Learning Structured Representations with Equivariant Contrastive Learning

Sharut Gupta [*1] **Joshua Robinson** [*1] **Derek Lim** [1] **Soledad Villar** [2] **Stefanie Jegelka** [1]

Abstract

Self-supervised learning converts raw perceptual data to a compact space using Euclidean distances to measure variations in data. In this paper, we enhance the embedding space by enforcing transformations of input space to correspond to simple (i.e., linear) transformations of embedding space. Specifically, in the contrastive learning setting, we introduce an *equivariance* objective and theoretically prove and empirically demonstrate that its minima forces augmentations on inputs to correspond to *rotations* on the spherical embedding space. Our method, CARE: **C**ontrastive **A**ugmentation-induced **R**otational **E**quivariance, improves performance on downstream tasks by only allowing small rotations.

1. Introduction

Understanding the ideal structure of neural network representation spaces for intelligent behavior to emerge remains limited (Ma et al., 2022). Learning low-dimensional spaces where simple Euclidean distances effectively measure data similarity is a key factor. Recent advancements have successfully achieved this at web-scale using self-supervision (Chen et al., 2020; Radford et al., 2021). However, many use cases require richer structural relationships, such as encoding object relations (e.g., parent-child or treatment-object) through simple transformations of embeddings, which has driven learning in knowledge graphs. (Bordes et al., 2013; Yasunaga et al., 2022). But, similar capabilities have been notably absent from existing self-supervised learning recipes.

Recent contrastive self-supervised learning approaches have explored ways to close this gap by ensuring input transformations $a \in \mathcal{A}$ correspond to predictable transformations T_a in embedding space i.e., $f(a(x)) \approx T_a f(x)$, a notion called equivariance (Dangovski et al., 2022; Devillers & Lefort, 2023; Garrido et al., 2023; Bhardwaj et al., 2023). Typically, a learnable feed-forward network is used as T_a, resulting in complex and hard-to-interpret relationships between the embeddings of x and $a(x)$. It also suffers from geometric pathologies, such as inconsistency under compositions: $T_{a_2 \circ a_1} f(x) \neq T_{a_2} T_{a_1} f(x)$.

To address these concerns, we propose CARE, an equivariant contrastive learning framework that learns to approximately translate augmentations in the input space (such as cropping, blurring, and jittering) into simple local *linear* transformations in feature space. Here, we use the sphere as our feature space (the standard space for contrastive learning), so we specifically consider transformations that are isometries of the sphere: rotations and reflections, i.e., orthogonal transformations. CARE trains f to preserve angles, i.e., $f(a(x))^\top f(a(x')) \approx f(x)^\top f(x')$, a property that must hold if f is orthogonally equivariant. We show that achieving low error on this seemingly weaker property also implies approximate equivariance and enjoys consistency under compositions. Critically, we can easily integrate CARE into contrastive learning workflows since both operate by comparing pairs of data.

2. Rethinking how augmentations are used in self supervised learning

This work introduces CARE, an equivariant contrastive learning approach respecting two key design principles:

Principle 1. *The map T_a satisfying $f(a(x)) = T_a f(x)$ should be linear, where $f : \mathcal{X} \to \mathbb{S}^{d-1}$ is a feature extracting model mapping to the unit sphere.*

Principle 2. *Equivariance should be learned from* pairs *of data, as in invariant contrastive learning.*

The first principle asks that f converts complex perturbations a of input data into much simpler (i.e., linear) transformations in embedding space. Specifically, we constrain the complexity of T_a by considering isometries of the sphere, $O(d) = \{Q \in \mathbb{R}^{d \times d} : QQ^T = Q^T Q = I\}$, containing all rotations and reflections. Throughout this paper we define $f(a(x)) = T_a f(x)$ for $T_a \in O(d)$ to be *orthogonal equivariance*. This approach draws heavily from ideas in linear representation theory (Curtis & Reiner, 1966; Serre et al., 1977), which studies how to convert abstract group

[*]Equal contribution [1]MIT CSAIL [2]Johns Hopkins University. Correspondence to: Sharut Gupta <sharut@mit.edu>.

Presented at the 2^{nd} Annual Workshop on Topology, Algebra, and Geometry in Machine Learning (TAG-ML) at the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. 2023.

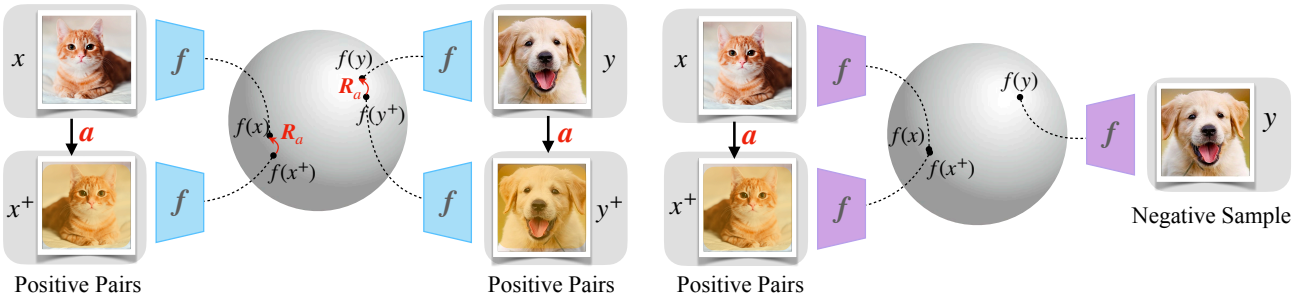

Figure 1: CARE is an equivariant contrastive learning approach that trains augmentations (cropping, blurring, etc.) of input data to correspond to orthogonal transformations of embedding space.

structures into matrix spaces equipped with standard matrix multiplication as the group operation.

The second principle stipulates *how* we want to learn orthogonal equivariance. Our method, CARE, explicitly learns T_a by training f so that an augmentation a applied to two different inputs $x, x^+ \in \mathcal{X}$ produces the same change in embedding space.

encodes data augmentations (cropping, blurring, jittering, etc.) as $O(d)$ transformations of embeddings using an equivariance-promoting objective function. CARE can be viewed as an instance of *symmetry regularization*, a term introduced by (Shakerinava et al., 2022).

3. CARE: Contrastive Augmentation-induced Rotational Equivariance

This section introduces a simple and practical approach for training a model $f : \mathcal{X} \to \mathbb{S}^{d-1}$ so that f is orthogonally equivariant: i.e., a data augmentation $a \sim \mathcal{A}$ (cropping, blurring, jittering, etc.) applied to any input $x \in \mathcal{X}$ causes the embedding $f(x)$ to transformed by the same $R_a \in O(d)$ for all $x \in \mathcal{X}$: $f(a(x)) = R_a f(x)$.

To achieve this, we consider the loss $\mathcal{L}_{\text{equi}}(f) = \mathbb{E}_{a\sim\mathcal{A}}\mathbb{E}_{x,x'\sim\mathcal{X}}\left[f(a(x'))^\top f(a(x)) - f(x)^\top f(x')\right]^2$

This is necessarily true if f is orthogonally equivariant or, more generally, $R_a \in O(d)$ exists. But the converse—that $\mathcal{L}_{\text{equi}} = 0$ implies orthogonal equivariance—is non-obvious and is theoretically analyzed in Section 3.1.

A trivial but undesirable solution that minimizes $\mathcal{L}_{\text{equi}}$ is to collapse the embeddings of all points to be the same (see Figure 2). One natural approach to avoiding trivial solutions is to combine the equivariance loss with a non-collapse term such as the uniformity $\mathcal{L}_{\text{unif}}(f) = \log \mathbb{E}_{x,x'\sim\mathcal{X}} \exp\left(f(x)^\top f(x')\right)$ (Wang & Isola, 2020) whose optima f distribute points uniformly over the sphere $\mathcal{L}(f) = \mathcal{L}_{\text{equi}}(f) + \mathcal{L}_{\text{unif}}(f)$. This is directly comparable to the InfoNCE loss, which can similarly be decomposed into two terms $\mathcal{L}_{\text{InfoNCE}}(f) = \mathcal{L}_{\text{inv}}(f) + \mathcal{L}_{\text{unif}}(f)$ where $\mathcal{L}_{\text{inv}}(f) = \mathbb{E}_{a,a'\sim\mathcal{A}}\|f(a(x)) - f(a'(x))\|$ is minimized when f is invariant to $\mathcal{A}$—i.e., $f(a(x)) = f(x)$. Figure 2 shows that training using $\mathcal{L}_{\text{equi}}+\mathcal{L}_{\text{unif}}$ yields non-trivial representations. However, the performance is below that of invariance-based contrastive learning approaches. We hypothesize that this is because data augmentations—which make small perceptual changes to data—should correspond to *small* perturbations of embeddings, which $\mathcal{L}_{\text{equi}}$ does not enforce.

To rule out this possibility, we introduce CARE: **C**ontrastive **A**ugmentation-induced **R**otational **E**quivariance. CARE additionally enforces the orthogonal transformations in embedding space to be *localized* by reintroducing an invariance loss term $\mathcal{L}_{\text{inv}}$ to encourage f to be approximately invariant. Doing so breaks the indifference of $\mathcal{L}_{\text{equi}}$ between large and small rotations, biasing towards small. Specifically, we propose the following objective that combines our equivariant loss with InfoNCE $\mathcal{L}_{\text{CARE}}(f) = \mathcal{L}_{\text{inv}}(f) + \mathcal{L}_{\text{unif}}(f) + \lambda\mathcal{L}_{\text{equi}}(f)$ where λ weights the equivariant loss.

3.1. Theoretical properties of the orthogonally equivariant loss

Proposition 1. *Suppose $\mathcal{L}_{equi}(f) = 0$. Then for almost every $a \in \mathcal{A}$, there is an orthogonal matrix $R_a \in O(d)$ such that $f(a(x)) = R_a f(x)$ for almost all $x \in \mathcal{X}$.*

Figure 1 illustrates this result. This result can be expressed as the existence of a mapping $\rho : \mathcal{A} \to O(d)$ that encodes the space of augmentations within $O(d)$. This raises a natural question: how much of the structure of $\mathcal{A}$ does this encoding preserve?

Corollary 1. *If $\mathcal{L}_{equi}(f) = 0$, then $\rho : \mathcal{A} \to O(d)$ given by $\rho(a) = R_a$ satisfies $\rho(a' \circ a) = \rho(a')\rho(a)$ for almost all a, a'. That is, ρ defines a group action on $\mathbb{S}^{d-1}$ up to a set of measure zero.*

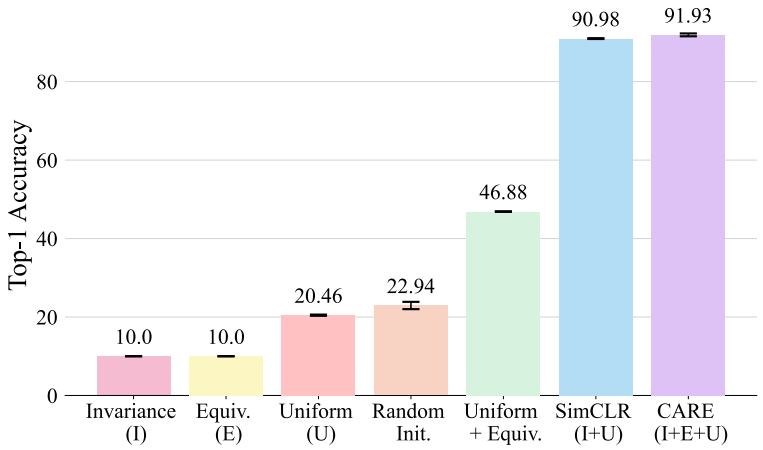

Figure 2: Ablating different loss terms. Combining $\mathcal{L}_{\text{equi}}$ with a uniformity promoting non-collapses term suffices to learn non-trivial features. However, optimal performance is achieved when encouraging *smaller* rotations, as in CARE. ResNet-50 models pretrained on CIFAR10 and evaluated with linear probes.

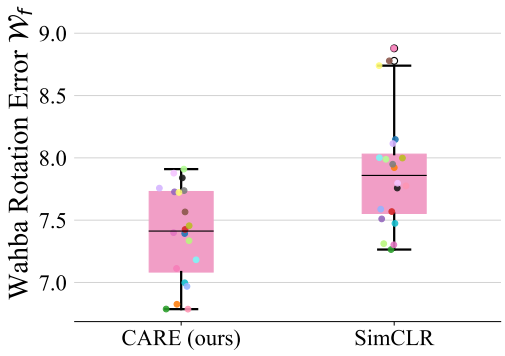

Figure 3: CARE learns a representation space with better rotational equivariance. We compare the models by the error of optimally rotating a set of embeddings to match the embeddings of augmented inputs, known as Wahba's problem (Section 4).

Formally, this result states that if $\mathcal{A}$ is a semi-group, then $\rho : \mathcal{A} \to O(d)$ defines a group homomorphism, or a linear group representation of $\mathcal{A}$ (Curtis & Reiner, 1966). This property does not hold for non-linear actions (Devillers & Lefort, 2023).

3.2. Extensions to other groups

Notably, the computation of $\mathcal{L}_{\text{equi}}$ solely relies on pairwise data instances $x, x' \in \mathcal{X}$, so it naturally aligns with the contrastive learning paradigm that already works with pairs of data. By changing the inner product, our method applies to other groups that are defined as stabilizers of bilinear forms, such as the Lorentz group, or the symplectic group.

Such extensions to other groups also allow us to use CARE for different embedding space geometries, such as hyperbolic space for self-supervised learners (Ge et al., 2022). If we constrain our embedding to a hyperboloid model of hyperbolic space, then linear isometries of this space are precisely the Lorentz group. Hence, using our equivariance loss with the Minkowski inner product replacing the Euclidean inner product would allow us to learn hyperbolic representations that transform the embeddings according to the action of the Lorentz group. Further discussions on extensions to other groups and geometries are given in Appendix D.

4. Measuring orthogonal action on embedding space

Wahba's problem. We sample a batch of data $\{x_i\}_{i=1}^n$ and an augmentation a and measure how applying a transforms the embeddings of each x_i consistently. Let F and $F_a \in \mathbb{R}^{d\times n}$ have ith columns $f(x_i)$ and $f(a(x_i))$ respectively, then we compute the error $\mathcal{W}_f = \min_{R\in SO(d)} \|RF - F_a\|_{\text{Fro}}$. Here, $\|\cdot\|_{\text{Fro}}$ represents the Frobenius norm. If $\mathcal{W}_f = 0$, it means that $f(a(x_i)) = R_a f(x_i)$ holds for all i. This problem is widely known as *Wahba's problem*.

Relative rotational equivariance. We define a metric for measuring the equivariance *relative* to the invariance of f, $\gamma_f = \mathbb{E}_{a\sim\mathcal{A}}\mathbb{E}_{x,x'\sim\mathcal{X}}\left\{\frac{(\|f(a(x'))-f(a(x))\|^2-\|f(x')-f(x)\|^2)^2}{(\|f(a(x'))-f(x')\|^2+\|f(a(x))-f(x)\|^2)^2}\right\}$. Details about the metric and the corresponding experimental results are provided in Appendix H.2.1

5. Experiments

We examine the representations learned by CARE, as well as those obtained from purely invariance-based contrastive approaches. We study three aspects of our model: 1) qualitative measures of orthogonal equivariance, 2) quantitative evaluation of the effect of equivariance on sensitivity to data transforms, and 3) performance of features learned by CARE on image classification tasks. Results for qualitative measures and detailed experiment configurations are presented in Appendix H.2.1 and G, respectively.

5.1. Quantitative measures for orthogonal equivariance

Wahba's Problem We compare ResNet-18 models pretrained with CARE and with SimCLR on CIFAR10. For each model, we compute the optimal value $\mathcal{W}_f$ of Wahba's problem, as introduced in Section 4, over repeated trials. In each trial, we sample a single augmentation $a \sim \mathcal{A}$ at

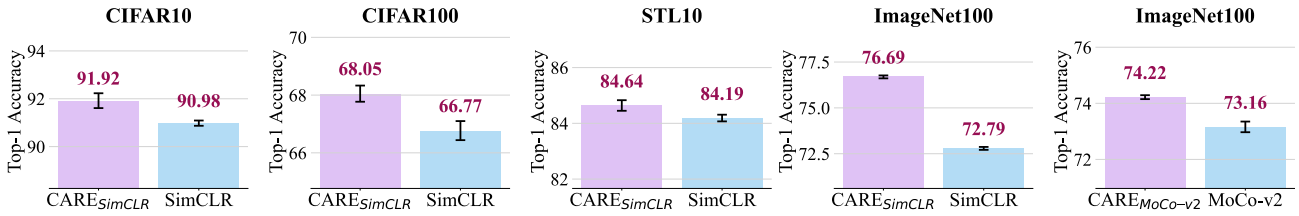

Figure 4: Top-1 linear readout accuracy (%) on CIFAR10, CIFAR100, STL10 and ImageNet100. All results are from 5 independent seed runs for the linear probe. We refer to the model trained using CARE with SimCLR or MoCo-v2 backbone as $\text{CARE}_{\text{SimCLR}}$ and $\text{CARE}_{\text{MoCo-v2}}$ respectively.

random and compute $\mathcal{W}_f$ for $f = f_{\text{CARE}}$ and $f = f_{\text{SimCLR}}$ over the test data. We repeat this process 20 times and plot the results in Figure 3, where the colors of dots indicate the sampled augmentation. Results show that CARE has a lower average error and worst-case error. Furthermore, comparing point-wise for a single augmentation, CARE achieves lower error in nearly all cases.

Results for relative rotational equivariance metric are reported in Appendix H.2.1

5.2. Linear probe for image classification

We examine the quality of features learned by CARE for solving image classification tasks on four benchmarks: CIFAR10, CIFAR100, STL10, and ImageNet100 using CARE, SimCLR and MoCo-v2 (see Appendix G for details). Figure 4 shows consistent improvements in performance using CARE, showing the benefits of our structured embedding approach for image recognition tasks.

Detailed discussion about the limitations and broader impact of our work is provided in Appendix I.

6. Acknowledgements

This research was supported by NSF award CCF-2112665. Sharut Gupta is supported by MIT Presidential Fellowship. Derek Lim is supported by National Science Foundation Graduate Research Fellowship. We acknowledge MIT SuperCloud and Lincoln Laboratory Supercomputing Center (Reuther et al., 2018) for providing HPC resources that have contributed to this work. We wish to thank Michael Murphy for insightful discussions on extensions of our method to biology.

References

Bhardwaj, S., McClinton, W., Wang, T., Lajoie, G., Sun, C., Isola, P., and Krishnan, D. Steerable equivariant representation learning. *preprint arXiv:2302.11349*, 2023.

Bordes, A., Usunier, N., Garcia-Duran, A., Weston, J., and Yakhnenko, O. Translating embeddings for modeling multi-relational data. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 26, 2013.

Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning (ICML)*, pp. 1597–1607. PMLR, 2020.

Curtis, C. W. and Reiner, I. *Representation theory of finite groups and associative algebras*, volume 356. American Mathematical Soc., 1966.

Dangovski, R., Jing, L., Loh, C., Han, S., Srivastava, A., Cheung, B., Agrawal, P., and Soljačić, M. Equivariant contrastive learning. In *International Conference on Learning Representations (ICLR)*, 2022.

Devillers, A. and Lefort, M. Equimod: An equivariance module to improve self-supervised learning. In *International Conference on Learning Representations (ICLR)*, 2023.

Garrido, Q., Najman, L., and Lecun, Y. Self-supervised learning of split invariant equivariant representations. *preprint arXiv:2302.10283*, 2023.

Ge, S., Mishra, S., Kornblith, S., Li, C.-L., and Jacobs, D. Hyperbolic contrastive learning for visual representations beyond objects. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

Ma, Y., Tsao, D., and Shum, H.-Y. On the principles of parsimony and self-consistency for the emergence of intelligence. *Frontiers of Information Technology & Electronic Engineering*, 23(9):1298–1323, 2022.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*, pp. 8748–8763. PMLR, 2021.

Reuther, A., Kepner, J., Byun, C., Samsi, S., Arcand, W., Bestor, D., Bergeron, B., Gadepally, V., Houle, M., Hubbell, M., et al. Interactive supercomputing on 40,000 cores for

machine learning and data analysis. In *2018 IEEE High Performance extreme Computing Conference (HPEC)*, pp. 1–6. IEEE, 2018.

Serre, J.-P. et al. *Linear representations of finite groups*, volume 42. Springer, 1977.

Shakerinava, M., Mondal, A. K., and Ravanbakhsh, S. Structuring representations using group invariants. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Wang, T. and Isola, P. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *International Conference on Machine Learning (ICML)*, pp. 9929–9939. PMLR, 2020.

Yasunaga, M., Bosselut, A., Ren, H., Zhang, X., Manning, C. D., Liang, P. S., and Leskovec, J. Deep bidirectional language-knowledge graph pretraining. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pp. 37309–37323, 2022.