Use case determines the validity of neural systems comparisons

Erin Grant (erin.grant@ucl.ac.uk)
Gatsby Unit & Sainsbury Wellcome Centre,
University College London

Brian Cheung (cheungb@mit.edu)
Center for Brains, Minds and Machines,
Massachusetts Institute of Technology

Tomaso Poggio (tp@csail.mit.edu)
Center for Brains, Minds and Machines,
Massachusetts Institute of Technology

Andrew Saxe (a.saxe@ucl.ac.uk)
Gatsby Unit & Sainsbury Wellcome Centre,
University College London

Abstract

Deep learning provides new data-driven tools to relate neural activity to perception and cognition, aiding scientists in developing theories of neural computation that increasingly resemble biological systems both at the level of behavior (Geirhos et al., 2021) and of neural activity (Conwell et al., 2023). But what in a deep neural network should correspond to what in a biological system? This question is addressed implicitly in the use of comparison methodologies that relate specific neural or behavioral dimensions via a particular functional form. However, distinct comparison methodologies can give conflicting results in recovering even a known ground-truth model in an idealized setting (Han et al., 2023), leaving open the question of what to conclude from the outcome of a comparison using any given methodology.

Here, we develop a framework to make explicit and quantitative the effect of *both* hypothesis-driven aspects—such as the architecture of a deep neural network—*as well as* methodological choices—such as the input stimuli or similarity measure—in a systems comparison setting. We demonstrate via both simulated and analytic learning dynamics of deep neural networks that, while the role of the comparison methodology is often de-emphasized relative to hypothesis-driven aspects, this choice can impact and even invert the conclusions to be drawn from a comparison between neural systems. We provide evidence that the right way to adjudicate a comparison depends on the use case—the scientific hypothesis under investigation—which could range from identifying single-neuron or circuit-level correspondences to capturing generalizability to new stimulus dimensions.

Keywords: deep learning; computational modeling; representational simliarity; system identification

Idealizing neural systems comparison

When using deep neural networks as computational models of brain and behavior, researchers commonly use **comparison methodologies** such as representational similarity analysis (RSA; Kriegeskorte et al., 2008) or behavioral extrapolation tests (Geirhos et al., 2021) to quantify the goodness-of-fit of a given network to neural or behavioral data. We aim to evaluate the validity of such comparisons in an idealized setting in which the reference system is not a biological brain or mind but instead a known, white-box system—another deep neural network. Namely, we consider pairs of *teacher* (target) and *student* (model) neural networks, f^* and f, and neural or behavioral similarity measures defined over these networks, $\text{sim}(f^*, f)$; see Fig. 1**A**. Negative results in this idealized case suggest caution when interpreting real-world comparisons between neural systems.

A theory of neural systems comparison

To obtain *analytical* insight into the impact of **comparison methodologies**, we can make use of theoretical results about the learning dynamics of deep neural networks. For simplicity of presentation, we present here deep *linear* neural networks analyzed with RSA, and vary their hyperparameters systematically to control the comparison; however, we can extend a The exact learning dynamics of a deep linear feed-forward network, $f(\mathbf{x}) = \mathbf{W}^L \cdots \mathbf{W}^1 \mathbf{x}$, can be described analytically in terms of the singular value decomposition (SVD) of the task-dependent input-output correlation matrix, $\Sigma^{yx} = \mathbf{U}\mathbf{S}\mathbf{V}^T$ (Saxe et al., 2019). Critically, certain parameters that must be set by the modeler (such as the initial conditions of learning, including the initialization scale of the weights ρ) control the representational structure that the network learns to solve a given task, which, in turn, can determine the outcome of a systems comparison.

We demonstrate in Fig. 1**B**-**C** how this dependency influences the outcome of a systems comparison based on representational similarity, with a resulting double dissociation: Functionally similar (dissimilar) systems can be made to appear similar (dissimilar) via a methodological parameter. In Fig. 1**B**, all four networks are trained on the same task to convergence and thus implement (on average and with respect to the training objective) identical input-output functions, yet their representational similarity matrices are similar or different based on the initialization scale, ρ.

Conversely, in Fig. 1**C**, we consider wide (N large) networks trained on one of two tasks. In this setting, networks trained with small initializations will be judged dissimilar, but networks trained with large initializations will be judged similar. These results show that representational similarity as a **comparison methodology** does not separate networks trained on similar tasks from those trained on different tasks without further assumptions on methodological parameters.

Simulated neural systems comparison

We simulate learning in deep non-linear feed-forward networks, $f(\mathbf{x}) = \mathbf{W}^L \sigma(\mathbf{W}^{L-1} \cdots \sigma(\mathbf{W}^1 \mathbf{x}))$ for a non-linear activation function σ (hyperbolic tangent in Fig. 1**D**-**F**), in teacher-students settings (Fig. 1**D**-**F**), and compare systems using different methodologies. In concordance with the analytical results, we can attenuate and even reverse representational similarity for functionally similar networks; see Fig. 1**F**.

The confounding role of representational regime

Rich and lazy learning—distinct representational regimes attested via models of biological learning (Farrell et al., 2023)—can be controlled via methodological parameters already demonstrated (ρ and N Woodworth et al., 2020). Our results suggest representational regimes should be treated as hypothesis-relevant variables like architecture, and reveal even more granular representational regimes than rich and lazy that control representational comparisons.

The principal role of use case

Our framework allows us to idealize scientific use cases as parametric interventions $f^* \mapsto f$, allowing us to make explicit the causal effect of **model** or **task misspecification**. For

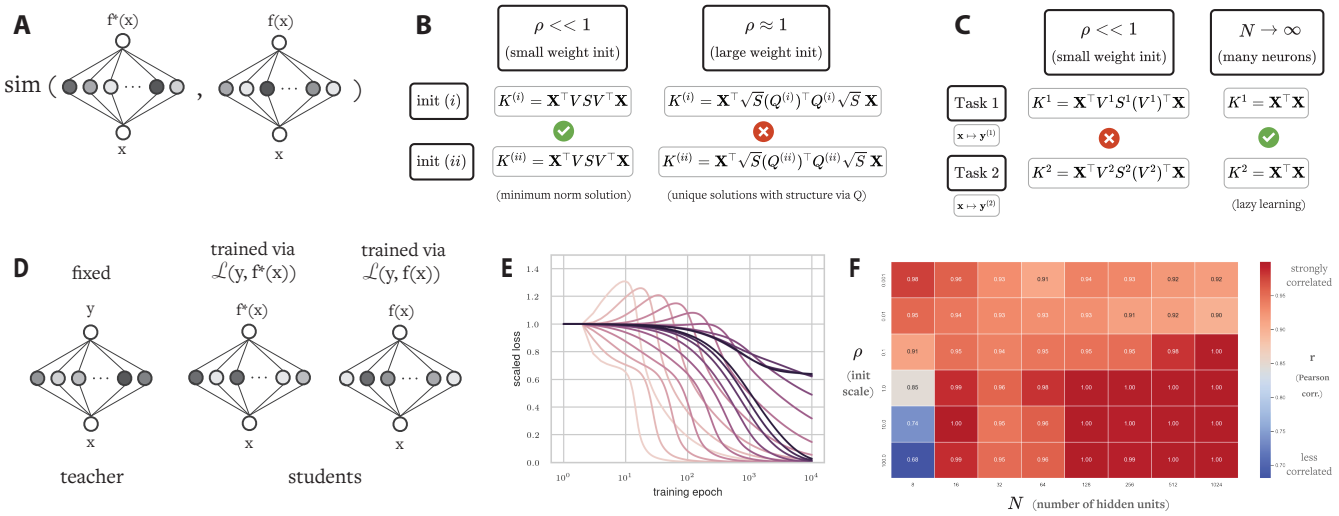

Figure 1: **A** The idealized systems comparison setting. Given two systems f^* are f, are the two systems judged as similar per sim(f^*, f)? **B**-**C** Analytical results reveal a double dissociation between functional and representational similarity in a trained two-layer linear network, where representational similarity is captured by a simple kernel matrix $K = \left(\mathbf{W}^1\mathbf{X}\right)^\top \left(\mathbf{W}^1\mathbf{X}\right)$. **B** Functionally similar systems can be made to appear similar or dissimilar: initial weight variance ρ^2 determines the learned representational structure in distinct regimes (incl. rich & lazy). **C** Functionally *dis*similar systems can be made to appear similar or dissimilar: A large number of neurons N prevents the network from learning a task-specific representation. **D** The teacher-students setting defines a learning environment when simulating learning dynamics. **E** Training and generalization dynamics of two-layer non-linear networks in the teacher-students setting of **D**, demonstrating student convergence of training loss (lower bound) to the teacher's behavior and thus functional similarity, but distinct learning dynamics and final generalization error (upper bound) as the initialization scale ρ (color) is varied, evidencing distinct representational regimes. **F** Simulations of pairs of deep non-linear networks that are functionally identical (here, trained to convergence on a simple XOR task) but whose representational comparison (here, RSA) depends non-monotonically on ρ, N.

example, we investigate the robustness of conclusions about representational or functional similarity to measurement noise in neural activity, and to employing a surrogate activation function such as rectified linear activation (ReLU) with more favorable analytical properties in a computational model. We also investigate the effect of task misspecification via correlated but distinct teachers in a teachers-students setting, which aims to answer questions about the impact of training on a surrogate task, say image classification, instead of a biological objective.

Prior work

Prior work has empirically demonstrated negative results in specific settings when employing a specific **comparison methodology**. For example, Han et al. (2023) demonstrate difficulty matching neural networks of the same hyperparameters and architecture when using linear probing, and Dujmović et al. (2023) demonstrate cases in which second-order correlations between stimuli confound representational similarity analyses. However, these prior works present isolated counterexamples, and the generalizability of such conclusions to other **comparison methodologies** is not made clear.

More critically, observational results cannot provide a strong causal link between a component of the comparison methodology—say, the complexity of the candidate model as realized by the number of hidden units N—directly through learning to the outcome of the comparison—say a given linear probe score. In contrast, our theory-based approach can analytically link such components to the outcome of the comparison via their causal effect on the representational learning dynamics of a student model, enabling us to derive precise results about a *continuity* of **comparison methodologies**, including the **representational regime** and **use case**.

Conclusions

We demonstrate positive and negative results in relating representational and functional similarity via simulated and analytical learning dynamics of deep neural networks. Our *idealized* setting, where the ground truth model is known and representable by the candidate model, brings into question conclusions from neural systems comparisons performed in more complex settings in the wild. We provide a path forward: Contextualizing comparisons within a use case, such as aiming to capture correspondence at a granular (*e.g.*, circuit) level under model or task misspecification, can provide sufficient constraints to guarantee the outcome of a systems comparison.

Acknowledgments

This work was supported by a Schmidt Science Polymath Award to A.S., and the Sainsbury Wellcome Centre Core Grant from Wellcome (219627/Z/19/Z) and the Gatsby Charitable Foundation (GAT3850). A.S. is a CIFAR Azrieli Global Scholar in the Learning in Machines & Brains program.

References

Conwell, C, Prince, JS, Kay, KN, Alvarez, GA, & Konkle, T. (2023). *What can 1.8 billion regressions tell us about the pressures shaping high-level visual representation in brains and machines?*

Dujmović, M, Bowers, JS, Adolfi, F, & Malhotra, G. (2023). *Obstacles to inferring mechanistic similarity using representational similarity analysis*.

Farrell, M, Recanatesi, S, & Shea-Brown, E. (2023). From lazy to rich to exclusive task representations in neural networks and neural codes. *Curr. Opin. Neurobio.*

Geirhos, R, Narayanappa, K, Mitzkus, B, Thieringer, T, Bethge, M, Wichmann, FA, & Brendel, W. (2021). Partial success in closing the gap between human and machine vision. *Adv. NeurIPS*.

Han, Y, Poggio, TA, & Cheung, B. (2023). System identification of neural systems: If we got it right, would we know? *Proc. ICML*.

Kriegeskorte, N, Mur, M, & Bandettini, PA. (2008). Representational similarity analysis: Connecting the branches of systems neuro. *Front. Sys. Neuro.*

Saxe, AM, McClelland, JL, & Ganguli, S. (2019). A mathematical theory of semantic development in deep neural networks. *Proc. National Acad. Sciences*.

Woodworth, B, Gunasekar, S, Lee, JD, Moroshko, E, Savarese, P, Golan, I, Soudry, D, & Srebro, N. (2020). Kernel and rich regimes in overparametrized models. *Proc. CoLT*.