Accounting for the reliability of deep neural networks in representational modeling

Zirui Chen (zchen160@jhu.edu)
Department of Cognitive Science, Johns Hopkins University
3400 N. Charles Street, Baltimore, MD 21218

Michael F. Bonner (mfbonner@jhu.edu)
Department of Cognitive Science, Johns Hopkins University
3400 N. Charles Street, Baltimore, MD 21218

Abstract

In neuroscience, a critical goal is to develop computational models for explaining cortical responses to sensory stimuli. It has been widely recognized that, when evaluating the similarity between brain and model representations, it is necessary to estimate the noise ceiling of cortical activity measurements. However, one important source of noise that has been neglected is the reliability of the models themselves. For deep neural networks, a natural criterion is the consistency of representations learned across different random initializations. Here we demonstrate how to account for the reliability of both brains and models when assessing their similarity, using a metric called integrated reliability integrated reliability. We used simulated data to validate integrated reliability as a more accurate measure for evaluating the limitations in representational modeling compared with conventional noise ceiling estimates based on brain reliability alone. Furthermore, through analyses on actual neural networks and brain representations, we show that model reliability is a key constraint on representational modeling results in neuroscience. Our findings underscore the need to identify and mitigate model variability for improving computational models of cortical representation.

Keywords: deep neural network; representational modeling; vision; noise ceiling; visual representation; fMRI

Introduction

Deep neural networks (DNNs) have demonstrated remarkable success in modeling the representations of biological sensory systems (Kriegeskorte, 2015; Richards et al., 2019; Yamins & DiCarlo, 2016). However, the evaluation of model-brain similarity faces challenges due to internal noise in the measured representations. This has traditionally been addressed by estimating a noise ceiling of cortical activity measurements, often defined as the reliability of representations across trials or subjects, which is thought to be the maximum similarity any model can theoretically achieve (Kriegeskorte, Mur, & Bandettini, 2008; Conwell, Prince, Kay, Alvarez, & Konkle, 2023). Meanwhile, the variability in DNN representations has been largely overlooked, since individual models, when assessed in isolation, generate deterministic outputs. Nevertheless, the feature learning process in DNNs is inherently nondeterministic, influenced by factors such as initialization and dropout. Research has demonstrated that models differing only in their initialization seeds can exhibit significant representational variations (Mehrer, Spoerer, Kriegeskorte, & Kietzmann, 2020). Further studies have revealed that the individual dimensions of neural network representations vary greatly in their reliability, with some learned universally across all initializations and others exhibiting no reliability whatsoever across different initializations (Chen & Bonner, 2023).

In this work, we describe a metric, integrated reliability, that accounts for the reliability of both human brains and DNNs. We validated integrated reliability by simulating model and brain representations with ground-truth similarity, confirming that our metric accurately reflects the true underlying relationship across datasets. Moreover, we characterized the complex nature of reliability in actual DNNs and reveal its impact on patterns of representational similarity to the human brain. Our findings highlight the importance of assessing the reliability of model representations to gain a deeper understanding of the limitations present in current representational models of the human cortex.

Methods

We generated simulated data for models (X) and brains (Y) and manipulated both their internal reliability (r_{xx} and r_{yy}) and their ground-truth representational similarity to one another. The simulated data had power-law eigenspectra matching the eigenvalue distributions of actual neural networks and human brain representations (Gauthaman, Ménard, & Bonner, 2023), and we manipulated the reliability and representational similarity of each principal component. We computed observed similarity scores by taking the correlation (r_{xy}) between paired dimensions of (X) and (Y).

The integrated reliability of (X) and (Y) is the maximum correlation that can be observed between (X) and (Y) given their noise ceilings, and it can be computed as the square root of the product of their internal reliabilities: $\sqrt{r_{xx}r_{yy}}$. Note that for simplicity, we assume that the representations of (X) and (Y) have already been aligned along putative shared dimensions, which is typically accomplished by first fitting encoding or decoding models. Nonetheless, this approach readily generalizes to real data that have been analyzed with encoding or decoding models.

Results & Discussion

Figure 1 shows the observed similarity scores, r_{xy}, and integrated reliability scores, $\sqrt{r_{xx}r_{yy}}$, for instances of (X) and (Y) with varied levels of ground-truth similarity and internal reliability. Figure 1b illustrates a scenario in which the observed similarity scores, r_{xy}, have an upper bound that is determined by the ground-truth representational similarity of (X) and (Y). In this scenario, our simulations yield a range of integrated reliability scores, but even for highly reliable dimensions on the right side of the x-axis, r_{xy} remains below the diagonal. Because r_{xy} is substantially less than the integrated reliability in this scenario, we know that there are fundamental differences between the representations of (X) and (Y) that cannot be attributed to limitations in model and data reliability. In contrast, Figure 1c illustrates a scenario in which the observed similarity scores, r_{xy}, have an upper bound that is determined by the integrated reliability of (X) and (Y). In this scenario, we cannot determine whether any observed differences between (X) and (Y) are due to fundamental differences in their representations or instead to the limited reliability of model instances and brain recordings.

These scenarios highlight important implications for how we interpret representational modeling analyses. For example, if

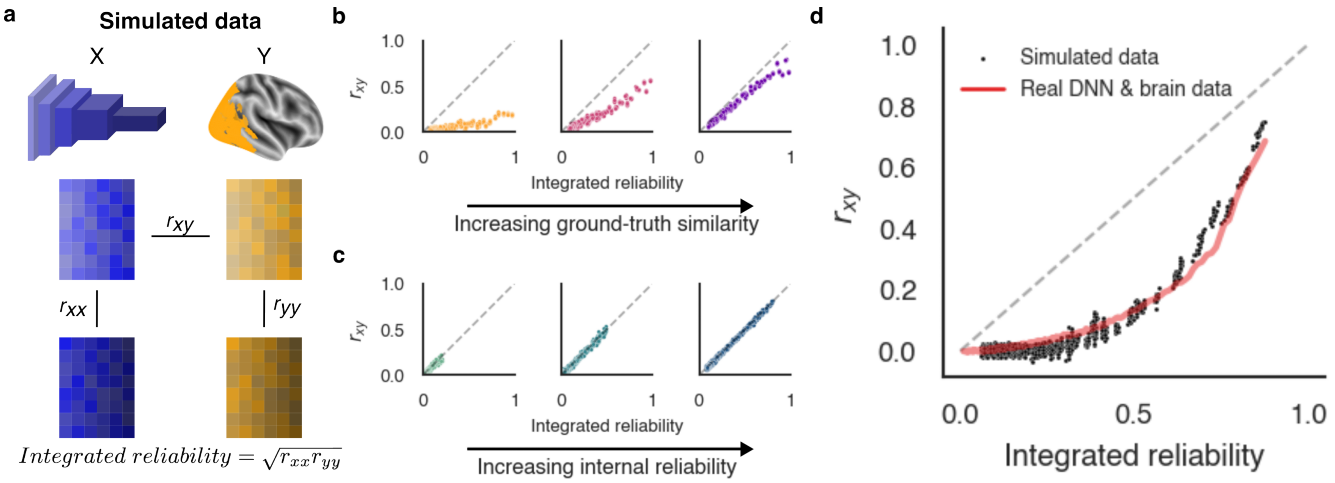

Figure 1: (a) Similarity scores were computed across simulated representations of models and brains (r_{xy}) and reliability scores were computed across different instances of these simulated representations (r_{xx} and r_{yy}). (b-c) These panels show the relationship between integrated reliability and similarity scores for simulations using different levels of ground-truth similarity and internal reliability. In panel b, the simulations have varied ground-truth similarity, and this sets the upper bound on observed similarity scores. In panel c, the simulations have varied internal reliability distributions, and in this case, it is reliability that sets the upper bound on observed similarity scores. (d) Simulated representations in which both ground-truth similarity and internal reliability vary as a function of principal-component rank. This simulation closely matches the findings for real DNNs and human fMRI data, using ResNets trained with different random seeds (Schürholt et al., 2022) and visual cortex fMRI data from the Natural Scenes Dataset (Allen et al., 2022).

observed similarity scores reach the integrated reliability ceiling, it raises the possibility that the model is already fundamentally correct and that the observed similarity scores are simply limited by noise in the data. In this scenario, the only way to determine whether a model has shortcomings is to increase the reliability of both the model and the brain data.

We next explored a scenario in which ground-truth similarity and internal reliability exponentially decreased as a function of principal-component rank. We were specifically interested in this scenario because we observed related trends in DNNs and fMRI data and because many natural datasets have signals that decay exponentially at high-rank principal components. As shown in Figure 1d, this scenario yields a nonlinear relationship between observed similarity and integrated reliability. Here the most reliable dimensions also have high ground-truth similarity across (X) and (Y), while less reliable dimensions have lower ground-truth similarity. We further compared this scenario with real data from DNNs and fMRI recordings, and we found that this simulation closely reproduces the relationship between similarity and reliability observed in real models and brains. This suggests that the similarities between DNNs and brains decrease rapidly as a function of principal-component rank, even though these representations are, nonetheless, reliable across many ranks in both brains and DNNs.

In sum, our work demonstrates the importance of accounting for the reliability of both brain recordings and computational models when assessing their representational similarity. The integrated reliability metric illustrated here can help to properly interpret the successes and failures of representational models and to distinguish between cases where models are fundamentally misaligned with the brain or where improvements are still possible by increasing the reliability of both models and brain recordings.

References

Allen, E. J., St-Yves, G., Wu, Y., Breedlove, J. L., Prince, J. S., Dowdle, L. T., ... others (2022). A massive 7t fmri dataset to bridge cognitive neuroscience and artificial intelligence. *Nature neuroscience*, *25*(1), 116–126.

Chen, Z., & Bonner, M. (2023). Canonical dimensions of vision. In *2023 conference on cognitive computational neuroscience.* Cognitive Computational Neuroscience. doi: 10.32470/ccn.2023.1588-0

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*.

Gauthaman, R. M., Ménard, B., & Bonner, M. (2023). Revealing the high-dimensional latent structure in visual cortical representations. In *2023 conference on cognitive computational neuroscience.* Cognitive Computational Neuroscience. doi: 10.32470/ccn.2023.1652-0

Kriegeskorte, N. (2015). Deep neural networks: a new framework for modeling biological vision and brain information processing. *Annual review of vision science*, *1*, 417–446.

Kriegeskorte, N., Mur, M., & Bandettini, P. A. (2008). Representational similarity analysis-connecting the branches of systems neuroscience. *Frontiers in systems neuroscience*, 4.

Mehrer, J., Spoerer, C. J., Kriegeskorte, N., & Kietzmann, T. C. (2020). Individual differences among deep neural network models. *Nature communications*, *11*(1), 5725.

Richards, B. A., Lillicrap, T. P., Beaudoin, P., Bengio, Y., Bogacz, R., Christensen, A., ... others (2019). A deep learning framework for neuroscience. *Nature neuroscience*, *22*(11), 1761–1770.

Schürholt, K., Taskiran, D., Knyazev, B., Giró-i Nieto, X., & Borth, D. (2022). Model zoos: A dataset of diverse populations of neural network models. *Advances in Neural Information Processing Systems*, *35*, 38134–38148.

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