Representational dynamics in the human ventral stream captured in deep recurrent neural nets

Tim C Kietzmann (tim.kietzmann@mrc-cbu.cam.ac.uk)
MRC Cognition and Brain Sciences Unit, University of Cambridge,
15 Chaucer Road, Cambridge, CB2 7EF, United Kingdom

Courtney J Spoerer (courtney.spoerer@mrc-cbu.cam.ac.uk)
MRC Cognition and Brain Sciences Unit, University of Cambridge,
15 Chaucer Road, Cambridge, CB2 7EF, United Kingdom

Lynn Sörensen (lynn.sorensen@mrc-cbu.cam.ac.uk)
Psychology Department, University of Amsterdam
Nieuwe Achtergracht 129, 1018 WS Amsterdam, Netherlands

Radoslaw M Cichy (rmcichy@zedat.fu-berlin.de)
Department of Education and Psychology, Freie Universität Berlin
Habelschwerdter Allee 45, 14195 Berlin, Germany

Olaf Hauk (olaf.hauk@mrc-cbu.cam.ac.uk)
MRC Cognition and Brain Sciences Unit, University of Cambridge,
15 Chaucer Road, Cambridge, CB2 7EF, United Kingdom

Nikolaus Kriegeskorte (n.kriegeskorte@columbia.edu)
Department of Psychology, Columbia University
3227 Broadway, New York, NY 10027, USA

Abstract:

Feedforward models of visual processing provide human-level object-recognition performance and state-of-the-art predictions of temporally averaged neural responses. However, the primate visual system processes information through dynamic recurrent signaling. Here we characterize and model the representational dynamics of visual processing along multiple areas of the human ventral stream by combining source-reconstructed magnetoencephalography data and deep learning. Our analyses of the empirical data revealed neural responses that traverse distinct encoding schemes across time and space, in line with signatures of recurrent signaling. Next, we estimated the ability of different deep network architectures to capture the neural dynamics by using neural representational trajectories as space- and time-varying target functions. Feedforward models, with units that ramp-up their activity over time, predicted nonlinear representational dynamics, but failed to account for the neural effects. Recurrent models of matched parametric complexity significantly better explained the held-out data. We next optimised the recurrent networks for a classification objective only. While performing significantly better than random networks, the variance explained fell short of the architecture's capacity. This paves the way for the search for additional objectives that the ventral stream may optimize, including category-orthogonal objectives, noise, occlusion, manipulability, and semantics.

Keywords: deep learning, recurrence, visual processing, MEG

Introduction

Our ability to derive meaning from retinal input relies on an intricate network of interconnected cortical regions along the ventral visual pathway. Although neuronal selectivity has been characterized across a variety of visual areas, we still do not have a mechanistic understanding of the underlying computations. To this end, deep neural networks (DNNs) provide a promising tool, enabling us to obtain image-computable models that directly test theoretical predictions (Kietzmann, McClure, & Kriegeskorte, 2017; Kriegeskorte, 2015; Yamins & DiCarlo, 2016). To date, feed-forward DNNs optimized for categorization accuracy provide the best model of time-averaged neural responses to novel stimuli across visual areas (Cadena et al., 2017; Güçlü

& van Gerven, 2014; Khaligh-Razavi & Kriegeskorte, 2014; Wen et al., 2017; Yamins et al., 2014). The primate visual system, however, contains abundant lateral and feedback connections. These give rise to rapid recurrent interactions, which are thought to contribute to visual inference (Freiwald & Tsao, 2010; Sugase, Yamane, Ueno, & Kawano, 1999).

Results and Discussion

To better understand how neural representations change across space and time, we here combine MEG source reconstruction with representational similarity analysis (RSA; Figure 1A). We focus our analyses on the human ventral stream, a network of cortical regions known to process object information. Three regions of interest (ROIs) were defined in each participant, including early- (V1-V3), intermediate- (V4t,LO1-3), and high-level (IT/PHC) visual areas. Across two sessions each, 15 participants viewed 92 stimuli from a diverse set of natural object categories (Figure 1B) while we recorded stimulus evoked activation patterns across MEG 306 channels. Following source projection using individual anatomical MRIs and minimum norm estimates (MNE), we used correlation distance to estimate all pairwise dissimilarities in the neural population activity (summarized in representational dissimilarity matrices (RDMs, Kriegeskorte, Mur, & Bandettini, 2008)). RDMs were computed separately for each ROI and time-point, yielding representational trajectories that describe how the neural code changes across time in each area. The empirical data indicate substantial changes across the first 300ms of processing (Figure 1D, row 1 for an example), suggesting that the underlying neural populations traverse multiple distinct representational schemes.

Given the time-varying representational trajectories for all ROIs, we used deep learning as a framework to gain insight into the principles that may underlie the corresponding neural computations. Deep learning has the advantage of yielding image-computable, task-performing models, while enabling researchers to explicitly test for different hypotheses about the computational objectives that the brain may optimize. This normative approach asks what functions need to be optimized in an artificial system to result in internal representations that best align with neural data. Before addressing the ultimate question of neural objectives, however, we first tested different DNN architectures for their ability to model 300 ms of representational dynamics observed across all measured ROIs in the brain. We derived a stochastic gradient descent method, an extension to representational distance learning (RDL, Mcclure & Kriegeskorte, 2016), to inject the time-varying MEG RDM trajectories of the three different ROIs as separate target functions into different layers of the deep networks (Figure 1C). Each trained DNN thereby attempts to model all representational dynamics observed across time in all human ventral stream ROIs. The networks were trained using a separate set of 141k images (RDL-61), matching the categorical structure of the 92 stimuli shown during the human MEG data acquisition. In addition to the time-varying RDL objective, the networks were optimized using a time-decaying categorization objective to boost RDL generalization. After training, the representational trajectories of the DNNs were extracted for the previously unseen 92 experimental stimuli, and tested against unseen data using cross-validation procedures (split-half across MEG measurement sessions).

Two architecture types were tested. First, we extended the feed-forward model class to exhibit non-linear dynamics by allowing each neural network layer to ramp-up its activity over time (models B_{K9}, and B_{K11}). This was contrasted with a fully recurrent convolutional architecture that contained bottom-up, top-down and lateral connections (BLT network, Spoerer, McClure, & Kriegeskorte, 2017). BLT and ramp-up networks were matched in the number of free parameters. Figure 1D shows the model predictions derived from the different architectures for region V4t,LO1-3, together with empirical MEG RDMs. To quantitatively evaluate the match between empirical and model representations, we first compared the average distance across all stimulus conditions for each time-point (Figure 2A). Ramp-up feed-forward networks exhibit non-linear representational dynamics, but they were not able to closely follow the empirical data (trajectory correlations with held-out data: .83, .58, .48 for V1-3, V4t,LO1-3, and IT/PHC). Fully recurrent DNNs, however, closely matched the ventral stream dynamics (trajectory correlations: .95, .93, .97 for V1-3, V4t,LO1-3, and IT/PHC). Next, we evaluated the detailed patterns of representational distances beyond changes in mean distance, by computing the correlation between the model RDMs and the held-out empirical RDMs for each time-point. The model fit was computed by averaging these correlations across time (Figure 2B). For each ventral stream area, the recurrent models significantly outperformed the feed-forward ramp-up models in predicting the detailed distance patterns (Wilcoxon sign-rank test, FDR corrected across all tests at $q<0.05$).

Having established that fully recurrent BLT models are capable of capturing the empirically observed representational dynamics, we next estimated how

much of the explanatory power can be achieved by following the normative approach of optimizing solely on a categorization objective (565 categories, 1.5 million images; Mehrer, Kietzmann, & Kriegeskorte, 2017). In terms of categorization performance, the recurrent architecture clearly outperformed the feed-forward ramp-up model (50.70% validation accuracy for BLT, vs 40.73 % for B_{K11}). Using the same RDL training set as before, we optimised the read out from these category networks to best explain the empirical data by selecting the best layers, weighting the features in each layer, and optimising the read out timing. Matching previous results, the category-optimised networks clearly outperformed random control networks. However, the correlations were reduced by about 40% compared to the networks' capacity as demonstrated when optimised via RDL training. Our approach of combining MEG source-based RSA and DNN modelling paves the way for the search for additional objectives beyond categorisation that govern neural activity in the first 300 ms of computation.

Acknowledgments

This research was funded by the UK Medical Research Council (Programme MC-A060- 5PR20), by a European Research Council Starting Grant (ERC-2010-StG 261352), by the Human Brain Project (EU grant 604102), and a DFG research fellowship to TCK.

References

Cadena, S. A., Denfield, G. H., Walker, E. Y., Gatys, L. A., Tolias, A. S., Bethge, M., & Ecker, A. S. (2017). Deep convolutional models improve predictions of macaque V1 responses to natural images. *BioRxiv*, (doi: 10.1101/201764), 1–16. https://doi.org/10.1101/201764

Freiwald, W. A., & Tsao, D. Y. (2010). Functional compartmentalization and viewpoint generalization within the macaque face-processing system. *Science*, *330*(6005), 845–51. https://doi.org/10.1126/science.1194908

Güçlü, U., & van Gerven, M. A. J. (2014). Unsupervised Feature Learning Improves Prediction of Human Brain Activity in Response to Natural Images. *PLoS Computational Biology*, *10*(8). https://doi.org/10.1371/journal.pcbi.1003724

Khaligh-Razavi, S.-M., & Kriegeskorte, N. (2014). Deep Supervised, but Not Unsupervised, Models May Explain IT Cortical Representation. *PLoS Computational Biology*, *10*(11), 1–29. https://doi.org/10.1371/journal.pcbi.1003915

Kietzmann, T. C., McClure, P., & Kriegeskorte, N. (2017). Deep Neural Networks In Computational Neuroscience. *BioRxiv*, 133504. https://doi.org/10.1101/133504

Kriegeskorte, N. (2015). Deep Neural Networks: A New Framework for Modeling Biological Vision and Brain Information Processing. *Annual Review of Vision Science*, *1*(1), 417–446. https://doi.org/10.1146/annurev-vision-082114-035447

Kriegeskorte, N., Mur, M., & Bandettini, P. (2008). Representational similarity analysis - connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, *2*(November), 4. https://doi.org/10.3389/neuro.06.004.2008

Mcclure, P., & Kriegeskorte, N. (2016). Representational Distance Learning for Deep Neural Networks. *Frontiers in Computational Neuroscience*, *10*, 131.

Mehrer, J., Kietzmann, T. C., & Kriegeskorte, N. (2017). Deep neural networks trained on ecologically relevant categories better explain human IT. In *Cognitive Computational Neuroscience Meeting* (Vol. 1, pp. 1–2).

Spoerer, C. J., McClure, P., & Kriegeskorte, N. (2017). Recurrent Convolutional Neural Networks: A Better Model of Biological Object Recognition. *Frontiers in Psychology*, *8*(1551), 1–14.

Sugase, Y., Yamane, S., Ueno, S., & Kawano, K. (1999). Global and fine information coded by single neurons in the temporal visual cortex. *Nature*, *400*(6747), 869–73. https://doi.org/10.1038/23703

Wen, H., Shi, J., Zhang, Y., Lu, K.-H., Cao, J., & Liu, Z. (2017). Neural Encoding and Decoding with Deep Learning for Dynamic Natural Vision. *Cerebral Cortex*, (May), 1–25. https://doi.org/10.1093/cercor/bhx268

Yamins, D. L., & DiCarlo, J. J. (2016). Using goal-driven deep learning models to understand sensory cortex. *Nature Neuroscience*, *19*(3), 356–365. https://doi.org/10.1038/nn.4244

Yamins, D. L., Hong, H., Cadieu, C., Solomon, E., Seibert, D., & DiCarlo, J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences of the United States of America*, *111*(23), 8619–24.

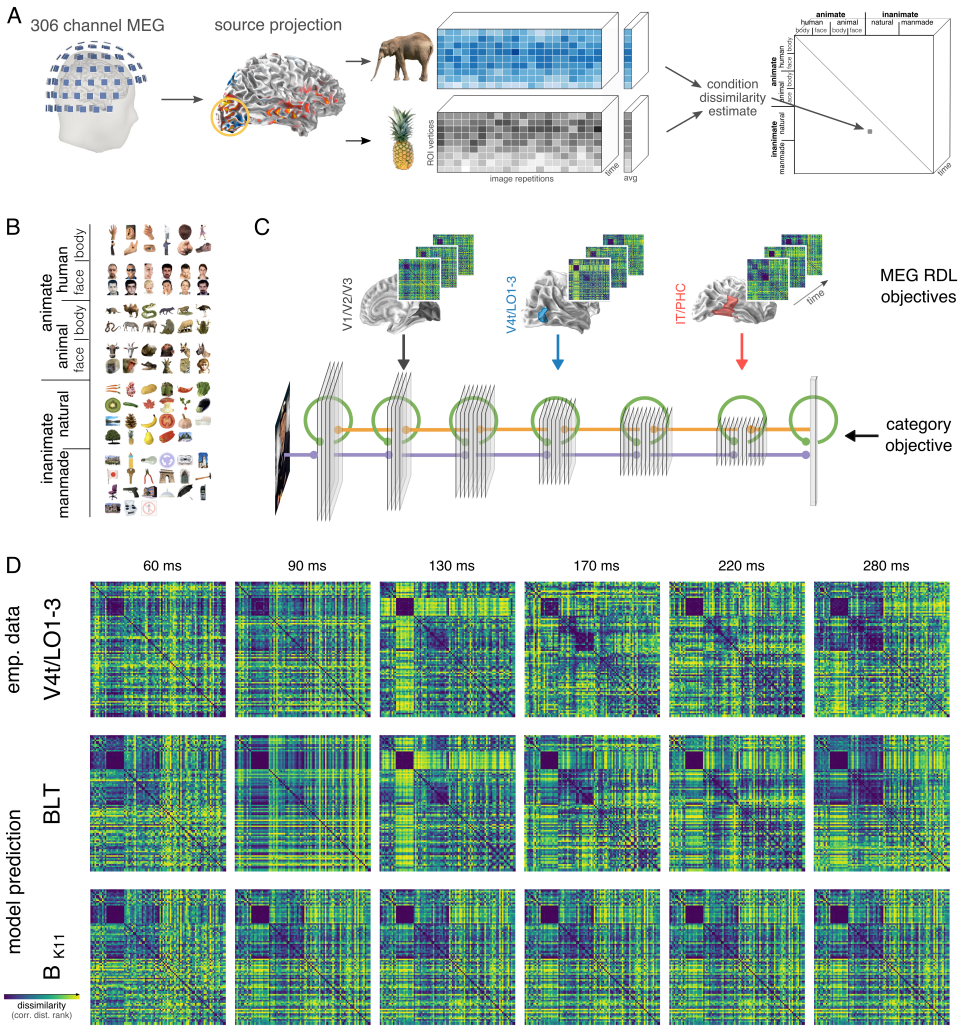

Figure 1: (**A**) MEG analysis pipeline to extract time-varying representational trajectories across multiple ROIs. (**B**) Stimulus set and category structure of experimental stimuli. Matching the category structure, a novel training set was created (141k images) based on which DNNs were trained. (**C**) The time-varying neural RDMs of three separate ROIs were used as target for deep learning. Given pairs of stimuli, the neural distance (for each given ROI and time-point) was used to derive a learning gradient, pushing networks to mirror the neural distances as close as possible. Networks thereby attempt to simultaneously model the time-varying RDMs across all measured ventral stream ROIs (**D**) Example RDM movies for a ventral stream ROI (V4t,LO1-3) and corresponding DNN model predictions, as derived from recurrent (middle row) and ramp-up feed-forward networks (bottom-row).

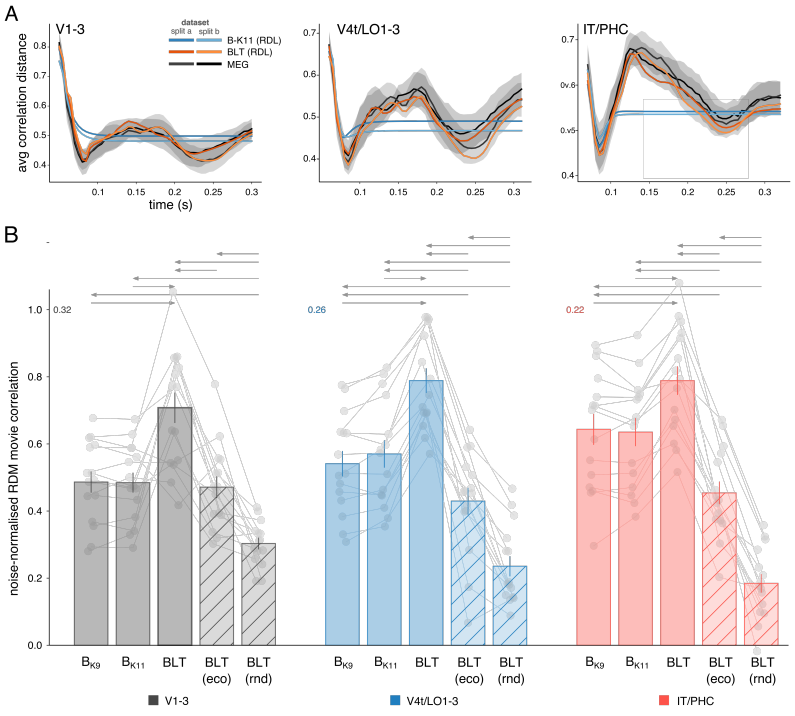

Figure 2: (A) Temporal development of the average pattern distance for MEG data as well as different DNN architectures. (B) Average RDM correlations between held-out empirical RDMs and model predictions, normalized by the predictive performance of the training data (latter shown at 1.0 for each ROI). Arrows indicate significant differences (non-parametric tests, FDR corrected for all tests at $q<0.05$). Uniform bars indicate networks optimised using deep learning on the empirical data directly (RDL). Striped bars indicate networks that were trained via external objectives (categorization, and random), but subsequently optimised using RDL set to allow for optimal read-out. Recurrent- outperform feed-forward DNNs, and classification objectives outperform random weights. Yet, a large gap remains between the recurrent categorization networks, and the explanatory capacity of the underlying architecture.