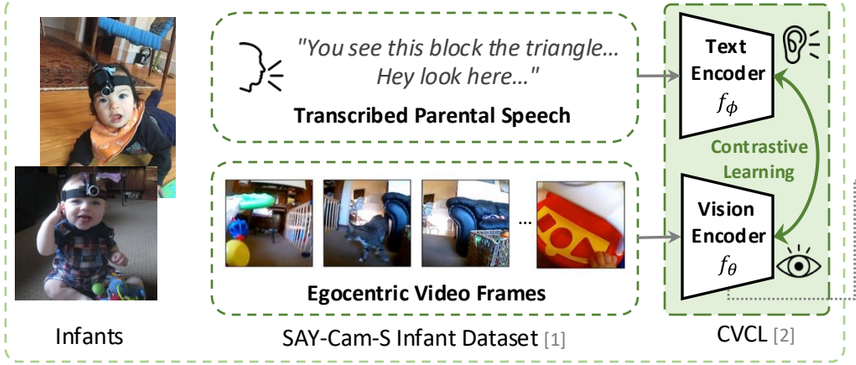
Discovering Hidden Visual Concepts Beyond Linguistic Input in Infant Learning

Xueyi Ke[1], Satoshi Tsutsui[1], Yayun Zhang[2], Bihan Wen[1]

[1]Nanyang Technological University [2]The Max Planck Institute for Psycholinguistics

Scan here for project page!

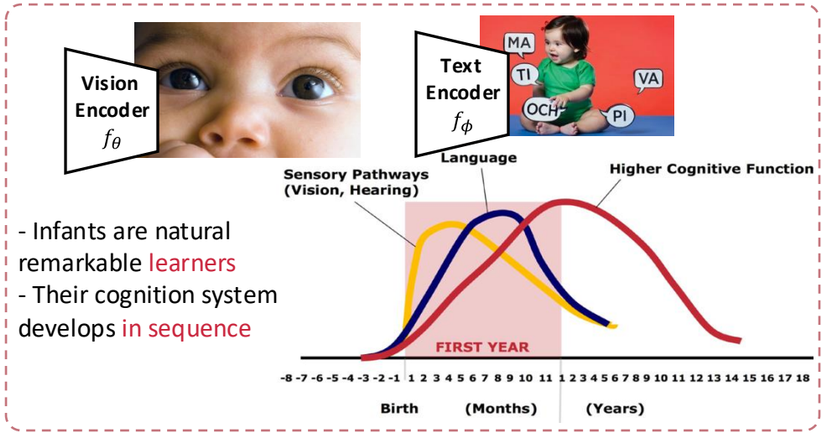
Prior Work: Training Models from Infant Data



Computational Model

Our Research Question:

Does the vision encoder of a model trained on infant egocentric data develop visual representations beyond its linguistic training data, like human infants?

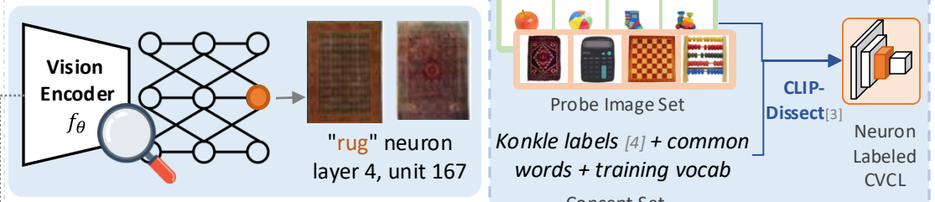


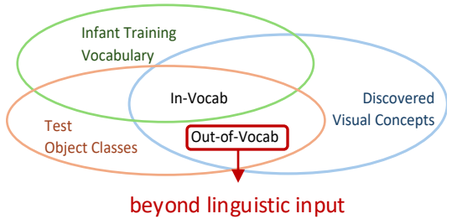
- Infants are natural remarkable learners
- Their cognition system develops in sequence

Does Vision Develop Earlier?

Vision encoder may recognize patterns beyond text in representation space:

To verify, we apply "neuron labeling"[3] : assign an object label to each neuron.

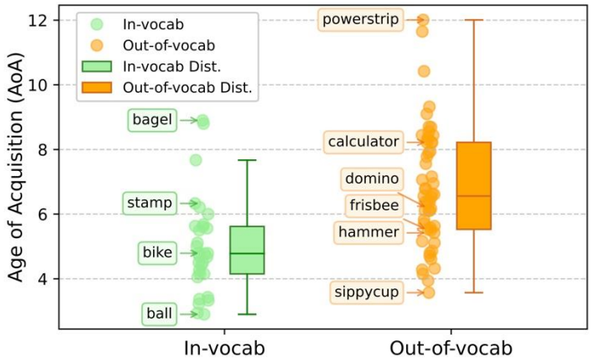




Discovered visual concepts from visual representation are categorized as in- or out-of-vocabulary.

Many out-of-vocabulary visual concepts emerge in specific neurons.

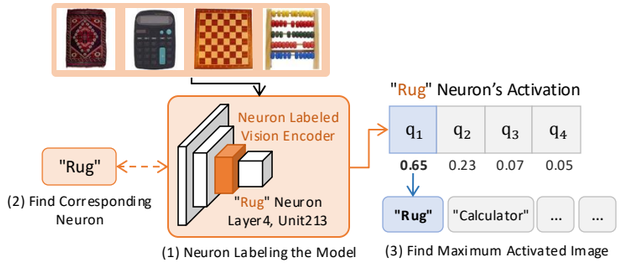
We rate visual concept from cognitive perspective: Age of Acquisition (AoA) [5]



Complex patterns in out-of-vocabulary concepts suggest visual learning **beyond text supervision**.

Neuron-based Classification

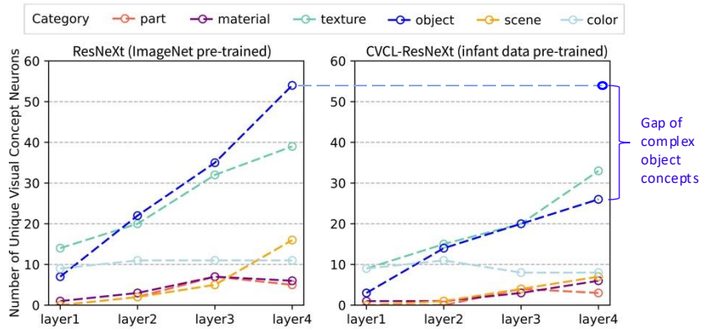
Model can classify images using only a single visual concept neuron, **without any additional training**.



Method	Model	In-vocab	Out-of-vocab	All
Vanilla	CLIP-ResNet50	$98.81_{\pm 0.16}$	$96.93_{\pm 0.06}$	$97.42_{\pm 0.05}$
	ResNeXt50	✗	✗	✗
	CVCL-ResNeXt50 (*Baseline*)	$\mathbf{36.18}_{\pm 0.91}$	✗	✗
Neuron Classifier	CLIP-ResNet50	$91.59_{\pm 0.52}$	$88.66_{\pm 0.35}$	$89.79_{\pm 0.38}$
	ResNeXt50	$88.17_{\pm 0.45}$	$93.28_{\pm 0.36}$	$91.88_{\pm 0.15}$
	CVCL-ResNeXt50	$\mathbf{79.50}_{\pm 0.78}$	$\mathbf{76.81}_{\pm 0.35}$	$\mathbf{77.79}_{\pm 0.40}$

ImageNet vs Infant Pretrained Representations

A significant gap in representation richness exists across final layers.



[1] Jessica S., et al. (2021). Say-Cam: A large, longitudinal audiovisual dataset recorded from the infant's perspective. *Open mind*.
[2] Vong, W. K., Wang, W., Orhan, A. E., & Lake, B. M. (2024). Grounded language acquisition through the eyes and ears of a single child. *Science*.
[3] Oikarinen, T., & Weng, T. W. (2023). CLIP-Dissect: Automatic Description of Neuron Representations in Deep Vision Networks. *ICLR*.
[4] Konkle, T., et al. (2010). Conceptual distinctiveness supports detailed visual long-term memory for real-world objects. *Journal of experimental psychology*.
[5] Kuperman, V., Stadthagen-Gonzalez, H., & Brysbaert, M. (2012). Age-of-acquisition ratings for 30,000 English words. Behavior research methods, 44, 978-990.

Acknowledgements: This work was partially supported by the National Research Foundation Singapore Competitive Research Program (award number CRP29-2022-0003). We thank Wai Keen Vong for invaluable discussion and CVCL's pretrained weights. We thank Jingyi Lin for figure discussions.