Understanding Feature Learning in Neural Networks via Manifold Capacity and Effective Geometry

Chi-Ning Chou (cchou@flatironinstitute.org)
Center for Computational Neuroscience, Flatiron Institute, New York, USA

Hang Le (nguyethang2205@gmail.com)
Center for Computational Neuroscience, Flatiron Institute, New York, USA

Yichen Wang (yichenwang@flatironinstitute.org)
Center for Computational Neuroscience, Flatiron Institute, New York, USA

SueYeon Chung (schung@flatironinstitute.org)
Center for Computational Neuroscience, Flatiron Institute, New York, USA

Abstract

Humans learn to perform complicated tasks through incorporating task-relevant features into neural representations in the brain. This ability, known as feature learning, has been widely demonstrated in various brain areas as well as artificial neural networks. However, fundamental questions, such as quantifying the degree of feature learning and gaining mechanistic understanding of feature learning, remain elusive. In this work, we propose the utilization of manifold capacity theory to understand feature learning. Manifold capacity has been shown to quantify task-relevant coding efficiency of neural representations beyond training and testing accuracy. The increase in capacity alongside learning can thus be considered a signature of task-relevant feature learning. Moreover, capacity is analytically linked to effective geometric measures such as manifold radius and dimension. As a consequence, the dynamics of effective manifold geometry can further elucidate the underlying mechanisms of feature learning. We demonstrate the applicability of using manifold capacity and effective geometry to understand feature learning though artificial neural networks. Concretely, we use these quantitative measures as mesoscopic descriptors to describe different learning strategies and stages throughout learning. Moreover, we use these understanding to explain how neural networks generalize to other tasks with a distribution shift.

Keywords: feature learning; neural representations; population geometry; manifold capacity

Introduction

From navigating in a new city, adapting novel motor skills, to learning new cognitive tasks, our brain undergoes changes in its circuitry. Specifically, learning is reflected through incorporating task-relevant information and features into neural representations (Hubel & Wiesel, 1959). Furthermore, feature learning lends a new facet to study the underlying learning mechanism in artificial neural networks (Farrell, Recanatesi, & Shea-Brown, 2023). Despite extensive theoretical endeavors (Geiger, Spigler, Jacot, & Wyart, 2020),(Ba et al., 2022) to pinpoint and elucidate the genesis of feature learning in simplified models, several fundamental research inquiries persist: (i) How can we quantify the degree of feature learning in realistic models? (ii) How to summarize the high-dimensional changes of features into interpretable and task-relevant descriptors?

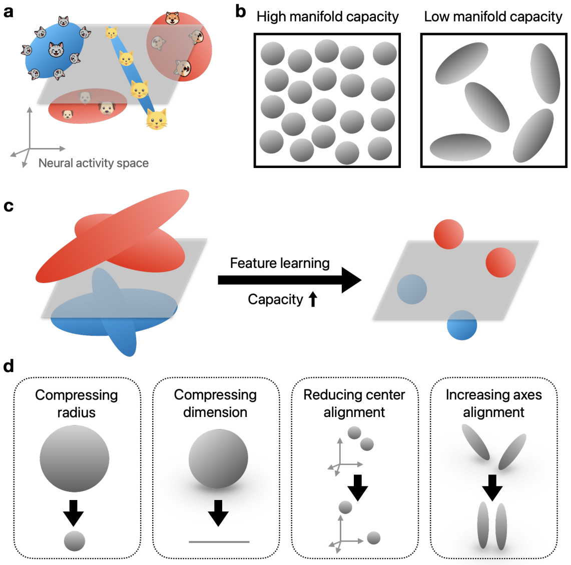

Figure 1: **Understand feature learning via manifold capacity and effective manifold geometry.** **a**, Neural manifolds are collections of neural activities corresponding to the same task-condition in the neural state space. For example, the neural responses to all cat stimuli constitute the cat manifold. **b**, Manifold capacity quantifies the packing efficiency (i.e., the number of manifolds per neurons) of neural representations. The manifolds on the left have higher capacity than that on the right because one can pack more manifolds in the neural state space. **c**, We propose the use of the increment of manifold capacity to quantify the degree of feature learning. **d**, The effective geometric measures in GCMC further provide intermediate-level descriptors for different learning strategies that lead to feature learning.

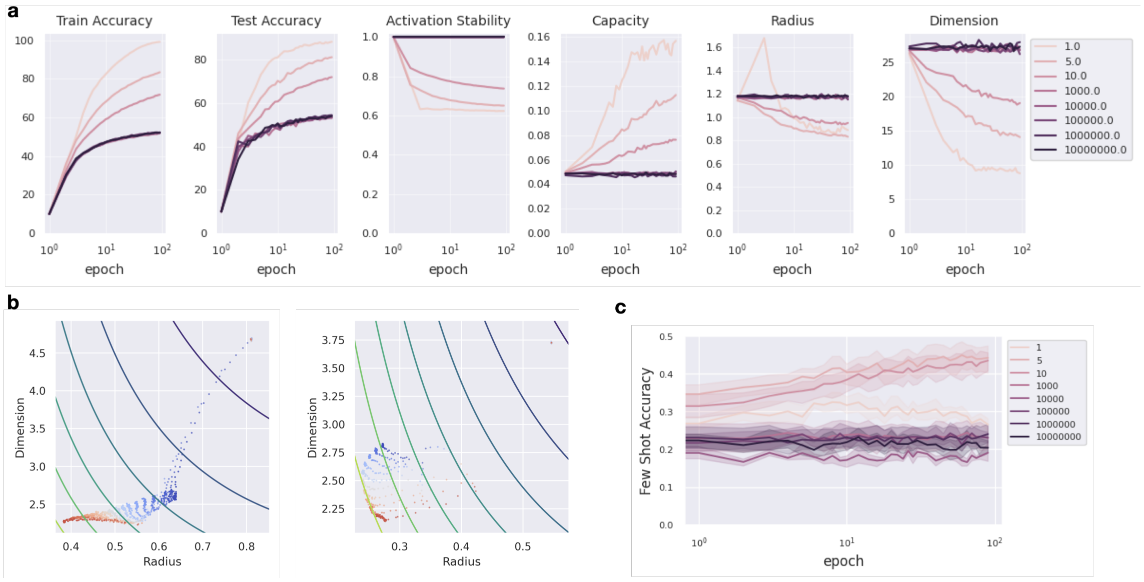

Figure 2: **Manifold capacity and effective manifold geometry quantify feature learning. a**, We train a VGG-11 on CIFAR-10 with various scaling factor α (the larger the scaling factor, the lazier the training as in Ref. (Chizat et al., 2019). We show that capacity tracks the degree of feature learning as consistent with a heuristic measure (i.e., activation stability, the percentage of neurons over ReLU layers that, after training, are activated for the same inputs at initialization) used in Ref. (Chizat et al., 2019). **b**, Effective manifold geometric measures describe the underlying learning strategy of feature learning. In these plots, the x-axis is the effective radius and the y-axis is the effective dimension. The contour is the manifold capacity, which can be approximated by a function of radius and dimension as shown in (Chung et al., 2018; Chou et al., 2024). We train 2-layer neural networks on synthetic data with different learning rates and data generative models. The color from blue to red represents learning rate from low to high. The dots with the same color correspond to the same 2-layer network during different training epoch. Left: We train 2-layer neural networks with multiple random binary readouts. Here, by increasing the degree of feature learning (from blue to red), the vanilla gradient descent focuses on compressing the radius. Right: We train 2-layer neural networks with Gaussian point clouds with larger signal-to-noise ratio. Here, by increasing the degree of feature learning (from blue to red), the vanilla gradient descent focuses on compressing the dimension. **c**, Using GCMC to understand a few-shot learning task. Left: We consider the same scenario as in part (a). As scaling factor decreases (dark to light), the capacity monotonically increases while both radius and dimension decrease, suggesting the transition from lazy to rich training. Interestingly, the radius bounces back when the scaling factor closes to 1. Right: We measure the few-shot learning accuracy (Snell et al., 2017) of CIFAR-100 on each model at each epoch and plot the result. The few-shot accuracy improves upon the transition to feature learning. As the scaling factor approaches 1, the drop of few-shot accuracy is explained by the growth of radius.

What are the underlying learning strategies in different models and learning stages? (iii) How does the improvement of features contribute to computational benefits beyond training and test accuracy? In particular, does improved features counteract a distribution shift?

The Manifold Capacity Theory (MCT) (Chung et al., 2018) quantifies the neural manifold's representational efficiency through the classification capacity (Gardner, 1988), which measures the amount of linearly decodable information per neuron (Fig. 1a). The MCT analytically characterizes the classification capacity as a function of the shape of a manifold (Chung et al., 2018; Wakhloo, Sussman, & Chung, 2023). The effective Geometric measures from Correlated Manifold Capacity theory (GCMC) (Chou et al., 2024) further suggests the definition of computationally relevant geometric terms such as effective dimension and effective radius of neural manifolds. GCMC as well as its predecessor have been shown to capture the task-relevant structures in neural representations in both biological datasets (Chou et al., 2024; Yao et al., 2023; Paraouty et al., 2023; Froudarakis et al., 2020) and artificial neural networks (Cohen, Chung, Lee, & Sompolinsky, 2020; Dapello et al., 2021; Kuoch et al., 2023).

Results

In this work, we propose the usage of GCMC as a quantification method for understanding feature learning. First, we show that manifold capacity quantifies the degree of feature learning (Fig. 2a). Intuitively, manifold capacity measures how efficient information is stored in neural representations for downstream readout. Hence, it serves as a well-normalized mathematical definition for feature learning. In artificial neural networks, we demonstrate that capacity, as a quantification for lazy to rich learning, is well-aligned with previous methods, such as generalization accuracy, optimization path, and kernel methods (Chizat et al., 2019; Geiger et al., 2020). Second, the analytical theory in GCMC has induced effective geometric measures to summarize the different geometric attributes to efficient neural representations at the mesoscopic level (Fig. 1d). We use these effective geometric measures to characterize the task-relevant geometric changes in feature learning (Fig. 2b). For example, in two-layer networks, we demonstrate how the differences in training data and/or task can lead to different learning strategies during learning. Finally, we use effective geometry to explain computational consequences of feature learning. As an example, we consider a few-shot learning task to study how feature learning facilitates generalization in the presence of distribution shifts. We find that capacity and effective radius explain the performance of a few-shot learning task (Fig. 2c). In summary, manifold capacity and effective geometry open the door to study feature learning across the task and representational level.

References

Ba, J., Erdogdu, M. A., Suzuki, T., Wang, Z., Wu, D., & Yang, G. (2022). High-dimensional asymptotics of feature learning: How one gradient step improves the representation. *Advances in Neural Information Processing Systems*, *35*, 37932–37946.

Chizat, L., Oyallon, E., & Bach, F. (2019). On lazy training in differentiable programming. *Advances in neural information processing systems*, *32*.

Chou, C.-N., Arend, L., Wakhloo, A. J., Kim, R., Slatton, W., & Chung, S. (2024). Neural manifold capacity captures representation geometry, correlations, and task-efficiency across species and behaviors. *bioRxiv*.

Chung, S., Lee, D. D., & Sompolinsky, H. (2018). Classification and geometry of general perceptual manifolds. *Physical Review X*.

Cohen, U., Chung, S., Lee, D. D., & Sompolinsky, H. (2020). Separability and geometry of object manifolds in deep neural networks. *Nature communications*.

Dapello, J., Feather, J., Le, H., Marques, T., Cox, D., McDermott, J., ... Chung, S. (2021). Neural population geometry reveals the role of stochasticity in robust perception. *Advances in Neural Information Processing Systems*, *34*, 15595–15607.

Farrell, M., Recanatesi, S., & Shea-Brown, E. (2023). From lazy to rich to exclusive task representations in neural networks and neural codes. *Current opinion in neurobiology*, *83*, 102780.

Froudarakis, E., Cohen, U., Diamantaki, M., Walker, E. Y., Reimer, J., Berens, P., ... Tolias, A. S. (2020). Object manifold geometry across the mouse cortical visual hierarchy. *BioRxiv*.

Gardner, E. (1988). The space of interactions in neural network models. *Journal of physics A: Mathematical and general*, *21*(1), 257.

Geiger, M., Spigler, S., Jacot, A., & Wyart, M. (2020). Disentangling feature and lazy training in deep neural networks. *Journal of Statistical Mechanics: Theory and Experiment*, *2020*(11), 113301.

Hubel, D. H., & Wiesel, T. N. (1959). Receptive fields of single neurones in the cat's striate cortex. *Journal of Physiology*.

Kuoch, M., Chou, C.-N., Parthasarathy, N., Dapello, J., DiCarlo, J. J., Sompolinsky, H., & Chung, S. (2023). Probing biological and artificial neural networks with task-dependent neural manifolds. In *Conference on parsimony and learning (proceedings track).*

Paraouty, N., Yao, J. D., Varnet, L., Chou, C.-N., Chung, S., & Sanes, D. H. (2023). Sensory cortex plasticity supports auditory social learning. *Nature communications*, *14*(1), 5828.

Snell, J., Swersky, K., & Zemel, R. (2017). Prototypical networks for few-shot learning. *Advances in neural information processing systems*, *30*.

Wakhloo, A. J., Sussman, T. J., & Chung, S. (2023). Linear classification of neural manifolds with correlated variability. *Physical Review Letters*.

Yao, J. D., Zemlianova, K. O., Hocker, D. L., Savin, C., Constantinople, C. M., Chung, S., & Sanes, D. H. (2023). Transformation of acoustic information to sensory decision variables in the parietal cortex. *Proceedings of the National Academy of Sciences*, *120*(2), e2212120120.