Contrasting computational models of task-dependent readout from the ventral visual stream

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

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

Tim C. Kietzmann$^\top$ (tim.kietzmann@uni-osnabrueck.de)
Institute of Cognitive Science, University of Osnabrück
Osnabrück, Germany

Sushrut Thorat$^\top$ (sthorat@uni-osnabrueck.de)
Institute of Cognitive Science, University of Osnabrück
Osnabrück, Germany

$^\top$co-senior authors.

Abstract

Humans categorize visual information based on features of different complexity. While some tasks require high-level information, others rely on low-level cues. This raises the question of how these features, originating from different parts of the visual system, are integrated for perceptual decisions. Here, we test three potential mechanisms: single readout, direct access, and attentional routing. The mechanisms were implemented and contrasted based on ANN models, equipped with different readout strategies. These networks were trained to perform two tasks that require access to low or high-level visual features, respectively, and subsequently evaluated in terms of performance and their ability to predict human reaction times of participants performing the same tasks. We found that the direct access and attentional routing models performed as well as humans in both tasks, while the single readout model did not perform well. Importantly, the attentional routing model best predicted human reaction times overall. These results indicate that neither a readout only from high-level visual cortex nor direct access to upstream regions might be sufficient to explain human categorization behavior across tasks, and suggest attentional modulations along the ventral visual stream as a critical mechanism that enables flexible readout through high-level visual cortex.

Keywords: visual categorization; perceptual readout; deep neural networks; attentional modulation

Introduction

Human categorization behavior is based on visual features of different complexity, represented at distinct hierarchical stages along the ventral visual stream (Grill-Spector & Weiner, 2014; Op de Beeck, Haushofer, & Kanwisher, 2008). According to task demands, different feature representations must be accessed and transformed into behavior. What computational mechanism enables such flexible readout of task-relevant feature representations in the ventral visual stream?

One account posits that a linear readout from the final processing stage in the visual system, the inferior temporal (IT) cortex, is sufficient to account for categorization behavior across various tasks (Cohen, Alvarez, Nakayama, & Konkle, 2017; Majaj, Hong, Solomon, & DiCarlo, 2015). This view is in line with work showing that low-level visual features such as color, rotation, and pose can be decoded from IT in addition to category information (Hong, Yamins, Majaj, & DiCarlo, 2016).

In contrast, behaviorally relevant feature representations have been identified in distinct stages along the ventral visual stream (Contier, Baker, & Hebart, 2023; Singer, Karapetian, Hebart, & Cichy, 2023; Yeh, Thorat, & Peelen, 2024), suggesting that readout might directly access multiple stages. This view is supported by computational evidence for the functional relevance of readouts from earlier stages in the ventral stream before IT cortex (Birman & Gardner, 2019; Jagadeesh & Gardner, 2021).

A third possibility is that visual information is not directly accessed from early visual regions, but is locally modulated via attention (Gilbert & Li, 2013; Thorat, Aldegheri, van Gerven, & Peelen, 2019), and this altered code is passed on for subsequent readout from IT. Theories of feature-based attention, as well as recent advances in incorporating top-down attentional mechanisms in deep neural network models, support the view that attention may be a key component of visual categorization (Konkle & Alvarez, 2024; Lindsay & Miller, 2018; Thorat, van Gerven, & Peelen, 2019).

We contrasted these three readout hypotheses by expressing them in artificial neural network (ANN) models. We trained the models on two tasks requiring access to either low- or high-level visual features. To compare the models, we assessed their task performance and ability to predict the behavioral responses of humans performing the same tasks.

Methods

Neural network architectures and training.

We developed three ANN architectures, each expressing a different hypothesis about the readout from the ventral visual stream (Fig. 1): 1) readout only from IT (i.e. Single Readout), 2) direct access to all stages in the ventral stream (i.e. Direct Access), and 3) attentional modulations with a readout from IT (i.e. Attentional Routing). All architectures share an AlexNet backbone (Krizhevsky, Sutskever, & Hinton, 2012), pre-trained on ILSVRC2012 (Russakovsky et al., 2015), with frozen backbone parameters during subsequent model training. For task-specific readout, all models feed into a linear

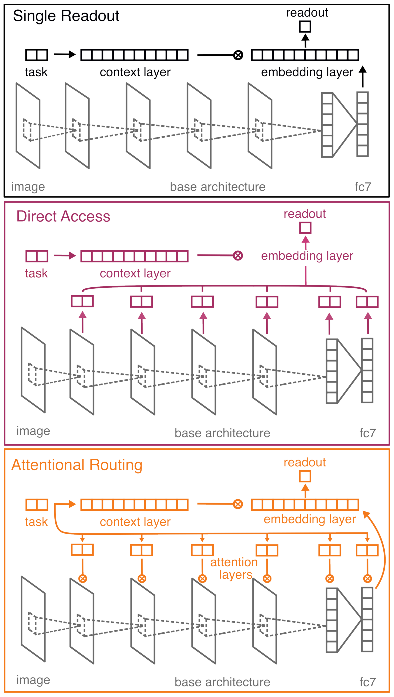

Figure 1: Neural network architectures expressing different readout mechanisms.

embedding layer which is multiplicatively modulated by a task context layer. The linear embedding layer feeds into a binary classification layer.

The Single Readout model feeds from AlexNet fc7 into the linear embedding layer. The Direct Access model projects all feature channel outputs (averaged across spatial dimensions) for each convolutional block into the linear embedding layer - each layer projects into a distinct and equal subset of the layer. In the Attentional Routing model the feedforward pass through the base architecture is modulated with multiplicative feature-based attention (attention layers in Fig. 1) based on the task (Lindsay & Miller, 2018). After the feedforward pass, the network projects from fc7 into the linear embedding layer.

The projection from the fully connected layers in AlexNet into the linear embedding layers was regularized with dropout ($p = 0.5$) for all models. The training dataset consisted of segmented objects on random texture backgrounds with colored outlines. Objects were segmented and outlined using the segmentation masks and images from MSCOCO (Lin et al., 2014) and placed on random texture backgrounds retrieved from `https://github.com/abin24/Textures-Dataset`. All models underwent interleaved training on two tasks: content classification (animate vs. inanimate) and color classification (red outline vs. blue outline). All results are averages across model instances trained with 5 random seeds.

Behavioural experiment

We collected reaction times and accuracies from 29 participants for the same tasks the networks were trained and evaluated on. In each trial, a participant was presented with one of 120 segmented object images (from the test set of the network (Fig. 2A) with colored outlines for 200ms (half animate/inanimate, half red/blue outline) and was instructed to either report if the object was inanimate/animate or had a blue/red outline in separate blocks.

Results

To contrast different hypotheses of the readout from the ventral visual stream, we compared ANN models expressing these hypotheses in terms of task performance on a high-level content task and low-level color task, and in terms of their alignment with human behavioral responses in the same tasks. Since the number of units in the embedding layers is a key hyperparameter in the models, serving as the informational bottleneck before classification, we trained all models with different sizes of the embedding layers (from 10-40 in steps of 10) to assess the robustness of the results. For the content task, we found that all models performed similarly well (pair-wise $p = 0.990$, McNemar test, FDR-corrected) and better than humans (pair-wise $p < 0.001$, one-sided t-test, FDR-corrected) across all embedding sizes. For the color task, both the Direct Access as well as the Attentional Routing models outperformed the Single Readout model (pair-wise $p < 0.006$, McNemar test, FDR-corrected) and only the Single Readout model performed worse than humans ($p < 0.001$, one-sided t-test, FDR-corrected).

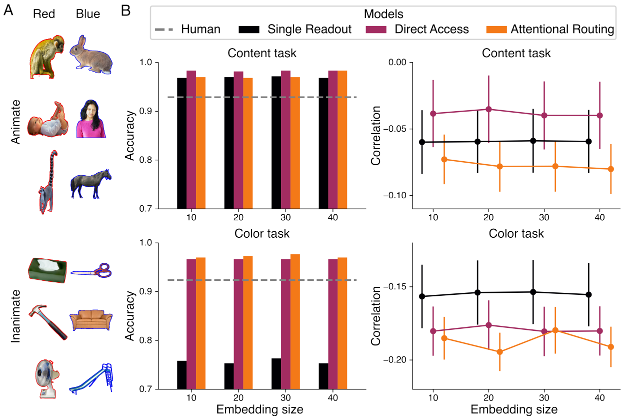

Figure 2: A) Example stimuli from the test set. B) Task performance and reaction time correlations with varying embedding layer size for different readout models. Error bars represent the standard error of the mean across 29 human participants.

Next, we compared the models in terms of their alignment with human behavior by correlating the image-specific output of the classifier layer ($p(animacy)$ or $p(color)$, depending on the task) with the corresponding reaction times for both tasks. A 3-way ANOVA with the factors “Task”, “Model” and “Embedding Size” comparing the correlations revealed a significant main effect of “Task” and of “Model” (both $p < 0.030$), but no significant two or three-way interactions between the factors (all $p > 0.077$). Therefore, we averaged correlations across embedding sizes and tasks and performed pairwise tests between models (FDR-corrected). This revealed significantly stronger negative correlations for the Attentional Routing model than the Direct Access model ($p = 0.018$), marginally stronger negative correlations for the Attentional Routing compared to the Single Readout model ($p = 0.055$), but no significant difference between the Single Readout and Direct Access models ($p = 0.882$).

Conclusion

By contrasting three potential mechanisms of task-dependent readout from the ventral visual stream, we gained two main insights. First, readout from the final stage of processing is not sufficient to account for human performance in a color categorization task that requires sensitivity to low-level visual features. Second, even though direct access to low-level visual features improves performance, it does not explain human reaction times better than readout from the final stage of processing. A model that reads out from the final stage of processing and, additionally, allows for attentional modulation of the feedforward pass performs at a human level in both tasks and explains human reaction times better than the other two models. In sum, this suggests that readout from the final processing stage is not sufficient for explaining human behavior across tasks, and that task-specific attentional routing, combined with readout from the final stage of processing, might support task-dependent human categorization behavior.

Acknowledgments

This work was supported by the German Research Council grants (CI241/1-1, CI241/3-1, CI241/7-1,INST 272/297-1) awarded to RMC, a European Research Council grant (ERC-StG-2018-803370) awarded to RMC, and a European Union grant (ERC-StG, TIME, Project 101039524) awarded to TCK.

References

Birman, D., & Gardner, J. L. (2019). A flexible readout mechanism of human sensory representations. *Nature communications*, *10*(1), 3500.

Cohen, M. A., Alvarez, G. A., Nakayama, K., & Konkle, T. (2017). Visual search for object categories is predicted by the representational architecture of high-level visual cortex. *Journal of neurophysiology*, *117*(1), 388–402.

Contier, O., Baker, C. I., & Hebart, M. N. (2023). Distributed representations of behaviorally-relevant object dimensions in the human visual system. *bioRxiv*.

Gilbert, C. D., & Li, W. (2013). Top-down influences on visual processing. *Nature Reviews Neuroscience*, *14*(5), 350–363.

Grill-Spector, K., & Weiner, K. S. (2014). The functional architecture of the ventral temporal cortex and its role in categorization. *Nature Reviews Neuroscience*, *15*(8), 536–548.

Hong, H., Yamins, D. L., Majaj, N. J., & DiCarlo, J. J. (2016). Explicit information for category-orthogonal object properties increases along the ventral stream. *Nature neuroscience*, *19*(4), 613–622.

Jagadeesh, A. V., & Gardner, J. L. (2021). V1-and it-like representations are directly accessible to human visual perception. In *Svrhm 2021 workshop@ neurips.*

Konkle, T., & Alvarez, G. (2024). Cognitive steering in deep neural networks via long-range modulatory feedback connections. *Advances in Neural Information Processing Systems*, *36*.

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).

Lindsay, G. W., & Miller, K. D. (2018). How biological attention mechanisms improve task performance in a large-scale visual system model. *ELife*, *7*, e38105.

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.

Op de Beeck, H. P., Haushofer, J., & Kanwisher, N. G. (2008). Interpreting fmri data: maps, modules and dimensions. *Nature Reviews Neuroscience*, *9*(2), 123–135.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., ... others (2015). Imagenet large scale visual recognition challenge. *International journal of computer vision*, *115*, 211–252.

Singer, J. J., Karapetian, A., Hebart, M. N., & Cichy, R. M. (2023). The link between visual representations and behavior in human scene perception. *bioRxiv*, 2023–08.

Thorat, S., Aldegheri, G., van Gerven, M. A., & Peelen, M. V. (2019). Modulation of early visual processing alleviates capacity limits in solving multiple tasks. *arXiv preprint arXiv:1907.12309*.

Thorat, S., van Gerven, M. A., & Peelen, M. V. (2019). The functional role of cue-driven feature-based feedback in object recognition. *arXiv preprint arXiv:1903.10446*.

Yeh, L.-C., Thorat, S., & Peelen, M. V. (2024). Predicting cued and oddball visual search performance from fmri, meg, and dnn neural representational similarity. *Journal of Neuroscience*, *44*(12).