Equivariant Self-Supervised Learning Improves IT Predictivity

Thomas Yerxa (tey214@nyu.edu)
Center for Neural Science, New York University

Jenelle Feather (jfeather@flatironinstitute.org)
Center for Computational Neuroscience, Flatiron Institute

Eero Simoncelli (eps2@nyu.edu)
Center for Computational Neuroscience, Flatiron Institute
and Center for Neural Science, New York University

SueYeon Chung (schung@flatironinstitute.org)
Center for Computational Neuroscience, Flatiron Institute
and Center for Neural Science, New York University

Abstract

We present a novel method for self-supervised learning of representations that are equivariant to a set of transformations. When trained on images, we demonstrate that the learned representations effectively factorize sources of variability in their inputs, and provide improved prediction of responses of cells in macaque visual area IT across four different datasets.

Keywords: self-supervised learning; factorized representations; brain-model alignment

Introduction

Task-trained deep neural networks have emerged as leading models of neural responses in the primate visual system, especially for later stages of the ventral stream such as inferotemporal (IT) cortex. One major criticism of this approach is that the tasks used to train such networks (predominantly object recognition) rely on large numbers of labeled examples, and are thus not ecologically plausible. Recent approaches in representation learning circumvent the need for labeled examples, and match or surpass supervised learning methods on a variety of tasks. These approaches generally rely on supervision signals extracted from the data, rather than from human annotations, and are thus called “self-supervised”. Many such methods utilize an objective function that encourages invariance to a particular set of image transformations, while simultaneously enforcing that distinct images are mapped to distinct representations (thereby preventing “collapse” to a trivial solution that is invariant across all inputs). However, this training is not well aligned with known characteristics of visual perception: transformations for which the network is encouraged to be invariant are generally quite perceptible to humans (Feather, Leclerc, Madry, & McDermott, 2023). Moreover, recent work has shown that the *factorization* of variability due to image transformations is more closely related to neural predictivity than the *lack* of variability (Lindsey & Issa, 2023).

In this work we introduce a novel self-supervsed learning method that trades off invariance (which discards information about the input transformation) and equivariance (which maintains information about the input transformation). Specifically, a representation is said to be equivariant with respect to some transformation of the inputs if the same transformation, applied to different inputs, results in the same change in the representation. We demonstrate that our equivariant learning approach produces representations that contain more ”category orthogonal” information, better factorize the sources of variability in the datasets, and better predict neural activity in visual area IT.

Method

Transformation Invariant Self-Supervised Learning (TiSSL)

Denote by $X \in \mathrm{R}^{B\times D}$ a dataset of images (i.e. ImageNet), and let $\tau(\cdot;\rho) : \mathrm{R}^D \rightarrow \mathrm{R}^D$ be a function parameterized by ρ that maps images to images (for example τ might be the random crop operation, in which case ρ could specify the region to be cropped). The goal of TiSSL algorithms is to learn the parameters W of some function $f(\cdot;W) : \mathrm{R}^D \rightarrow \mathrm{R}^d$ such that variability over ρ is minimal. Many methods achieve this by observing pairs of randomly augmented views of a batch of images: $X^A = \tau(X;\rho_1)$, $X^B = \tau(X;\rho_2)$, with $\rho_1, \rho_2 \sim p(\rho)$ where $p(\rho)$ is a pre-chosen probability distribution over augmentation parameters. Generally TiSSL frameworks employ an objective function that operates on the outputs of f, $Z^A = f(X^A;W)$, $Z^B = f(X^B;W)$. Here, we focus on the Barlow Twins objective (Zbontar, Jing, Misra, LeCun, & Deny, 2021): $\mathcal{L}_{BT} = \Sigma_i(1-\mathcal{C}_{ii})^2 + \lambda\Sigma_{i,i\neq j}(\mathcal{C}_{ij})^2$ where $\mathcal{C}$ is the cross-correlation matrix between Z^A and Z^B. The first term encourages the outputs in response to the same image subjected to different augmentations to be correlated, while the second encourages the outputs in response to distinct images to be uncorrelated.

Transformation Equivariant Self Supervised Learning (TeSSL)

Complete invariance to augmentations is often undesirable, and as such Transformation-invariant Self Supervised learning (TiSSL) methods commonly employ a technique known as ”guillotine regularization.” Concretely, the learned function is decomposed into two stages a ”feature extractor” and a ”projector”: $f(\cdot) = g(h(\cdot))$. The loss function is applied to the output of the projector during training, which is then discarded and the feature extractor is used as the learned representation for downstream tasks (such as object classification or predicting neural activity). While this allows for h to retain some variability to ρ, it is uncontrolled, and in practice the architecture of g must be carefully tuned in conjunction with the distribution $p(\rho)$ for any particular TiSSL objective in order to achieve strong performance on downstream tasks.

We propose a method to learn structured variability to augmentations by introducing a dual loss. Given two random non-overlapping equal-sized partitions of the dataset $X_1, X_2 \in \mathrm{R}^{\mathrm{B}/2\times\mathrm{D}}$, we apply the same random augmentations to both X_1 and X_2, so that the first row of $X_1^{A/B}$ and the first row of $X_2^{A/B}$ contain distinct images that have been subjected to the same random augmentation (and so on for subsequent rows). Additionally, we employ the use of two projectors, g_{inv} and g_{equi}, that will learn to extract invariant and equivariant features from the shared base representation h. Specifically we have, $Z_{1/2}^{A/B} = g_{inv}(h(X_{1/2}^{A/B})$ and $\tilde{Z}_{1/2}^{A/B} = g_{equi}(h(X_{1/2}^{A/B})$ and optimize all three functions jointly to minimize $\mathcal{L}_{TeSSL} = (1-\lambda)\mathcal{L}_{TiSSL}([Z_1^A, Z_2^A],[Z_1^B, Z_2^B]) + \lambda\mathcal{L}_{TiSSL}(\tilde{Z}_1^A - \tilde{Z}_1^B, \tilde{Z}_2^A - \tilde{Z}_2^B)$. The second loss term now encourages similar image augmentations applied to distinct images to correspond to similar transformations in the output space, and distinct augmentations to be coded independently. The hyperparameter λ allows us to adjust the relative importance of learning augmentation-orthogonal and augmentation-related information.

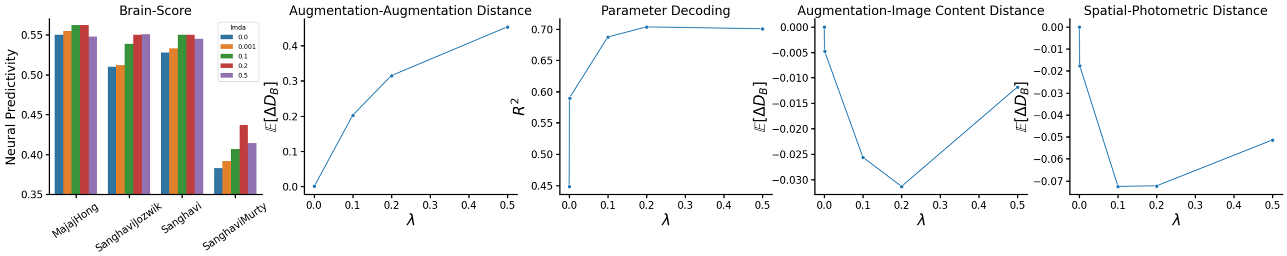

Figure 1: From left to right: (1) Neural predictivity for 4 datasets collected from macaque area IT, available in the Brain-Score database (Schrimpf et al., 2018). For all datasets, predictivity monotonically increases from $\lambda \in [0, .2]$ and decreases from 0.2 to 0.5 for 3 of 4. (2) Relative Bures distance between between pairs of distinct images subjected to many augmentations. This metric is closely related to the equivariance objective and so we see a monotonic increase in shared augmentation variability as λ is increased. (3) Accuracy (correlation) of augmentation parameters decoded from the representation. The equivariant objective increases the linearly accessible information about the input transformations (although there is a slight decrease from $\lambda = 0.2$ to $\lambda = 0.5$. (4) Negative values here indicate that there is less shared variability between image content and augmentations in the equivariant networks relative to the vanilla networks (i.e. the two sources of variability have been factorized). Note that the degree of factorization decreases from $\lambda = 0.2$ to $\lambda = 0.5$. (5) Shared variability arising from two types of augmentations shows a similar pattern as in (4).

Results

As the notation above suggests, our TeSSL method can be applied to any existing TiSSL framework. For simplicity we restrict ourselves to the Barlow Twins objective though we have observed identical trends for other choices of $\mathcal{L}_{TiSSL}$ (namely MMCR (Yerxa, Kuang, Simoncelli, & Chung, 2024) and SimCLR (Chen, Kornblith, Norouzi, & Hinton, 2020)). We parameterize h using a ResNet-50 architecture, and $g_{inv/equi}$ as MLPs with architectures described in the original Barlow Twins paper. We train for 100 epochs on the ImageNet-1k dataset and sweep the hyperparameter λ over $[0.0, 0.001, 0.1, 0.2, 0.5]$.

Representational Analyses

We conducted a series of experiments to measure the extent to which: (1) augmentation information was linearly decodable from h, (2) augmentation related variability was shared across images, (3) variability due to augmentations was factorized from variability across base images, and (4) variability due to different types of augmentations (random cropping vs. photometric distortions) was factorized (orthogonal).

For (1) we fit linear regressions to decode augmentation parameters from the outputs of h in response to clean and augmented images. For (2)-(4) we estimate the covariance of responses to various ensembles of inputs, and computed trace-normalized Bures distances between pairs of covariance matrices. For example for (2), we subjected two base images to many different random augmentations, estimated the covariance of responses to each image over augmentations, and computed the Bures metric between the two covariance matrices (a high distance indicates there is little shared variability and vice versa for a low distance). For (3) we followed a similar procedure but instead compared variability over augmentations to a single image to the global variability over all images in the dataset. For (4) we subjected a single image to many different random crops, and to many different photometric distortions. The results of these analyses are summarized in Figure 1. For (2)-(4) in all cases we computed the same distances for a particular TeSSL network (choice of λ) and the TiSSL network ($\lambda = 0$, and report the mean difference between the two, $E[\Delta D_B]$ where D_B is the Bures metric. So the leftmost point in each such plot by definition is zero (as the iSSL network is being compared to itself). Note that (1)-(4) correspond to panels (2)-(5) in Fig. 1 above.

Neural Predictivity

We utilized the BrainScore (Schrimpf et al., 2018) evaluation pipeline to measure the extent to which each learned representation can linearly predict neural responses measured in area IT in four different experimental datasets. For reference our highest performing model ($\lambda = 0.2$) has the 10th highest average predictivity of IT out of approximately 250 publicly available models currently on the Brain-Score leaderboard. Across a reasonably large range of values of λ, the equivariant model improves the neural predictivity relative to the invariant baseline for all four datasets. We note that predictivity seems to peak together with our factorization measurements, an observation in line with (Lindsey & Issa, 2023), but here we demonstrate for the first time that equivariance can be used as a learning signal to improve brain-model alignment.

References

Chen, T., Kornblith, S., Norouzi, M., & Hinton, G. (2020). A simple framework for contrastive learning of visual representations. In *International conference on machine learning* (pp. 1597–1607).

Feather, J., Leclerc, G., Madry, A., & McDermott, J. H. (2023). Model metamers reveal divergent invariances between biological and artificial neural networks. *Nature Neuroscience*, *26*(11), 2017–2034.

Lindsey, J. W., & Issa, E. B. (2023). Factorized visual representations in the primate visual system and deep neural networks. *bioRxiv*, 2023–04.

Schrimpf, M., Kubilius, J., Hong, H., Majaj, N. J., Rajalingham, R., Issa, E. B., ... DiCarlo, J. J. (2018). Brain-score: Which artificial neural network for object recognition is most brain-like? *bioRxiv preprint*. Retrieved from `https://www.biorxiv.org/content/10.1101/407007v2`

Yerxa, T., Kuang, Y., Simoncelli, E., & Chung, S. (2024). Learning efficient coding of natural images with maximum manifold capacity representations. *Advances in Neural Information Processing Systems*, *36*.

Zbontar, J., Jing, L., Misra, I., LeCun, Y., & Deny, S. (2021). Barlow twins: Self-supervised learning via redundancy reduction. In *International conference on machine learning* (pp. 12310–12320).