Differentiating image representations in terms of their local geometry

David Lipshutz[*,1,2], Jenelle Feather[*,1], Sarah E Harvey[1], Alex H Williams[1,3], Eero P Simoncelli[1,3]

[1]Center for Computational Neuroscience, Flatiron Institute [2]Department of Neuroscience, Baylor College of Medicine

[3]Center for Neural Science, New York University

Abstract

Similarity between neural representations is often quantified by measuring alignment of the representations over a set of natural stimuli that are relatively far apart in stimulus space. However, systems with similar global structure can have strikingly different sensitivities to local stimulus distortions, suggesting a need for metrics that compare local sensitivities of representations. We propose a framework for comparing a set of image representations in terms of their sensitivities to local distortions. We quantify the local geometry of a representation using the Fisher information matrix, a standard statistical tool for characterizing the sensitivity to local stimulus perturbations, and use this to define a metric on the local geometry of representations near a base image. This metric may then be used to differentiate a set of representations, by finding a pair of "principal distortions" that maximize the variance of the representations under the metric. We apply our method to models of the early visual system and to a set of deep neural network (DNN) models.

Keywords: representational similarity, Fisher information, stimulus generation

Introduction

Similarity between neural representations is often quantified by measuring alignment of neural responses to a set of stimuli that are relatively far apart in stimulus space (Kriegeskorte, Mur, & Bandettini, 2008). However, neurals systems with similar global structure can have strikingly different sensitivities to local stimulus distortions (Szegedy et al., 2013), suggesting a need for methods that compare local sensitivities of neural representations and, in particular, highlight differences between systems even when global structure seems similar.

We propose a framework for comparing a set of image representations in terms of their sensitivities to local distortions, which builds on existing methods (Berardino, Ballé, Laparra, & Simoncelli, 2017; Zhou, Chun, Subramanian, & Simoncelli, 2023). We quantify the local geometry of a representation using the Fisher information matrix (FIM), a standard statistical tool for characterizing the sensitivity to local stimulus perturbations, and use this to define a metric on the local geometry of representations near a base image. This metric may then be used to optimally discriminate a set of representations, by finding a pair of "principal distortions" (PDs) that maximize the variance of the representations under this metric.

We apply our method to a nested set of models of the early visual system to identify distortions that differentiate these models and can potentially be used to evaluate how well these models predict human visual sensitivities. We then apply our method to a set of visual deep neural networks (DNNs) with varying architectures and training procedures. We find distortions that allow for visualization of differences in the sensitivities between layers of the networks and DNN architectures. We further explore differences between standard ImageNet trained networks and their shape-bias enhanced counterparts, and between standard networks and their adversarially-trained counterparts. In all cases, we illustrate how the method generates novel image distortions that highlight differences between models.

Principal distortions (PDs)

Given a collection of stochastic image representations, we develop a method for comparing their local geometries around an image $\mathbf{s}$. We assume that each representation is defined by a conditional density $p(\mathbf{r}|\mathbf{s})$, where $\mathbf{r}$ is a stochastic response (e.g., spike counts or noisy model responses). The local sensitivity of a representation can be expressed in terms of the FIM $\boldsymbol{I}(\mathbf{s}) := \mathbb{E}_{\mathbf{r}\sim p(\mathbf{r}|\mathbf{s})}[\nabla_{\mathbf{s}} \log p(\mathbf{r}|\mathbf{s}) \nabla_{\mathbf{s}} \log p(\mathbf{r}|\mathbf{s})^\top]$, which has been used to link neural representations to perceptual discrimination. Specifically, the local sensitivity of a representation at a stimulus $\mathbf{s}$ to a distortion $\boldsymbol{u}$ is given by $d(\boldsymbol{u}) := \sqrt{\boldsymbol{u}^\top \boldsymbol{I}(\mathbf{s})\boldsymbol{u}}$. Due to the high-dimensionality of images, a comprehensive comparison of local sensitivities of representations is impractical, so it is useful to develop a method for choosing distortions along which to assess and compare representations.

Building on "eigen-distortions" (Berardino et al., 2017) for probing the local geometry of a *single* image representation, Zhou et al. (2023) proposed comparing *two* image representations $p_A(\mathbf{r}|\mathbf{s})$ and $p_B(\mathbf{r}|\mathbf{s})$ by choosing "generalized eigen-distortions" that extremize the ratio of their sensitivities: $\boldsymbol{\varepsilon}_1 = \arg\max_{\boldsymbol{u}} d_A(\boldsymbol{u})/d_B(\boldsymbol{u})$ and $\boldsymbol{\varepsilon}_2 = \arg\min_{\boldsymbol{v}} d_A(\boldsymbol{v})/d_B(\boldsymbol{v})$. The optima can be expressed in closed form, but the method is limited to pairwise model comparisons.

To compare $N > 2$ image representations, we re-express the generalized eigen-distortions as the solution to the optimization problem: $\{\boldsymbol{\varepsilon}_1, \boldsymbol{\varepsilon}_2\} = \arg\max_{\boldsymbol{u},\boldsymbol{v}} m_{\boldsymbol{u},\boldsymbol{v}}(\boldsymbol{I}_A(\mathbf{s}), \boldsymbol{I}_B(\mathbf{s}))$, where $m_{\boldsymbol{u},\boldsymbol{v}}(\cdot,\cdot)$ is a *(pseudo-)metric* on the FIMs (at image $\mathbf{s}$) defined in terms of the log sensitivity ratios to distortions $\boldsymbol{u}, \boldsymbol{v}$:

$$m_{\boldsymbol{u},\boldsymbol{v}}(\boldsymbol{I}_A, \boldsymbol{I}_B) := \left| \log \frac{d_A(\boldsymbol{u})}{d_A(\boldsymbol{v})} - \log \frac{d_B(\boldsymbol{u})}{d_B(\boldsymbol{v})} \right|. \tag{1}$$

We then optimize over distortions so as to maximize the *variance* of the N representations $A_1, \ldots, A_N$ under this metric:

$$\{\boldsymbol{\varepsilon}_1, \boldsymbol{\varepsilon}_2\} = \arg\max_{\boldsymbol{u},\boldsymbol{v}} \sum_{i=1}^{N} \sum_{j=1}^{N} m^2_{\boldsymbol{u},\boldsymbol{v}}(\boldsymbol{I}_{A_i}, \boldsymbol{I}_{A_j})$$

and refer them as "principal distortions" (Feather, Lipshutz, Harvey, Williams, & Simoncelli, 2025).

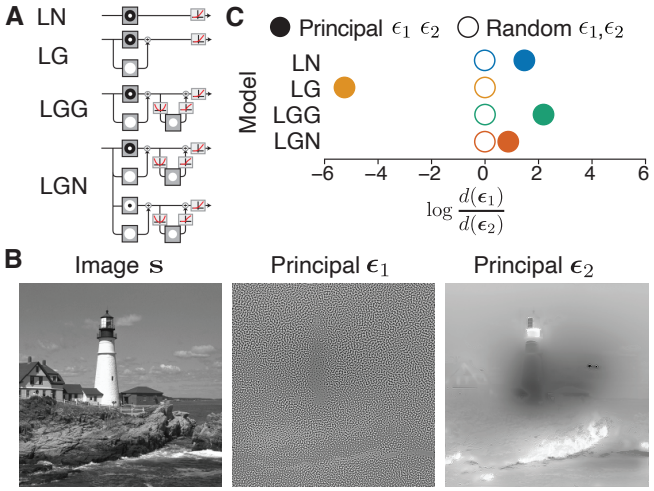

Figure 1: **A)** Four nested early visual models (LN is the most basic, LGN is the full model). **B)** Natural image $\mathbf{s}$ and principal distortions $\{\boldsymbol{\epsilon}_1, \boldsymbol{\epsilon}_2\}$ for differentiating the models. **C)** Log sensitivity ratios of the two principal distortions (filled circles) and two random distortions (hollow circles) for each model.

Early visual models

We generated PDs for a nested family of models of the early primate visual system (Fig. 1). The full model (LGN) contains two parallel cascades representing ON and OFF center-surround filter channels, rectification, and both luminance and contrast gain control nonlinearities (Fig. 1A). The other models are reduced versions of this model. The filter sizes, amplitudes, and normalization values of each model were previously fit separately to predict a dataset of human distortion ratings (Berardino et al., 2017).

The optimized PDs $\{\boldsymbol{\epsilon}_1, \boldsymbol{\epsilon}_2\}$ at a base image $\mathbf{s}$ visual distinct (Fig. 1B) and separate the models (in terms of their log sensitivity ratios) far more than random distortions (Fig. 1C). Future work with human perceptual experiments could quantify human sensitivities to the PDs and then compare them to the models via the metric defined in equation 1.

Deep neural networks (DNNs)

DNNs, originally developed for object recognition, have also been examined as models of the primate visual system (Yamins & DiCarlo, 2016). Numerous models have been proposed, but many perform quite similarly on behavioral tasks or neural benchmarks (Schrimpf et al., 2018).

We demonstrate that PDs can be used to probe for differences in the local geometry of the DNN representations and how these differences are due to architecture or training procedure (Fig. 2). Notably, depending on the training procedure, the PDs would either separate the DNN layers by architecture or training procedure. These examples demonstrate that PDs can be used to separate collections of similar models, and points to its utility in probing complex high-level representations.

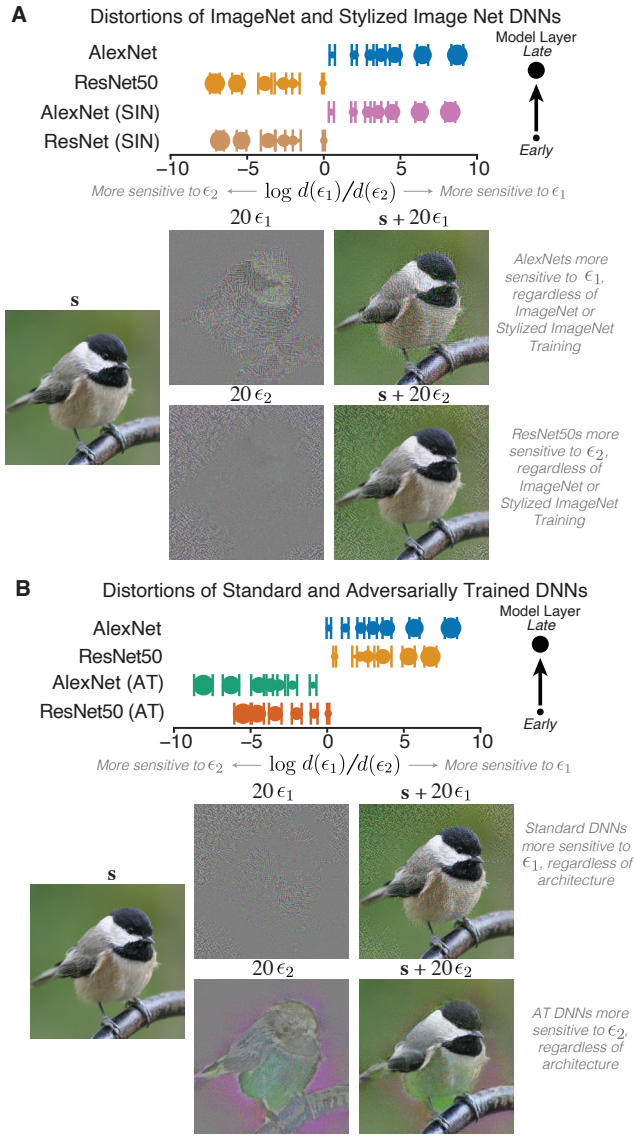

Figure 2: Comparison of $N = 26$ layers from AlexNet and ResNet50 variants trained to **(A)** increase "shape-bias" by training on Stylized Image Net (SIN) (Geirhos et al., 2019) or **(B)** reduce adversarial vulnerability via adversarial training (AT) (Feather et al., 2023). In each case, we plot the mean log sensitivity ratios of the PDs computed for 100 random images (top) and show an example base image and PDs (bottom). **(A)** When comparing DNNs trained on ImageNet versus SIN, the PDs separate DNNs by architecture, suggesting that changes in local geometry induced by SIN are small relative to those induced by the architecture. AlexNet was more sensitive to distortions in non-smooth regions of the image, while ResNet50 was more sensitive to distortions in smooth regions of the image. **(B)** When comparing DNNs and AT DNNs, the PDs reliably separated the model classes by training type rather than architecture, suggesting that AT notably influenced the local geometry of the DNN representations. The PDs were visually different from those shown in **(A)**: AT networks were more sensitive to distortions in color contrast across boundaries.

Acknowledgments

The Flatiron Institute is a division of the Simons Foundation. The computations reported in this paper were performed using resources made available by the Flatiron Institute.

References

Berardino, A., Ballé, J., Laparra, V., & Simoncelli, E. P. (2017). Eigen-distortions of hierarchical representations. *Advances in Neural Information Processing Systems*, *30*, 3531–3540.

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.

Feather, J., Lipshutz, D., Harvey, S. E., Williams, A. H., & Simoncelli, E. P. (2025). Discriminating image representations with principal distortions. In *International conference on learning representations*.

Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F. A., & Brendel, W. (2019). ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. In *International Conference on Learning Representations*.

Kriegeskorte, N., Mur, M., & Bandettini, P. A. (2008). Representational similarity analysis-connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, *2*, 249.

Schrimpf, M., Kubilius, J., Hong, H., Majaj, N. J., Rajalingham, R., Issa, E. B., ... others (2018). Brain-score: Which artificial neural network for object recognition is most brain-like? *BioRxiv*, 407007.

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., & Fergus, R. (2013). Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*.

Yamins, D. L., & DiCarlo, J. J. (2016). Using goal-driven deep learning models to understand sensory cortex. *Nature Neuroscience*, *19*(3), 356–365.

Zhou, J., Chun, C., Subramanian, A., & Simoncelli, E. P. (2023). Comparing neural models using their perceptual discriminability predictions. In *Proceedings of UniReps: the First Workshop on Unifying Representations in Neural Models* (pp. 170–181).