Uncovering Object Categories in Infant Views

Naiti S. Bhatt[1,2], Bria Long[3], Michael C. Frank[3]
[1]Keck Science Department, Scripps College; [2]Department of Psychology, University of Edinburgh; [3]Department of Psychology, Stanford University

The time is ripe to study infant object learning.

Recent advances in naturalistic data collection and computer vision models:

- SAYCam: longitudinal dataset of 3 infants' at-home headcam videos
  - 3 hours of video collected twice per week from 6 to 32 months of age
- Open-source computer vision (CV) models (e.g., Facebook AI Research (FAIR)'s Detectron2 perform well in object detection and recognition

Why characterize the inputs to infant object learning?

Quantifying the process of how infants learn to categorize objects helps:

1. better understand the visual system
2. inform early word learning
3. help develop better and more developmentally plausible computer vision models for object recognition

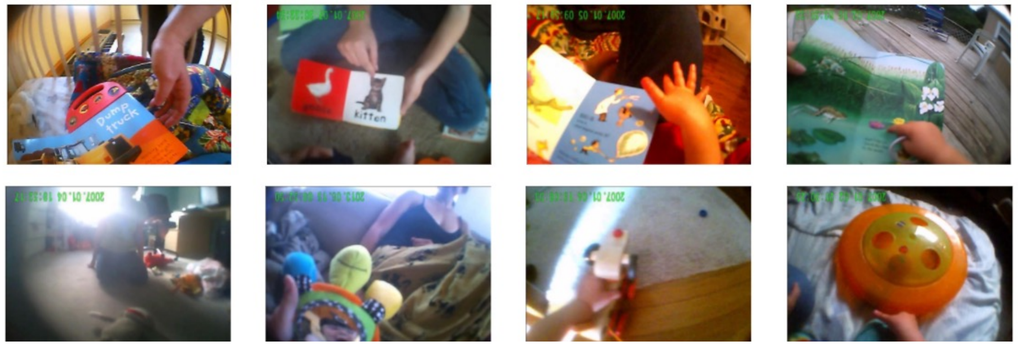
What objects do infants interact with?

Clerkin et al., 2017 used headcam videos from 8–10-month-olds during mealtime:

infants see a few objects frequently and most objects infrequently (Zipfian)

How generalizable is this finding?

Specifically: looking at **interactive** frames with child hands present

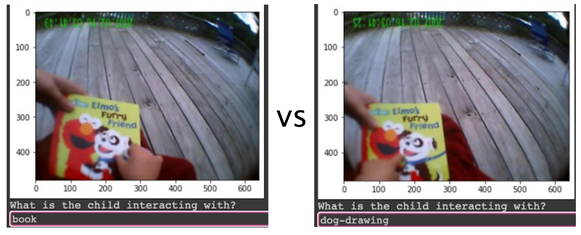


Manual Annotations

Randomly sampled ~3000 images

Labeling conventions included:

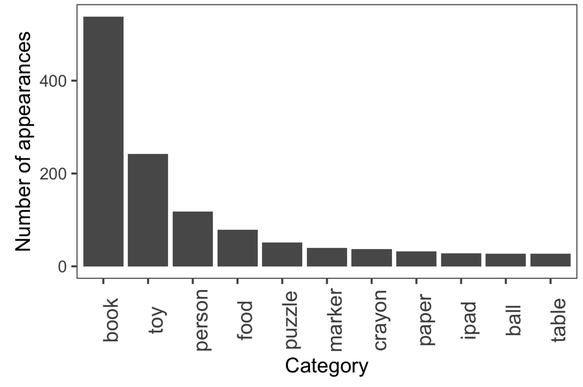
1. Basic level labels (i.e., "bird' not "robin")
2. When interacting with something unknown, label as "unknown object"
3. Label depictions with "-drawing" tag
4. Label main object that the child is interacting with (label according to if the child is pointing at a specific thing in the book or just at the general book)



Objects are Zipfian distributed

Indeed, infants interact with a few objects frequently and most objects infrequently (Zipfian).

Most frequent: book + toy



How well can existing computer vision models categorize objects in the infant view?

There are any more frames in the SAYCam dataset than can be reasonably manually labeled.

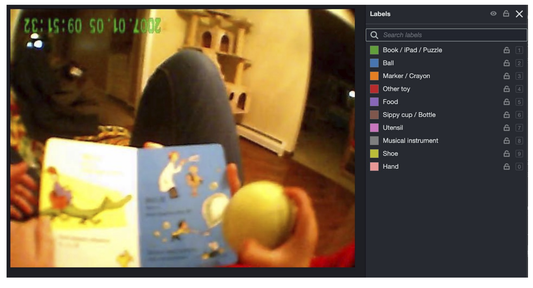
We look to computer vision models to automate object detection, recognition, and classification.

However, models and infants learn from very different views!

We tested FAIR's Detectron2 open-source CV models on SAYCam frames.
Only "person" was labeled well.

Crowdsourced Annotations

Got ground truth labels of "obvious objects" using Amazon Sagemaker.



Model training is promising

Pre-training: Poor detection & classification

Post-training: Improved performance for frequent categories