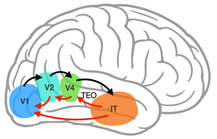
Representational Geometry Dynamics in Networks After Long-Range Modulatory Feedback

Cindy Kexin Luo, George A. Alvarez, Talia Konkle
Department of Psychology & Kempner Institute, Harvard University
Vision Sciences Lab
Kempner INSTITUTE at Harvard University

Introduction

Feedback in the Visual System

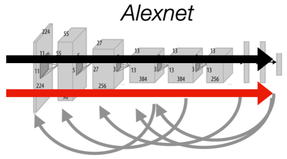


- Extensive feedback connections in the neural system (pathways from IT to V4, V4 to V1).
- Feedforward and recurrent processing iteratively refine perceptual interpretations.

Roles of Feedback:

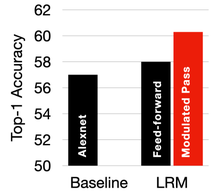
- Emergence of coherent visual objects (Di Lollo, 2012)
- Sharpening representation (Jehee et al., 2007)
- Predictive coding (Rao & Ballard, 1999)
- Functional equivalence to increased depth of processing (Kar et al., 2019)
- Figure-ground segregation (Heinen, Jolij, & Lamme, 2005)
- Construct new representations (Williams et al., 2008)

Model with Long-Range Modulatory (LRM) Feedback



(Konkle & Alvarez, 2023)

ImageNet Accuracy



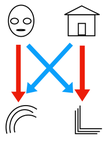
Feedback Improves:
✔ Image Recognition
✔ Adversarial Robustness
✔ Brain Alignment

Q: How Does Feedback Intrinsically Reshape Representational Geometry?

How the LRM Model Integrates Feedback

Key Design #1: Channel-to-Channel Influences

Intuition:

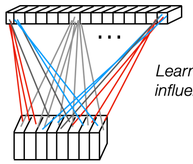


Learn these!

H x W x Feature channels

Source Layer: 1 x 1 x **1000**

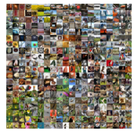
Destination Layer: 35 x 35 x **384**



Learnable influences

Key Design #2: Representation Learning

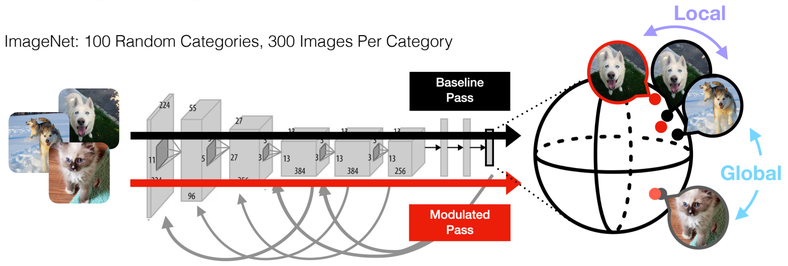
ImageNet: 1000-way categorization



$$\mathcal{L}_{total} = \frac{\mathcal{L}_{feedforward} + \mathcal{L}_{Modulated\ pass}}{2}$$

Probing Changes in Representational Geometry

ImageNet: 100 Random Categories, 300 Images Per Category

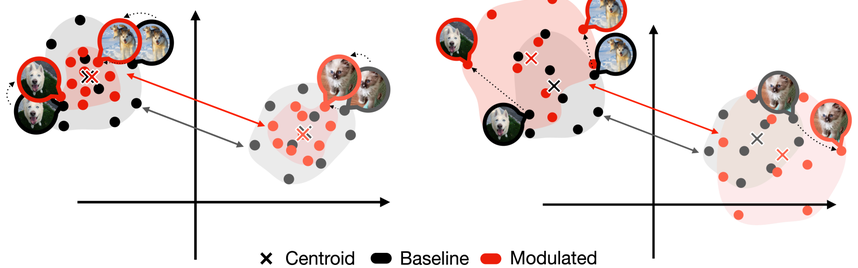


(Bowman, Iwashita, & Zeithamova, 2020; Muttenthaler et al. 2023)

Hypotheses: How Feedback Reshapes Representations

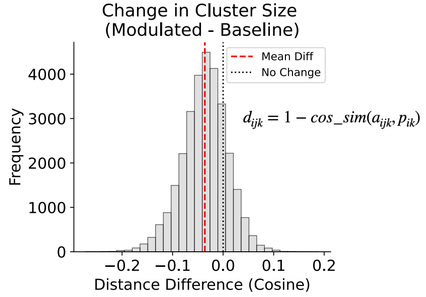
(a) **Shrink** + **Stay**

(b) **Expand** + **Shift**

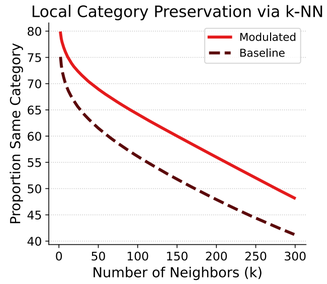


Local Effect

Feedback Compacts Category Clusters



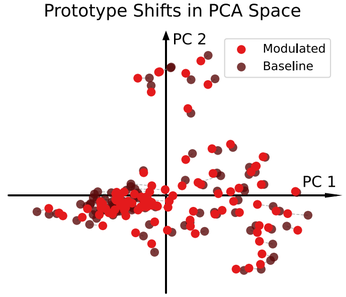
→ Exemplars move **significantly closer** to their category centroids.

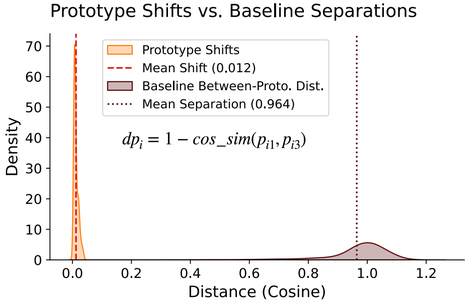


→ Local neighbors have **more of the same category** after modulation.

Global Effect

Overall Category Arrangement Remains Stable

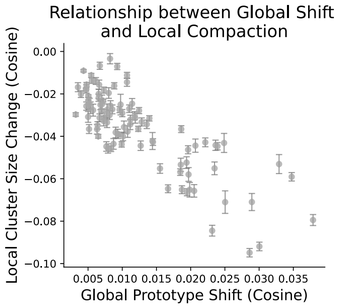




RSA: ρ = 0.95 between baseline and modulated RDMs.

→ The global structure of inter-category relationships **remains largely intact.**

Exploratory Analysis



Are Categories with Larger Global Shifts Also More Locally Refined?

→ **Yes!**

A robust correlation ($r = -.80$, $p < .001$) between local size changes and global prototype shifts.

Conclusions

Feedback modulation induces **an automatic "prototype effect"** on **representations (Shrink + Stay):**

1. It **compacts** and **refines local representations** within a category.
2. It largely **preserves global structure.**

These adjustments may enhance downstream category-based task efficiency.

>>> But what role does the training objective play?
Extending the analysis to the **Self-Supervised LRM**—stay tuned!

>>> Does this reflect what happens in the brain?
We also observed that the LRM model achieved higher brain alignment on the Brain-Score benchmark.

- Extra-Classical Receptive Field?
- Changes across layers -> Different stages of processing