Constructing Representational Similarity Metrics Through Linear Decoding

Sarah E Harvey[1], David Lipshutz[1,2], Alex H Williams[1,3]
[1]Flatiron Institute [2]Baylor College of Medicine [3]New York University

Abstract

Neural responses encode information that is useful for a variety of downstream tasks; however, many methods for comparing neural representations do not explicitly leverage this perspective and instead highlight geometric invariances. Here, we show that many representational similarity measures can be equivalently motivated from a decoding perspective. Specifically, measures like CKA and CCA are shown to quantify the average alignment between optimal linear readouts across a distribution of decoding tasks. This approach suggests a metric on neural representations in which the distance between representations directly quantifies differences in the decoding of neural data. We demonstrate this in an ensemble of DNNs trained for image classification and human fMRI representations from the Natural Scenes Dataset. Our work demonstrates a tight link between the geometry of neural representations and the ability to linearly decode information. This perspective suggests new ways of measuring similarity between neural systems and also provides novel, unifying interpretations of existing measures.

Summary

Developing methodologies for quantifying similarity between high-dimensional neural representations is an active research direction in computational neuroscience. The analyses enabled by these similarity measures have implications for understanding variability in neural computations across individuals and species (Kriegeskorte et al., 2008), as well as for comparisons between biological systems and computational models (Schrimpf et al., 2018). Concurrently, one common approach used to interpret neural systems is to build regression models or "decoders" that reconstruct features of the stimulus from neural responses. Here, we leverage this idea to quantify the similarity among different neural systems. Our approach is distinct from typical motivations behind representational (dis)similarity measures like representational similarity analysis (RSA), centered kernel alignment (CKA), canonical correlation analysis (CCA), and Procrustes shape distance, which highlight geometric intuition and invariances to orthogonal or affine transformations but lack interpretations in terms of the information encoded. However, we show that CKA, CCA, and other measures can be equivalently motivated as methods that compare similarity in decoding patterns.

Methods

Suppose we record the activity of N_X neurons in one animal and N_Y neurons in a second animal, both in response to M stimuli (e.g. the activity in response to a set of M natural images). These datasets can be represented as a pair

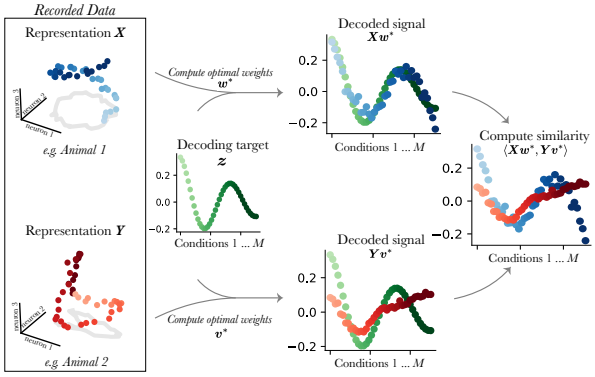

Figure 1: Schematic of the proposed framework for comparing representations X and Y (each dot represents mean neural responses to one of M conditions) in terms of a decoding target z. The *decoding similarity* is the inner product of their optimal predictors (far right panel): $\langle Xw^*, Yv^* \rangle$.

of matrices, $X \in \mathbb{R}^{M \times N_X}$ and $Y \in \mathbb{R}^{M \times N_Y}$, and we assume the columns of X and Y have been normalized to have zero mean. Here, we propose a family of metrics based on the similarity of *decoded information* and show that this framework connects many existing methods of representational similarity to notions of similarity in terms in terms of behavior on decoding tasks.

We consider the problem of decoding a target vector $z \in \mathbb{R}^M$ from neural population responses by a linear function $X \mapsto Xw$ where $w \in \mathbb{R}^{N_X}$ (Fig. 1). Specifically, we consider the following class of regression problems:

$$w^* = \arg\max_w \left\{ \frac{1}{M} z^\top X w - \frac{1}{2} w^\top G(X) w \right\}, \qquad (1)$$

where $G(\cdot)$ maps $\mathbb{R}^{M \times N}$ to the set of positive definite $N \times N$ matrices. When $G(X) = C_X + \lambda I$, where $C_X := M^{-1} X^\top X$, then w^* coincides with the optimum of the ridge regression problem: $\arg\min_w M^{-1} \|z - Xw\|_2^2 + \lambda \|w\|_2^2$. In general, we can interpret Eq. 1 as maximizing the similarity between the target z and the linear readout Xw, subject to a penalty term on w that depends on $G(\cdot)$. We let $v^* \in \mathbb{R}^{N_Y}$ denote the optimal decoding vector for Y.

A straightforward way to quantify the alignment between the decoded signals is to compute the inner product between the linear predictors for a particular decoding target z, which we will call the *decoding similarity*:

$$\langle Xw^*, Yv^* \rangle = z^\top K_{G,X} K_{G,Y} z = \mathrm{Tr}[K_{G,X} K_{G,Y} z z^\top] \qquad (2)$$

where $K_{G,X} := XG(X)^{-1}X^\top$ and $K_{G,Y} := YG(Y)^{-1}Y^\top$ are similarity matrices. This measures how similar the predictors Xw^* and Yv^* are across a sampling of M conditions (Fig. 1).

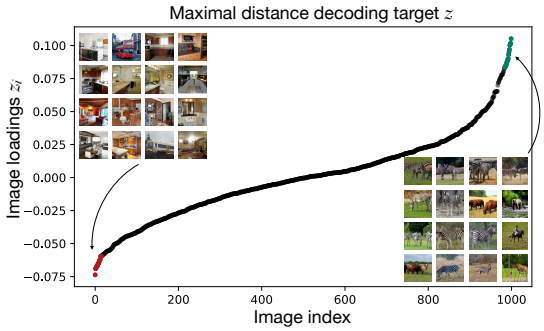

Figure 2: Decoding target z is chosen to minimize the similarity between the penultimate layer of AlexNet and human V4. Each element of z corresponds to an image in the COCO dataset. The elements of z are ordered with 16 example images shown for smallest/largest elements.

When the decoding target z is not known, one could consider the best/worst case scenario (i.e., maximize/minimize the similarity with respect to z). Fig. 2 shows the sorted maximally discriminative decoding target vector z using the penultimate layer representation formed by AlexNet and human subject 1 V4 fMRI representation taken from the Natural Scenes Dataset (NSD) (Allen et al., 2021), both in response to the same 1000 COCO images. The optimal z found is somewhat interpretable as sorting images based on color or semantic features, and investigating these maximally informative probes is an interesting direction for future work. A perhaps more intuitive and robust option–motivated by the idea that neural representations likely encode information that is useful for a variety of tasks–is to consider the average similarity over an ensemble of decoding tasks $z \sim P_z$ and quantify the *average decoding similarity* (ADS) by:

$$\mathbb{E}_{z\sim P_z}\langle Xw^*, Yv^*\rangle = \mathrm{Tr}[K_{G,X}K_{G,Y}K_z] \tag{3}$$

where $K_z = \mathbb{E}_{z\sim P_z}[zz^\top]$ is a measure of the correlation between decoding targets across conditions.

In the special case that $K_z = I$ we may choose $G(X)$ such that the ADS exactly corresponds with commonly used representational similarity measures (details omitted), such as $G(X) = I \Rightarrow$ **CKA** (Cortes, Mohri, & Rostamizadeh, 2012, def. 2) or $G(X) = C_X \Rightarrow$ **CCA** (Raghu, Gilmer, Yosinski, & Sohl-Dickstein, 2017). These representational similarity measures can then be interpreted as quantifying the average alignment between optimal linear readouts (subject to a particular regularization) across a distribution of decoding tasks, and can be interpolated between by varying the regularization parameters of the linear regression task. Other measures of representational (dis)similarity can also be directly reproduced in this framework with particular choices of regularization function $G(\cdot)$, such as **GULP** (Boix-Adsera, Lawrence, Stepaniants, & Rigollet, 2022), and we derive upper and lower bounds with the **Procrustes** distance.

This decoding setup suggests a simple method to construct an interpretable metric for neural representations in terms their decoding similarity on human V4 fMRI data (Fig. 3).

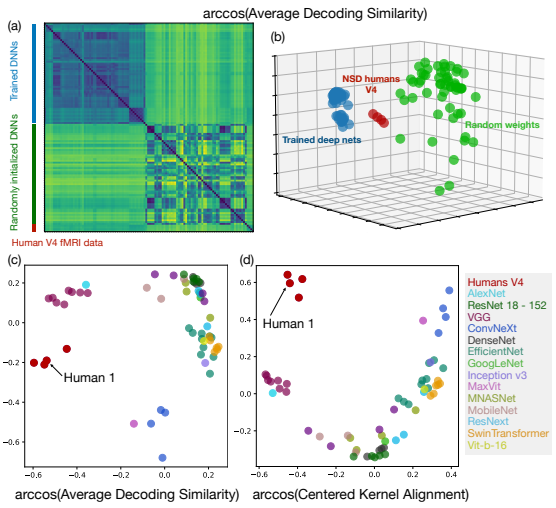

Figure 3: (a) Arccos(ADS) distance matrix for every pair of DNN representations and human fMRI representations. (b) Representations visualized as points in a low dimensional space using MDS and PCA. (c-d) Restricting to only trained DNN representations, the low-dimensional visualization shows similar clustering behavior between average decoding similarity and CKA.

Specifically, we *define* the decoding targets $z_i \in \mathbb{R}^M$ as the neural fMRI responses of voxel *i* in human subject 1 V4 to the $M = 1000$ input COCO images that were shown to all NSD experiment participants. We then can compute the similarity with respect to average decoding of human 1's V4 responses between every pair in an ensemble of 56 (trained and untrained) DNNs and other human individuals, and this score can be converted into an angular distance by taking the arccosine.

Fig. 3(a) shows the distance matrix formed by taking the arccosine of the ADS between the penultimate layer of the DNNs and four human V4 fMRI representations. Each neural representation is now a point in this metric space, which can be visualized in a low dimensional Euclidean space using multidimensional scaling (MDS) and principal components analysis (PCA) (b-c). Comparing panel (c) and (d) highlights the close connection between ADS and CKA described above–even in this relaxed case of $K_z \neq I$–perhaps motivating the use of CKA as a reasonable aggregate metric when the decoding target covariance is approximately identity, or when the choice of decoding target is not clear.

Our method differs from the commonly used approach of computing the linear predictivity of brain data from deep representations (Conwell, Prince, Kay, Alvarez, & Konkle, 2023), resulting in a one-dimensional ranking of how *well* the DNNs predict neural data. In particular, our method is informative of how well each network can predict neural activity (distances between each representation and human 1's representation) *and* how similar the representations are to each other in terms linear decoding of human 1's representation (pairwise distances between network representations). This allows us to observe clustering structure with architecture in Fig. 3 (a-d).

References

Allen, E. J., St-Yves, G., Wu, Y., Breedlove, J. L., Dowdle, L. T., Caron, B., ... Kay, K. (2021). A massive 7t fmri dataset to bridge cognitive and computational neuroscience. *Nature Neuroscience*. doi: 10.1101/2021.02.22.432340

Boix-Adsera, E., Lawrence, H., Stepaniants, G., & Rigollet, P. (2022). Gulp: a prediction-based metric between representations. In *Neurips* (Vol. 35, pp. 7115–7127).

Conwell, C., Prince, J. S., Kay, K. N., Alvarez, G. A., & Konkle, T. (2023). What can 1.8 billion regressions tell us about the pressures shaping high-level visual representation in brains and machines? *bioRxiv*. doi: 10.1101/2022.03.28.485868

Cortes, C., Mohri, M., & Rostamizadeh, A. (2012). Algorithms for learning kernels based on centered alignment. *JMLR*, *13*, 795–828.

Kriegeskorte, N., Mur, M., Ruff, D. A., Kiani, R., Bodurka, J., Esteky, H., ... Bandettini, P. A. (2008). Matching categorical object representations in inferior temporal cortex of man and monkey. *Neuron*, *60*(6), 1126–1141.

Raghu, M., Gilmer, J., Yosinski, J., & Sohl-Dickstein, J. (2017). Svcca: Singular vector canonical correlation analysis for deep learning dynamics and interpretability. *NeurIPS*, *30*.

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.