Neural networks learn to magnify areas near decision boundaries

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

[1]Harvard University [2]Yale University

Introduction

- Geometric deep learning seeks to embed strong geometric inductive biases in neural networks [BBCV21].
- Here, we investigate the learned geometry in neural networks without a strong prior to gain insights in the data-dependent representational geometry over training.
- We focus on the feature representation and measure geometric quantities using the pullback metric induced by a flat metric in the feature space.

Prerequisites: Pullback Metric and Volume Element

We assume Einstein summation convention throughout. Consider a d dimensional input $\boldsymbol{x} \in \mathcal{D} \subseteq \mathbb{R}^d$ being fed to a feature map $\boldsymbol{\Phi} : \mathbb{R}^d \to \mathcal{H}$ where $\mathcal{H}$ is a separable Hilbert space of possibly infinite dimension n, then $\Phi(\mathcal{D}) = \mathcal{M}$ 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 test the following hypothesis, inspired by the work of [AW99] 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.*

Our primary contributions are:

- We analytically derive volume element and scalar curvature for feature map of a shallow neural network with finite and infinite hidden units, Gaussian weights and smooth activations.
- We provide empirical evidence of the hypothesis by training multiple dataset (simulated, MNIST, CIFAR10) using both shallow and deep neural networks and visualize the volume element landscape.
- We extend the framework to self-supervised learning methods such as Barlow-Twins, showing that expansion emerges also in the absence of supervision.

Shallow Network: Analytic Computations

Consider a single-hidden layer neural network with activation $\phi(.)$ and feature map given by $\Phi_j(x) = \frac{1}{\sqrt{n}}\phi(w_j \cdot x + b_j)$. Exploiting the symmetry under permutation of indices, the input space metric and volume element expansion factor are given by

$$g_{\mu\nu} = \frac{1}{n}\phi'(z_j) w_{j\mu} w_{j\nu}, \quad \det g = \frac{1}{n^d d!} M^2_{j_1 \cdots j_d} \phi'(z_{j_1})^2 \cdots \phi'(z_{j_d})^2 \tag{3}$$

where $z_j = w_j x + b_j$ are the preactivation patterns and $M_{j_1 \cdots j_d} = \det([w_{j_i i}]_{1 \leq i \leq d})$ is the minor of the weight matrix obtained by selecting units $j_1, \ldots, j_d$. Below shows an example trained on 2D sinusoidal dataset.

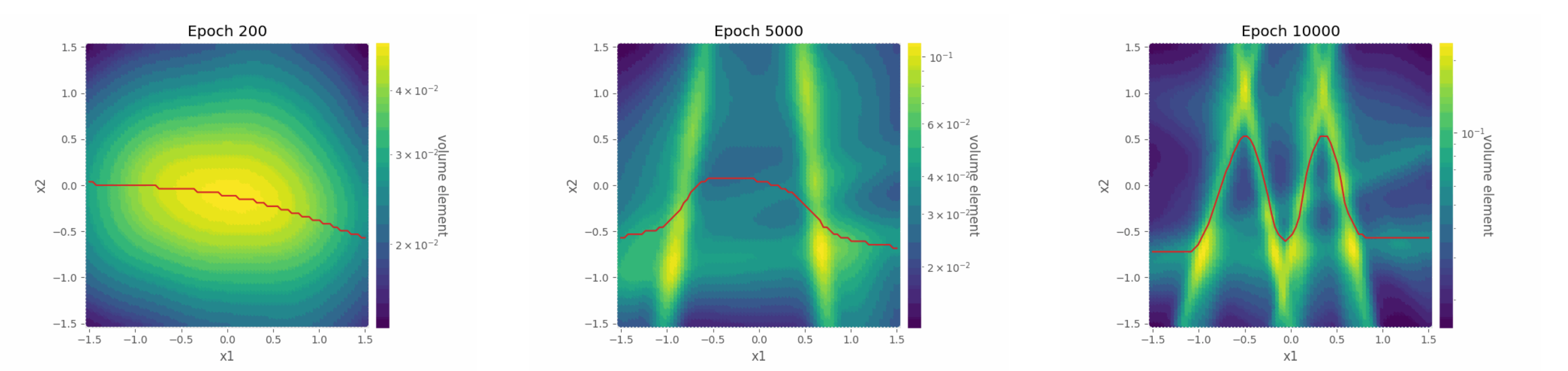

Figure 1. Volume element at start (left), mid (mid), end (right) of training for a single-hidden layer network with 250 hidden units and Sigmoid activation.

Shallow Network: Empirical Evidence with MNIST

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 Network: Empirical Evidence with CIFAR10

Our observation that volume element expand near the decision boundary is consistent with deep network. We use ResNet34 with GELU activation trained on CIFAR10 to demonstrate the findings. Below shows the decision space and volume element progression across 4 blocks of ResNet34 model. Similar pattern has been observed on non-smooth activation such as ReLU.

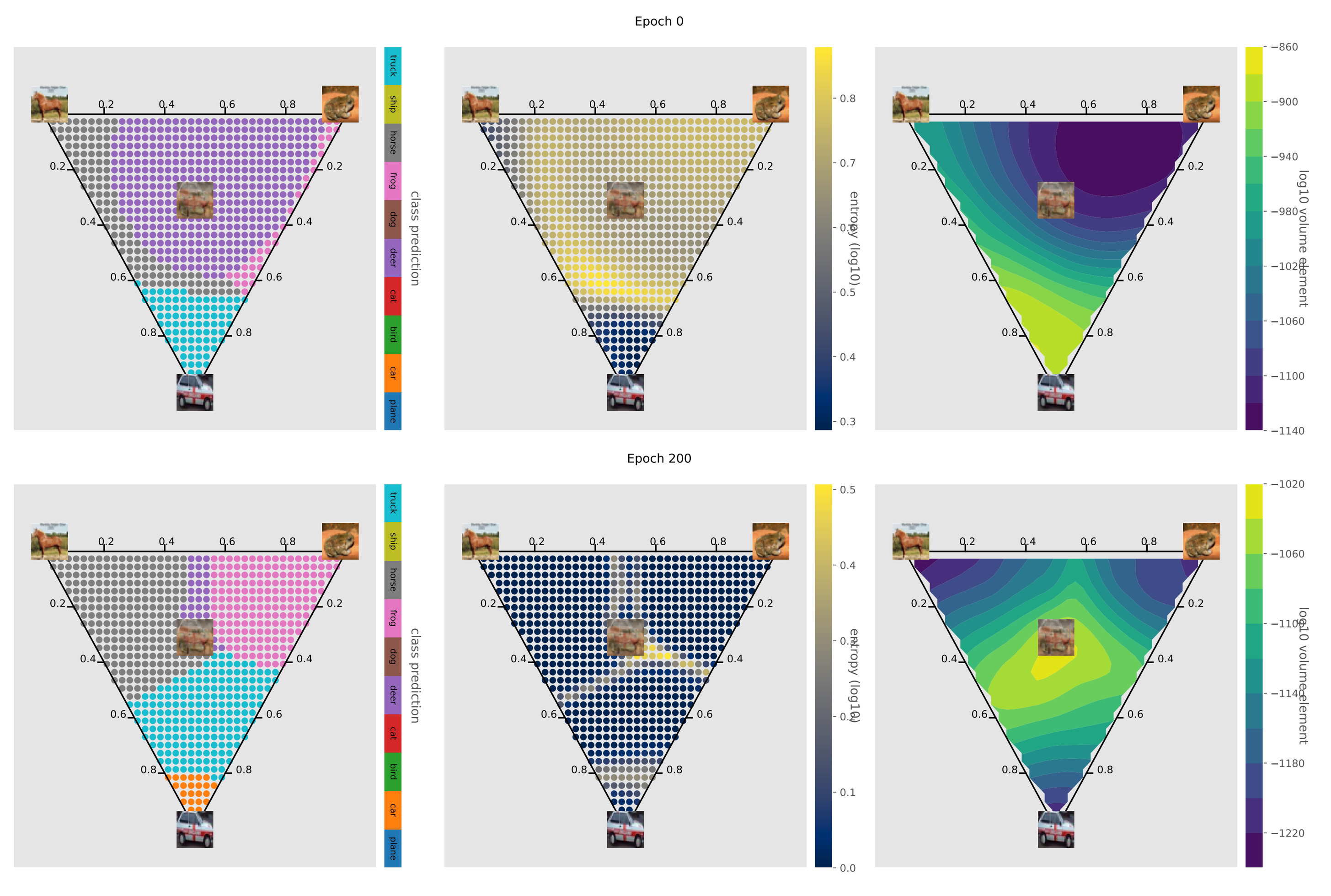

Figure 3. Digit predictions, $\log_{10}$(entropy), and $\log_{10}(\sqrt{\det g})$ for the hyperplane spanned by three randomly sampled training point a horse, a frog, and a car across different epochs.

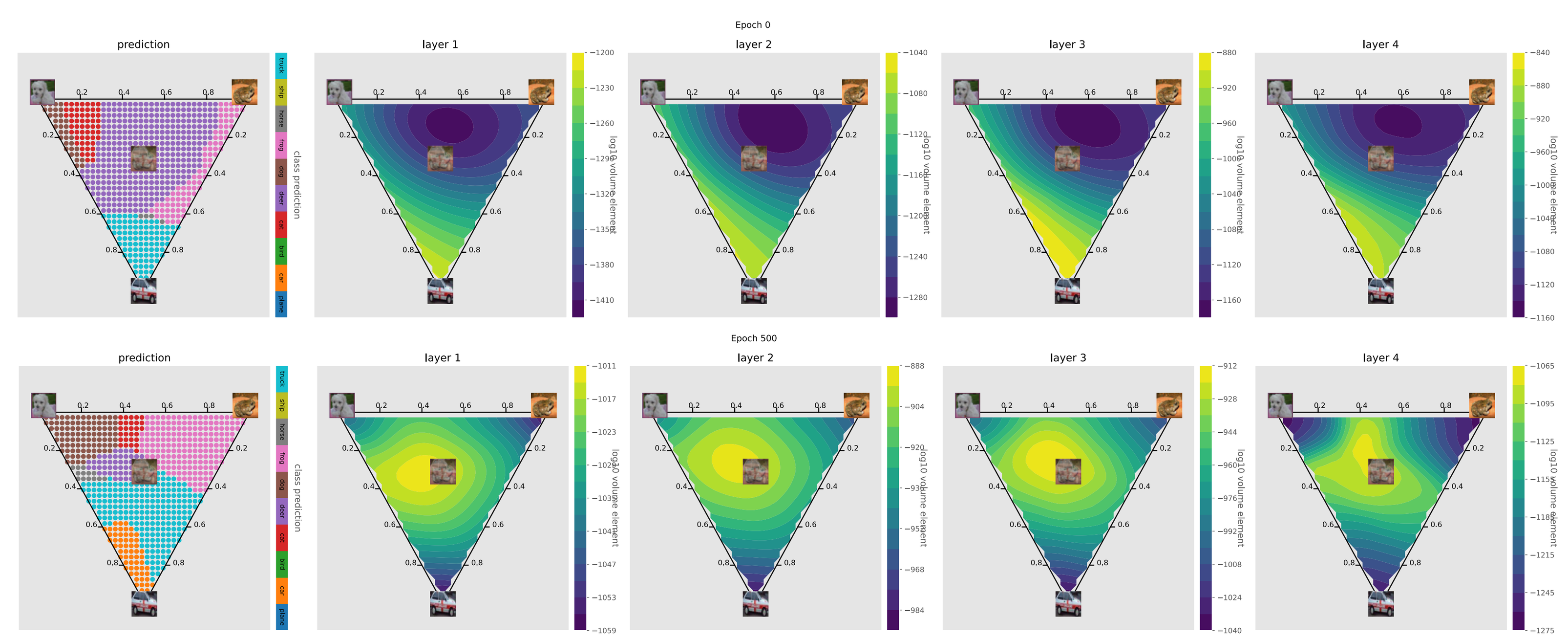

Figure 4. Log Volume elements at different stages of training on CIFAR10 across all 4 representation blocks.

Extension to Self-Supervised Learning: example 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[ZJM+21].

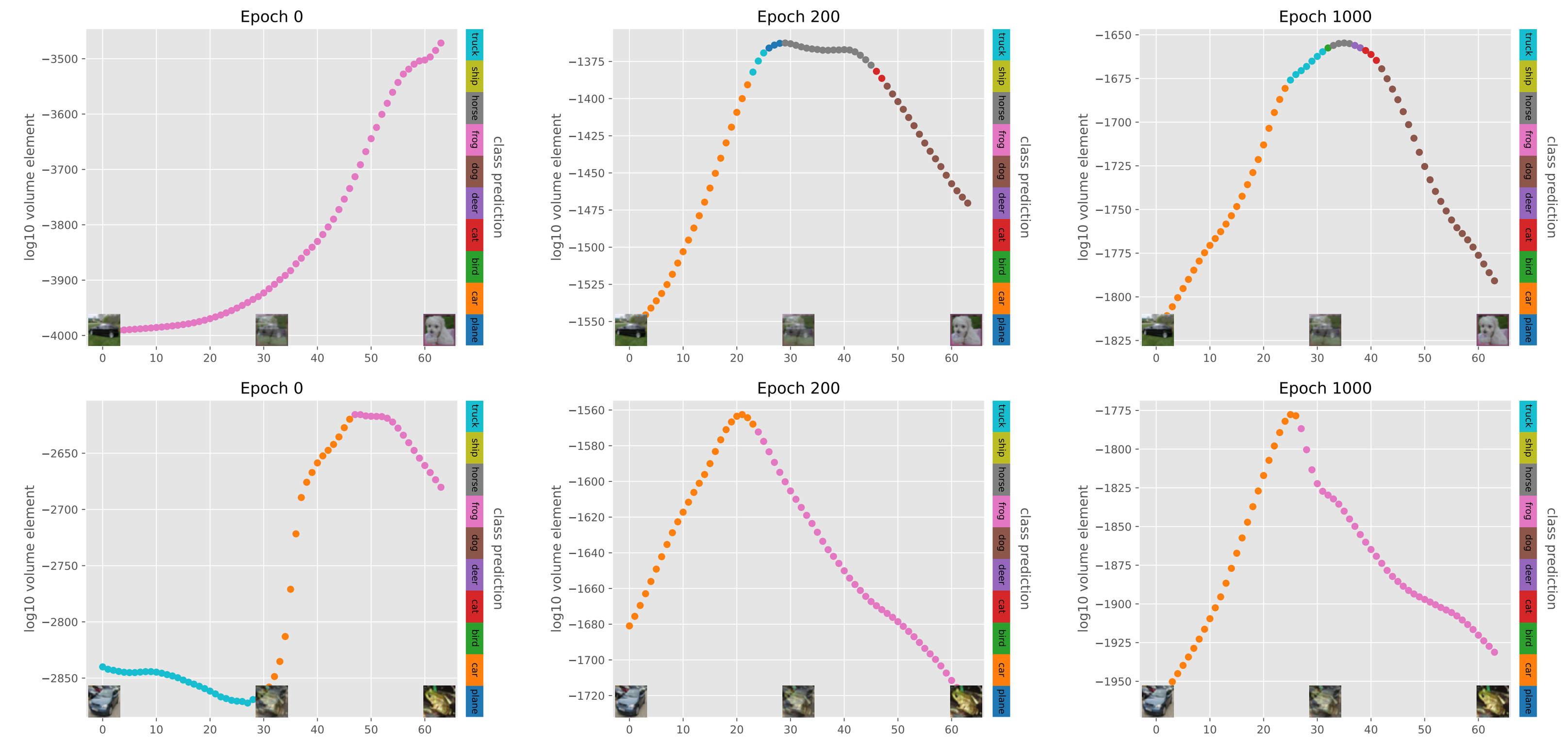

Figure 5. $\log(\sqrt{\det g})$ induced at interpolated images between a car and a dog (top row) and between a car and a frog (bottom row) by Barlow Twins with ResNet-34 backbone and a GELU activation.

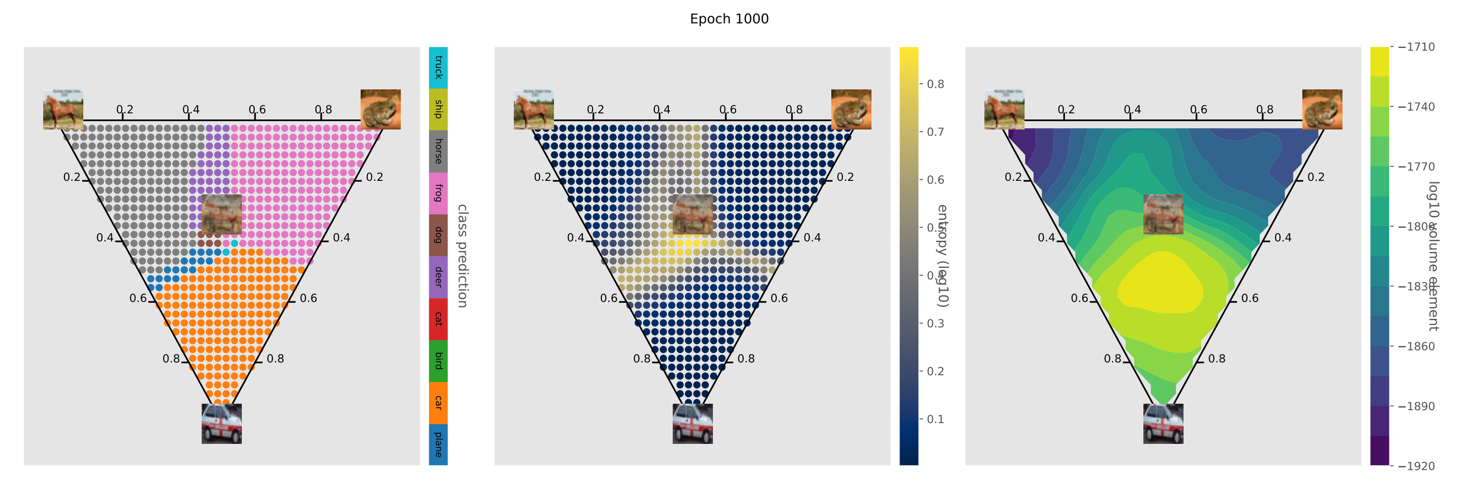

Figure 6. Digit predictions, $\log_{10}$(entropy), and $\log_{10}(\sqrt{\det g})$ for the hyperplane spanned by three randomly sampled training point a horse, a frog, and a car across different epochs for Barlow Twins with ResNet-34 backbone using GELU activation.

Future Work: Can this be used to improve generalization?

We believe the following lines of works are crucial in continuing these geometric investigation of the neural network;

- use Riemannian metric to provide adversarial robustness guarantee (in progress);
- geometrically characterize different regimes of feature learning;
- apply Riemannian metric computations to semantic data deduplications;
- a thorough evaluation of pretrained neural network without access to labels.

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.

References

[AW99] Shun-ichi Amari and Si Wu. Improving support vector machine classifiers by modifying kernel functions. *Neural Networks*, 12(6):783–789, 1999.

[BBCV21] Michael M Bronstein, Joan Bruna, Taco Cohen, and Petar Veličković. Geometric deep learning: Grids, groups, graphs, geodesics, and gauges. *arXiv preprint arXiv:2104.13478*, 2021.

[ZJM+21] Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction. In *International Conference on Machine Learning*, pages 12310–12320. PMLR, 2021.