Riemannian geometry of neural object representations

Jacob A. Zavatone-Veth [1] Sheng Yang [1] Julian A. Rubinfien [2]
Cengiz Pehlevan [1]

[1]Harvard University [2]Yale University

Introduction

- Identifying the correct set of observables to describe distributed neural representations is an important problem in both machine learning and in neuroscience.
- Here, we propose that the Riemannian geometry of neural representations provides a compelling candidate framework.
- This is inspired by work on geometric deep learning, which seeks to embed strong geometric inductive biases in neural networks [1]. In contrast, we investigate the geometry learned by neural networks without a strong prior.

Prerequisites: Pullback Metric and Volume Element

Consider a d dimensional input $\boldsymbol{x} \in \mathcal{D} \subseteq \mathbb{R}^d$ being fed to a feature map $\boldsymbol{\Phi} : \mathbb{R}^d \to \mathcal{H}$, with $\dim(\mathcal{H}) = n$. $\Phi(D) = \mathcal{M}$ constitutes a submanifold in $\mathcal{H}$. The flat metric in $\mathcal{H}$ can be pulled back to input space by

$$g_{\mu\nu} = \frac{\partial \Phi_i}{\partial x^\mu} \frac{\partial \Phi_i}{\partial x^\nu} \tag{1}$$

for $\mu, \nu \in [d]$, $i \in [n]$. Assuming sufficient smoothness of Φ, the metric is nonsingular iff $n \geqslant d$. The local volume rate of change is measured by the volume element:

$$dV = \sqrt{\det g}\, d^d x \tag{2}$$

Our main interest in this work is the volume element expansion factor $\sqrt{\det g}$, which measures how local areas in the input space are magnified by the feature map Φ.

Main Contribution

We perform a preliminary investigation of the geometry learned by deep vision models, focusing on the volume element as the simplest observable to measure.

We test the following hypothesis, inspired by the work of [2] on adapting SVM kernels:

Hypothesis: *Deep neural networks trained to perform supervised classification tasks using standard gradient-based methods learn to magnify areas near decision boundaries.*

Shallow Network: Toy Task

Consider a single-hidden layer neural network with activation $\phi(.)$ and feature map given by $\Phi_j(x) = \frac{1}{\sqrt{n}} \phi(w_j \cdot x + b_j)$. Below shows an example trained on 2D sinusoidal dataset.

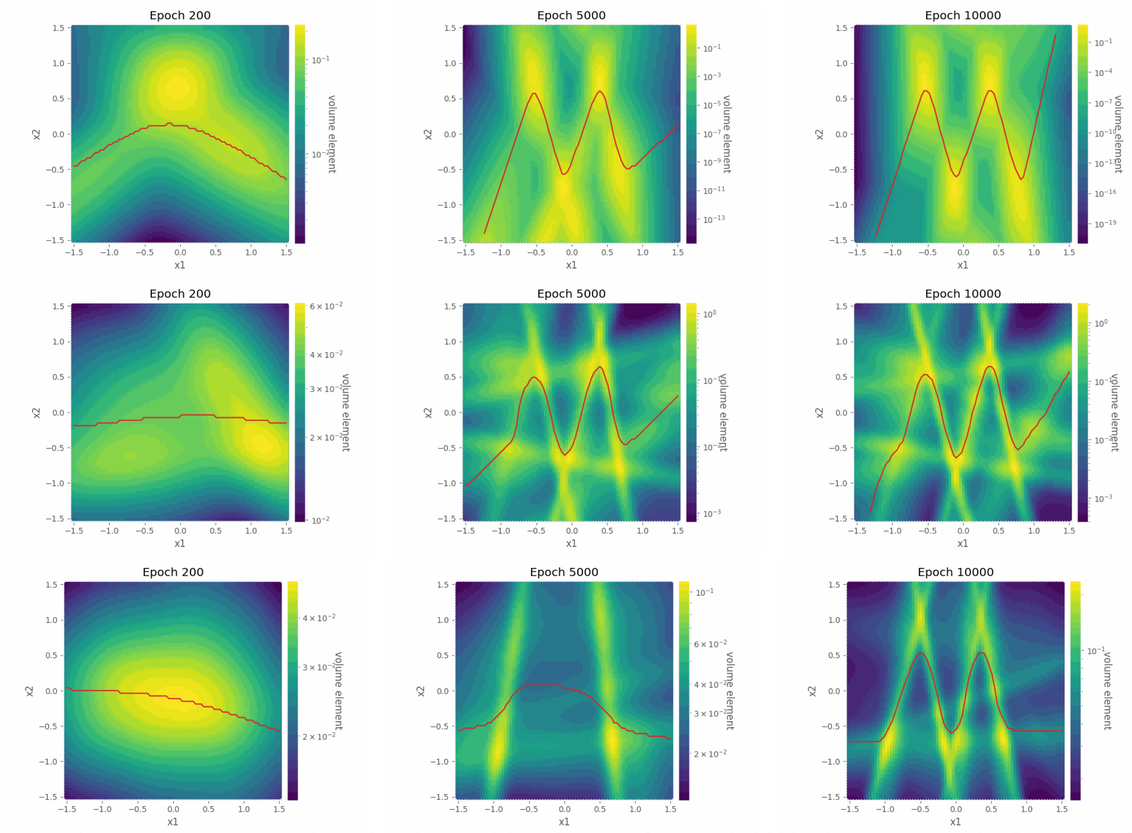

Figure 1. Volume element at start (left), mid (mid), end (right) of training for a single-hidden layer network with 5 (top row), 25 (mid row), 200 (bottom row) hidden units and Sigmoid activation.

Shallow networks trained on MNIST images

Figure 2. Volume elements and decisions at different stages of training on MNIST. Top panel: linear interpolation by 7 and 6; bottom panel: plane interpolation by 7, 6, and 1

Deep ResNets

Our observation that volume elements expand near the decision boundary holds for deep networks. We use ResNet34 with GELU activation trained on CIFAR10. Below shows the decision space and volume element progression across 4 blocks of ResNet34 model. Similar patterns hold for ReLU activations.

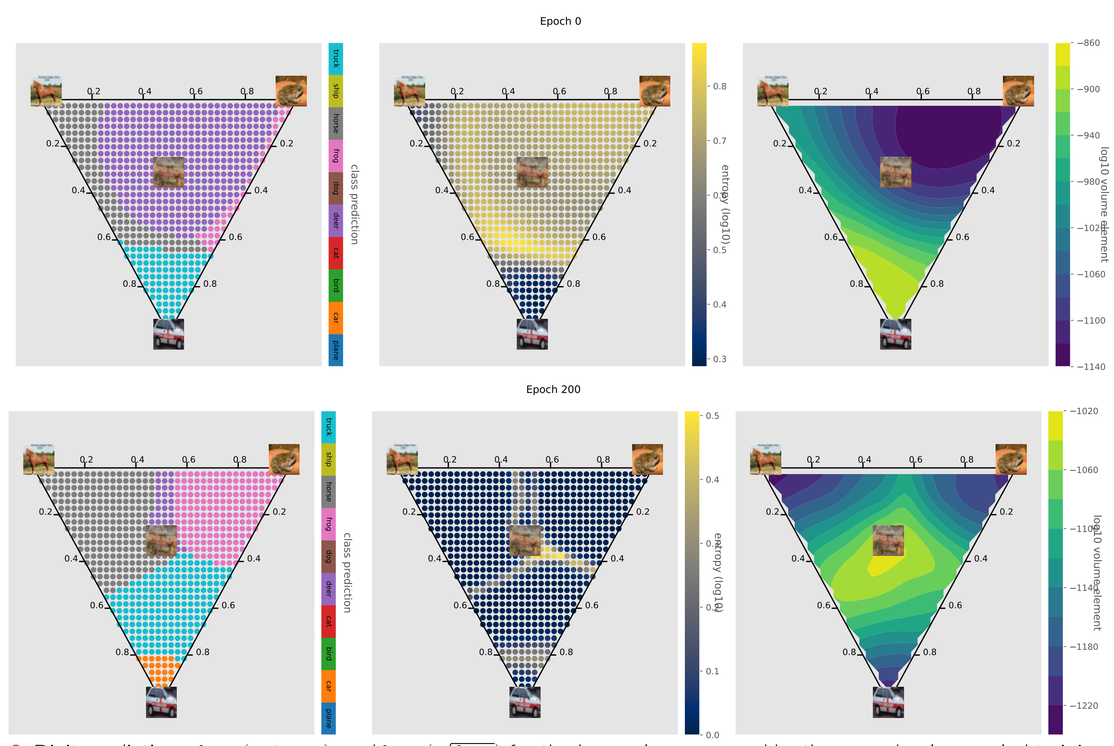

Figure 3. Digit predictions, $\log_{10}$(entropy), and $\log_{10}(\sqrt{\det g})$ for the hyperplane spanned by three randomly sampled training point a horse, a frog, and a car across different epochs.

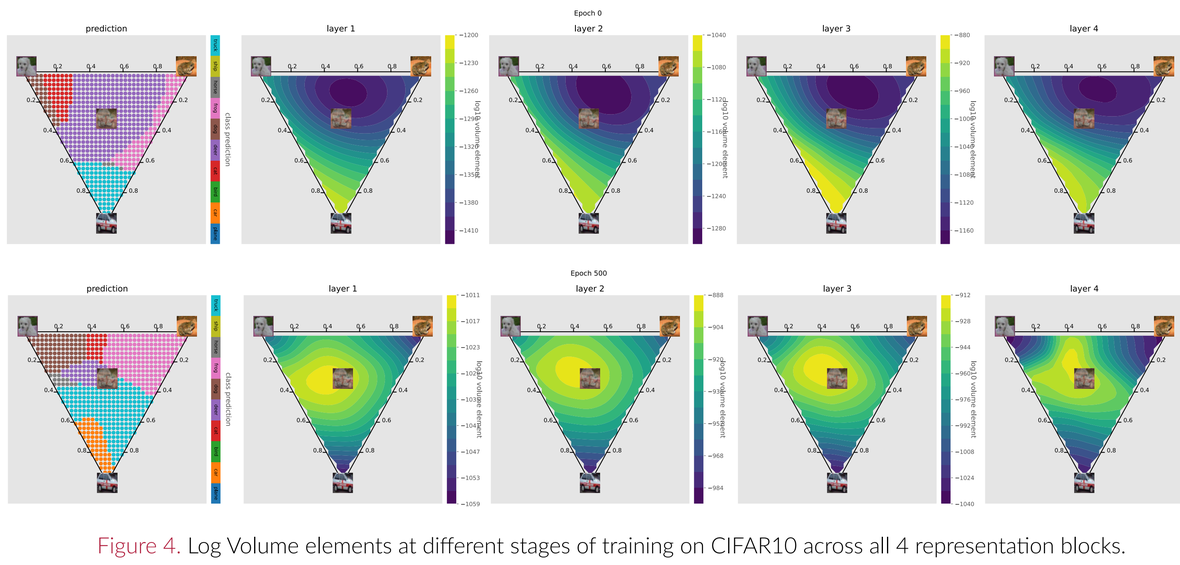

Figure 4. Log Volume elements at different stages of training on CIFAR10 across all 4 representation blocks.

Self-supervised learning with Barlow-Twins

To demonstrate the broader utility of visualizing the induced volume element, we consider ResNet feature maps trained with the self-supervised learning (SSL) method Barlow Twins[3].

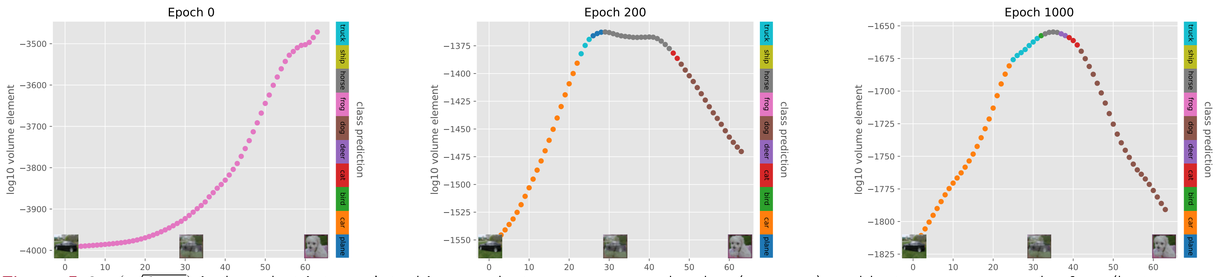

Figure 5. $\log(\sqrt{\det g})$ induced at interpolated images between a car and a dog (top row) and between a car and a frog (bottom row) by Barlow Twins with ResNet-34 backbone and a GELU activation.

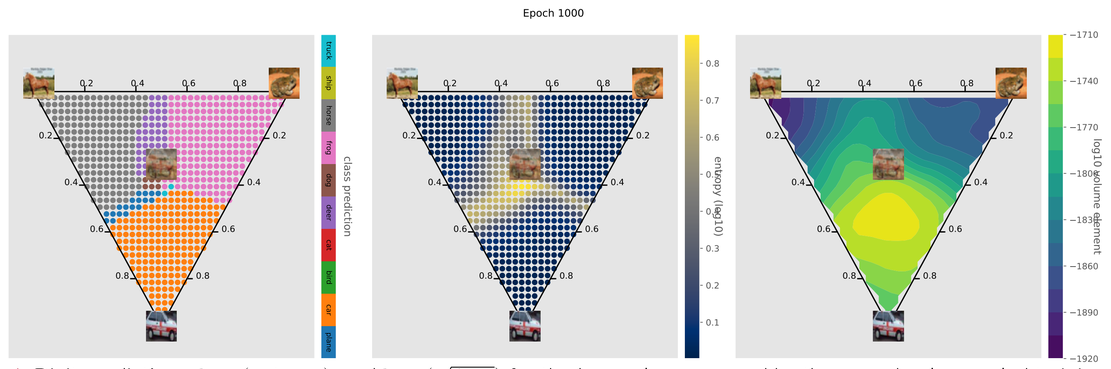

Figure 6. Digit predictions, $\log_{10}$(entropy), and $\log_{10}(\sqrt{\det g})$ for the hyperplane spanned by three randomly sampled training point a horse, a frog, and a car across different epochs for Barlow Twins with ResNet-34 backbone using GELU activation.

Can this be applied to biological neural representations?

As task-optimized artificial models that are believed to capture certain salient features of the ventral visual stream[4–6] magnify areas near object class boundaries, does the ventral stream itself do so? Recent works have probed certain aspects of local variation in tuning of small populations of neurons in ventral visual areas,[7,8] but a systematic geometric analysis of visual population codes is lacking. The crucial technical challenge in analyzing real neural data from a Riemannian perspective is obtaining reliable estimates of the response derivatives that define the induced metric. Fortunately, recent works have made progress towards computing such estimates.[7–9] Therefore, we propose that such an analysis is possible in the near future.

References

1. Bronstein, M. M. *et al. arXiv preprint arXiv:2104.13478* (2021).
2. Amari, S.-i. & Wu, S. *Neural Networks* (1999).
3. Zbontar, J. *et al.* in *International Conference on Machine Learning* (2021).
4. Yamins, D. L. K. & DiCarlo, J. J. *Nat. Neurosci.* ISSN: 1546-1726 (2016).
5. Richards, B. A. *et al. Nat. Neurosci.* ISSN: 1546-1726 (2019).
6. Feather, J. *et al. Nat. Neurosci.* ISSN: 1546-1726 (2023).
7. Wang, B. & Ponce, C. R. *Cell Rep.* ISSN: 2211-1247 (2022).
8. Wang, B. & Ponce, C. R. in *NeurReps* (2023).
9. Acosta, F. E. *et al. arXiv* (2022).

Acknowledgements

CP and JZV were supported by NSF Award DMS-2134157 and NSF CAREER Award IIS-2239780. CP is further supported by a Sloan Research Fellowship. This work has been made possible in part by a gift from the Chan Zuckerberg Initiative Foundation to establish the Kempner Institute for the Study of Natural and Artificial Intelligence.

jzavatoneveth@g.harvard.edu

pehlevan.seas.harvard.edu