Towards Task-Appropriate Readout Mechanisms For Physical Scene Understanding

Khaled Jedoui Al-Karkari (thekej@stanford.edu)
Stanford University
Stanford, CA

Rahuk Venkatesh (rmvenkat@stanford.edu)
Stanford University
Stanford, CA

Haoliang Wang (haw027@ucsd.edu)
University of California, San Diego
San Diego, CA

Thomas O'Connell (tpo@mit.edu)
Massachusetts Institute of Technology, Cambridge
Cambridge, MA

Yoon H. Bai (yhb@mit.edu)
Massachusetts Institute of Technology, Cambridge
Cambridge, MA

Joshua B. Tenenbaum (jbt@mit.edu)
Massachusetts Institute of Technology, Cambridge
Cambridge, MA

Judith E. Fan (jefan@stanford.edu)
Stanford University
Stanford, CA

Kevin A. Smith (k2smith@mit.edu)
Massachusetts Institute of Technology, Cambridge
Cambridge, MA

Daniel L.K. Yamins (yamins@stanford.edu)
Stanford University
Stanford, CA

Abstract

Establishing robust readouts is essential in the study and interpretation of both artificial intelligence models and the brain. These linking functions extract relevant information that allows us to understand model behavior, as well as brain activity patterns. However, building appropriate readouts for any given task has been challenging due to the lack of clear strategies and methods to design such functions. In this paper, we propose an approach to derive a readout model of least complexity using an idealized data representation (with all information needed to solve a task). We investigate the ability of our model to decode representations for a simple physics understanding task; object contact detection. We demonstrate that our approach provides a better qualitative signal about AI models, compared to the traditional linear readout. Our readout promises to not only improve the benchmarking of AI models, but also provides a path forward for building more powerful neural decoders and gaining insight into how different brain regions represent and reason about the physical world.

Keywords: AI, Neuroscience, Cognitive Benchmarking.

Introduction

Understanding the neural mechanisms that give rise to cognition and behavior is a central goal of both neuroscience and artificial intelligence research (Macpherson et al., 2021). Achieving this goal requires developing quantitative methods, and “linking functions”, to relate patterns of brain activity to the external stimuli, behaviors, and internal representations involved. Readout techniques have emerged as powerful tools for achieving this, enabling us to decode, or map, features of interest onto recorded neural (or computational model) responses. By providing an explicit test of hypothesized links between brain activity and computational processes, readout approaches have become indispensable for understanding how the brain processes information (Koren, Bondanelli, & Panzeri, 2023). Similarly, readout models have also become crucial benchmarks for evaluating computational models of brain function against empirical datasets (Bear et al., 2021).

The use of readout models has progressed significantly, from early pioneering work introducing classifier-based readouts to decode object information from IT neurons (Hung, Kreiman, Poggio, & DiCarlo, 2005), to using learned weighted sums of IT activity to precisely predict human object recognition behavior patterns (Majaj, Hong, Solomon, & DiCarlo, 2015). As deep neural network models became prominent, readouts took on a new role, applying large-scale primate behavioral readouts to reveal the limitations of feedforward networks in capturing image-level discrimination patterns (Rajalingham et al., 2018). Building on this foundation, the scope of readout models has expanded further, enabling the decoding of both seen and imagined objects from brain activity (Horikawa & Kamitani, 2017) and enabling the reconstruction of visual illusions from brain activity (Cheng et al., 2023). Moreover, readout models have also been adapted to study the semantic reconstruction of language from brain recordings (Tang, LeBel, Jain, & Huth, 2023), as well as improving EEG signal decoding (de Oliveira & Rodrigues, 2023). Most recently, scientists worked to expand readout approaches beyond the linear vs. nonlinear divide that traditionally constrained mapping models (Ivanova et al., 2022), putting emphasis on measure of model complexity over linearity when selecting and evaluating mapping models. Despite the proposed conceptual advances (Ivanova et al., 2022), the lack of a concrete strategy for selecting the appropriate readout given a task, remains an ongoing challenge in the field.

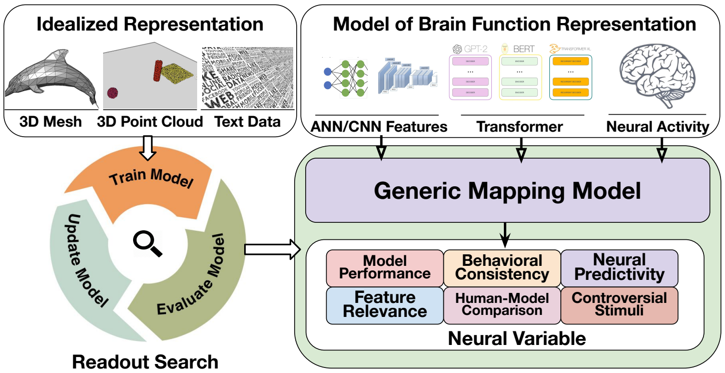

Figure 1: We propose a generalizable strategy that leverages idealized representations and performs a search to find a generic minimum complexity model capable of handling different neural representations.

In this paper, we propose a procedure to identify a suitable readout for a given task using idealized data representation (**Figure 1**). Our approach represents a generalizable strategy that provides readout models capable of handling different neural representations, variable input sizes, and complex tasks beyond categorization, crucial for understanding higher-order cognitive processes beyond simple object recognition. We demonstrate the efficacy of this systematic approach on a physics understanding task; object contact detection.

Methods

This study proposes a novel strategy for designing task-appropriate readout models that leverage idealized representations. Our method aims to achieve the following objectives:

1. Identify the smallest parameter readout model capable of effectively solving the idealized representation problem.
2. Ensure that our readout extends to any neural representation and accommodates variable input sizes.

Dataset & Task We investigate the efficacy of our approach using the Physion v1.5 Benchmark (Bear et al., 2021). This benchmark comprises realistic simulations of diverse physical scenarios, challenging models to reason about stability, rolling motion, containment, object linkage, and other physical concepts. Instead of relying on 2D video inputs, we use an augmented point cloud structure as our ground truth data **(Figure 2a)**. This provides a richer physics representation, encoding spatial coordinates **(x, y, z)**, time **(t)**, and color **(r, g, b)** information for each object in the scene. Our logic stems from the belief that this is a sufficient representation to solve this task

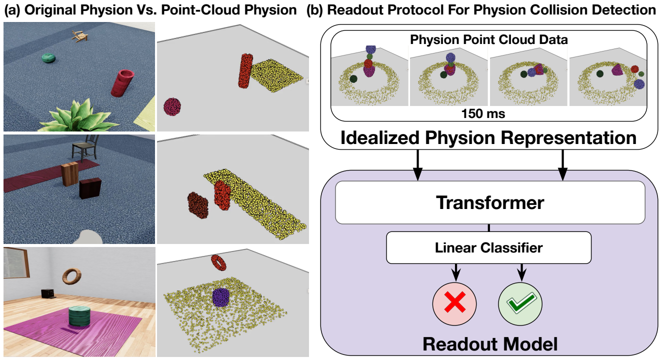
(a) Original Physion Vs. Point-Cloud Physion
(b) Readout Protocol For Physion Collision Detection
Physion Point Cloud Data
150 ms
Idealized Physion Representation
Transformer
Linear Classifier
Readout Model

Figure 2: **(a)** We use an augmented object-centric point cloud structure as our idealized representation. **(b)** We use a readout protocol that combines Physion Point-Cloud data with a transformer encoder that learns to predict contact labels.

and that the smallest model that solves it should be considered a valid readout. We train our readouts on an object contact detection task, where the model must detect contact events within a 150ms video stimulus. Objects of interest are highlighted with red and yellow to guide the model's attention.

Readout Model Architecture Our approach relies on a readout architecture that combines a transformer encoder (Vaswani et al., 2023) with a linear classifier that predicts contact labels. The ability of the transformer for handling variable input sizes, makes it suitable for tasks with a temporal nature, while keeping a relatively small number of parameters (compared to a linear model). For our point-cloud data, we combine our transformer readout with a two-layer multi-layer perceptron (MLP) that serves as a feature extractor, similar to the PointNet approach (Qi, Su, Mo, & Guibas, 2017). The MLP pre-processes our input to generate a higher-level representation which is then fed into the transformer **(Figure 2b)**. We remove the MLP component when probing other models.

Readout Search Strategy We adopt an iterative approach to model complexity, starting with a single-layer transformer encoder with one head attention. We progressively increase the number of layers/heads until satisfactory performance on the held-out Physion Test set is reached **(Figure 1)**. Models are tested by splitting videos into 150ms snippets and deducing a global contact label from local predictions. Once we reach the desired model, we adapt our readout and data to be applicable to other representations, enforcing a same parameter readout for all representations.

Human-model evaluation We collect human responses (25 per test stimuli) on the same contact detection task (full video stimuli). By analyzing the inter-rater reliability between model predictions and human behavior, we can establish a baseline for human performance and assess how closely models align with human judgments using Cohen κ's coefficient **(Figure 3)**.

Results

Optimal Readout Model Our readout search strategy explores various parameter sizes and identifies an optimal configuration with 4 attention heads, 1 encoder layer and a 128-dimensional embedding. This results in a model with $200k$ parameters, significantly less than the $2.3M$ parameters required by a linear readout for the same task.

Performance Comparison We compare the performance of our transformer readout against a linear readout trained on the same idealized (point-cloud) representation. The linear model fails to achieve satisfactory performance, reaching only 82% accuracy. In contrast, our model achieves human-level performance (94.22%) with an accuracy of 94.83% **(Figure 3a)**. Furthermore, our model aligns better with human behavior showing a Cohen κ score of 0.81 (vs. 0.63 for linear).

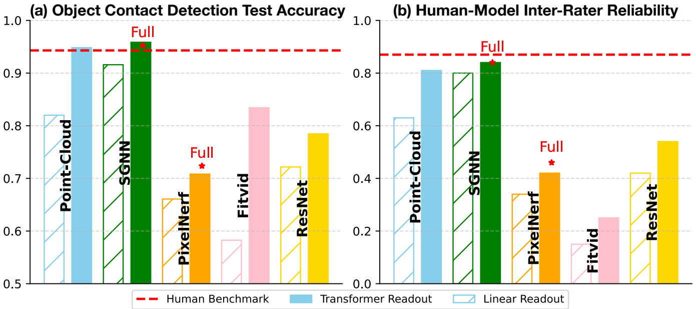

Figure 3: Our readout procedure finds a transformer readout that **(a)** outperforms the linear readout and achieves human performance on contact detection and **(b)** aligns better with human behavior.

Additionally, we show the versatility of our approach by adapting our readout to 4 different representations: particle-based representation with **SGNN** (Han et al., 2022) trained on Physion Particles(Bear et al., 2021), a 3D spatio-temporal representation model, **PixelNERF** (Yu, Ye, Tancik, & Kanazawa, 2021) trained on shapenet (Chang et al., 2015) and augmented with a LSTM layer (finetuned on the Physion v1.5 train dataset), a forward prediction video model, **FitVid** (Babaeizadeh et al., 2021) trained on Ego4d(Grauman et al., 2021) and a static image representation model, **ResNet50** (He, Zhang, Ren, & Sun, 2015) trained on ImageNet (Russakovsky et al., 2014). We constrain the dimension of our features to be the same using PCA, in order to use the same readout for evaluation (non-PCA results are also reported on **Figure 3a and 3b**). Our results showcase the effectiveness of each model in physics understanding and demonstrate the superiority of our new readout as a benchmark for any model.

Conclusion

By developing a generic readout model selection procedure, we aim to provide a principled approach for mapping neural representations to computational outputs, irrespective of system and task complexity. Our approach demonstrates that discovering a readout mechanism based on idealized representations can better map network representations onto human performance as compared to linear readouts. While currently only demonstrated on Physion,this method holds promise for advancing neural decoding in physics and other higher-order cognitive processes, enabling a deeper understanding of how brains and AI models represent and reason about the world.

References

Babaeizadeh, M., Saffar, M. T., Nair, S., Levine, S., Finn, C., & Erhan, D. (2021). *Fitvid: Overfitting in pixel-level video prediction.*

Bear, D. M., Wang, E., Mrowca, D., Binder, F. J., Tung, H. F., Pramod, R. T., … Fan, J. E. (2021). Physion: Evaluating physical prediction from vision in humans and machines..

Chang, A. X., Funkhouser, T., Guibas, L., Hanrahan, P., Huang, Q., Li, Z., … Yu, F. (2015). *ShapeNet: An Information-Rich 3D Model Repository* (Tech. Rep. No. arXiv:1512.03012 [cs.GR]). Stanford University — Princeton University — Toyota Technological Institute at Chicago.

Cheng, F. L., Horikawa, T., Majima, K., Tanaka, M., Abdelhack, M., Aoki, S. C., … Kamitani, Y. (2023). Reconstructing visual illusory experiences from human brain activity. *Science Advances*, *9*(46), eadj3906.

de Oliveira, I. H., & Rodrigues, A. C. (2023). Empirical comparison of deep learning methods for eeg decoding. *Frontiers in Neuroscience*, *16*.

Grauman, K., Westbury, A., Byrne, E., Chavis, Z., Furnari, A., Girdhar, R., … Malik, J. (2021). Ego4d: Around the world in 3, 000 hours of egocentric video. *CoRR*, *abs/2110.07058*. Retrieved from `https://arxiv.org/abs/2110.07058`

Han, J., Huang, W., Ma, H., Li, J., Tenenbaum, J. B., & Gan, C. (2022). Learning physical dynamics with subequivariant graph neural networks. *arXiv preprint arXiv:2210.06876*.

He, K., Zhang, X., Ren, S., & Sun, J. (2015). Deep residual learning for image recognition. *CoRR*, *abs/1512.03385*. Retrieved from `http://arxiv.org/abs/1512.03385`

Horikawa, T., & Kamitani, Y. (2017). Generic decoding of seen and imagined objects using hierarchical visual features. *Nature Communications*, *8*(1), 15037.

Hung, C. P., Kreiman, G., Poggio, T., & DiCarlo, J. J. (2005). Fast readout of object identity from macaque inferior temporal cortex. *Science*, *310*(5749), 863–866.

Ivanova, A. A., Schrimpf, M., Anzellotti, S., Zaslavsky, N., Fedorenko, E., & Isik, L. (2022, August). Beyond linear regression: mapping models in cognitive neuroscience should align with research goals. *Neurons, Behavior, Data analysis, and Theory*, *1*.

Koren, V., Bondanelli, G., & Panzeri, S. (2023). Computational methods to study information processing in neural circuits. *Computational and Structural Biotechnology Journal*, *21*, 910–922.

Macpherson, T., Churchland, A., Sejnowski, T., DiCarlo, J., Kamitani, Y., Takahashi, H., & Hikida, T. (2021). Natural and artificial intelligence: A brief introduction to the interplay between ai and neuroscience research. *Neural Networks*, *144*, 603-613.

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

Qi, C. R., Su, H., Mo, K., & Guibas, L. J. (2017). *Pointnet: Deep learning on point sets for 3d classification and segmentation.*

Rajalingham, R., Issa, E. B., Bashivan, P., Kar, K., Schmidt, K., & DiCarlo, J. J. (2018). Large-scale, high-resolution comparison of the core visual object recognition behavior of humans, monkeys, and state-of-the-art deep artificial neural networks. *Journal of Neuroscience*, *38*(33), 7255–7269.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., … Fei-Fei, L. (2014). Imagenet large scale visual recognition challenge. *CoRR*, *abs/1409.0575*. Retrieved from `http://arxiv.org/abs/1409.0575`

Tang, J., LeBel, A., Jain, S., & Huth, A. G. (2023). Semantic reconstruction of continuous language from non-invasive brain recordings. *Nature Neuroscience*, *26*(5), 858–866.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., … Polosukhin, I. (2023). *Attention is all you need.*

Yu, A., Ye, V., Tancik, M., & Kanazawa, A. (2021). pixelNeRF: Neural radiance fields from one or few images. In *Cvpr.*