Beyond feedforward: Leveraging discrepancies between humans and convolutional neural networks reveals recurrent dynamics during object recognition

Pablo Oyarzo (pablo.oyarzo@fu-berlin.de)
Department of Education and Psychology, Freie Universität Berlin
Berlin, Germany

Johannes J.D. Singer (johannes.singer@arcor.de)
Department of Education and Psychology, Freie Universität Berlin
Berlin, Germany

Kohitij Kar (k0h1t1j@yorku.ca)
Department of Biology, Centre for Vision Research, York University
Toronto, Canada

Radoslaw M. Cichy (rmcichy@zedat.fu-berlin.de)
Department of Education and Psychology, Freie Universität Berlin
Berlin, Germany

Abstract

Convolutional neural networks (CNNs) have emerged as leading models for primate object recognition, yet humans often outperform them, revealing misalignments with human behavior and brain responses. This discrepancy indicates unique brain-specific computations engaged when object recognition is challenging. Here, we leverage this gap to identify the human neural mechanisms driving these computations. Specifically, we compared EEG and fMRI responses to images on which a feedforward CNN (AlexNet) and humans perform on par versus images on which the CNN performs worse. We find that for images where the CNN performs worse, humans show delayed information processing and the specific recruitment of frontal brain areas, suggesting the involvement of additional top-down recurrent computations. These results pinpoint the neural mechanisms beyond feedforward processing engaged for robust object perception when vision is challenging.

Keywords: object recognition; deep neural networks; multivariate pattern analysis; EEG; fMRI.

Introduction

Primates can rapidly and accurately recognize objects (DiCarlo, Zoccolan, & Rust, 2012) despite varying viewing conditions. Currently, the best family of models of biological vision are CNNs (Cichy, Khosla, Pantazis, Torralba, & Oliva, 2016; Cichy & Kaiser, 2019; Schrimpf et al., 2018; Yamins et al., 2014), yet they are far from being fully aligned with human behavioral and brain measures (Geirhos, Meding, & Wichmann, 2020; Rajalingham et al., 2018; Wichmann & Geirhos, 2023). Further, even seemingly identical behavior might be driven by vastly different mechanisms in CNNs and primate brains (Kar, Kubilius, Schmidt, Issa, & DiCarlo, 2019; Kar & DiCarlo, 2021; Kietzmann et al., 2019; Kreiman & Serre, 2020). However, such instances of model-brain misalignment present the opportunity to experimentally probe the engagement of specific mechanisms in human brains that current models lack, allowing for a more precise characterization of human object recognition. Here, we compare how the human brain processes objects in response to image stimuli that are behaviorally aligned versus misaligned between a CNN and humans, following the approach developed by Kar et al. (2019). Specifically, we selected two sets of images: one on which the recognition performance of both systems was comparable (control images), and another on which humans largely outperformed the CNN (challenge images). We then measured EEG and fMRI responses to both image sets. We used multivariate pattern analysis to compare object information in the brain across time and space for the two image sets. Our results revealed a delayed emergence of object information in EEG sensor space and engagement of additional frontal brain regions, specifically in the case of challenge images. This additional time might be due to the recurrent processing of information involving top-down interactions between frontal regions and the sensory cortices.

Methods

Stimuli

We used a set of 1320 images, each containing one of 10 objects (Kar et al., 2019). Each image was either a photograph (from MS COCO; (Lin et al., 2014)) or contained a synthetic object rendered with a combination of transformations (i.e., size, rotation, and position) on a natural background.

Behavioral experiment

We used behavioral data collected from 88 participants performing a binary object discrimination task on Amazon Mechanical Turk (Kar et al., 2019). In each trial, the target image appeared for 100 ms, followed by a 100 ms blank screen. Then, a response screen showed a canonical version of the target object alongside an alternative distractor object. The subjects' task was to correctly identify the target object.

Selection of challenge and control images

We compared participants' image-by-image performance with AlexNet (Krizhevsky, Sutskever, & Hinton, 2012) to determine the challenge and control images. First, we identified challenge images as those for which humans exceeded AlexNet's performance by more than 1.5 d', and control images as those for which the absolute difference in performance did not exceed 0.4 d' (Fig. 1A). Next, we created a smaller, final stimulus set for the EEG and fMRI experiments. For this, we matched each challenge image with a control image that best matched the human behavioral scores. This process yielded two sets of 121 images each, with no significant difference in human performance between them (u(120)=8155, p=0.13) but a significant difference in CNN performance (u(120)=41, $p<0.001$).

EEG experiment

35 participants (23.5 ± 4.1 years old, 28 female) took part in the EEG study. We recorded brain responses to each image in both the challenge and control sets using 64 channels (68 repetitions per image). We presented images using a rapid serial visual presentation paradigm: trains of 14 images were presented in random order (200 ms on, 100 ms off). The task was to report after each image train whether a catch image (a paper clip) was presented (probability 0.42).

fMRI experiment

31 participants (27 ± 4.8 years old, 21 female) took part in the fMRI study. As in the EEG, we recorded brain responses to all images (8-10 repetitions each). On each trial, we presented an image for 500 ms, followed by a 2500 ms blank screen. The task was to respond to a change in fixation cross color interspersed between the main trials (probability 0.25).

Statistical analysis

We conducted multivariate pattern analysis (Haxby, Connolly, & Guntupalli, 2014) on both EEG and fMRI data to investigate the temporal and spatial dynamics of object information, respectively. To assess the significance of the identified patterns, we performed sign-flipping tests (10,000 iterations). We

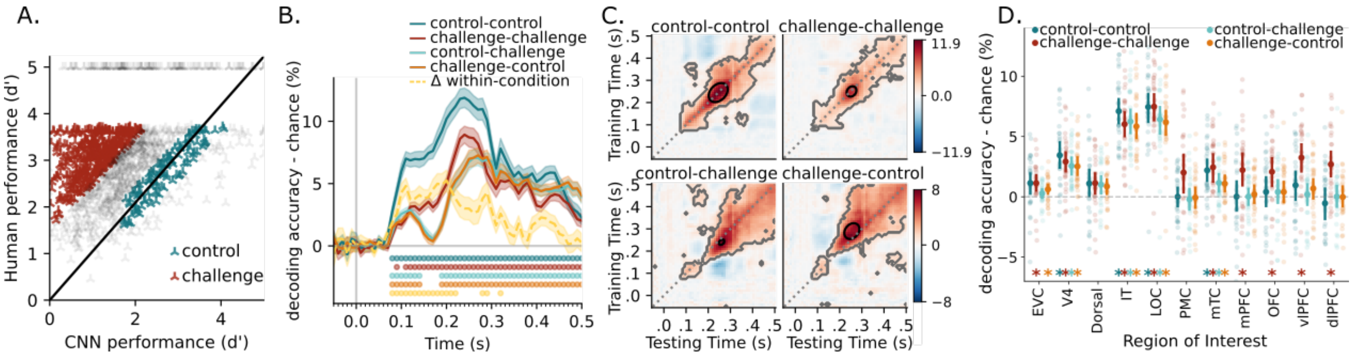

Figure 1: **Main results.** A) Comparison of Human and AlexNet's performance. B) Time course of object decoding. C) Temporal generalization of decoders. D) Object decoding in different brain regions.

adjusted all results for multiple comparisons using Bonferroni correction.

Results

EEG analysis reveals that object representations are delayed for challenge vs. control images.

To assess the temporal dynamics with which visual representations emerge we performed a decoding analysis of the EEG data. We tested cross-validated support vector machine (SVM) classifiers to discriminate the 10 objects within a temporal window of -50 ms to +500 ms around stimulus onset with a bin size of 5 ms.

We first conducted the decoding analyses separately for the challenge and control images and compared the results. Decoders trained and tested on the same time points (Fig. 1B, blue and red lines) accurately identified objects in both sets starting at 90 ms after stimulus onset (all ps$<$.001). However, the time courses also differed significantly from 90 to 220 ms (yellow line, all ps$<$.001), indicating differences in processing.

Next, we cross-decoded across stimulus sets. This was possible in a similar time frame as within-set decoding, except between 140-190 ms (Fig. 1B, orange and cyan lines), revealing a divergent pattern of object processing during this intermediate interval.

Finally, we performed temporal generalization analyses (King & Dehaene, 2014) by training and testing classifiers at all possible time point combinations. In both challenge and control sets, within-set decoding revealed symmetric patterns predominantly along the diagonal, as expected. In addition, limited off-diagonal significant generalization indicates a highly dynamic sequence of object representations (Fig. 1C, above). In contrast, across-set decoding revealed an asymmetrical pattern: decoders trained on control images generalized better to later time points when tested on challenge images (Fig. 1C, lower-left; peak at 240 ms training time, 265 ms testing time). This pattern is reversed when cross-validation is done in the opposite direction (Figure 1C, lower right; peak at 275 ms training time, 250 ms testing time). This effect is detectable starting around 150 ms after stimulus presentation. In both cases, peak decoding was located off-diagonal (p$<$.05). This shows that similar object representations emerge later for the challenge images than for the control images.

fMRI analysis reveals object information in frontal areas only for challenge images.

To determine how object representations differ for challenge versus control images spatially, we analyzed fMRI data focusing on 11 key areas from the HCP atlas (Glasser et al., 2016), including visual and frontal regions. Akin to the EEG analysis, we used within and across-set decoding to analyze the data. We found significant object decoding within and across both image sets in the visual ventral stream areas (V4, IT, LOC) and the medial temporal cortex (all ps$<$.001), as expected. However, in frontal regions (mPFC, OFC, vlPFC and dlPFC) object information was only present for challenge images (all ps$<$.001), and cross-set decoding was not possible (Fig. 1D, all ps$>$.05). These results suggest that frontal brain regions are recruited for object recognition when processing challenge images.

Conclusion

By utilizing the misalignment in object recognition performance between humans and a baseline feedforward CNN model, we identified stimuli sets that are processed differently by the brain. Our findings support four key conclusions. First, the differences in information processing between these sets reflect the engagement of distinct brain mechanisms characterized by unique temporal dynamics and the recruitment of specific brain regions. Second, these mechanisms may involve recurrent computations for challenging images, as processing diverges at intermediate stages and later converges into similar representations, albeit with a delay. Third, in such cases, information processing extends beyond the ventral visual stream, including frontal regions, suggesting the involvement of long-range feedback. Finally, the brain demonstrates flexibility in engaging these mechanisms, depending on the sufficiency of feedforward processing alone to accomplish object recognition tasks.

Acknowledgements

PO was supported by the Scholarship Program of the National Agency for Research and Development (ANID; 2019-72200281). KK was supported by the Canada Research Chair Program. RMC was supported by The German Research Council grants (CI241/1-1, CI241/3-1, CI241/7-1, INST 272/297-1) and the European Research Council grant (ERC-StG-2018-803370).

References

Cichy, R. M., & Kaiser, D. (2019). Deep neural networks as scientific models. *Trends in cognitive sciences*, *23*(4), 305–317.

Cichy, R. M., Khosla, A., Pantazis, D., Torralba, A., & Oliva, A. (2016). Comparison of deep neural networks to spatio-temporal cortical dynamics of human visual object recognition reveals hierarchical correspondence. *Scientific reports*, *6*(1), 27755.

DiCarlo, J. J., Zoccolan, D., & Rust, N. C. (2012). How does the brain solve visual object recognition? *Neuron*, *73*(3), 415–434.

Geirhos, R., Meding, K., & Wichmann, F. A. (2020). Beyond accuracy: quantifying trial-by-trial behaviour of cnns and humans by measuring error consistency. *Advances in Neural Information Processing Systems*, *33*, 13890–13902.

Glasser, M. F., Coalson, T. S., Robinson, E. C., Hacker, C. D., Harwell, J., Yacoub, E., … others (2016). A multi-modal parcellation of human cerebral cortex. *Nature*, *536*(7615), 171–178.

Haxby, J. V., Connolly, A. C., & Guntupalli, J. S. (2014). Decoding neural representational spaces using multivariate pattern analysis. *Annual review of neuroscience*, *37*, 435–456.

Kar, K., & DiCarlo, J. J. (2021). Fast recurrent processing via ventrolateral prefrontal cortex is needed by the primate ventral stream for robust core visual object recognition. *Neuron*, *109*(1), 164–176.

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.

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.

King, J.-R., & Dehaene, S. (2014). Characterizing the dynamics of mental representations: the temporal generalization method. *Trends in cognitive sciences*, *18*(4), 203–210.

Kreiman, G., & Serre, T. (2020). Beyond the feedforward sweep: feedback computations in the visual cortex. *Annals of the New York Academy of Sciences*, *1464*(1), 222–241.

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

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., … Zitnick, C. L. (2014). Microsoft coco: Common objects in context. In *Computer vision–eccv 2014: 13th european conference, zurich, switzerland, september 6-12, 2014, proceedings, part v 13* (pp. 740–755).

Rajalingham, R., Issa, E. B., Bashivan, P., Kar, K., Schmidt, K., & DiCarlo, J. J. (2018). Large-scale, high-resolution comparison of the core visual object recognition behavior of humans, monkeys, and state-of-the-art deep artificial neural networks. *Journal of Neuroscience*, *38*(33), 7255–7269.

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.

Wichmann, F. A., & Geirhos, R. (2023). Are deep neural networks adequate behavioral models of human visual perception? *Annual Review of Vision Science*, *9*, 501–524.

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.