherical latent spaces for stable variational autoencoders, 2018.

[71] Y. Yang and Z. Xu. Rethinking the value of labels for improving class-imbalanced learning. *Advances in neural information processing systems*, 33:19290–19301, 2020.

[72] Y. You, I. Gitman, and B. Ginsburg. Large batch training of convolutional networks. *arXiv preprint arXiv:1708.03888*, 2017.

[73] J. Zbontar, L. Jing, I. Misra, Y. LeCun, and S. Deny. Barlow twins: Self-supervised learning via redundancy reduction. In *International Conference on Machine Learning*, pages 12310–12320. PMLR, 2021.

[74] Y. Zhong, H. Tang, J. Chen, J. Peng, and Y.-X. Wang. Is self-supervised learning more robust than supervised learning? *arXiv preprint arXiv:2206.05259*, 2022.