Representational Geometry Dynamics in Networks After Long-Range Modulatory Feedback

Kexin Cindy Luo (kexinluo@g.harvard.edu)
Department of Psychology & Kempner Institute, Harvard University, Cambridge, MA, USA.

George A. Alvarez (alvarez@wjh.harvard.edu)
Department of Psychology & Kempner Institute, Harvard University, Cambridge, MA, USA.

Talia Konkle (talia_konkle@harvard.edu)
Department of Psychology & Kempner Institute, Harvard University, Cambridge, MA, USA.

Abstract

The human visual system employs extensive long-range feedback circuitry, where feedforward and feedback connections iteratively refine interpretations through reentrant loops (Di Lollo, 2012). Inspired by this neuroanatomy, a recent computational model incorporated long-range modulatory feedback into a convolutional neural network (Konkle & Alvarez, 2023). While this prior work focused on injecting an external goal signal to leverage feedback for category-based attention, here we investigated its default operation: how learned feedback intrinsically reshapes representational geometry without top-down goals. Analyzing activations from this model across two passes—feedforward versus modulated—on ImageNet data, we examined local (within-category) and global (between-category) structure. Our results demonstrate that feedback significantly compacts category clusters: exemplars move closer to prototypes, and the local structure improves as more near neighbors fall within the same category. Notably, this occurs while largely preserving global structure, as between-category distances remain relatively stable. An exploratory analysis linking local and global changes suggested a positive relationship between local compaction and prototype shifts. These findings reveal an emergent "prototype effect" where fixed long-range feedback automatically refines local representations, potentially enhancing categorical processing efficiency without disrupting overall representational organization. This suggests intrinsic feedback dynamics might contribute fundamentally to perceptual organization.

Introduction

Human visual perception is not merely a passive, feedforward process—it is active and dynamic, shaped by the interaction of incoming sensory inputs with higher-order cognitive processes. This active nature is supported by extensive feedback circuits that span the visual hierarchy (Gilbert & Li, 2013), including prominent pathways from high-level areas such as the inferotemporal cortex (IT) to intermediate areas like V4, and from V4 back to early visual cortex such as V1. These feedforward and feedback pathways are thought to engage in iterative, reentrant loops that progressively refine perceptual interpretations (Lamme & Roelfsema, 2000; Di Lollo, 2012).

Despite the prevalence of feedback in biological vision, most deep learning models for visual object categorization are predominantly feedforward. While achieving remarkable performance, these models often lack mechanisms analogous to biological feedback, limiting their capacity to model top-down influences such as attentional steering. Addressing this, a recent computational model introduced fixed long-range modulatory (LRM) feedback pathways into a convolutional neural network (Konkle & Alvarez, 2023). In this LRM model, feedback is implemented via a two-pass process: a feedforward pass generates initial representations from input images, and a subsequent modulated pass enables later-stage activations to influence earlier-stage representations through learned channel-to-channel connections. Although previous research has demonstrated the effectiveness of leveraging these feedback connections for externally guided, category-based attention, the default operation of this feedback circuitry remains less explored. Specifically, how does this learned, fixed feedback mechanism intrinsically reshape representations without top-down task goals?

Here, we investigate the representational dynamics induced by default feedback modulation in the LRM model, focusing on how it alters the geometry of stimulus representations. Recognizing that representational structure can be characterized at multiple scales in both humans and models (Bowman, Iwashita, & Zeithamova, 2020; Muttenthaler et al., 2023), we examine changes at both local levels (e.g., the structure within categories) and global levels (e.g., the relationships between categories) following feedback modulation.

Model & Data

We used the LRM3 model with its defined feedback connections as detailed in Konkle and Alvarez (2023) to investigate feedback-driven representational changes. LRM3 achieves a higher Top-1 ImageNet classification accuracy after feedback modulation compared to its initial feedforward pass or a vanilla AlexNet. Additionally, when evaluated on the Brain-Score benchmark (Schrimpf et al., 2020), both the feedforward and modulated LRM3 representations exhibit stronger correspondence to human neural data than those of standard AlexNet.

For our analyses[1], we extracted and compared activations from two passes: the initial feedforward pass before modulation (the baseline pass) and the modulated pass after two cycles of feedback (the modulated pass). Input stimuli comprised 300 images from each of 100 randomly selected categories from the ImageNet dataset (Deng et al., 2009).

Results

Feedback Effects on Local Geometry

First, we investigated whether feedback modulation impacts local representational geometry by altering the cluster size, quantified using the distance between exemplars and their prototypes. For each exemplar image embedding a_{ijk} (activation for image j in category i at pass k), we computed its cosine distance to the category prototype p_{ik} — defined as the mean embedding of all images within category i at pass k. We defined the distance as $d_{ijk} = 1 - \cos\_\text{sim}(a_{ijk}, p_{ik})$.

We compared these exemplar-to-prototype distances between the baseline pass and the modulated pass. A reduction in d_{ijk} following modulation would indicate that exemplars are clustering more tightly around their category prototype. A linear mixed-effects model revealed a significant decrease in these distances after feedback modulation ($M = -.036$,

[1]Code available at: https://github.com/cindyLuo99/reprGeo_LRM.

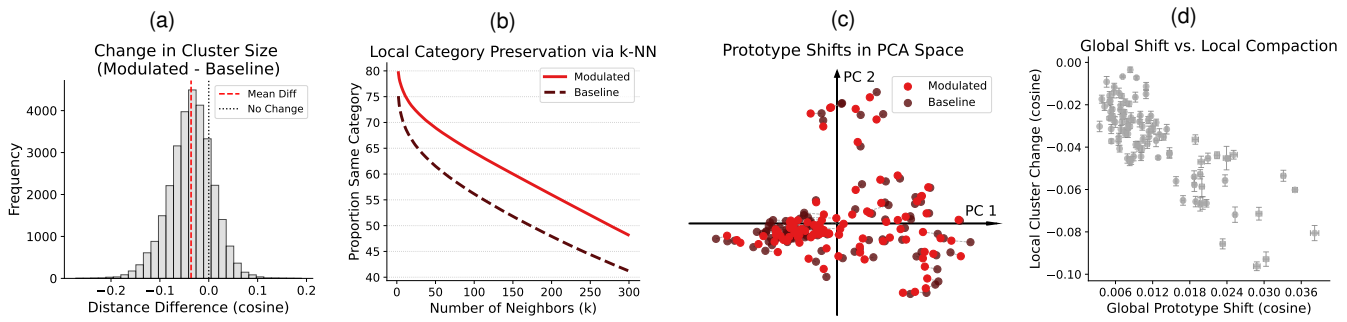

Figure 1: Effects of default feedback modulation on representational geometry. (**a**) Significant cluster shrinkage indicated by negative change in exemplar-prototype distance (Modulated - Baseline). (**b**) Increased proportion of neighbors belonging to the same category in the modulated pass, indicating improved local structure. (**c**) Category prototype shifts induced by feedback modulation, visualized in the principal component space. Baseline and modulated prototypes for each category are connected by a gray dashed line. (**d**) Correlation between global prototype shifts and local compaction. Error bars indicate standard error from 5-fold cross-validation (some horizontal error bars are negligible due to small prototype shift variance, $\sim e^{-6}$).

$p < .001$; Fig. 1a). Specifically, exemplar embeddings shifted closer to their category prototypes, demonstrating that the feedback process enhances local representational cohesion and creates more compact category clusters.

Given this observed shrinkage in cluster size, we next assessed whether modulation improves local category preservation. For each embedding, we computed the proportion of its k-nearest neighbors (k-NN) belonging to the same category as itself, systematically varying k from 1 to 300. An increase in this proportion in the modulated condition (relative to baseline) reflects improved local categorical structure. As shown in Fig. 1b, feedback modulation led to a systematic increase in the proportion of within-category neighbors across all k, indicating enhanced local category consistency.

Feedback Effects on Global Geometry

We then investigated how feedback modulation affects global representational structure, focusing on the geometry of category prototypes. For each category i, we quantify its prototype shift as $\Delta p_i = 1 - \cos\_\text{sim}(p_{i,1}, p_{i,3})$, where $p_{i,1}$ and $p_{i,3}$ are the category's prototype before and after modulation, respectively. These shifts were small (average distance $= .012$), particularly relative to the average pairwise cosine distance between different category prototypes in the baseline pass (average between-category distance $= .964$). Furthermore, the average change in these between-prototype distances after modulation was minimal (average change $= .0025$). Consistent with this stability, Representational Similarity Analysis showed a strong correlation between the baseline and modulated prototype RDMs ($\rho = .95, p < .001$), indicating high stability in the overall representational geometry. To visualize this global structural stability, we conducted PCA on baseline activations, obtained the first two principal components (PCs), and plotted all prototypes projected onto these PCs (Fig. 1c). As indicated in PCA space, category prototypes showed no consistent drift after modulation. These findings suggest feedback modulation largely preserves the global structure of inter-category relationships.

Relationship Between Local and Global Changes

While the analyses above indicated that feedback modulation primarily impacts local geometry, we conducted an exploratory analysis to examine any potential relationship between the observed local cluster compaction and modest shifts in the global positions of category prototypes. Specifically, do categories experiencing greater global repositioning also exhibit greater local refinement? To address this, we performed a 5-fold cross-validation procedure where, for each fold, training set activations were used to compute category prototype shifts and test set activations were used to compute local cluster size differences. We found a robust negative correlation ($r = -.80, p < .001$), indicating that categories with larger global prototype shifts also exhibited greater local compaction (Fig. 1d).

Discussion

Our findings show that feedback modulation intrinsically induces representational adjustments in the visual system, even without explicit external goals. Locally, feedback robustly enhanced categorical structure, leading to more compact category clusters. Globally, category prototypes shifted modestly, and inter-category distances remained largely stable. Together, these results suggest that fixed long-range feedback connections induce an automatic 'prototype effect', compacting within-category representations while preserving global structures. Importantly, this local sharpening helps explain the consistent boost in classification accuracy we observe after feedback modulation. More generally, these emergent feedback dynamics might naturally refine local representations without disrupting overall structure, thereby improving downstream category-based task efficiency.

Acknowledgments

This work was supported by Kempner Graduate Fellowship to KCL. We thank all members of the Harvard Vision Lab for their insightful comments on this project and their helpful feedback on earlier versions of the draft.

References

Bowman, C. R., Iwashita, T., & Zeithamova, D. (2020). Tracking prototype and exemplar representations in the brain across learning. *eLife*, *9*, e59360. doi: 10.7554/eLife.59360

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *2009 ieee conference on computer vision and pattern recognition* (p. 248-255). doi: 10.1109/CVPR.2009.5206848

Di Lollo, V. (2012). The feature-binding problem is an ill-posed problem. *Trends Cogn. Sci.*, *16*(6), 317–321.

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

Konkle, T., & Alvarez, G. (2023). Cognitive steering in deep neural networks via long-range modulatory feedback connections. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, & S. Levine (Eds.), *Advances in neural information processing systems* (Vol. 36, pp. 21613–21634). Curran Associates, Inc.

Lamme, V. A., & Roelfsema, P. R. (2000). The distinct modes of vision offered by feedforward and recurrent processing. *Trends Neurosci.*, *23*(11), 571–579.

Muttenthaler, L., Linhardt, L., Dippel, J., Vandermeulen, R. A., Hermann, K., Lampinen, A. K., & Kornblith, S. (2023). Improving neural network representations using human similarity judgments. In *Thirty-seventh conference on neural information processing systems.*

Schrimpf, M., Kubilius, J., Hong, H., Majaj, N. J., Rajalingham, R., Issa, E. B., ... DiCarlo, J. J. (2020). Brain-score: Which artificial neural network for object recognition is most brain-like? *bioRxiv*. doi: 10.1101/407007