Constraining vision models to predict image memorability yields significant gains in producing more brain-aligned models of the primate ventral stream

Ram Ahuja (ramahuja@my.yorku.ca)
Department of Biology, York University
Toronto, Ontario, M3J1P3, Canada

Soroush Ziaee (soroush1@yorku.ca)
Department of Biology, York University
Toronto, Ontario, M3J1P3, Canada

Ezgi Fide (ezgifide@yorku.ca)
Department of Psychology, York University
Toronto, Ontario, M3J1P3, Canada

Shayna Rosenbaum (shaynar@yorku.ca)
Department of Psychology, York University
Toronto, Ontario, M3J1P3, Canada

Kohitij Kar (k0h1t1j@yorku.ca)
Department of Biology, York University
Toronto, Ontario, M3J1P3, Canada

Abstract

The primate ventral visual stream that culminates in the inferior temporal (IT) cortex supports critical functions, including object recognition and visual memory. Previous work has demonstrated that artificial neural networks (ANNs) optimized for object categorization exhibit unprecedented but partial alignment with ventral stream representations. However, it remains unknown whether ANNs constrained to predict human image memorability could explain a unique part of the neural variance – potentially bridging the remaining explanatory gap. We observed that models trained to predict image memorability predict unique variances of the IT neural responses. Interestingly, joint categorization and memorability training yielded networks that captured significantly more variance in neural responses than models trained on either objective alone. Our results suggest that incorporating diverse, functionally relevant objectives leads to ANNs more closely aligned with the primate ventral visual stream's representational geometry and functional properties.

Introduction

The IT cortex, which lies at the apex of the ventral visual pathway, has been implicated in high-level visual representations that support core object recognition (Hung, Kreiman, Poggio, & DiCarlo, 2005; Majaj, Hong, Solomon, & DiCarlo, 2015). However, object perception is deeply intertwined with other visual functions like image memorability, whereby some images are more memorable than others. Prior studies suggest IT cortex population activity magnitudes are predictive of image memorability (Jaegle et al., 2019), hinting at this area's potential role in object and memory encoding. Recently, specific ANNs (**Figure 1A**) have emerged as promising models of the primate ventral stream (Yamins & DiCarlo, 2016; Kar & DiCarlo, 2023). When trained on object recognition tasks, hierarchical layers of ANNs exhibit response patterns similar to neurons in corresponding cortical hierarchy (Yamins et al., 2014). This partial brain-ANN alignment has generated excitement for leveraging ANNs as encoding models of the visual cortex. Building on this, we hypothesized that explicitly training ANNs to predict image memorability (similar to (Khosla, Raju, Torralba, & Oliva, 2015)), in addition to object categorization might constrain the training procedure to develop internal representations that are more similar to the brain. To test this, we compared neural recordings from macaque IT with representations in the ANNs trained on 1) object categorization, 2) memorability prediction, and 3) a joint objective of both tasks. Our results show that ANNs explain unique variance in macaque IT neurons when constrained to predict memorability. This suggests this behavioral objective plays an important role in shaping IT representations and should be considered in future ventral stream models.

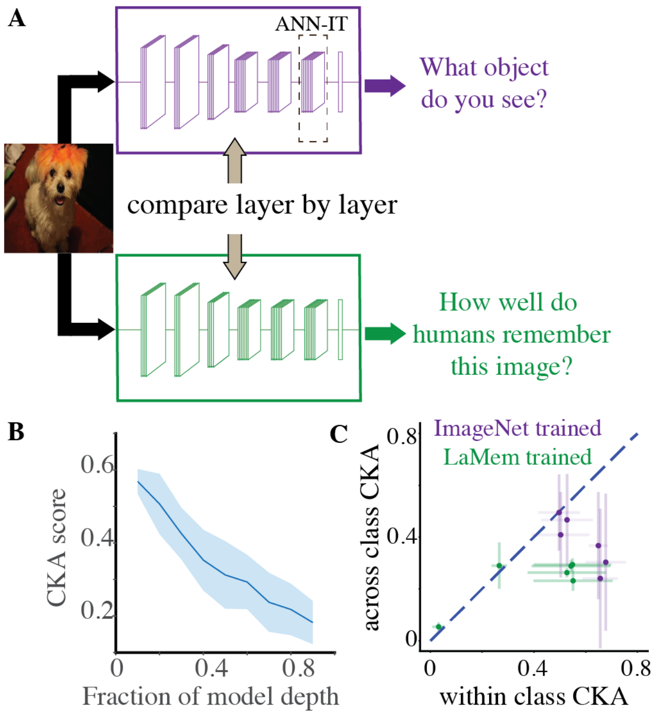

Figure 1: **A.** ANNs can be trained to perform object categorization and memorability prediction. Do these models develop similar internal representations? **B.** A layer-by-layer comparison of ANNs trained explicitly on the two objectives reveals a progressive difference in representation based on CKA scores. Error bar denotes s.e.m across 5 ANN architectures. **C.** ANN-IT of the models trained with object categorization are similar to each other than ANN-IT of models trained on memorability (paired t-test, $p<0.05$).

Results

ANNs trained to only predict image memorability explain significant IT response variance

As a first step, we explicitly trained ANNs (architectures: e.g., AlexNet (Krizhevsky, Sutskever, & Hinton, 2012), ResNets (He, Zhang, Ren, & Sun, 2016) etc.) to predict memorability scores of images, using the Lamem dataset (Khosla et al., 2015). We asked whether these ANNs can explain any amount of variance in the neural response patterns of the macaque IT cortex (Kar, Kubilius, Schmidt, Issa, & DiCarlo, 2019) using linear regression methods ((Kar et al., 2019; Yamins et al., 2014)). We observed that the activations of IT-like layers, as retrieved from Brain-Score (Schrimpf et al., 2018), were able to significantly predict the firing rates of individual IT neurons (**Figure 2A,C - y-axis values significantly** $>$**0**). Interestingly, the Lamem-trained networks better predicted IT neurons than ImageNet-trained ANNs (**Figure 2A**), only for the early phase of the IT responses. While LaMem-trained ANNs still predicted significant variance of the late-phase IT responses, they showed lower neural predictivity than the ImageNet-trained ANNs (**Figure 2C**).

LaMem-trained ANNs develop different internal representations than ImageNet-trained ANNs

Having established that memorability-trained ANNs can explain some IT neural variance, we next asked how their internal representations compare to standard object recognition-trained ANNs. Despite having identical architectures, models trained on memorability develop markedly different representational geometries (compared within and across architectures), as revealed by representational similarity analysis, specifically Centered Kernel Alignment (CKA) scores (Kornblith, Norouzi, Lee, & Hinton, 2019) (**Figure 1B, C**). This suggests that the task objectives of memorability prediction vs. object categorization drive ANNs to learn distinct features and representational organizations, even when operating on the same input images. But do these lead to capturing unique variances in the IT neurons?

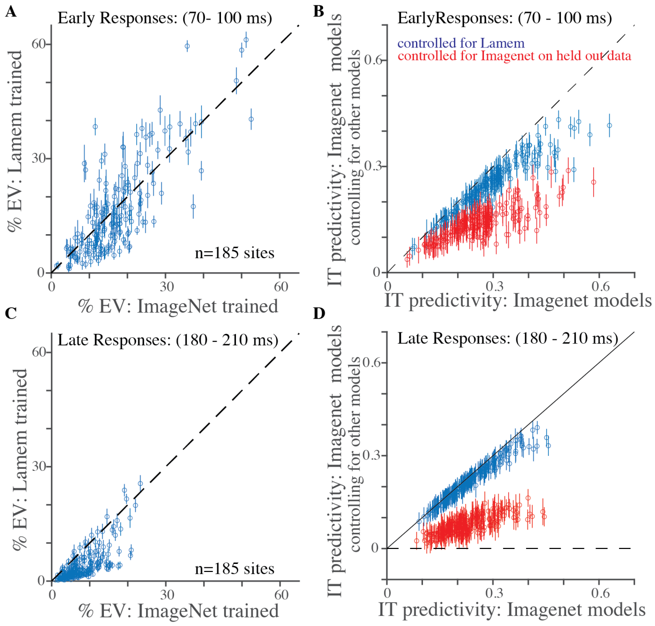

Figure 2: **A.** Comparison of % EV on early-phase (70-100 ms) IT responses between ANNs trained on LaMem (y-axis) and ImageNet (x-axis). **B.** Comparison of early-phase IT response predictivity between ANNs trained on ImageNet (x-axis) and ImageNet while controlling for LaMem-based model predictions (blue) or held-out ImageNet-trained model-based predictions (y-axis). **C.** Same as A, but for late-phase (180-210 ms) IT responses. **D.** Same as B, but for late-phase IT responses.

LaMem-trained ANNs predict unique variance of macaque IT responses compared to ImageNet-trained ANNs

We compared how well IT responses correlate with predictions from ANNs trained on ImageNet with and without controlling for the LaMem-based model predictions. This partial correlation analysis revealed that the LaMem-trained and ImageNet-trained ANNs predicted largely non-overlapping portions of the IT neural variance (**Figure 2B and D**). These results further predict that jointly optimizing models on both image memorability and object categorization (which was, in fact, what was originally implemented in MemNet (Khosla et al., 2015)) might significantly improve neural predictivity beyond ANNs trained on either one of the objectives.

ANNs trained simultaneously on LaMem and ImageNet predict more IT variance than models trained on these objectives individually

We trained an ANN (ResNet-50 architecture) to predict both memorability and object categories across images. We hypothesized that this would improve its ability to predict IT neural responses since a combined model would develop a richer representation capturing multiple ethologically relevant features encoded in the IT cortex. Interestingly, this jointly optimized ANN model was able to account for a significantly higher fraction of IT neural variance compared to models trained on either memorability or object recognition alone, but only for early phases of the IT responses (**Figure 3**).

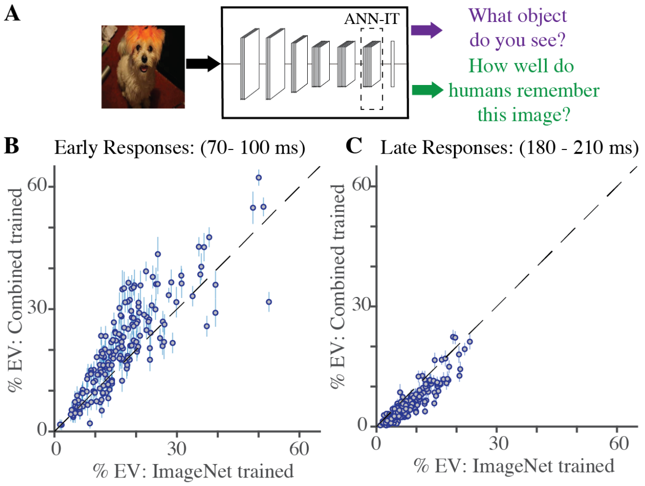

Figure 3: **A.** ANNs trained on dual objectives (image memorability and object categorization. **B.** Comparison of % EV on early-phase (70-100 ms) IT responses between ANNs trained on combined goals (y-axis) and ImageNet (x-axis). **C.** Same as B, but for late-phase IT responses.

Conclusion

Our results demonstrate that incorporating visual memorability as a training objective improves the alignment of ANNs with the macaque IT cortex, suggesting that constraints produced by memory encoding are critical factors shaping the ventral stream. Investigating the time-dependent variations observed in IT predictivity gains could reveal how IT employs a flexible, temporally multiplexed coding strategy to independently facilitate object recognition and memory formation.

Acknowledgments

KK has been supported by funds from the Canada Foundation for Innovation (CFI), the Canada Research Chair Program, the Simons Foundation Autism Research Initiative (SFARI, 967073), the Canada First Research Excellence Funds (VISTA Program), and a Google Research Award.

References

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 770–778).

Hung, C. P., Kreiman, G., Poggio, T., & DiCarlo, J. J. (2005). Fast readout of object identity from macaque inferior temporal cortex. *Science*, *310*(5749), 863–866.

Jaegle, A., Mehrpour, V., Mohsenzadeh, Y., Meyer, T., Oliva, A., & Rust, N. (2019). Population response magnitude variation in inferotemporal cortex predicts image memorability. *Elife*, *8*, e47596.

Kar, K., & DiCarlo, J. J. (2023). The quest for an integrated set of neural mechanisms underlying object recognition in primates. *arXiv preprint arXiv:2312.05956*.

Kar, K., Kubilius, J., Schmidt, K., Issa, E. B., & DiCarlo, J. J. (2019). Evidence that recurrent circuits are critical to the ventral stream's execution of core object recognition behavior. *Nature neuroscience*, *22*(6), 974–983.

Khosla, A., Raju, A. S., Torralba, A., & Oliva, A. (2015). Understanding and predicting image memorability at a large scale. In *Proceedings of the ieee international conference on computer vision* (pp. 2390–2398).

Kornblith, S., Norouzi, M., Lee, H., & Hinton, G. (2019). Similarity of neural network representations revisited. In *International conference on machine learning* (pp. 3519–3529).

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

Majaj, N. J., Hong, H., Solomon, E. A., & DiCarlo, J. J. (2015). Simple learned weighted sums of inferior temporal neuronal firing rates accurately predict human core object recognition performance. *Journal of Neuroscience*, *35*(39), 13402–13418.

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.

Yamins, D. L., & DiCarlo, J. J. (2016). Using goal-driven deep learning models to understand sensory cortex. *Nature neuroscience*, *19*(3), 356–365.

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.