A single spatial transform improves predictions of neural responses by deep neural network models

Niklas Müller (n.muller@uva.nl)
Psychology Research Institute, University of Amsterdam, The Netherlands

H. Steven Scholte[*] (h.s.scholte@uva.nl)
Psychology Research Institute, University of Amsterdam, The Netherlands

Iris I. A. Groen[*] (i.i.a.groen@uva.nl)
Informatics Institute, University of Amsterdam, The Netherlands
Psychology Research Institute, University of Amsterdam, The Netherlands

[*] Shared senior author

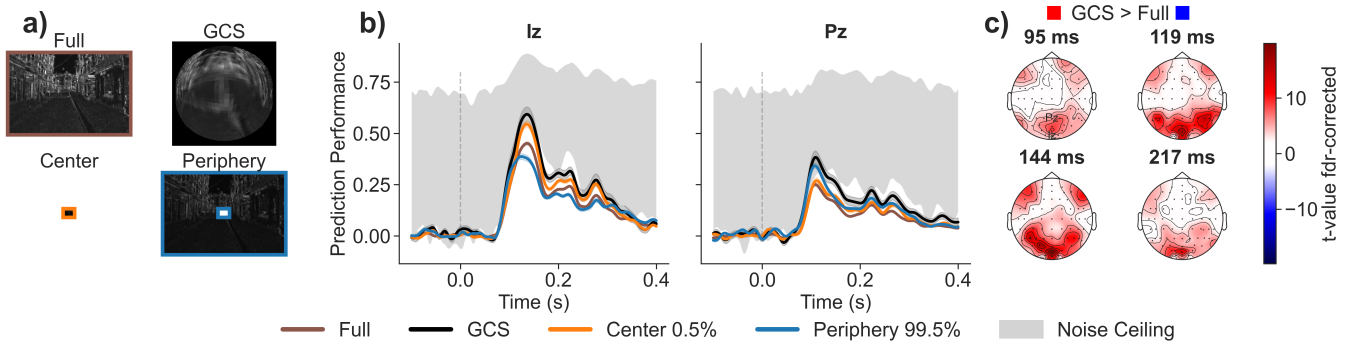

Figure 1: **a)** Transformations of DNN activation maps. **b)** Prediction timecourse for encoding models using transforms in a) at a posterior (Iz) and central electrode (Pz). **c)** Topoplots of t-tests for GCS vs. full transform at 4 time points (FDR-corrected).

Abstract

Encoding models are a powerful tool for predicting neural responses on a per-image basis using the features of deep neural networks (DNNs). Efforts to improve prediction performance have largely focused on changing aspects of DNN training or model architecture. Here, we take a pre-trained DNN and explore whether a fixed, spatial reweighting of features can improve neural predictions without the need for retraining the neural network. We find that spatially distinct areas of visual input (center versus periphery) uniquely contribute to the temporal dynamics of human EEG recordings. These dynamics are unified when transforming feature maps based on ganglion cell sampling (GCS). The same GCS transform improved predictions of both monkey electrophysiology and human fMRI recordings.

Keywords: Encoding models; Deep neural network; EEG; Visual Processing

Introduction

Deep neural networks (DNNs) have recently emerged as state-of-the-art models of primate visual processing. Linear encoding models that regress neural data onto convolutional features of task-optimized DNNs achieve high performance across multiple modalities, including EEG (Gifford, Dwivedi, Roig, & Cichy, 2022), MEG (Seeliger et al., 2018), BOLD (Storrs, Kietzmann, Walther, Mehrer, & Kriegeskorte, 2021), and electrophysiology (Yamins et al., 2014). While substantial research has studied effects of training strategy (e.g., supervised vs. self-supervised), training dataset, or model architecture (e.g., CNNs vs. Transformers) (Conwell, Prince, Kay, Alvarez, & Konkle, 2022), less is known about the optimal way of mapping DNN features into a neural space.

Unlike standard DNNs, primate brains do not uniformly sample visual input. The eye has a foveal region with a high density of cones and retinal ganglion cells (RGCs), and a periphery with low cone and RGC density (but high rod density) (Oyster, Takahashi, & Hurst, 1981; Kreiman, 2021). The non-uniform distribution of RGCs yields an over-representation of foveal input which projects to substantially more brain volume than peripheral input (cortical magnification, (Cowey & Rolls, 1974)). While several studies have explored effects of applying retina-like transformations on DNN performance (Lindsey, Ocko, Ganguli, & Deny, 2019; Deza & Konkle, 2020), this non-uniform differential sampling of foveal and peripheral information is usually not taken into account when fitting neural data with convolutional features from deep neural networks.

Here, we show that applying a simple spatial transformation inspired by retinal sampling to DNN feature maps improves encoding model performance across multiple datasets.

Methods

Neural datasets Human subjects (n=31) were presented with 702 large, high-resolution images (2155x1440 pixels, 50x29.5° va) in a rapid-serial-visual-presentation (RSVP) paradigm while EEG was collected. One trial consisted of a series of 20 stimuli (100 ms) interspersed by gray screens (300 ms). Subjects were instructed to fixate and detect target images presented in 50% of trials (targets were excluded from analysis). 66% of the images were repeated 5 times (the training split) and 33% 10 times (test split) across the experiment. EEG data were preprocessed by creating 500 ms epochs, demeaning, filtering, and removing bad trials, applying ocular correction and converting to Current Source Density responses. Epochs were averaged across repetitions, creating ERPs specific to each subject, electrode and image. We also used two datasets from Brain-Score (Schrimpf et al., 2018) with electrophysiology recordings in V1, V2, V4, and IT in macaques viewing textures and natural images, and fMRI data from the Algonauts challenge 2023 (Gifford et al., 2023).

Encoding models Each image in the EEG dataset was manually annotated with bounding boxes for 16 common object classes. Cropping objects from the full images yielded a new set of 82.236 images used to train an Alexnet (Krizhevsky, Sutskever, & Hinton, 2012) on 16-way object classification. After training, features were extracted for full scene images from the three max-pooling layers. The following spatial transforms were applied on each kernel's activation map: cropping the central 0.5% and keeping either (1) the central part ("Center") or (2) the surrounding part ("Periphery"), or (3) applying a ganglion cell sampling (GCS) to the full image that magni-

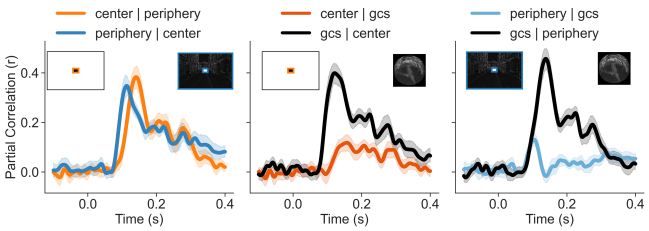

Figure 2: Partial correlations of encoding models using other models' predictions as covariate.

fies central information and degrades peripheral information (da Costa, Kornemann, Goebel, & Senden, 2023). A baseline of no transformation ("Full") was also included **(Fig. 1a)**. The transformed features were concatenated across maps and a 100-component PCA was applied. For each condition, a linear encoding model was fitted on the train split and prediction performance was computed on the test split (Pearson correlation). For encoding models on other datasets, we used an ImageNet-trained Alexnet and code and default train/test splits from Brain-Score and Algonauts.

Results

Compared to baseline, the GCS transform improves predictions of EEG responses for the entire ERP time course (see example electrodes in **(Fig. 1b)**). A whole-scalp comparison between baseline and GCS **(Fig. 1c)** shows significant improvements across a large swath of posterior electrodes.

The GCS transform also outperforms models using center and periphery crops, which themselves enhance encoding model performance relative to baseline: center-crops show better predictions than full for posterior Iz while periphery-crops outperform full and center-crop at central Pz for the full time course **(Fig. 1b)**. The periphery-crop model also appears to show an earlier rise and peak than center-crop **(Fig. 1b)**.

To better understand how including, excluding, or reweighting central and peripheral information predicts EEG responses at specific time points, we computed partial correlations for our different encoding model predictions using another model's predictions as a covariate **(Fig. 2)**. Partial correlations of center-crop encoding compared to periphery-crop and vice versa **(Fig. 2a)** show a drastic difference in timing, with periphery-crop uniquely explaining early variance and center-crop uniquely explaining later variance.

Importantly, the GCS transform captures both the early peripherally-dominated and late centrally-dominated processing in the EEG signal **(Fig. 2b-c)**. The center-crop predictions (orange line) explain almost no unique variance when using GCS predictions as covariate, while the opposite test (black line) shows that GCS predictions explain substantially more variance than center-crop information **(Fig. 2b)**. A similar pattern holds when comparing periphery-crop (blue line) against GCS predictions and vice versa **(Fig. 2c)**. Together, these results show that the GCS transform captures both dynamics using a single, biologically-inspired transformation.

Given the clear improvement resulting from applying the GCS transform on DNN predictions in our EEG data, we also investigated if it improves DNN predictions on two publicly available neural benchmarks **(Fig. 3)**. We find clear benefits in a subset of brain regions for both datasets, with strongest effects in monkey V1 and V4 and in human V3 and hV4.

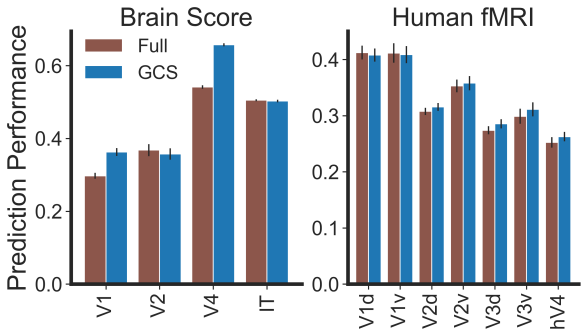

Figure 3: Comparison of Full vs GCS on monkey electrophysiology (Brain-Score) and human fMRI (Algonauts).

Discussion

Our findings show that inductive biases about retinal sampling can be operationalized to reweight DNN features to improve encoding models of primate neural responses.

The high temporal resolution of EEG recordings revealed unique contributions of central and peripheral visual information at distinct time points. These findings support models of visual perception that suggest fast/early holistic scene processing that is largely informed by peripheral information followed by slow/late object-centered processing of foveal information (Rosenholtz, Huang, Raj, Balas, & Ilie, 2012; Larson & Loschky, 2009; Deza & Konkle, 2020). Both dynamics are unified in the retinally-inspired GCS transform.

Importantly, we also tested the effect of applying GCS on input images directly, which shows similar improvements. At the same time, the transform shifts the input out-of-distribution, compared to the images used during training, thereby potentially impacting task performance and decreasing behavioural alignment. Applying GCS on DNN features instead, yields stronger improvements of predictions while preserving the input distribution and task performance.

It is remarkable that a simple spatial reweighting of features yields prediction improvements across multiple modalities. The differing stimulus sizes across datasets yield varying amounts of foveal and peripheral stimulation. Further studies are needed to investigate the dependency on large field stimulation for an improvement of GCS for predicting neural data.

Conclusion

Without the need for re-training or fine-tuning, applying GCS during feature extraction is a simple transformation that better captures the dynamics of primate visual processing.

Acknowledgments

This work is supported by the Interdisciplinary PhD Programme of University of Amsterdam Data Science Center.

References

Conwell, C., Prince, J. S., Kay, K. N., Alvarez, G. A., & Konkle, T. (2022). What can 1.8 billion regressions tell us about the pressures shaping high-level visual representation in brains and machines? *BioRxiv*, 2022–03.

Cowey, A., & Rolls, E. (1974). Human cortical magnification factor and its relation to visual acuity. *Experimental Brain Research*, *21*, 447–454.

da Costa, D., Kornemann, L., Goebel, R., & Senden, M. (2023). Unlocking the secrets of the primate visual cortex: A cnn-based approach traces the origins of major organizational principles to retinal sampling. *bioRxiv*, 2023–04.

Deza, A., & Konkle, T. (2020). Emergent properties of foveated perceptual systems. *arXiv preprint arXiv:2006.07991*.

Gifford, A. T., Dwivedi, K., Roig, G., & Cichy, R. M. (2022). A large and rich eeg dataset for modeling human visual object recognition. *NeuroImage*, *264*, 119754.

Gifford, A. T., Lahner, B., Saba-Sadiya, S., Vilas, M. G., Lascelles, A., Oliva, A., ... Cichy, R. M. (2023). The algonauts project 2023 challenge: How the human brain makes sense of natural scenes. *arXiv preprint arXiv:2301.03198*.

Kreiman, G. (2021). *Biological and computer vision*. Cambridge University Press.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, *25*.

Larson, A. M., & Loschky, L. C. (2009). The contributions of central versus peripheral vision to scene gist recognition. *Journal of vision*, *9*(10), 6–6.

Lindsey, J., Ocko, S. A., Ganguli, S., & Deny, S. (2019). A unified theory of early visual representations from retina to cortex through anatomically constrained deep cnns. *arXiv preprint arXiv:1901.00945*.

Oyster, C. W., Takahashi, E. S., & Hurst, D. C. (1981). Density, soma size, and regional distribution of rabbit retinal ganglion cells. *Journal of Neuroscience*, *1*(12), 1331–1346.

Rosenholtz, R., Huang, J., Raj, A., Balas, B. J., & Ilie, L. (2012). A summary statistic representation in peripheral vision explains visual search. *Journal of vision*, *12*(4), 14–14.

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.

Seeliger, K., Fritsche, M., Güçlü, U., Schoenmakers, S., Schoffelen, J.-M., Bosch, S. E., & Van Gerven, M. (2018). Convolutional neural network-based encoding and decoding of visual object recognition in space and time. *NeuroImage*, *180*, 253–266.

Storrs, K. R., Kietzmann, T. C., Walther, A., Mehrer, J., & Kriegeskorte, N. (2021). Diverse deep neural networks all predict human inferior temporal cortex well, after training and fitting. *Journal of cognitive neuroscience*, *33*(10), 2044–2064.

Yamins, D. L., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., & DiCarlo, J. J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the national academy of sciences*, *111*(23), 8619–8624.