Recurrent models optimized for face recognition exhibit dynamic representational signatures resembling the primate brain

Hossein Adeli[1] {ha2366@columbia.edu}, Nikolaus Kriegeskorte[1]
[1]Zuckerman Mind Brain Behavior Institute, Columbia University, New York, USA

Abstract

Understanding the dynamics of neural representations is crucial for elucidating the mechanisms of visual recognition in the primate brain. Here we investigate the representational dynamics of recurrent convolutional neural networks (RCNNs) optimized for face-identification and object-recognition tasks. Using representational similarity analysis (RSA), we observed that only models that were trained for face identification showed a late-emerging prominent distinction of identities as seen in the monkey face patch AM. Interestingly, early model responses (to a diverse set of images including human faces, monkey faces, and non-face objects) strongly separated the objects from faces. Our results also show that models that were trained simultaneously on both face identification and object recognition were more likely to show the signature of mirror symmetric viewpoint tuning in their intermediate representations as has been reported for monkey face patch AL. These findings suggest that the dynamics of face recognition that emerges in a hierarchical recurrent neural network prioritizes category-level recognition at early stages, triggering category-specific computations that enable individual-level recognition.

Keywords: Face perception; Recurrent processing; Deep Neural Networks

Introduction

Neurons with selectivity for different categories are found in higher areas of the primate ventral visual pathway. A well characterized network of ventral cortical areas are the face patches, where neurons respond more to faces than non-faces (Hesse & Tsao, 2020). The face patches form a coarse hierarchy of areas that become progressively more identity-selective and view-invariant (Freiwald & Tsao, 2010). A fundamental ongoing debate in the field is whether (a) features for all categories span a more domain-general multivariate representational space with different categories forming clusters in that space (Vinken, Prince, Konkle, & Livingstone, 2023) or (b) category-selective cells are specialized for detecting features associated with a specific category (Dobs, Yuan, Martinez, & Kanwisher, 2023) (Note that (a) and (b) are not mutually exclusive, but (b) makes a strong claim that certain neurons contribute exclusively or primarily to the representation of certain categories.)

These two views can be combined into a more dynamic view of category processing (Sugase, Yamane, Ueno, & Kawano, 1999; Shi et al., 2023) where the representational geometry evolves from face detection (separating faces from non-faces) to face identification (separating different face identities). In this work we set to examine the plausibility of this view in recurrent convolutional neural networks (RCNNs).

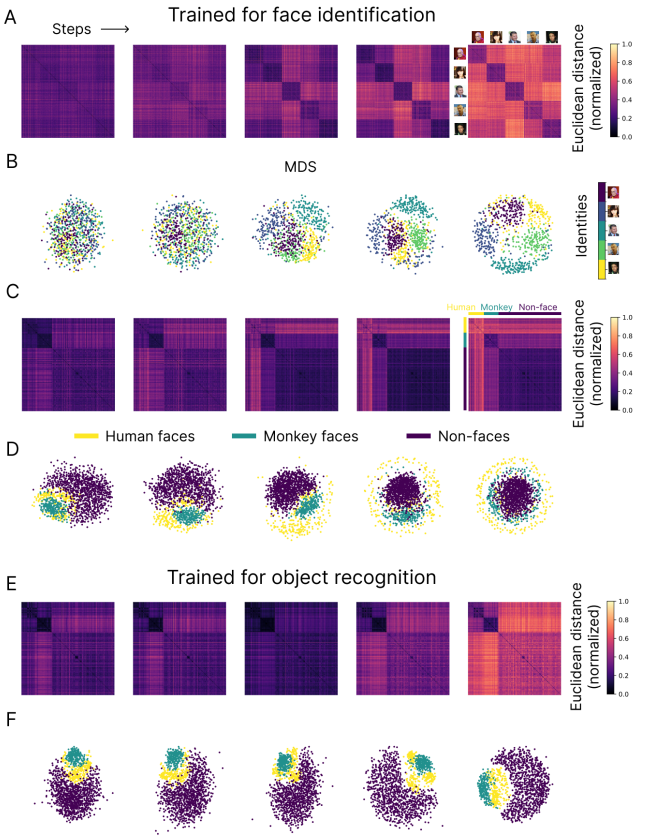

Figure 1: **A)** RDMs for five identities with 150 images for each and **B)** MDS showing face-identity selectivity resembling face patch AM. **C)** RDMs for the same model on a dataset including human faces, monkey faces, and non-faces with **D)** MDS showed differentiation among face identities emerging only late in the process. **E)** and **F)** Models trained on object recognition did not exhibit the late response.

Feedforward DNNs correspond in their feature selectivity across layers to the brain areas in the hierarchy of the ventral pathway (V1-V2-V4-IT) (Khaligh-Razavi & Kriegeskorte, 2014; Yamins et al., 2014). However these models cannot capture the dynamics of neural response within each layer due to a lack of recurrent connections (Kietzmann et al., 2019). In order to study the dynamic signatures of recognition, we test RCNNs trained with different objectives.

Methods

We built four-layer fully convolutional neural networks inspired by both the BL model of (Spoerer, McClure, & Kriegeskorte, 2017) and the CORNet-RT model (Kubilius et al., 2018). The four convolutional layers coarsely correspond to the stages in the ventral pathway. Each network layer receives both bottom-up (B) input and lateral (L) input representing the state of the layer at the previous time step. All layers go through 8 steps of recurrence. The computational graph follows the principle of biological unrolling (Kubilius et al., 2018; Spoerer, Kietzmann, Mehrer, Charest, & Kriegeskorte, 2020), so the input reaches the final layer after 3 steps of processing. Global average pooling is applied to the convolutional maps in the last layer which is then linearly read out for recognition. We trained the model for face identification (using the VGGFace2 dataset), object-category recognition (using the Imagenet dataset), or a combination of both tasks. We matched the total number of training images across the three training conditions.

Results

First, using representational similarity analysis, we found that models trained on face recognition, but not models trained on general object recognition, when tested on a held-out set, show face-identity selectivity resembling primate face patch AM (Fig. 1)(Freiwald & Tsao, 2010). Visualization of the representational space using multidimensional scaling (MDS, metric stress objective) further showed the emergence of identity clusters over timesteps. Next we tested the same models on a dataset including human faces, monkey faces, and non-face objects (Vinken et al., 2023). The representational dissimilarity matrices (RDMs, Euclidean distance) showed distinctions among human face identities emerging only late in the process. In earlier steps, the representational geometry separates the objects from the faces, and in later steps the differences among face identities come to be prevalently represented.

Next we tested whether mirror symmetric viewpoint tuning as observed in monkey face patch AL (Freiwald & Tsao, 2010) is present in these models. Fig. 2A left shows a synthetic RDM for such a response. We selected 25 identities from FEI dataset (do Amaral, Fígaro-Garcia, Gattas, & Thomaz, 2016) and clustered them by face orientation. Faces that have the same orientation (e.g. 90 deg) or mirror symmetric orientation (-90 deg) would have smaller dissimilarity, hence darker colors in the RDM. Fig. 2A right shows the expected geometry if the neurons are purely identity selective. We looked at the RDMs for two models, one trained only for face identification (Fig. 2B) and another trained on both face identification and object recognition (Fig. 2C). For both models the identity selectivity becomes more prominent in higher layers and later steps with the model trained only for face identification showing a stronger effect. The model trained for both face identification and object recognition, however, shows a stronger signature of mirror symmetric response in its intermediate representation. Models trained for only object recognition do not show either of these signature patterns (not visualized).

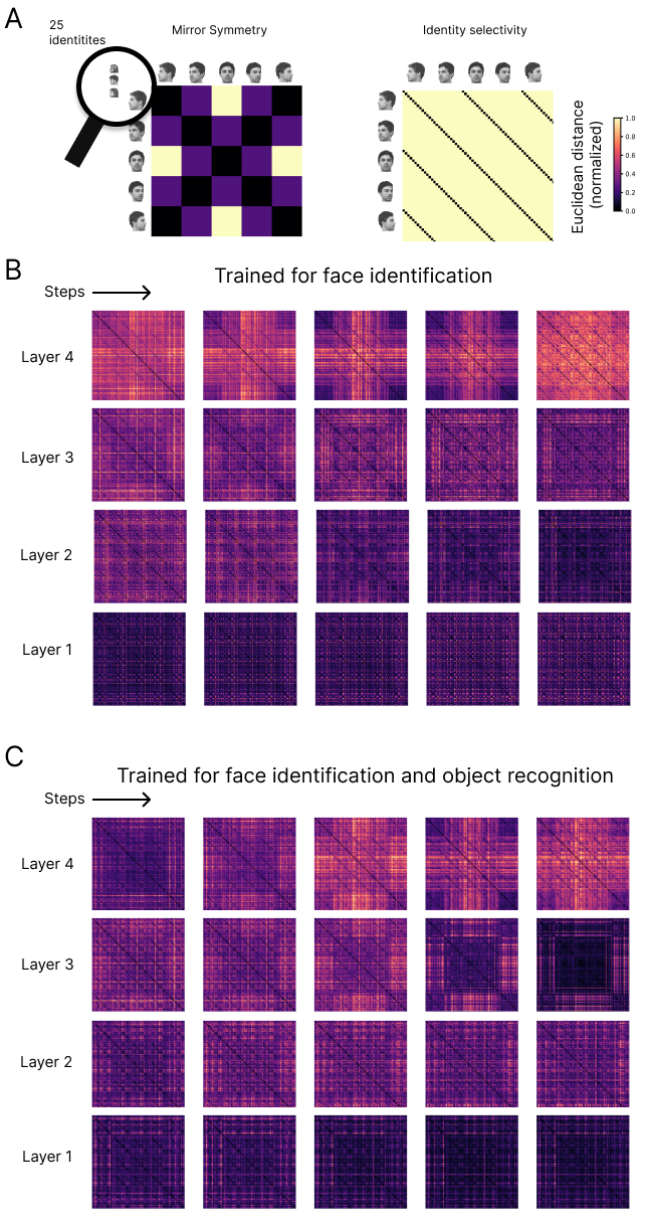

Figure 2: **A)** Synthetic RDMs showing representation geometry for mirror symmetric and identity selective cells. Faces from 25 different identities are clustered by the their orientation. **B)** RDMS over 5 timesteps for the 4 layers of a network trained for face identification. **C)** RDMs for a network trained for both face identification and and object recognition.

Discussion

We found that recurrent convolutional models trained on both face recognition and object-category recognition show a dynamic response that first emphasizes categorical distinctions and later individuates faces. This suggests that early domain-general processing establishes the category of the objects and provides the basis for engagement of domain-specific computations supporting identification (Sugase et al., 1999; Freiwald & Tsao, 2010; Kriegeskorte, Formisano, Sorger, & Goebel, 2007).

Acknowledgments

Research reported in this publication was supported in part by the National Institute of Neurological Disorders and Stroke of the National Institutes of Health under award number [RF1NS128897]. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

References

do Amaral, V., Fígaro-Garcia, C., Gattas, G. J. F., & Thomaz, C. E. (2016). Normalização espacial de imagens frontais de face em ambientes controlados e não-controlados. *FaSCi-Tech*, *1*(1).

Dobs, K., Yuan, J., Martinez, J., & Kanwisher, N. (2023). Behavioral signatures of face perception emerge in deep neural networks optimized for face recognition. *Proceedings of the National Academy of Sciences*, *120*(32), e2220642120.

Freiwald, W. A., & Tsao, D. Y. (2010). Functional compartmentalization and viewpoint generalization within the macaque face-processing system. *Science*, *330*(6005), 845–851.

Hesse, J. K., & Tsao, D. Y. (2020). The macaque face patch system: a turtle's underbelly for the brain. *Nature Reviews Neuroscience*, *21*(12), 695–716.

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

Kietzmann, T. C., Spoerer, C. J., Sörensen, L. K., Cichy, R. M., Hauk, O., & Kriegeskorte, N. (2019). Recurrence is required to capture the representational dynamics of the human visual system. *Proceedings of the National Academy of Sciences*, *116*(43), 21854–21863.

Kriegeskorte, N., Formisano, E., Sorger, B., & Goebel, R. (2007). Individual faces elicit distinct response patterns in human anterior temporal cortex. *Proceedings of the National Academy of Sciences*, *104*(51), 20600–20605.

Kubilius, J., Schrimpf, M., Nayebi, A., Bear, D., Yamins, D. L., & DiCarlo, J. J. (2018). Cornet: Modeling the neural mechanisms of core object recognition. *BioRxiv*, 408385.

Shi, Y., Bi, D., Hesse, J. K., Lanfranchi, F. F., Chen, S., & Tsao, D. Y. (2023). Rapid, concerted switching of the neural code in inferotemporal cortex. *bioRxiv*, 2023–12.

Spoerer, C. J., Kietzmann, T. C., Mehrer, J., Charest, I., & Kriegeskorte, N. (2020). Recurrent neural networks can explain flexible trading of speed and accuracy in biological vision. *PLoS computational biology*, *16*(10), e1008215.

Spoerer, C. J., McClure, P., & Kriegeskorte, N. (2017). Recurrent convolutional neural networks: a better model of biological object recognition. *Frontiers in psychology*, *8*, 278016.

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), 869–873.

Vinken, K., Prince, J. S., Konkle, T., & Livingstone, M. S. (2023). The neural code for "face cells" is not face-specific. *Science Advances*, *9*(35), eadg1736.

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.