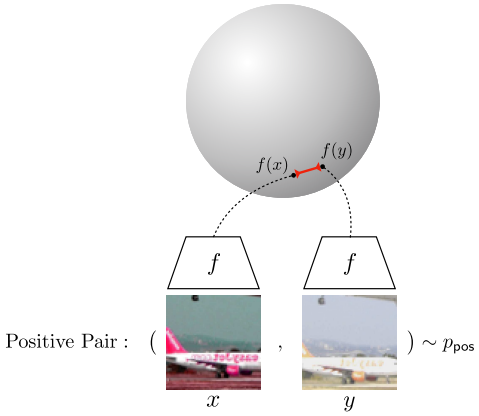
Understanding Contrastive Representation Learning through Alignment and Uniformity on the Hypersphere

Tongzhou Wang [1] **Phillip Isola** [1]

Abstract

Contrastive representation learning has been outstandingly successful in practice. In this work, we identify two key properties related to the contrastive loss: (1) *alignment* (closeness) of features from positive pairs, and (2) *uniformity* of the induced distribution of the (normalized) features on the hypersphere. We prove that, asymptotically, the contrastive loss optimizes these properties, and analyze their positive effects on downstream tasks. Empirically, we introduce an optimizable metric to quantify each property. Extensive experiments on standard vision and language datasets confirm the strong agreement between *both* metrics and downstream task performance. Directly optimizing for these two metrics leads to representations with comparable or better performance at downstream tasks than contrastive learning.

Project Page: `ssnl.github.io/hypersphere`.
Code: `github.com/SsnL/align_uniform`.



Alignment: Similar samples have similar features.
(Figure inspired by Tian et al. (2019).)

Uniformity: Preserve maximal information.

Figure 1: Illustration of alignment and uniformity of feature distributions on the output unit hypersphere. STL-10 (Coates et al., 2011) images are used for demonstration.

1. Introduction

A vast number of recent empirical works learn representations with a unit ℓ_2 norm constraint, effectively restricting the output space to the unit hypersphere (Parkhi et al., 2015; Schroff et al., 2015; Liu et al., 2017; Hasnat et al., 2017; Wang et al., 2017; Bojanowski & Joulin, 2017; Mettes et al., 2019; Hou et al., 2019; Davidson et al., 2018; Xu & Durrett, 2018), including many recent unsupervised contrastive representation learning methods (Wu et al., 2018; Bachman et al., 2019; Tian et al., 2019; He et al., 2019; Chen et al., 2020).

Intuitively, having the features live on the unit hypersphere leads to several desirable traits. Fixed-norm vectors are known to improve training stability in modern machine learning where dot products are ubiquitous (Xu & Durrett, 2018; Wang et al., 2017). Moreover, if features of a class are sufficiently well clustered, they are linearly separable with the rest of feature space (see Figure 2), a common criterion used to evaluate representation quality.

While the unit hypersphere is a popular choice of feature space, not all encoders that map onto it are created equal. Recent works argue that representations should additionally be invariant to unnecessary details, and preserve as much information as possible (Oord et al., 2018; Tian et al., 2019; Hjelm et al., 2018; Bachman et al., 2019). Let us call these two properties *alignment* and *uniformity* (see

[1]MIT Computer Science & Artificial Intelligence Lab (CSAIL). Correspondence to: Tongzhou Wang <tongzhou@mit.edu>.

Proceedings of the 37^{th} International Conference on Machine Learning, Online, PMLR 119, 2020.

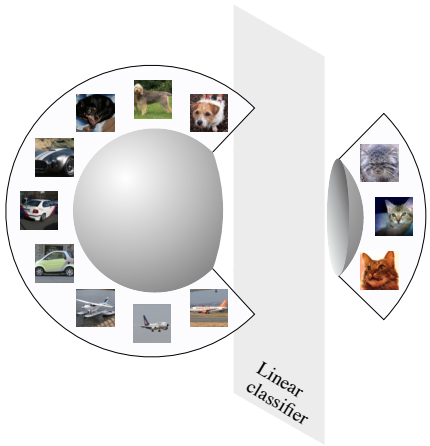


Figure 2: **Hypersphere:** When classes are well-clustered (forming spherical caps), they are linearly separable. The same does not hold for Euclidean spaces.

Figure 1). *Alignment* favors encoders that assign similar features to similar samples. *Uniformity* prefers a feature distribution that preserves maximal information, i.e., the uniform distribution on the unit hypersphere.

In this work, we analyze the *alignment* and *uniformity* properties. We show that a currently popular form of contrastive representation learning in fact directly optimizes for these two properties in the limit of infinite negative samples. We propose theoretically-motivated metrics for alignment and uniformity, and observe strong agreement between them and downstream task performance. Remarkably, directly optimizing for these two metrics leads to comparable or better performance than contrastive learning.

Our main contributions are:

- We propose quantifiable metrics for *alignment* and *uniformity* as two measures of representation quality, with theoretical motivations.
- We prove that the contrastive loss optimizes for alignment and uniformity asymptotically.
- Empirically, we find strong agreement between *both* metrics and downstream task performance.
- Despite being simple in form, our proposed metrics, when directly optimized with no other loss, empirically lead to comparable or better performance at downstream tasks than contrastive learning.

2. Related Work

Unsupervised Contrastive Representation Learning has seen remarkable success in learning representations for image and sequential data (Logeswaran & Lee, 2018; Wu et al., 2018; Oord et al., 2018; Hénaff et al., 2019; Tian et al., 2019; Hjelm et al., 2018; Bachman et al., 2019; Tian et al., 2019; He et al., 2019; Chen et al., 2020). The common motivation behind these work is the InfoMax principle (Linsker, 1988), which we here instantiate as maximizing the mutual information (MI) between two views (Tian et al., 2019; Bachman et al., 2019; Wu et al., 2020). However, this interpretation is known to be inconsistent with the actual behavior in practice, e.g., optimizing a tighter bound on MI can lead to worse representations (Tschannen et al., 2019). What the contrastive loss exactly does remains largely a mystery. Analysis based on the assumption of latent classes provides nice theoretical insights (Saunshi et al., 2019), but unfortunately has a rather large gap with empirical practices: the result that representation quality suffers with a large number of negatives is inconsistent with empirical observations (Wu et al., 2018; Tian et al., 2019; He et al., 2019; Chen et al., 2020). In this paper, we analyze and characterize the behavior of contrastive learning from the perspective of alignment and uniformity properties, and empirically verify our claims with standard representation learning tasks.

Representation learning on the unit hypersphere. Outside contrastive learning, many other representation learning approaches also normalize their features to be on the unit hypersphere. In variational autoencoders, the hyperspherical latent space has been shown to perform better than the Euclidean space (Xu & Durrett, 2018; Davidson et al., 2018). Directly matching uniformly sampled points on the unit hypersphere is known to provide good representations (Bojanowski & Joulin, 2017), agreeing with our intuition that uniformity is a desirable property. Mettes et al. (2019) optimizes prototype representations on the unit hypersphere for classification. Hyperspherical face embeddings greatly outperform the unnormalized counterparts (Parkhi et al., 2015; Liu et al., 2017; Wang et al., 2017; Schroff et al., 2015). Its empirical success suggests that the unit hypersphere is indeed a nice feature space. In this work, we formally investigate the interplay between the hypersphere geometry and the popular contrastive representation learning.

Distributing points on the unit hypersphere. The problem of uniformly distributing points on the unit hypersphere is a well-studied one. It is often defined as minimizing the total pairwise potential w.r.t. a certain kernel function (Borodachov et al., 2019; Landkof, 1972), e.g., the Thomson problem of finding the minimal electrostatic potential energy configuration of electrons (Thomson, 1904), and minimization of the Riesz s-potential (Götz & Saff, 2001; Hardin & Saff, 2005; Liu et al., 2018). The uniformity metric we propose is based on the Gaussian potential, which can be used to represent a very general class of kernels and is closely related to the universally optimal point configurations (Borodachov et al., 2019; Cohn & Kumar, 2007). Additionally, the best-packing problem on hyperspheres (often called the Tammes problem) is also well studied (Tammes, 1930).

3. Preliminaries on Unsupervised Contrastive Representation Learning

The popular unsupervised contrastive representation learning method (often referred to as *contrastive learning* in this paper) learns representations from unlabeled data. It assumes a way to sample *positive pairs*, representing similar samples that should have similar representations. Empirically, the positive pairs are often obtained by taking two independently randomly augmented versions of the same sample, e.g. two crops of the same image (Wu et al., 2018; Hjelm et al., 2018; Bachman et al., 2019; He et al., 2019; Chen et al., 2020).

Let $p_{\mathsf{data}}(\cdot)$ be the data distribution over $\mathbb{R}^n$ and $p_{\mathsf{pos}}(\cdot,\cdot)$ the distribution of positive pairs over $\mathbb{R}^n \times \mathbb{R}^n$. Based on empirical practices, we assume the following property.

Assumption. Distributions p_{data} and p_{pos} should satisfy

- Symmetry: $\forall x, y,\ p_{\mathsf{pos}}(x,y) = p_{\mathsf{pos}}(y,x)$.
- Matching marginal: $\forall x,\ \int p_{\mathsf{pos}}(x,y)\,\mathrm{d}y = p_{\mathsf{data}}(x)$.

We consider the following specific and widely popular form of contrastive loss for training an encoder $f\colon \mathbb{R}^n \to \mathcal{S}^{m-1}$, mapping data to ℓ_2 normalized feature vectors of dimension m. This loss has been shown effective by many recent representation learning methods (Logeswaran & Lee, 2018; Wu et al., 2018; Tian et al., 2019; He et al., 2019; Hjelm et al., 2018; Bachman et al., 2019; Chen et al., 2020).

$$\mathcal{L}_{\mathsf{contrastive}}(f;\tau,M) \triangleq \mathop{\mathbb{E}}_{\substack{(x,y)\sim p_{\mathsf{pos}} \\ \{x_i^-\}_{i=1}^M \overset{\text{i.i.d.}}{\sim} p_{\mathsf{data}}}} \left[-\log \frac{e^{f(x)^\mathsf{T} f(y)/\tau}}{e^{f(x)^\mathsf{T} f(y)/\tau} + \sum_i e^{f(x_i^-)^\mathsf{T} f(y)/\tau}} \right], \tag{1}$$

where $\tau > 0$ is a scalar temperature hyperparameter, and $M \in \mathbb{Z}_+$ is a fixed number of negative samples.

The term *contrastive loss* has also been generally used to refer to various objectives based on positive and negative samples, e.g., in Siamese networks (Chopra et al., 2005; Hadsell et al., 2006). In this work, we focus on the specific form in Equation (1) that is widely used in modern unsupervised contrastive representation learning literature.

Necessity of normalization. Without the norm constraint, the `softmax` distribution can be made arbitrarily sharp by simply scaling all the features. Wang et al. (2017) provided an analysis on this effect and argued for the necessity of normalization when using feature vector dot products in a cross entropy loss, as is in Eqn. (1). Experimentally, Chen et al. (2020) also showed that normalizing outputs leads to superior representations.

The InfoMax principle. Many empirical works are motivated by the InfoMax principle of maximizing $I(f(x); f(y))$ for $(x,y) \sim p_{\mathsf{pos}}$ (Tian et al., 2019; Bachman et al., 2019; Wu et al., 2020). Usually they interpret $\mathcal{L}_{\mathsf{contrastive}}$ in Eqn. (1) as a lower bound of $I(f(x); f(y))$ (Oord et al., 2018; Hjelm et al., 2018; Bachman et al., 2019; Tian et al., 2019). However, this interpretation is known to have issues in practice, e.g., maximizing a tighter bound often leads to worse downstream task performance (Tschannen et al., 2019). Therefore, instead of viewing it as a bound, we investigate the exact behavior of directly optimizing $\mathcal{L}_{\mathsf{contrastive}}$ in the following sections.

4. Feature Distribution on the Hypersphere

The contrastive loss encourages learned feature representation for positive pairs to be similar, while pushing features from the randomly sampled negative pairs apart. Conventional wisdom says that representations should extract the most shared information between positive pairs and remain invariant to other noise factors (Linsker, 1988; Tian et al., 2019; Wu et al., 2020; Bachman et al., 2019). Therefore, the loss should prefer two following properties:

- *Alignment*: two samples forming a positive pair should be mapped to nearby features, and thus be (mostly) invariant to unneeded noise factors.
- *Uniformity*: feature vectors should be roughly uniformly distributed on the unit hypersphere $\mathcal{S}^{m-1}$, preserving as much information of the data as possible.

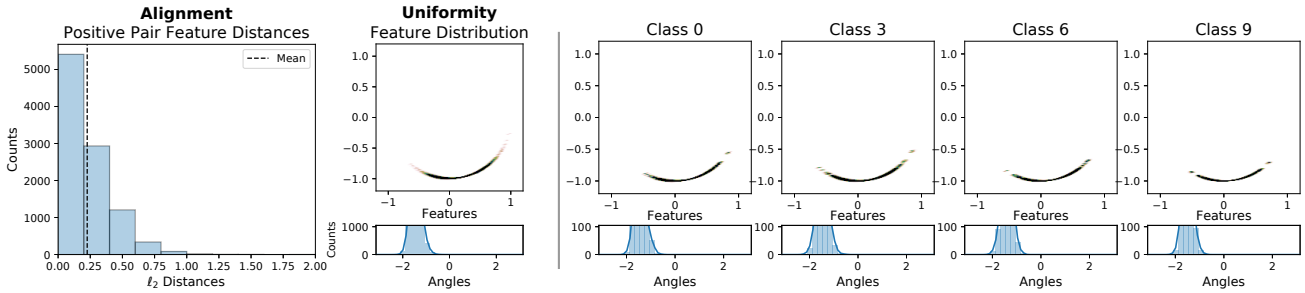
To empirically verify this, we visualize CIFAR-10 (Torralba et al., 2008; Krizhevsky et al., 2009) representations on $\mathcal{S}^1$ ($m = 2$) obtained via three different methods:

- Random initialization.
- Supervised predictive learning: An encoder and a linear classifier are jointly trained from scratch with cross entropy loss on supervised labels.
- Unsupervised contrastive learning: An encoder is trained w.r.t. $\mathcal{L}_{\mathsf{contrastive}}$ with $\tau = 0.5$ and $M = 256$.

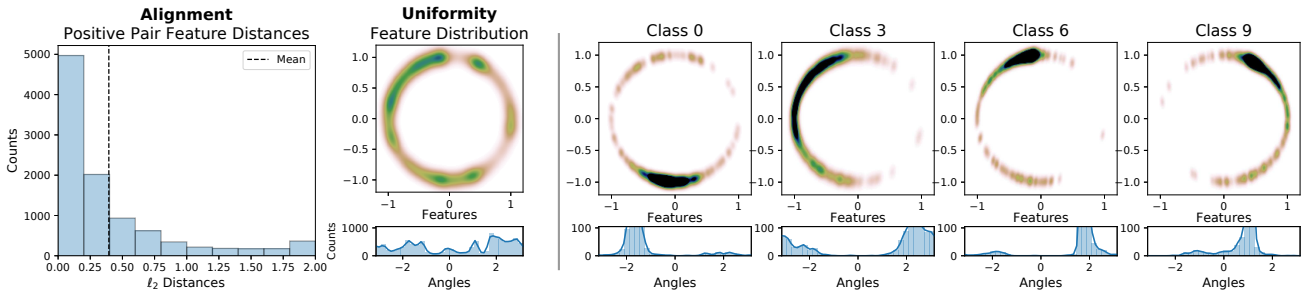
All three encoders share the same AlexNet based architecture (Krizhevsky et al., 2012), modified to map input images to 2-dimensional vectors in $\mathcal{S}^1$. Both predictive and contrastive learning use standard data augmentations to augment the dataset and sample positive pairs.

Figure 3 summarizes the resulting distributions of validation set features. Indeed, features from unsupervised contrastive learning (bottom in Figure 3) exhibit the most uniform distribution, and are closely clustered for positive pairs.

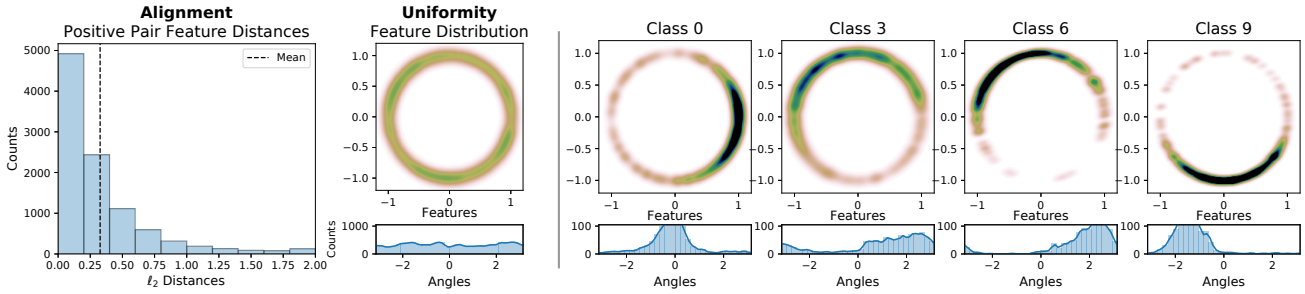
The form of the contrastive loss in Eqn. (1) also suggests this. We present informal arguments below, followed by more formal treatment in Section 4.2. From the symmetry



(a) **Random Initialization.** Linear classification validation accuracy: 12.71%.



(b) **Supervised Predictive Learning.** Linear classification validation accuracy: 57.19%.



(c) **Unsupervised Contrastive Learning.** Linear classification validation accuracy: 28.60%.

Figure 3: Representations of CIFAR-10 validation set on $\mathcal{S}^1$. **Alignment analysis:** We show distribution of distance between features of positive pairs (two random augmentations). **Uniformity analysis:** We plot feature distributions with Gaussian kernel density estimation (KDE) in $\mathbb{R}^2$ and von Mises-Fisher (vMF) KDE on angles (i.e., $\arctan2(y, x)$ for each point $(x, y) \in \mathcal{S}^1$). **Four rightmost plots** visualize feature distributions of selected specific classes. Representation from contrastive learning is both *aligned* (having low positive pair feature distances) and *uniform* (evenly distributed on $\mathcal{S}^1$).

of p, we can derive

$$\mathcal{L}_{\text{contrastive}}(f;\tau,M) = \mathop{\mathbb{E}}_{(x,y)\sim p_{\text{pos}}}\left[-f(x)^\top f(y)/\tau\right] + \mathop{\mathbb{E}}_{\substack{(x,y)\sim p_{\text{pos}} \\ \{x_i^-\}_{i=1}^M \overset{\text{i.i.d.}}{\sim} p_{\text{data}}}}\left[\log\left(e^{f(x)^\top f(y)/\tau} + \sum_i e^{f(x_i^-)^\top f(x)/\tau}\right)\right].$$

Because the $\sum_i e^{f(x_i^-)^\top f(x)/\tau}$ term is always positive and bounded below, the loss favors smaller $\mathbb{E}\left[-f(x)^\top f(y)/\tau\right]$, i.e., having more aligned positive pair features. Suppose the encoder is perfectly aligned, i.e., $\mathbb{P}\left[f(x) = f(y)\right] = 1$, then minimizing the loss is equivalent to optimizing

$$\mathop{\mathbb{E}}_{\substack{x\sim p_{\text{data}} \\ \{x_i^-\}_{i=1}^M \overset{\text{i.i.d.}}{\sim} p_{\text{data}}}}\left[\log\left(e^{1/\tau} + \sum_i e^{f(x_i^-)^\top f(x)/\tau}\right)\right],$$

which is akin to maximizing pairwise distances with a `LogSumExp` transformation. Intuitively, pushing all features away from each other should indeed cause them to be roughly uniformly distributed.

4.1. Quantifying Alignment and Uniformity

For further analysis, we need a way to measure alignment and uniformity. We propose the following two metrics (losses).

4.1.1. Alignment

The alignment loss is straightforwardly defined with the expected distance between positive pairs:

$$\mathcal{L}_{\text{align}}(f;\alpha) \triangleq \mathop{\mathbb{E}}_{(x,y)\sim p_{\text{pos}}}\left[\|f(x) - f(y)\|_2^\alpha\right], \quad \alpha > 0.$$

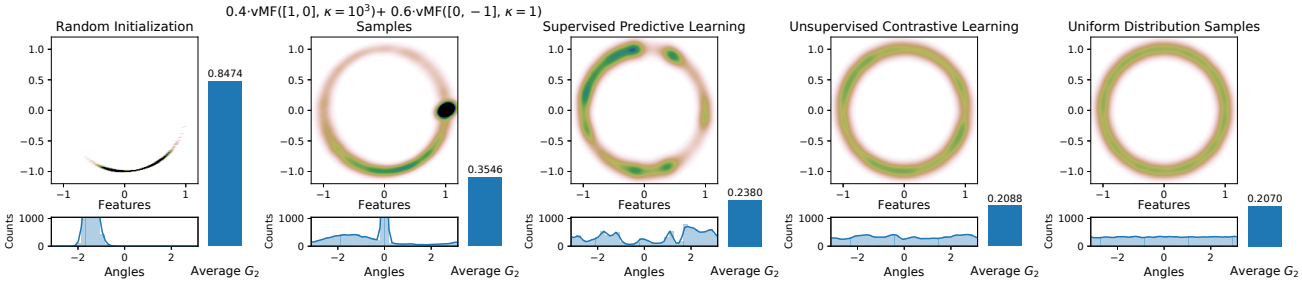


Figure 4: Average pairwise G_2 potential as a measure of uniformity. Each plot shows 10000 points distributed on $\mathcal{S}^1$, obtained via either applying an encoder on CIFAR-10 validation set (same as those in Figure 3) or sampling from a distribution on $\mathcal{S}^1$, as described in plot titles. We show the points with Gaussian KDE and the angles with vMF KDE.

4.1.2. Uniformity

We want the uniformity metric to be both asymptotically correct (i.e., the distribution optimizing this metric should converge to uniform distribution) and empirically reasonable with finite number of points. To this end, we consider the Gaussian potential kernel (also known as the Radial Basis Function (RBF) kernel) $G_t \colon \mathcal{S}^d \times \mathcal{S}^d \to \mathbb{R}_+$ (Cohn & Kumar, 2007; Borodachov et al., 2019):

$$G_t(u, v) \triangleq e^{-t\|u-v\|_2^2} = e^{2t \cdot u^\top v - 2t}, \quad t > 0,$$

and define the uniformity loss as the logarithm of the average pairwise Gaussian potential:

$$\mathcal{L}_{\mathsf{uniform}}(f; t) \triangleq \log \mathop{\mathbb{E}}_{\substack{x, y \overset{\text{i.i.d.}}{\sim} p_{\mathsf{data}}}} \left[G_t(u, v)\right], \quad t > 0,$$

where t is a fixed parameter.

The average pairwise Gaussian potential is nicely tied with the uniform distribution on the unit hypersphere.

Definition (Uniform distribution on $\mathcal{S}^d$)**.** σ_d denotes the normalized surface area measure on $\mathcal{S}^d$.

First, we show that the uniform distribution is the unique distribution that minimize the expected pairwise potential.

Proposition 1. *For $\mathcal{M}(\mathcal{S}^d)$ the set of Borel probability measures on $\mathcal{S}^d$, σ_d is the unique solution of*

$$\min_{\mu \in \mathcal{M}(\mathcal{S}^d)} \int_u \int_v G_t(u, v) \, \mathrm{d}\mu \, \mathrm{d}\mu.$$

Proof. See supplementary material. □

In addition, as number of points goes to infinity, distributions of points minimizing the average pairwise potential converge weak* to the uniform distribution. Recall the definition of the weak* convergence of measures.

Definition (Weak* convergence of measures)**.** A sequence of Borel measures $\{\mu_n\}_{n=1}^\infty$ in $\mathbb{R}^p$ converges weak* to a Borel measure μ if for all continuous function $f \colon \mathbb{R}^p \to \mathbb{R}$, we have

$$\lim_{n \to \infty} \int f(x) \, \mathrm{d}\mu_n(x) = \int f(x) \, \mathrm{d}\mu(x).$$

Proposition 2. *For each $N > 0$, the N point minimizer of the average pairwise potential is*

$$\mathbf{u}_N^* = \mathop{\arg\min}_{u_1, u_2, \ldots, u_N \in \mathcal{S}^d} \sum_{1 \le i < j \le N} G_t(u_i, u_j).$$

The normalized counting measures associated with the $\{\mathbf{u}_N^*\}_{N=1}^\infty$ sequence converge weak to σ_d.*

Proof. See supplementary material. □

Designing an objective minimized by the uniform distribution is in fact nontrivial. For instance, average pairwise dot products or Euclidean distances is simply optimized by any distribution that has zero mean. Among kernels that achieve uniformity at optima, the Gaussian kernel is special in that it is closely related to the universally optimal point configurations and can also be used to represent a general class of other kernels, including the Riesz s-potentials. We refer readers to Borodachov et al. (2019) and Cohn & Kumar (2007) for in-depths discussion on these topics. Moreover, as we show below, $\mathcal{L}_{\mathsf{uniform}}$, defined with the Gaussian kernel, has close connections with $\mathcal{L}_{\mathsf{contrastive}}$.

Empirically, we evaluate the average pairwise potential of various finite point collections on $\mathcal{S}^1$ in Figure 4. The values nicely align with our intuitive understanding of uniformity.

4.2. Limiting Behavior of Contrastive Learning

In this section, we formalize the intuition that contrastive learning optimizes alignment and uniformity, and characterize its asymptotic behavior. We consider optimization problems over all measurable encoder functions from the p_{data} measure in $\mathbb{R}^n$ to the Borel space $\mathcal{S}^{m-1}$.

We first define the notion of optimality for these metrics.

Definition (Perfect Alignment)**.** We say an encoder f is *perfectly aligned* if $f(x) = f(y)$ a.s. over $(x, y) \sim p_{\mathsf{pos}}$.

Definition (Perfect Uniformity)**.** We say an encoder f is *perfectly uniform* if the distribution of $f(x)$ for $x \sim p_{\mathsf{data}}$ is the uniform distribution σ_{m-1} on $\mathcal{S}^{m-1}$.

Realizability of perfect uniformity. We note that it is not always possible to achieve perfect uniformity, e.g., when the data manifold in $\mathbb{R}^n$ is lower dimensional than the feature space $\mathcal{S}^{m-1}$. Moreover, in the case that p_{data} and p_{pos} are formed from sampling augmented samples from a finite dataset, there cannot be an encoder that is *both* perfectly aligned and perfectly uniform, because perfect alignment implies that all augmentations from a single element have the same feature vector. Nonetheless, perfectly uniform encoder functions do exist under the conditions that $n \geq m-1$ and p_{data} has bounded density.

We analyze the asymptotics with infinite negative samples. Existing empirical work has established that larger number of negative samples consistently leads to better downstream task performances (Wu et al., 2018; Tian et al., 2019; He et al., 2019; Chen et al., 2020), and often uses very large values (e.g., $M = 65536$ in He et al. (2019)). The following theorem nicely confirms that optimizing w.r.t. the limiting loss indeed requires both alignment and uniformity.

Theorem 1 (Asymptotics of $\mathcal{L}_{\mathsf{contrastive}}$)**.** *For fixed $\tau > 0$, as the number of negative samples $M \to \infty$, the (normalized) contrastive loss converges to*

$$\begin{aligned}\lim_{M\to\infty} \mathcal{L}_{\mathsf{contrastive}}(f;\tau,M) - \log M = \\ -\frac{1}{\tau}\mathop{\mathbb{E}}_{(x,y)\sim p_{\mathsf{pos}}}\left[f(x)^\mathsf{T} f(y)\right] \\ + \mathop{\mathbb{E}}_{x\sim p_{\mathsf{data}}}\left[\log \mathop{\mathbb{E}}_{x^-\sim p_{\mathsf{data}}}\left[e^{f(x^-)^\mathsf{T} f(x)/\tau}\right]\right].\end{aligned} \tag{2}$$

We have the following results:

1. *The first term is minimized iff f is perfectly aligned.*
2. *If perfectly uniform encoders exist, they form the exact minimizers of the second term.*
3. *For the convergence in Equation (2), the absolute deviation from the limit decays in $\mathcal{O}(M^{-2/3})$.*

Proof. See supplementary material. □

Relation with $\mathcal{L}_{\mathsf{uniform}}$. The proof of Theorem 1 in the supplementary material connects the asymptotic $\mathcal{L}_{\mathsf{contrastive}}$ form with minimizing average pairwise Gaussian potential, i.e., minimizing $\mathcal{L}_{\mathsf{uniform}}$. Compared with the second term of Equation (2), $\mathcal{L}_{\mathsf{uniform}}$ essentially pushes the log outside the outer expectation, without changing the minimizer (perfectly uniform encoders). However, due to its pairwise nature, $\mathcal{L}_{\mathsf{uniform}}$ is much simpler in form and avoids the computationally expensive `softmax` operation in $\mathcal{L}_{\mathsf{contrastive}}$ (Goodman, 2001; Bengio et al.; Gutmann & Hyvärinen, 2010; Grave et al., 2017; Chen et al., 2018).

Relation with feature distribution entropy estimation. When p_{data} is uniform over finite samples $\{x_1, x_2, \ldots, x_N\}$ (e.g., a collected dataset), the second term in Equation (2) can be alternatively viewed as a resubstitution entropy estimator of $f(x)$ (Ahmad & Lin, 1976), where x follows the underlying distribution p_{nature} that generates $\{x_i\}_{i=1}^N$, via a von Mises-Fisher (vMF) kernel density estimation (KDE):

$$\begin{aligned}&\mathop{\mathbb{E}}_{x\sim p_{\mathsf{data}}}\left[\log \mathop{\mathbb{E}}_{x^-\sim p_{\mathsf{data}}}\left[e^{f(x^-)^\mathsf{T} f(x)/\tau}\right]\right] \\ &= \frac{1}{N}\sum_{i=1}^N \log\left(\frac{1}{N}\sum_{j=1}^N e^{f(x_i)^\mathsf{T} f(x_j)/\tau}\right) \\ &= \frac{1}{N}\sum_{i=1}^N \log \hat{p}_{\mathsf{vMF\text{-}KDE}}(f(x_i)) + \log Z_{\mathsf{vMF}} \\ &\triangleq -\hat{H}(f(x)) + \log Z_{\mathsf{vMF}}, && x\sim p_{\mathsf{nature}} \\ &\triangleq -\hat{I}(x; f(x)) + \log Z_{\mathsf{vMF}}, && x\sim p_{\mathsf{nature}},\end{aligned}$$

where

- $\hat{p}_{\mathsf{vMF\text{-}KDE}}$ is the KDE based on samples $\{f(x_j)\}_{j=1}^N$ using a vMF kernel with $\kappa = \tau^{-1}$,
- Z_{vMF} is the vMF normalization constant for $\kappa = \tau^{-1}$,
- $\hat{H}$ denotes the resubstitution entropy estimator,
- $\hat{I}$ denotes the mutual information estimator based on $\hat{H}$, since f is a deterministic function.

Relation with the InfoMax principle. Many empirical works are motivated by the InfoMax principle, i.e., maximizing $I(f(x); f(y))$ for $(x, y) \sim p_{\mathsf{pos}}$. However, the interpretation of $\mathcal{L}_{\mathsf{contrastive}}$ as a lower bound of $I(f(x); f(y))$ is known to be inconsistent with its actual behavior in practice (Tschannen et al., 2019). Our results instead analyze the properties of $\mathcal{L}_{\mathsf{contrastive}}$ itself. Considering the identity $I(f(x); f(y)) = H(f(x)) - H(f(x) \mid f(y))$, we can see that while uniformity indeed favors large $H(f(x))$, alignment is stronger than merely desiring small $H(f(x) \mid f(y))$. Instead, our above analysis suggests that $\mathcal{L}_{\mathsf{contrastive}}$ optimizes for *aligned* and *information-preserving* encoders.

Finally, even for the case where only a single negative sample is used (i.e., $M = 1$), we can still prove a weaker result, which we describe in details in the supplementary material.

5. Experiments

In this section, we empirically verify the hypothesis that alignment and uniformity are desired properties for representations. Recall that our two metrics are

$$\begin{aligned}\mathcal{L}_{\mathsf{align}}(f;\alpha) &\triangleq \mathbb{E}_{(x,y)\sim p_{\mathsf{pos}}}\left[\|f(x) - f(y)\|_2^\alpha\right] \\ \mathcal{L}_{\mathsf{uniform}}(f;t) &\triangleq \log \mathop{\mathbb{E}}_{x,y\overset{\text{i.i.d.}}{\sim} p_{\mathsf{data}}}\left[e^{-t\|f(x)-f(y)\|_2^2}\right].\end{aligned}$$

We conduct extensive experiments with convolutional neural network (CNN) and recurrent neural network (RNN) based

```
# bsz : batch size (number of positive pairs)
# d   : latent dim
# x   : Tensor, shape=[bsz, d]
#       latents for one side of positive pairs
# y   : Tensor, shape=[bsz, d]
#       latents for the other side of positive pairs
# lam : hyperparameter balancing the two losses

def lalign(x, y, alpha=2):
    return (x - y).norm(dim=1).pow(alpha).mean()

def lunif(x, t=2):
    sq_pdist = torch.pdist(x, p=2).pow(2)
    return sq_pdist.mul(-t).exp().mean().log()

loss = lalign(x, y) + lam * (lunif(x) + lunif(y)) / 2
```

Figure 5: PyTorch implementation of $\mathcal{L}_{\mathsf{align}}$ and $\mathcal{L}_{\mathsf{uniform}}$.

encoders on four popular representation learning benchmarks with distinct types of downstream tasks:

- STL-10 (Coates et al., 2011) classification on AlexNet-based encoder outputs or intermediate activations with a linear or k-nearest neighbor (k-NN) classifier.
- NYU-DEPTH-V2 (Nathan Silberman & Fergus, 2012) depth prediction on CNN encoder intermediate activations after convolution layers.
- IMAGENET-100 (100 randomly selected classes from IMAGENET) classification on CNN encoder penultimate layer activations with a linear classifier.
- BOOKCORPUS (Zhu et al., 2015) RNN sentence encoder outputs used for Moview Review Sentence Polarity (MR) (Pang & Lee, 2005) and Customer Product Review Sentiment (CR) (Wang & Manning, 2012) binary classification tasks with logisitc classifiers.

For image datasets, we follow the standard practice and choose positive pairs as two independent augmentations of the same image. For BOOKCORPUS, positive pairs are chosen as neighboring sentences, following Quick-Thought Vectors (Logeswaran & Lee, 2018).

We perform majority of our analysis on STL-10 and NYU-DEPTH-V2 encoders, where we calculate $\mathcal{L}_{\mathsf{contrastive}}$ with negatives being other samples within the minibatch following the standard practice (Hjelm et al., 2018; Bachman et al., 2019; Tian et al., 2019; Chen et al., 2020), and $\mathcal{L}_{\mathsf{uniform}}$ as the logarithm of average pairwise feature potentials also within the minibatch. Due to their simple forms, these two losses can be implemented in PyTorch (Paszke et al., 2019) with less than 10 lines of code, as shown in Figure 5.

To investigate *alignment* and *uniformity* properties on recent contrastive representation learning variants and larger datasets, we also analyze IMAGENET-100 encoders trained with Momentum Contrast (MoCo) (He et al., 2019) and BOOKCORPUS encoders trained with Quick-Thought Vectors (Logeswaran & Lee, 2018), with these methods modified to also allow $\mathcal{L}_{\mathsf{align}}$ and $\mathcal{L}_{\mathsf{uniform}}$.

We optimize a total of 306 STL-10 encoders, 64 NYU-DEPTH-V2 encoders, 45 IMAGENET-100 encoders, and 108 BOOKCORPUS encoders without supervision. The encoders are optimized w.r.t. weighted combinations of $\mathcal{L}_{\mathsf{contrastive}}$, $\mathcal{L}_{\mathsf{align}}$, and/or $\mathcal{L}_{\mathsf{uniform}}$, with varying

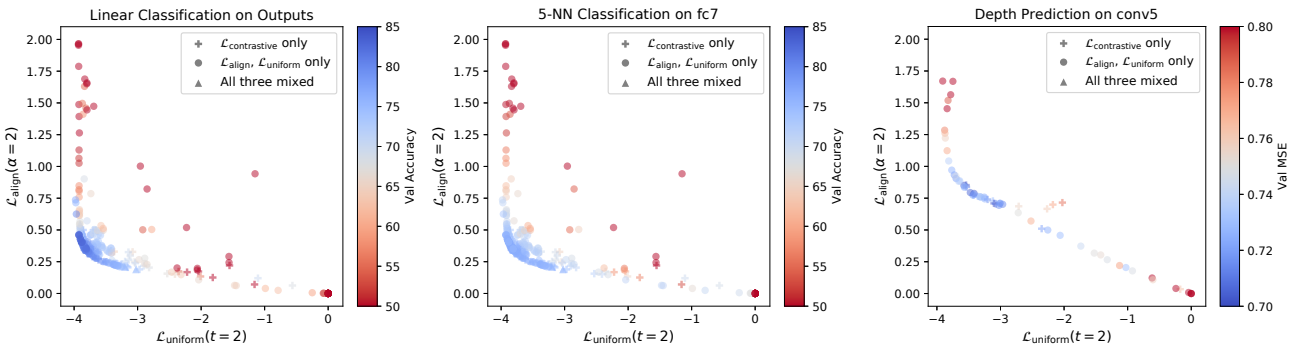
- (possibly zero) weights on the three losses,
- loss hyperparameters: τ for $\mathcal{L}_{\mathsf{contrastive}}$, α for $\mathcal{L}_{\mathsf{align}}$, and t for $\mathcal{L}_{\mathsf{uniform}}$,
- batch size (affecting the number of (negative) pairs for $\mathcal{L}_{\mathsf{contrastive}}$ and $\mathcal{L}_{\mathsf{uniform}}$),
- embedding dimension,
- number of training epochs and learning rate,
- initialization (from scratch vs. a pretrained encoder).

See the supplementary material for more experiment details and the exact configurations used.

Both $\mathcal{L}_{\mathsf{align}}$ and $\mathcal{L}_{\mathsf{uniform}}$ strongly agree with downstream task performance. For each encoder, we measure the downstream task performance, and the $\mathcal{L}_{\mathsf{align}}$, $\mathcal{L}_{\mathsf{uniform}}$ metrics on the validation set. Figure 6 visualizes the trends between both metrics and representation quality. We observe that the two metrics strongly agrees the representation quality overall. In particular, the best performing encoders are exactly the ones with low $\mathcal{L}_{\mathsf{align}}$ and $\mathcal{L}_{\mathsf{uniform}}$, i.e., the lower left corners in Figure 6. In the supplementary material, we observe that as long as the ratio between weights on $\mathcal{L}_{\mathsf{align}}$ and $\mathcal{L}_{\mathsf{uniform}}$ is not too large (e.g., < 4), the representation quality remains relatively good and insensitive to the exact weight choices.

Directly optimizing only $\mathcal{L}_{\mathsf{align}}$ and $\mathcal{L}_{\mathsf{uniform}}$ can lead to better representations. As shown in Table 1, encoders trained with only $\mathcal{L}_{\mathsf{align}}$ and $\mathcal{L}_{\mathsf{uniform}}$ consistently outperform their $\mathcal{L}_{\mathsf{contrastive}}$-trained counterparts, for both tasks. Theoretically, Theorem 1 showed that $\mathcal{L}_{\mathsf{contrastive}}$ optimizes alignment and uniformity asymptotically with infinite negative samples. This empirical performance gap suggests that directly optimizing these properties can be superior in practice, when we can only have finite negatives.

$\mathcal{L}_{\mathsf{align}}$ and $\mathcal{L}_{\mathsf{uniform}}$ causally affect downstream task performance. We take an encoder trained with $\mathcal{L}_{\mathsf{contrastive}}$ using a suboptimal temperature $\tau = 2.5$, and finetune it according to $\mathcal{L}_{\mathsf{align}}$ and/or $\mathcal{L}_{\mathsf{uniform}}$. Figure 7 visualizes the finetuning trajectories. When only one of alignment and uniformity is optimized, the corresponding metric improves, but both the other metric and performance degrade. However, when both properties are optimized, the representation quality steadily increases. These trends confirm the causal effect of alignment and uniformity on the representation quality, and suggest that directly optimizing them can be a reasonable choice.



(a) 306 STL-10 encoders are evaluated with linear classification on output features and 5-nearest neighbor (5-NN) on fc7 activations. Higher accuracy (blue color) is better.

(b) 64 NYU-DEPTH-V2 encoders are evaluated with CNN depth regressors on conv5 activations. Lower MSE (blue color) is better.

Figure 6: Metrics and performance of STL-10 and NYU-DEPTH-V2 experiments. Each point represents a trained encoder, with its x- and y-coordinates showing $\mathcal{L}_{\text{align}}$ and $\mathcal{L}_{\text{uniform}}$ metrics and color showing the performance on validation set. **Blue** is better for both tasks. Encoders with low $\mathcal{L}_{\text{align}}$ and $\mathcal{L}_{\text{uniform}}$ are consistently the better performing ones (lower left corners).

	STL-10 Validation Set Accuracy ↑				NYU-DEPTH-V2 Validation Set MSE ↓	
	Output + Linear	Output + 5-NN	fc7 + Linear	fc7 + 5-NN	conv5	conv4
Best $\mathcal{L}_{\text{contrastive}}$ only	80.46%	78.75%	83.89%	76.33%	0.7024	**0.7575**
Best $\mathcal{L}_{\text{align}}$ and $\mathcal{L}_{\text{uniform}}$ only	**81.15%**	78.89%	**84.43%**	**76.78%**	**0.7014**	0.7592

Table 1: Encoder evaluations. **STL-10:** Numbers show linear and 5-nearest neighbor (5-NN) classification accuracies. The best result is picked by encoder outputs linear classifier accuracy from a 5-fold training set cross validation, among all 150 encoders trained from scratch with 128-dimensional output and 768 batch size. **NYU-DEPTH-V2:** Numbers show depth prediction mean squared error (MSE). The best result is picked based on conv5 layer MSE from a 5-fold training set cross validation, among all 64 encoders trained from scratch with 128-dimensional output and 128 batch size.

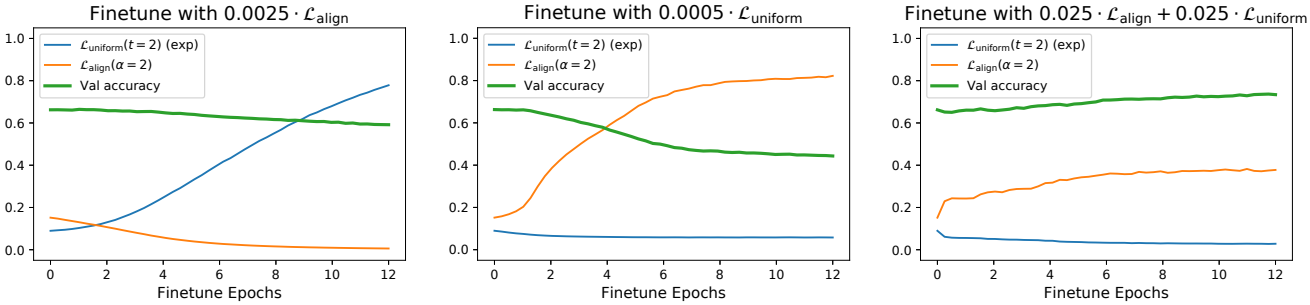
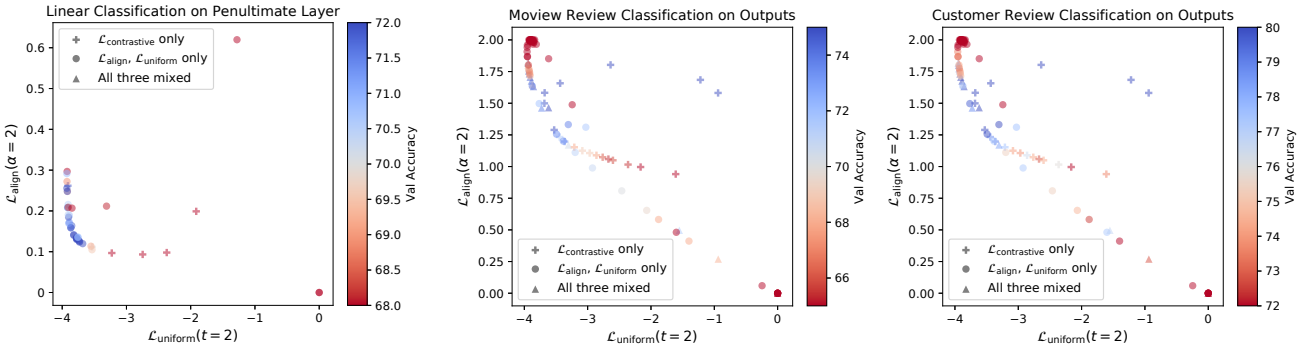


Figure 7: Finetuning trajectories from a STL-10 encoder trained with $\mathcal{L}_{\text{contrastive}}$ using a suboptimal temperature $\tau = 2.5$. Finetuning objectives are weighted combinations of $\mathcal{L}_{\text{align}}(\alpha{=}2)$ and $\mathcal{L}_{\text{uniform}}(t{=}2)$. For each intermediate checkpoint, we measure $\mathcal{L}_{\text{align}}$ and $\mathcal{L}_{\text{uniform}}$ metrics, as well as validation accuracy of a linear classifier trained from scratch on the encoder outputs. $\mathcal{L}_{\text{uniform}}$ is exponentiated for plotting purpose. **Left and middle:** Performance degrades if only one of alignment and uniformity is optimized. **Right:** Performance improves when both are optimized.

Alignment and uniformity also matter in other contrastive representation learning variants. MoCo (He et al., 2019) and Quick-Thought Vectors (Logeswaran & Lee, 2018) are contrastive representation learning variants that have nontrivial differences with directly optimizing $\mathcal{L}_{\text{contrastive}}$ in Equation (1). MoCo introduces a memory queue and a momentum encoder. Quick-Thought Vectors uses two different encoders to encode each sentence in a positive pair, only normalizes encoder outputs during evaluation, and does not use random sampling to obtain mini-batches. After modifying them to also allow $\mathcal{L}_{\text{align}}$ and $\mathcal{L}_{\text{uniform}}$, we train these methods on IMAGENET-100 and BOOKCORPUS, respectively. Figure 8 shows that $\mathcal{L}_{\text{align}}$ and $\mathcal{L}_{\text{uniform}}$ metrics are still correlated with the downstream task performances. Table 2 shows that directly optimizing them also leads to comparable or better representation quality. These results suggest that alignment and uniformity are indeed desirable properties for representations, for *both* image and text modalities, and are likely connected with general contrastive representation learning methods.



(a) 45 IMAGENET-100 encoders are trained with MoCo-based methods, and evaluated with linear classification.

(b) 108 BOOKCORPUS encoders are trained with Quick-Thought-Vectors-based methods, and evaluated with logistic binary classification on Movie Review Sentence Polarity (MR) and Customer Product Review Sentiment (CR) tasks.

Figure 8: Metrics and performance of IMAGENET-100 and BOOKCORPUS experiments. Each point represents a trained encoder, with its x- and y-coordinates showing $\mathcal{L}_{\text{align}}$ and $\mathcal{L}_{\text{uniform}}$ metrics and color showing the validation accuracy. **Blue** is better. Encoders with low $\mathcal{L}_{\text{align}}$ and $\mathcal{L}_{\text{uniform}}$ consistently perform well (lower left corners), even though the training methods (based on MoCo and Quick-Thought Vectors) are different from directly optimizing the contrastive loss in Equation (1).

	IMAGENET-100 MoCo-based Encoders		BOOKCORPUS Quick-Though-Vectors-based Encoders	
	top1 Val. Accuracy ↑	top5 Val. Accuracy ↑	MR Val. Accuracy ↑	CR Val. Accuracy ↑
Best $\mathcal{L}_{\text{contrastive}}$ only	72.80%	91.64%	**77.51%**	**83.86%**
Best $\mathcal{L}_{\text{align}}$ and $\mathcal{L}_{\text{uniform}}$ only	**74.60%**	**92.74%**	73.76%	80.95%

Table 2: Encoder evaluations. **IMAGENET-100:** Numbers show linear classifier accuracies on encoder penultimate layer activations.The best result is picked based on top1 accuracy from a 3-fold training set cross validation, among all 45 encoders trained from scratch with 128-dimensional output and 128 batch size. **BOOKCORPUS:** Numbers show Movie Review Sentence Polarity (MR) and Customer Product Sentiment (CR) classification accuracies of logistic classifiers fit on encoder outputs. The best result is picked based on accuracy from a 5-fold training set cross validation, individually for MR and CR, among all 108 encoders trained from scratch with 1200-dimensional output and 400 batch size.

6. Discussion

Alignment and *uniformity* are often alluded to as motivations for representation learning methods (see Figure 1). However, a thorough understanding of these properties is lacking in the literature.

Are they in fact related to the representation learning methods? Do they actually agree with the representation quality (measured by downstream task performance)?

In this work, we have presented a detailed investigation on the relation between these properties and the popular paradigm of contrastive representation learning. Through theoretical analysis and extensive experiments, we are able to relate the contrastive loss with the alignment and uniformity properties, and confirm their strong connection with downstream task performances. Remarkably, we have revealed that directly optimizing our proposed metrics often leads to representations of better quality.

Below we summarize several suggestions for future work.

Niceness of the unit hypersphere. Our analysis was based on the empirical observation that representations are often ℓ_2 normalized. Existing works have motivated this choice from a manifold mapping perspective (Liu et al., 2017; Davidson et al., 2018) and computation stability (Xu & Durrett, 2018; Wang et al., 2017). However, to our best knowledge, the question of why the unit hypersphere is a nice feature space is not yet rigorously answered. One possible direction is to formalize the intuition that connected sets with smooth boundaries are nearly linearly separable in the hyperspherical geometry (see Figure 2), since linear separability is one of the most widely used criteria for representation quality and is related to the notion of disentanglement (Higgins et al., 2018).

Beyond contrastive learning. Our analysis focused on the relationship between contrastive learning and the alignment and uniformity properties on the unit hypersphere. However, the ubiquitous presence of ℓ_2 normalization in the representation learning literature suggests that the connection may be more general. In fact, several existing empirical methods are directly related to uniformity on the hypersphere (Bojanowski & Joulin, 2017; Davidson et al., 2018; Xu & Durrett, 2018). We believe that relating a broader class of representations to uniformity and/or alignment on the hypersphere will provide novel insights and lead to better empirical algorithms.

Acknowledgements

We thank Philip Bachman, Ching-Yao Chuang, Justin Solomon, Yonglong Tian, and Zhenyang Zhang for many helpful comments and suggestions. Tongzhou Wang was supported by the MIT EECS Merrill Lynch Graduate Fellowship.

References

Ahmad, I. and Lin, P.-E. A nonparametric estimation of the entropy for absolutely continuous distributions (corresp.). *IEEE Transactions on Information Theory*, 22(3):372–375, 1976.

Bachman, P., Hjelm, R. D., and Buchwalter, W. Learning representations by maximizing mutual information across views. In *Advances in Neural Information Processing Systems*, pp. 15509–15519, 2019.

Bengio, Y. et al. Quick training of probabilistic neural nets by importance sampling.

Bojanowski, P. and Joulin, A. Unsupervised learning by predicting noise. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pp. 517–526. JMLR. org, 2017.

Borodachov, S. V., Hardin, D. P., and Saff, E. B. *Discrete energy on rectifiable sets*. Springer, 2019.

Chen, P. H., Si, S., Kumar, S., Li, Y., and Hsieh, C.-J. Learning to screen for fast softmax inference on large vocabulary neural networks. 2018.

Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. A simple framework for contrastive learning of visual representations. *arXiv preprint arXiv:2002.05709*, 2020.

Chopra, S., Hadsell, R., and LeCun, Y. 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, pp. 539–546. IEEE, 2005.

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

Cohn, H. and Kumar, A. Universally optimal distribution of points on spheres. *Journal of the American Mathematical Society*, 20 (1):99–148, 2007.

Davidson, T. R., Falorsi, L., De Cao, N., Kipf, T., and Tomczak, J. M. Hyperspherical variational auto-encoders. *34th Conference on Uncertainty in Artificial Intelligence (UAI-18)*, 2018.

Goodman, J. Classes for fast maximum entropy training. In *2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No. 01CH37221)*, volume 1, pp. 561–564. IEEE, 2001.

Götz, M. and Saff, E. B. Note on d—extremal configurations for the sphere in r d+1. In *Recent Progress in Multivariate Approximation*, pp. 159–162. Springer, 2001.

Grave, E., Joulin, A., Cissé, M., Jégou, H., et al. Efficient softmax approximation for gpus. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pp. 1302–1310. JMLR. org, 2017.

Gutmann, M. and Hyvärinen, A. 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.

Hadsell, R., Chopra, S., and LeCun, Y. 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.

Hardin, D. and Saff, E. Minimal riesz energy point configurations for rectifiable d-dimensional manifolds. *Advances in Mathematics*, 193(1):174–204, 2005.

Hasnat, M., Bohné, J., Milgram, J., Gentric, S., Chen, L., et al. von mises-fisher mixture model-based deep learning: Application to face verification. *arXiv preprint arXiv:1706.04264*, 2017.

He, K., Fan, H., Wu, Y., Xie, S., and Girshick, R. Momentum contrast for unsupervised visual representation learning. *arXiv preprint arXiv:1911.05722*, 2019.

Hénaff, O. J., Razavi, A., Doersch, C., Eslami, S., and Oord, A. v. d. Data-efficient image recognition with contrastive predictive coding. *arXiv preprint arXiv:1905.09272*, 2019.

Higgins, I., Amos, D., Pfau, D., Racaniere, S., Matthey, L., Rezende, D., and Lerchner, A. Towards a definition of disentangled representations. *arXiv preprint arXiv:1812.02230*, 2018.

Hjelm, R. D., Fedorov, A., Lavoie-Marchildon, S., Grewal, K., Bachman, P., Trischler, A., and Bengio, Y. Learning deep representations by mutual information estimation and maximization. *arXiv preprint arXiv:1808.06670*, 2018.

Hou, S., Pan, X., Loy, C. C., Wang, Z., and Lin, D. Learning a unified classifier incrementally via rebalancing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 831–839, 2019.

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

Krizhevsky, A., Sutskever, I., and Hinton, G. E. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pp. 1097–1105, 2012.

Landkof, N. S. *Foundations of modern potential theory*, volume 180. Springer, 1972.

Linsker, R. Self-organization in a perceptual network. *Computer*, 21(3):105–117, 1988.

Liu, W., Wen, Y., Yu, Z., Li, M., Raj, B., and Song, L. Sphereface: Deep hypersphere embedding for face recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 212–220, 2017.

Liu, W., Lin, R., Liu, Z., Liu, L., Yu, Z., Dai, B., and Song, L. Learning towards minimum hyperspherical energy. In *Advances in Neural Information Processing Systems*, pp. 6222–6233. 2018.

Logeswaran, L. and Lee, H. An efficient framework for learning sentence representations. In *International Conference on Learning Representations*, 2018.

Mettes, P., van der Pol, E., and Snoek, C. Hyperspherical prototype networks. In *Advances in Neural Information Processing Systems*, pp. 1485–1495, 2019.

Nathan Silberman, Derek Hoiem, P. K. and Fergus, R. Indoor segmentation and support inference from rgbd images. In *ECCV*, 2012.

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

Pang, B. and Lee, L. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *Proceedings of the 43rd annual meeting on association for computational linguistics*, pp. 115–124. Association for Computational Linguistics, 2005.

Parkhi, O. M., Vedaldi, A., and Zisserman, A. Deep face recognition. 2015.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., and Chintala, S. Pytorch: An imperative style, high-performance deep learning library. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 32*, pp. 8026–8037. 2019.

Saunshi, N., Plevrakis, O., Arora, S., Khodak, M., and Khandeparkar, H. A theoretical analysis of contrastive unsupervised representation learning. In *International Conference on Machine Learning*, pp. 5628–5637, 2019.

Schroff, F., Kalenichenko, D., and Philbin, J. 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.

Tammes, P. M. L. On the origin of number and arrangement of the places of exit on the surface of pollen-grains. *Recueil des travaux botaniques néerlandais*, 27(1):1–84, 1930.

Thomson, J. J. Xxiv. on the structure of the atom: an investigation of the stability and periods of oscillation of a number of corpuscles arranged at equal intervals around the circumference of a circle; with application of the results to the theory of atomic structure. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 7(39):237–265, 1904.

Tian, Y., Krishnan, D., and Isola, P. Contrastive multiview coding. *arXiv preprint arXiv:1906.05849*, 2019.

Torralba, A., Fergus, R., and Freeman, W. T. 80 million tiny images: A large data set for nonparametric object and scene recognition. *IEEE transactions on pattern analysis and machine intelligence*, 30(11):1958–1970, 2008.

Tschannen, M., Djolonga, J., Rubenstein, P. K., Gelly, S., and Lucic, M. On mutual information maximization for representation learning. *arXiv preprint arXiv:1907.13625*, 2019.

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

Wang, S. and Manning, C. D. Baselines and bigrams: Simple, good sentiment and topic classification. In *Proceedings of the 50th annual meeting of the association for computational linguistics: Short papers-volume 2*, pp. 90–94. Association for Computational Linguistics, 2012.

Wu, M., Zhuang, C., Yamins, D., and Goodman, N. On the importance of views in unsupervised representation learning. 2020.

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

Xu, J. and Durrett, G. Spherical latent spaces for stable variational autoencoders. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 4503–4513, 2018.

Zhu, Y., Kiros, R., Zemel, R., Salakhutdinov, R., Urtasun, R., Torralba, A., and Fidler, S. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *arXiv preprint arXiv:1506.06724*, 2015.