Representational constraints underlying similarity between task-optimized neural systems

Tahereh Toosi
Center for Theoretical Neuroscience
Zuckerman Mind Brain Behavior Institute
Columbia University
New York, NY
tahereh.toosi@columbia.edu

Abstract

Neural systems, artificial and biological, show similar representations of inputs when optimized to perform similar tasks. In visual systems optimized for tasks similar to object recognition, we propose that representation similarities arise from the constraints imposed by the development of abstractions in the representation across the processing stages. To study the effect of abstraction hierarchy of representations across different visual systems, we constructed a two-dimensional space in which each neural representation is positioned based on its distance from the pixel space and the class space. Trajectories of representations in all the task-optimized visual neural networks start close to the pixel space and gradually move towards higher abstract representations, such as object categories. We also observe that proximity in this abstraction space predicts the similarity of neural representations between visual systems. The gradual similar change of the representations suggests that the similarity across different task-optimized systems could arise from constraints on representational trajectories.

Introduction

Recent progress in neuroscience and AI has revealed that biological and artificial learning systems tend to form similar representations at various stages of processing when exposed to similar stimuli. This similarity has been observed in different modalities such as vision (Yamins et al., 2014; Khaligh-Razavi and Kriegeskorte, 2014; Schrimpf et al., 2020), auditory (Kell et al., 2018), and language (Schrimpf et al., 2021). This similarity in representations, particularly the similarity between intermediate levels of processing, has led to optimism about the potential of such similarities, between artificial systems and the brain, to help in understanding the mechanisms of brain function. However, as models with diverse objective functions and architectures are examined for these similarities, doubts arise regarding the inferences of mechanistic similarity in neural processing: If representational similarity indicates that the neural mechanisms underlying those representations are similar, then why is a comparable level of similarity observed despite significant structural and optimization differences (Schrimpf et al., 2018; Conwell et al., 2022)?

In this work, we investigate the similarity of representations between biological and artificial visual systems that are optimized for tasks similar to object classification, such as recognition, localization, and segregation. To study the neural representations between visual systems, as representations are developed across processing stages, we introduce an abstraction space. This is a conceptual framework that allows measurement of the similarity of each representation both from the input (pixel) and output (class) representations. This abstraction space has two coordinates: on one axis,

UniReps Workshop, Neural Information Processing Systems (NeurIPS 2023).

we measure the similarity of the representation relative to the pixel space and on the other axis, we measure the similarity of the representation relative to the object class space (Fig. 1A). We compare the trajectories of neural representations across processing stages for different visual systems in this abstraction space between the pixel input and the category output. We also examine how the proximity in this two-dimensional space is related to the similarity of representations between visual systems. The results of this study suggest that similar underlying constraints on the abstraction trajectory lead to the similarity of representations between visual systems.

Results

Abstraction space

Similarity between neural representations is measured when the networks receive the same stimuli. Here, we focus on a widely used stimulus set to study the similarity between biological vision and deep neural networks for object recognition (Majaj et al., 2015; Yamins et al., 2014). This stimulus set consists of 5760 grayscale images constructed from 64 objects from 8 categories with varying positions, views, and sizes superimposed on random natural backgrounds (Majaj et al., 2015; Yamins et al., 2014; Hong et al., 2016) (see Figure 1A for an example of a rotated face on top of a natural background). The pixel space is defined by the pixel values of the images, and the class space is defined by one-hot vectors representing each category. We used Centered Kernel Analysis (CKA, Kornblith et al. (2019)) to compute the similarity between each representation and these two spaces—pixel and class. However, similar results are obtained using other representational similarity metrics, such as Representation Similarity Analysis (RSA).

Independent of neural representations, we first assess the abstraction hierarchy in this abstraction space for features extracted directly from the image set (e.g., object position, background, etc.). After extracting each feature, we construct a corresponding feature space within the image data set (Toosi et al., 2023). For example, the space of the object positions (detection in 1A) is constructed by retaining only the pixels within a circular area around the object in each image. Similarly, the space of the segmented objects (segmentation in 1A) is defined by keeping the pixels inside the object boundary. We then measure the similarity of each feature space to both the original pixel space and the class space (Figure 1A). As expected, with this approach, we capture the progression of the abstraction for the features directly extracted from the images. This approach captures an array of both the high-dimensional features (e.g., pixels, edges, low-frequency or high-frequency parts of the image) and the low-dimensional features (e.g., detection labels, segmentation labels, object identification, and classification). When situated in this abstraction space, these features form a hierarchy of abstraction. For instance, background representations are less similar to the class space than to the pixel space, and detection (object position) is less similar to class representations than segmentation (object boundary).

Visual systems have similar trajectories in the abstraction hierarchy

To study the development of abstraction across different visual systems, we mapped both recorded neural activations from monkey visual cortices and DNN activations onto the defined abstraction space. The neural activity of the monkeys was recorded in the V4 and IT cortex. We included two timestamps for these recordings: early and late (for V4: 75ms and 105ms; for IT: 105ms and 145ms after stimulus onset, respectively). For DNNs, we used a family of ResNet50 architectures trained under various objective functions. These included standard classification (He et al., 2015), self-supervised, stylized (Geirhos et al., 2019), adversarially robust (Engstrom et al., 2019), robust to Gaussian noise (Cohen et al., 2019), and some custom optimized with varying degrees of data augmentation. Across all these systems, the trajectory of representations across layers begins near the pixel space representation and gradually follows a similar abstraction hierarchy toward more abstract representation culminating in classification (Figure 1B).

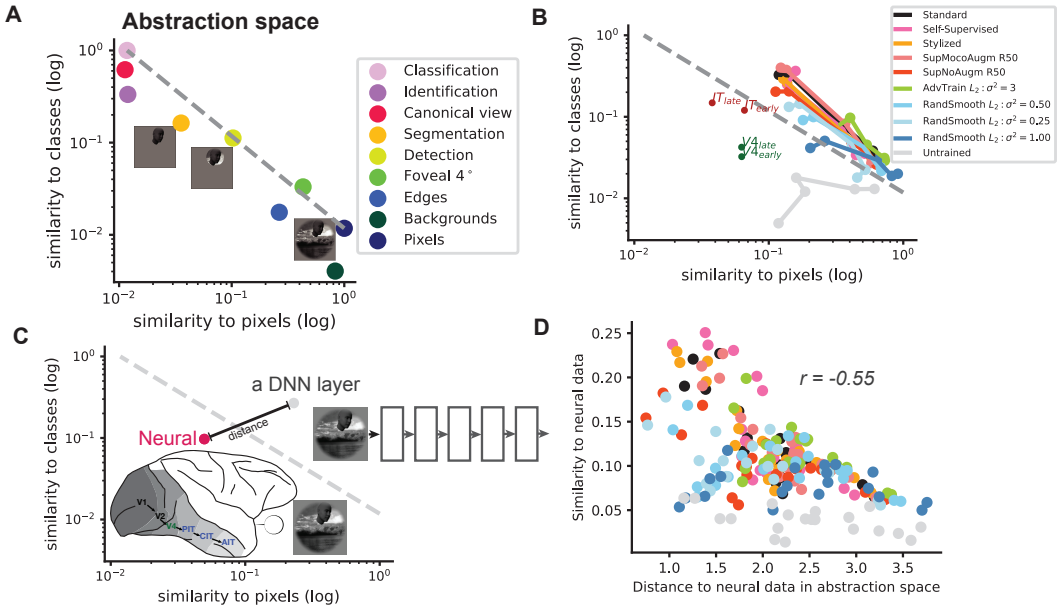

Figure 1: **A)** Similarity of each low-level and high-level feature spaces in images (e.g. edges and segmentation) to pixel space (x-axis) and to class space (y-axis), adapted from (Toosi et al., 2023). **B)** Neural data recorded from the V4 and IT cortex of monkeys (Majaj et al., 2015; Yamins et al., 2014) located in the abstraction space with green and red dots. Early and late denote the response latency after stimulus onset. Each trajectory shows the representations located in the abstraction space, for 5 layers of ResNet50 (layer1 to layer5 and avgpool), when trained under different objectives. **C)** The schematic of the Euclidean distance between each DNN layer and each neural data in the abstraction space. **D)** The x-axis shows the distance between each biological neural data and each layer in B in the abstraction space. The y-axis shows the similarity of each biological neural data and each DNN layer in B measured with CKA. Similar colors as in B are used to show the similarity of each DNN network layer with the biological neural data.

Proximity in the abstraction space predicts the representational similarity

We evaluated how well proximity in the abstraction coordinate system predicts the pairwise similarity of the representations. To assess this, we compared the Euclidean distance in the abstraction space between each DNN layer representation and the biological neural recordings (Figure 1C) against their pairwise representational similarity measured by CKA. The pairwise distance in the abstraction coordinate system is strongly correlated with the pairwise representational similarities (Pearson correlation, r=-0.55, p=1.9e-17, Figure 1D). Therefore, the proximity of two representations in the abstraction coordinate system serves as a reliable proxy for their representational similarity.

Discussion

Neural networks can be optimized for different objectives related to their input modality (e.g., object recognition for vision, music or speech recognition for audition, and odor recognition for olfaction). In vision, these objectives can include object classification, scene segmentation, object detection, or self-supervised objectives such as contrastive learning or autoencoders, which aim to create abstract representations of image pixels. In the stimulus set we studied, object classes represented the most abstract latent factor, underlying the generation of all 5760 images. Although the most abstract latent factor may vary depending on the images in a stimulus set, it remains, in any case, distant from the pixel space in the abstraction coordinate system.

Neural networks operating on pixels of the stimulus set form representations that are initially (early layers in DNNs, or early timestamps in recurrent neural networks) more similar to pixels, thus being located close to the pixel representation in the abstraction space (Figure 2A). The trajectory endpoint,

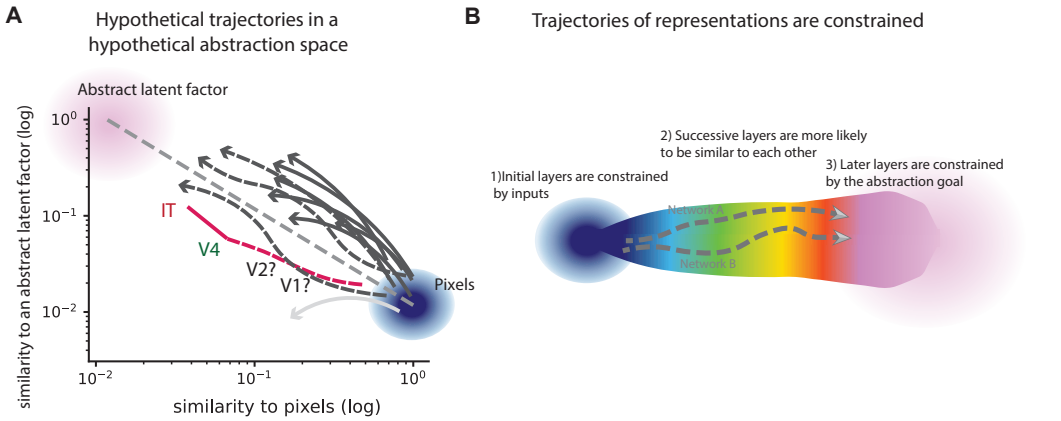

Figure 2: **A)** Each path illustrates how hypothetical deep neural networks (DNNs) and neural data progress from pixel-level representations (shown as a navy bulb) to more abstract latent factors (shown as a pink bulb). **B)** The similarity of intermediate representations between Network A and Network B might be caused by the fact that both networks have to path through similar abstraction stages.

depending on the objective function (e.g. detection, segmentation, classification), is located away from the pixel space and closer to the abstract latent factor, such as classification. Thus, the spectrum of intermediate representation trajectories is bounded at one end by the input (e.g., the stimulus set's pixels) and at the other by the latent generative factor (e.g., classes). The abstraction coordinates of successive layers are typically close to each other, indicating that the network's trajectory can be viewed as a gradually evolving continuous path. This holds even though each layer is represented as a discrete point in the abstraction coordinates (see Figure 2 B). With these limitations, the space that the trajectories of different networks can traverse is limited, leading to trajectories that are close together in the abstraction coordinates. As observed in this study, proximity in the abstraction coordinates reflects representational similarity, irrespective of the network's objective; networks optimized for different objectives exhibit similarities in their representations across the abstraction hierarchy.

In this work, we explored the representational similarity between biological and artificial neural visual systems by studying them in a new abstraction space. This analysis suggests that the observed similarity between these systems is shaped by the constrained trajectories these representations follow, transitioning from pixel-level to more abstract class-level representations. This understanding suggests inherent constraints in the evolution of neural representations and provides evidence for the utility of the abstraction space as a useful framework for analyzing this evolution. The correlation between spatial proximity in abstraction space and representational similarity showcases its potential to further explore the links between biological and artificial visual systems. This study invites further investigation to deepen our understanding of the inherent constraints on neural representations in task-optimized systems, potentially bridging insights between neuroscience and artificial intelligence in understanding visual recognition.

Acknowledgment

I would like to thank the DiCarlo lab at MIT for providing the neural data. T.T. is supported by NIH 1K99EY035357-01. This work was also supported by NIH RF1DA056397, NSF 1707398, and Gatsby Charitable Foundation GAT3708.

References

Cohen, J. M., Rosenfeld, E., and Kolter, J. Z. (2019). Certified adversarial robustness via randomized smoothing.

Conwell, C., Prince, J. S., Kay, K. N., Alvarez, G. A., and Konkle, T. (2022). What can 1.8 billion regressions tell us about the pressures shaping high-level visual representation in brains and machines?

Engstrom, L., Ilyas, A., Salman, H., Santurkar, S., and Tsipras, D. (2019). Robustness (python library).

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

He, K., Zhang, X., Ren, S., and Sun, J. (2015). Deep Residual Learning for Image Recognition. *arXiv:1512.03385 [cs]*. arXiv: 1512.03385.

Hong, H., Yamins, D. L. K., Majaj, N. J., and DiCarlo, J. J. (2016). Explicit information for category-orthogonal object properties increases along the ventral stream. *Nat. Neurosci.*, 19(4):613–622.

Kell, A. J., Yamins, D. L., Shook, E. N., Norman-Haignere, S. V., and McDermott, J. H. (2018). A task-optimized neural network replicates human auditory behavior, predicts brain responses, and reveals a cortical processing hierarchy. *Neuron*, 98(3):630–644.e16.

Khaligh-Razavi, S.-M. and Kriegeskorte, N. (2014). Deep supervised, but not unsupervised, models may explain IT cortical representation. *PLoS Comput. Biol.*, 10(11):e1003915.

Kornblith, S., Norouzi, M., Lee, H., and Hinton, G. (2019). Similarity of neural network representations revisited.

Majaj, N. J., Hong, H., Solomon, E. A., and DiCarlo, J. J. (2015). Simple learned weighted sums of inferior temporal neuronal firing rates accurately predict human core object recognition performance. *J. Neurosci.*, 35(39):13402–13418.

Schrimpf, M., Blank, I. A., Tuckute, G., Kauf, C., Hosseini, E. A., Kanwisher, N., Tenenbaum, J. B., and Fedorenko, E. (2021). The neural architecture of language: Integrative modeling converges on predictive processing. *Proc. Natl. Acad. Sci. U. S. A.*, 118(45):e2105646118.

Schrimpf, M., Kubilius, J., Hong, H., Majaj, N. J., Rajalingham, R., Issa, E. B., Kar, K., Bashivan, P., Prescott-Roy, J., Geiger, F., Schmidt, K., Yamins, D. L. K., and DiCarlo, J. J. (2018). Brain-score: Which artificial neural network for object recognition is most brain-like? *bioRxiv preprint*.

Schrimpf, M., Kubilius, J., Lee, M. J., Murty, N. A. R., Ajemian, R., and DiCarlo, J. J. (2020). Integrative benchmarking to advance neurally mechanistic models of human intelligence. *Neuron*.

Toosi, T., Kriegeskorte, N., and Issa, E. B. (2023). Object-enhanced and object-centered representations across primate ventral visual cortex. *Conference on Cognitive Computational Neuroscience*.

Yamins, D. L. K., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., and DiCarlo, J. J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proc. Natl. Acad. Sci. U. S. A.*, 111(23):8619–8624.