Deep Neural Network Models of Infant Visual Cortex

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

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

Anna Truzzi
Trinity College Institute of Neuroscience
Trinity College Dublin

Graham King
Trinity College Institute of Neuroscience
Trinity College Dublin

Lorijn Zaadnoordijk
Trinity College Institute of Neuroscience
Trinity College Dublin

Enna-Louse D'Arcy
Trinity College Institute of Neuroscience
Trinity College Dublin

Jessica White
Trinity College Institute of Neuroscience
Trinity College Dublin

Keelin Harrison
Trinity College Institute of Neuroscience
Trinity College Dublin

Chiara Caldinelli
Trinity College Institute of Neuroscience
Trinity College Dublin

Tamrin Holloway
Trinity College Institute of Neuroscience
Trinity College Dublin

Anna Kravchenko
Trinity College Institute of Neuroscience
Trinity College Dublin

Ailbhe Tarrant
The Rotunda Hospital
Children's Health Ireland at Temple Street

Angela T. Byrne
The Coombe Hospital
Chilren's Health Ireland at Crumlin

Adrienne Foran
The Rotunda Hospital
Children's Health Ireland at Temple Street

Eleanor J. 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

Abstract

Deep convolutional neural networks (DNNs) are now cemented as effective computational models in adult visual neuroscience. However, comparing the *learning* human brain to the *learning* models had not yet been possible due to the difficulty in collecting sufficient neuroimaging data from infants. To address this, we conducted longitudinal fMRI on 2-month-old infants (n=130), and again at 9-months-old (n=65), while they were awake and viewing a variety of visual stimuli. Multivariate pattern analysis (MVPA) revealed a complex representational structure in visual cortex already at 2-months. We show that fully-trained DNNs capture a significant proportion of this structure, and different learning algorithms can determine the developmental stage that a DNN best explains.

Keywords: visual representations; deep neural networks; ventral stream; fMRI; development

Introduction

DNNs are now commonly used as models of the adult ventral stream (Richards et al., 2019; Yamins et al., 2014; Zhuang et al., 2021), but even more than for adult vision research, they offer the potential for novel insight into visual development. There is considerable value in having a mechanistic model for a learning process isn't directly accessible, as infants cannot partake in typical cognitive experiments or report their experiences. Moreover, the parallel between infant and machine learning is of increasing interest for researchers in both fields (Zaadnoordijk, Besold, & Cusack, 2022; Smith & Slone, 2017). Recent work shows that data from an infant's perspective can efficiently train even large language models (Pandey, Wood, & Wood, 2024) and provides the necessary structure to learn word-visual referents (Vong, Wang, Orhan, & Lake, 2024). Our work brings this approach a step earlier in the developmental process, using DNNs to characterise the visual representations to which words are attached, and extending the neuroconnectionist (Doerig et al., 2023) research framework into developmental neuroscience.

Methods and Results

Awake infant fMRI

We acquired a large, longitudinal neuroimaging dataset of awake infants. Functional MRI (fMRI) was acquired from 2-month-old infants (n=130), and again when they were 9-months-old (n=65), while watching 12 categories of common objects. Each category had three exemplars across diverse viewpoints, totalling 36 images, and most infants participated for four repetitions of each stimulus to give 144 pictures across 10 minutes of scanning. The distribution of head motion was acceptable for infants (85% of runs had a median framewise displacement of less than 1.5mm at 2-months, and 97% at 9-months) and allowed for rigorous scrubbing (final sample included in MVPA n=103 2-month-olds and n=38 9-month-olds). A cohort of adults (n=17) was acquired for comparison. The BOLD response to each object exemplar was estimated with a generalised linear model and representational similarity analysis measured the representational geometry of early visual cortex (EVC) and ventrotemporal cortex (VTC). We observed distinct structure in both EVC and VTC from 2-months onwards (Fig. 1A,B).

Deep neural network modelling

Having successfully characterised the visual representations in infant visual cortex, we tested whether DNNs could model the developing brain. Analyses were written in Python v3.8 with PyTorch v2.0.1 and CUDA 11.7. We tested an untrained network with randomly initiated weights as well as two fully-trained networks. All models used AlexNet as the architectural backbone. Fully trained models were implemented with learning algorithms previously used in DNN models of adult VTC (Khaligh-Razavi & Kriegeskorte, 2014; Konkle & Alvarez, 2022). The supervised model used the default weights in torchvision v0.15.2, learned through an ImageNet recognition task. The self-supervised model was Instance Protocol Contrastive Learning, which learns from natural image structure by contrasting augmented versions of an image to an average prototype in the embedding space. Each of the three DNNs were presented with the 36 images from the fMRI study, and the activations used to construct an RDM for each layer.

Pre-trained DNNs model infant visual responses As infants receive tens-of-thousand times fewer visual samples than a pre-trained DNN (Frank, 2023), we expected that a model earlier in its training would better explain the infant visual cortex. Infant representations did correlate more to an untrained DNN than adult representations did at both 2-months and 9-months. However, the fully-trained network outperformed the untrained model across all age groups (Fig. 1C,G). Contrary to our expectations, visual input played a significant role in modelling the visual cortex from as early as 2-months. Despite less exposure to the world, infant visual representations are sophisticated enough to align well with those from a fully-trained neural network. This alignment increased with development, coupled with a decrease in correlation to the untrained network, demonstrating the continued influence of visual input as we age.

Learning algorithm modulates DNN-infant similarity Supervised learning is unlikely for preverbal infants, whereas self-supervised learning aligns better with infants' sensitivity to comparisons and patterns within the stream of sensory input (Fiser & Aslin, 2002). We found that infants correlated more than adults with the self-supervised DNN only in EVC and in shallow layers (Fig. 1F). We did not observe the expected hierarchical correspondence between early layers and adult EVC (Güçlü & Van Gerven, 2015), possibly due to the large cortical area covered by our ROIs. Nonetheless, the significantly higher correlations of infant EVC to early layers was not observed in the supervised model, revealing that the choice of learning algorithm can define the developmental stage that a DNN best explains. The influence of learning algorithm is

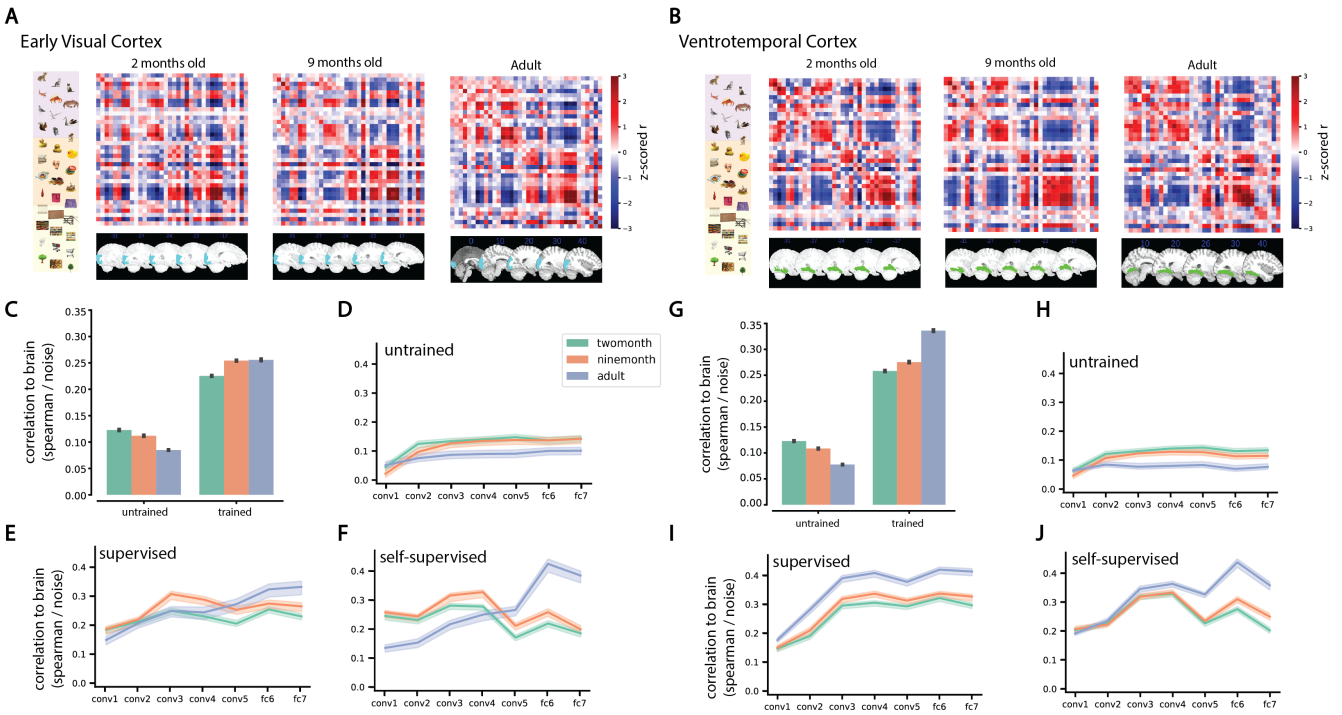

Figure 1: **A,B** Visual representations in EVC and VTC. Axes are the 12 categories spanning animate/inanimate classes, with 3 exemplars per category. **C,G** Spearman correlation, adjusted by the MRI noise ceiling, of the visual representations to untrained and fully-trained DNNs. **D-F, H-J** Layerwise spearman correlations to DNNs trained with different learning algorithms. Error bars/bands are the 95% confidence interval (CI) calculated using bootstrap resampling across pairs of subjects.

again evident in VTC. Adults correlated more than infants with every layer of the supervised model (Fig. 1I), but all age groups showed similar correlations to the early layers of the self-supervised DNN (Fig. 1J).

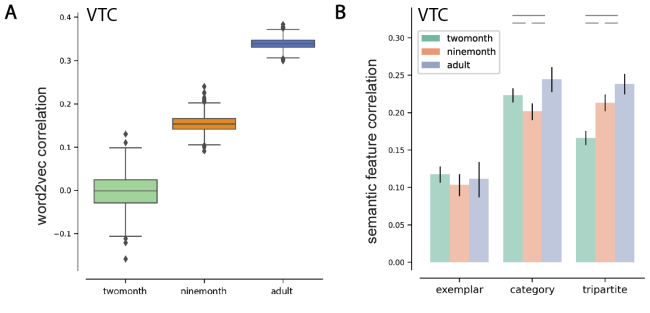

Figure 2: **A** Bootstrap distributions of VTC correlations to word2vec embeddings for the 12 categories tested. **B** Partial correlation of VTC representations to semantic models, controlling for the perceptual features size, elongation, colour and compactness (Spriet et al., 2022). Exemplar: tests for distinct responses to each image. Category: tests for generalisation across exemplars to form a category. Tripartite: tests for tripartite organisation by animate, inanimate small and inanimate big. Error bars: 95% CI calculated with bootstrapping across pairs of subjects.

Infant VTC contains semantic distinctions To determine the complexity of the structure in infant VTC, we examined correspondences to word-based models of semantic similarity. We obtained word2vec embeddings (Mikolov, Sutskever, Chen, Corrado, & Dean, 2013) for each of the 12 category labels, and used a GLM at the 12 category level to estimate category responses in the brain. At 2-months, we found no evidence for lexical semantics in VTC. However, this had strengthened by 9-months (Fig. 2A), a time when infants recognise a few nouns and begin associating words with visual categories (Bergelson & Swingley, 2012; Pomiechowska & Gliga, 2019), just prior to the development of speaking.

Is the structure in infant VTC entirely explained by perceptual features? Distinct patterns were evoked by different exemplars [Spearman correlation to an identity matrix 2-months: $\rho = 0.342$, CI=(0.333, 0.349); 9-months: $\rho = 0.351$, CI=(0.343, 0.359); Adults: $\rho = 0.336$, CI=(0.327, 0.345)]. VTC representations in all age-groups were correlated to a model that tested for generalisation across images to form a category, even when controlling for perceptual similarity (Fig. 2B). The tripartite organisation (Konkle & Caramazza, 2013) was present in VTC at 2-months, significantly increasing in its representational strength by 9-months and again into adulthood (Fig. 2B). This reveals that semantic distinctions, defined by category membership and learnable through visual input, are present from early infancy while the influence of labels semantic structure in VTC increases with age.

Acknowledgments

This work was funded by the ERC Advanced Grant ERC-2017-ADG, FOUNDCOG, 787981 and Irish Research Council grant GOIPG/2021/223. 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. 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

Bergelson, E., & Swingley, D. (2012). At 6–9 months, human infants know the meanings of many common nouns. *Proceedings of the National Academy of Sciences*, *109*(9), 3253–3258.

Doerig, A., Sommers, R. P., Seeliger, K., Richards, B., Ismael, J., Lindsay, G. W., ... others (2023). The neuroconnectionist research programme. *Nature Reviews Neuroscience*, *24*(7), 431–450.

Fiser, J., & Aslin, R. N. (2002). Statistical learning of new visual feature combinations by infants. *Proceedings of the National Academy of Sciences*, *99*(24), 15822–15826.

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.

Khaligh-Razavi, S.-M., & Kriegeskorte, N. (2014). Deep supervised, but not unsupervised, models may explain it cortical representation. *PLoS computational biology*, *10*(11), e1003915.

Konkle, T., & Alvarez, G. A. (2022). A self-supervised domain-general learning framework for human ventral stream representation. *Nature communications*, *13*(1), 491.

Konkle, T., & Caramazza, A. (2013). Tripartite organization of the ventral stream by animacy and object size. *Journal of Neuroscience*, *33*(25), 10235–10242.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, *26*.

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.

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.

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

Spriet, C., Abassi, E., Hochmann, J.-R., & Papeo, L. (2022). Visual object categorization in infancy. *Proceedings of the National Academy of Sciences*, *119*(8), e2105866119.

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.

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.

Zhuang, C., Yan, S., Nayebi, A., Schrimpf, M., Frank, M. C., DiCarlo, J. J., & Yamins, D. L. (2021). Unsupervised neural network models of the ventral visual stream. *Proceedings of the National Academy of Sciences*, *118*(3), e2014196118.