Models trained on infant views are more predictive of infant visual cortex

Cliona O'Doherty (odoherc1@tcd.ie)
Trinity College Institute of Neuroscience
Trinity College Dublin, Ireland

Áine T. Dineen
Trinity College Institute of Neuroscience
Trinity College Dublin, Ireland

Anna Truzzi
School of Psychology
Queen's University Belfast, U.K.

Graham King
Trinity College Institute of Neuroscience
Trinity College Dublin, Ireland

Enna-Louise D'Arcy
Trinity College Institute of Neuroscience
Trinity College Dublin, Ireland

Chiara Caldinelli
Trinity College Institute of Neuroscience
Trinity College Dublin, Ireland

Tamrin Holloway
Trinity College Institute of Neuroscience
Trinity College Dublin, Ireland

Eleanor Molloy
Paediatrics and Child Health, Trinity College Dublin
The Coombe Hospital
Children's Health Ireland at Crumlin

Rhodri Cusack (cusackrh@tcd.ie)
Trinity College Institute of Neuroscience
Trinity College Dublin, Ireland

Abstract

The perspective of a developing infant offers unique potential when training a neural network. Egocentric video from a young child can provide ample data for representation learning in vision and language models, to only some expense of model performance. It is known that pre-trained DNNs optimised for object classification are good models of the ventral visual stream in adults, but would the same be true prior to the onset of classification behaviour? Here, we explore whether models trained on infant views are more predictive of category responses in infant ventrotemporal cortex (VTC). Using awake fMRI in a large cohort of 2-month-olds, we find that - unlike adults - features from neural networks pre-trained on infant headcam data are better models of infant VVC.

Introduction

There is a noteable gap in the amount of data needed to train a modern artificial neural network compared to a child (Frank, 2023; Cusack, Ranzato, & Charvet, 2024). However, recent work has shown that relatively standard self-supervised models can learn from headcam data of young children (Orhan, Gupta, & Lake, 2020), and digital twin studies with newborn chicks suggest that vision transformers (ViT) don't necessarily need large quantities of data to perform well (Pandey, Wood, & Wood, 2024). Even with headcam recordings from only 1% of a single toddler's experience, Vong et al. (2024) report efficient learning of word-image mappings. The intuition driving this approach is that infants generate the optimal perspectives to facilitate their learning (Bambach, Crandall, Smith, & Yu, 2018) as they try to solve the task of attaching words to preverbal categories (Pomiechowska & Gliga, 2019). Studies of the adult ventral stream have used these kinds of vision models, typically trained with computer vision datasets like ImageNet. Model predictions of adult visual responses increase when optimised for classification performance (Yamins et al., 2014), but is the same true in a human that can't yet name the things they see? We predicted that a model trained on infant-like sensory input would learn features that are more predictive of visual responses in the developing brain.

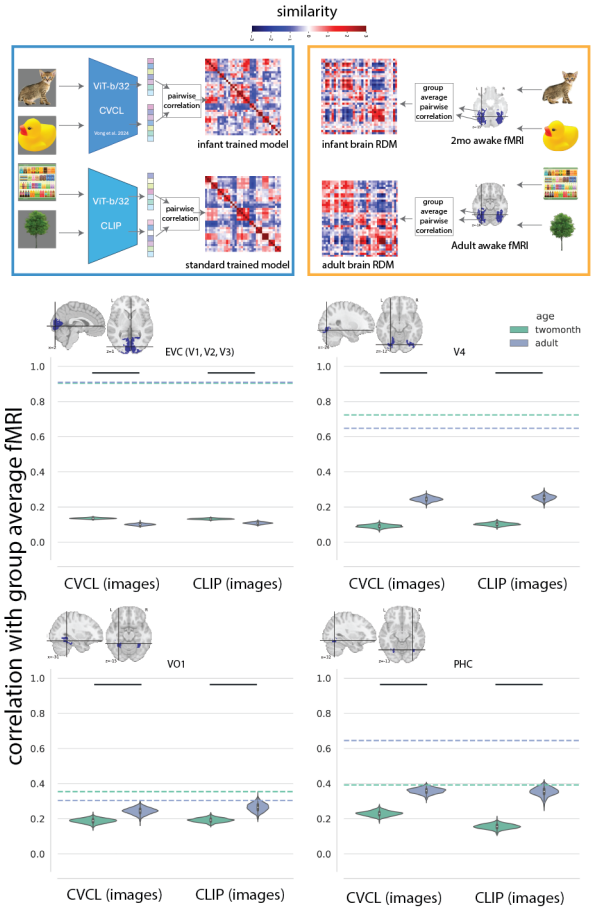

Figure 1: RDMs were obtained from two models and fMRI responses in infants and adults, using the same set of stimuli. CVCL was trained on infant views (Vong et al., 2024) and CLIP was trained using web-based data (Radford et al., 2021). Example RDMs are calculated from an aggregated VVC ROI. Violin plots show across-subject bootstrap distributions of Spearman's r between model and ROI RDMs within each age group. Dashed lines indicate the fMRI noise ceiling. Horizontal black lines denote significant differences across the age groups.

Methods

Awake infant fMRI

Awake fMRI was acquired from 112 2-month-old infants as they viewed images of 12 animate and inanimate categories. Each of the 12 categories had 3 exemplars, across diverse viewpoints. Pictures appeared in pseudo-random order against a black background for 3 s followed by a fixation cross, with a jittered inter-stimulus interval ranging between 3.5 – 4.5s. To maintain infants' engagement, images started at half of their final size and loomed towards them. Images were shown twice per 5 min functional run, with most infants participating for 10 min of awake scanning. Scans that were deemed unusable by the attending researcher were excluded, as well as scans with $>$1.5 mm framewise displacement (FWD) (85% of runs at 2-months had median FWD $<$1.5 mm). Runs with greater than 50% of rejected scans above 1.5 mm were not included in analyses. This resulted in a final sample of n=101 infants in the 2-month group (mean CGA=2.46 mo, 37 female). A dataset of healthy adults (n=17) viewing the same images was acquired for comparison. The BOLD response to each image was estimated with a general linear model, censoring high-motion frames. RDMs were calculated using correlations of voxelwise betas, across pairs of subjects. ROIs were defined using regions in the Julich atlas (Amunts, Mohlberg, Bludau, & Zilles, 2020) that overlap with domain-specific regions in VTC (Weiner et al., 2017).

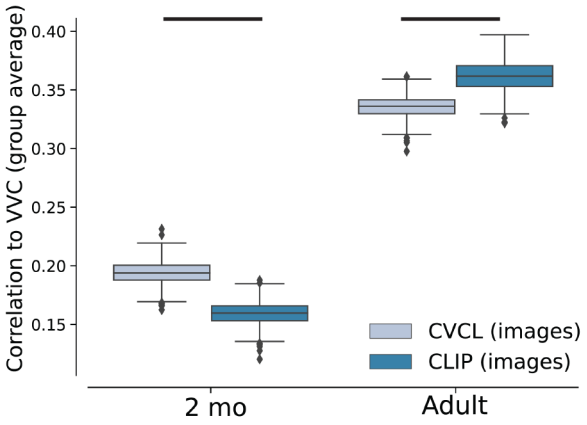

Figure 2: Bootstrap distributions (across subjects) of Spearman's r between pretrained models' feature embeddings and VVC of infants and adults. Horizontal bars indicate significant differences between brain-model correlations for the infant trained model, CVCL, and standard CLIP. [fMRI noise ceilings infants: 0.705, adults: 0.808].

Deep neural network modelling

Vision models were loaded with pretrained weights made publicly available from CLIP (Radford et al., 2021) and CVCL (Vong et al., 2024). For details of model training, please see original work. Briefly, both models were trained on a multimodal contrastive task, but differed critically in the training data. CLIP was input with 400 million image text-pairs from the web, whereas CVCL was trained on 600,000 images and 37,500 transcriptions from egocentric headcam data of a single child (Sullivan, Mei, Perfors, Wojcik, & Frank, 2021). This study focused on the vision encoders. Two ViT-b/32 architectures were loaded with the pretrained weights. 512-dimensional features were extracted from the output layer for our 36 fMRI stimuli, and pairwise correlations were calculated. This produced two model RDMs to which we could compare the brain: one with features driven by infant views, and another learned with typical training data for deep learning.

Representational similarity analysis

Within each age group and ROI, a group mean RDM was calculated using all pairwise combinations of subjects. Bootstrap distributions (1000 bootstraps across pairs of subjects) were used to calculate test statistics. To measure the bound on model performance, the split-half noise ceiling of the fMRI data was calculated using the Spearman-Brown prophecy formula (Lage-Castellanos, Valente, Formisano, & De Martino, 2019). A follow up searchlight analysis was conducted, using brain and model RDMs calculated using rsatoolbox (`https://github.com/rsagroup/rsatoolbox`). Searchlights of 5 voxel radius were calculated across the visual network (Schaefer et al., 2018) within each individual subject, and voxels were thresholded to correlations above the 95th percentile.

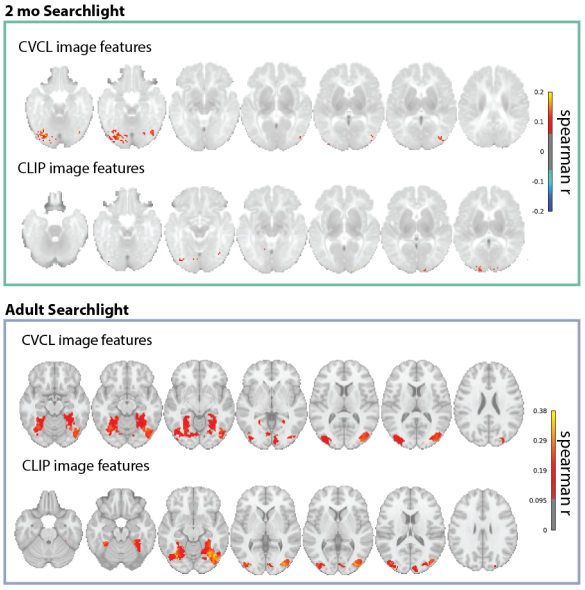

Figure 3: Visual network searchlight analysis for infants and adults. Searchlights were conducted in individual subjects, and the group average is shown. Searchlight RDMs used correlation distance as the metric, and Spearman's r for model/brain comparisons.

Results

We found that 2-month-old infants had distinct visual representations for object categories that were surprisingly mature (mean correlation of infants to adult group in EVC: 0.788 and VVC: 0.577). Individual ROIs were correlated to both models, with adults showing more similarity to the ViT in general (Fig.1). The exception to this was EVC, where infant representations corresponded more to both models. However, EVC would not be expected to show strong similarity to the output layer of a DNN (Güçlü & Van Gerven, 2015).

Fig. 2 displays the same brain-model correlations, emphasising instead the differences across CVCL and CLIP within a larger VVC region. We found that features learned by a model trained on infant views (Vong et al., 2024) were significantly more predictive of 2-month-old infants' VVC than a typically trained model. In adults, the inverse was true: features from CLIP trained on a standard dataset were more predictive of adult visual responses. When using these same models in a searchlight analysis, we recovered clusters of significant voxels in infant VVC for only the model trained on infant views, despite both models selectively predicting voxels in adult VVC.

These findings reveal that training an artificial neural network with infant egocentric views is important for modelling the infant brain. CVCL's more developmentally realistic task resulted in features that are relevant for infant learning. Although standard computer vision models are useful for adult studies, taking a developmental approach to machine learning (Smith & Slone, 2017; Zaadnoordijk, Besold, & Cusack, 2022) may be the way forward for studies of the developing mind.

Acknowledgments

This work was funded by the ERC Advanced Grant ERC-2017-ADG, FOUNDCOG, 787981 and Irish Research Council grant GOIPG/2021/223. Thank you to Vong et al. for making their weights available for use. We would like to thank Mr. Sojo Joseph, resident radiographer at Trinity College Institute of Neuroscience, for running the infant scans, and all contributing members of the FOUNDCOG scanning team: Jessica White, Anna Kravchenko, Claire Ambre, Katie Herbert, Anisha Wadhwa, Angela T. Byrne, Ailbhe Tarrant and Adrienne Foran. Finally, thank you to all the FOUNDCOG caregivers and infants who so generously dedicated their time to the study, without which this work would not be possible.

References

Amunts, K., Mohlberg, H., Bludau, S., & Zilles, K. (2020). Julich-brain: A 3d probabilistic atlas of the human brain's cytoarchitecture. *Science*, *369*(6506), 988–992.

Bambach, S., Crandall, D., Smith, L., & Yu, C. (2018). Toddler-inspired visual object learning. *Advances in neural information processing systems*, *31*.

Cusack, R., Ranzato, M., & Charvet, C. J. (2024). Helpless infants are learning a foundation model. *Trends in Cognitive Sciences*.

Frank, M. C. (2023). Bridging the data gap between children and large language models. *Trends in Cognitive Sciences*.

Güçlü, U., & Van Gerven, M. A. (2015). Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream. *Journal of Neuroscience*, *35*(27), 10005–10014.

Lage-Castellanos, A., Valente, G., Formisano, E., & De Martino, F. (2019). Methods for computing the maximum performance of computational models of fmri responses. *PLoS computational biology*, *15*(3), e1006397.

Orhan, E., Gupta, V., & Lake, B. M. (2020). Self-supervised learning through the eyes of a child. *Advances in Neural Information Processing Systems*, *33*, 9960–9971.

Pandey, L., Wood, S., & Wood, J. (2024). Are vision transformers more data hungry than newborn visual systems? *Advances in Neural Information Processing Systems*, *36*.

Pomiechowska, B., & Gliga, T. (2019). Lexical acquisition through category matching: 12-month-old infants associate words to visual categories. *Psychological Science*, *30*(2), 288–299.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., ... others (2021). Learning transferable visual models from natural language supervision. In *International conference on machine learning* (pp. 8748–8763).

Schaefer, A., Kong, R., Gordon, E. M., Laumann, T. O., Zuo, X.-N., Holmes, A. J., ... Yeo, B. T. (2018). Local-global parcellation of the human cerebral cortex from intrinsic functional connectivity mri. *Cerebral cortex*, *28*(9), 3095–3114.

Smith, L. B., & Slone, L. K. (2017). A developmental approach to machine learning? *Frontiers in psychology*, *8*, 296143.

Sullivan, J., Mei, M., Perfors, A., Wojcik, E., & Frank, M. C. (2021). Saycam: A large, longitudinal audiovisual dataset recorded from the infant's perspective. *Open mind*, *5*, 20–29.

Vong, W. K., Wang, W., Orhan, A. E., & Lake, B. M. (2024). Grounded language acquisition through the eyes and ears of a single child. *Science*, *383*(6682), 504–511.

Weiner, K. S., Barnett, M. A., Lorenz, S., Caspers, J., Stigliani, A., Amunts, K., ... Grill-Spector, K. (2017). The cytoarchitecture of domain-specific regions in human high-level visual cortex. *Cerebral cortex*, *27*(1), 146–161.

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.

Zaadnoordijk, L., Besold, T. R., & Cusack, R. (2022). Lessons from infant learning for unsupervised machine learning. *Nature Machine Intelligence*, *4*(6), 510–520.