Computational Cognitive Science

Lecture 15: Language Acquisition 2

Frank Keller

9 November 2026

School of Informatics
University of Edinburgh
keller@inf.ed.ac.uk

Overview

From Word Learning to Grounded Learning

Naturalistic Data

The CVCL Model

Evaluation

Results

Cognitive Interpretation

Reading: Vong et al. (2024).

From Word Learning to Grounded Learning

Connections to Earlier Lectures

Two ideas from earlier lectures come together here:

- Lecture 10: Battleday et al.'s (2020) model of categorization: Images are represented by high-dimensional vision embeddings. Similarity in these representations serves as psychological similarity.
- Lecture 14: Frank et al.'s (2009) model of word learning: Bayesian inference over speaker intentions is used to learn a lexicon from utterances paired with symbolic object labels.

Today: Can we learn vision representations and their links to words directly from the experience of a child?

A More Naturalistic Learning Problem

A child does not receive:

- a list of objects present in each scene;
- category labels for visual objects;
- explicit word-object alignments (a lexicon).

Instead, the learner receives a continuous, noisy stream of:

visual experience + linguistic utterances

Vong et al. (2024) ask how much can be learned using relatively generic mechanisms:

- representation learning to discover useful visual features;
- associative learning to connect language with visual experience.

How much word learning is possible without built-in constraints or social reasoning?

Naturalistic Data

SAYCam-S: A Child's-Eye Dataset

The model is trained on longitudinal head-mounted camera recordings from one child:

- age: 6–25 months;
- 61 hours of egocentric video;
- about 600,000 video frames;
- 37,500 transcribed utterances;
- visual and linguistic input are paired by temporal co-occurrence.

This is much more realistic than a laboratory word-learning dataset, but still only about 1% of the child's waking experience.

What Does the Training Input Look Like?



Time: 0:41
Utterance: You see this block the triangle



Time: 0:44
Utterance: It goes in



Time: 0:48
Utterance: Boop



Time: 0:52
Utterance: Hey look here



Time: 1:13
Utterance: Yeah

...



Time: 2:00
Utterance: You like the string



Time: 2:06
Utterance: You want the blocks too

What Does the Training Input Look Like?

The correspondence between visual experience and linguistic utterances is noisy and sparse:

- a word may occur when its referent is small or absent;
- many objects are visible but not mentioned;
- utterances contain many non-object words.

The CVCL Model

Child's View for Contrastive Learning

Vong et al. (2024) introduce CVCL: Child's View for Contrastive Learning

The model consists of a vision encoder f_θ and a language encoder f_ϕ:

$$\text{image } x \xrightarrow{f_\theta} \mathbf{v} \qquad\qquad \text{utterance } w \xrightarrow{f_\phi} \mathbf{u}$$

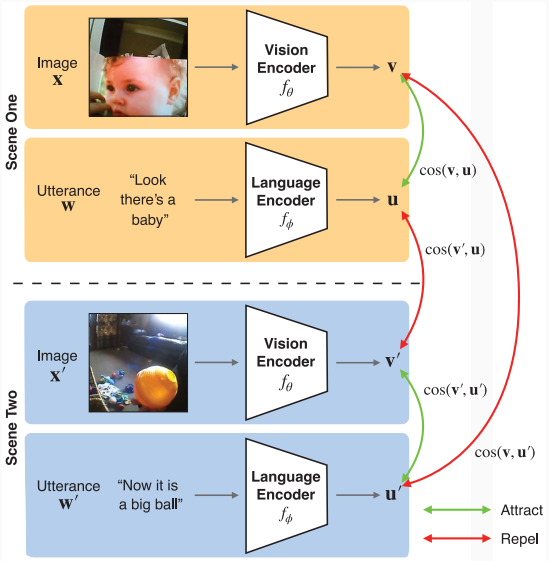
where $\mathbf{v}$ and $\mathbf{u}$ are vector representations (embeddings).

Images are represented in a learned, high-dimensional feature space (see lecture 10).

New: CVCL also learns a word embedding space and aligns the two spaces.

The language encoder learns one embedding per word; an utterance is represented by averaging its word embeddings.

Contrastive Learning Architecture



For matching image–utterance pairs:

$$\text{sim}(\mathbf{v}_i, \mathbf{u}_i) \uparrow$$

For mismatching pairs:

$$\text{sim}(\mathbf{v}_i, \mathbf{u}_j) \downarrow$$

Similarity is measured using cosine similarity.

Temporal co-occurrence supplies the learning signal.

Basic architecture similar to CLIP (Radford et al. 2021).

Contrastive Learning: Objective

For an instance i, we match each image with its corresponding utterance in the same batch:

$$P(w_i|x_i) = \frac{\exp(\text{sim}(\mathbf{v}_i, \mathbf{u}_i)/\tau)}{\sum_j \exp(\text{sim}(\mathbf{v}_i, \mathbf{u}_j)/\tau)}$$

The loss rewards the correct image–utterance pairing:

$$\mathcal{L}_{\text{frame}} = -\log P(w_i|x_i)$$

A symmetric second loss $\mathcal{L}_{\text{utterance}}$ rewards the correct utterance–image pairing.

We average $\mathcal{L}_{\text{frame}}$ and $\mathcal{L}_{\text{utterance}}$ to learn a shared multimodal embedding space.

The vision encoder is pre-trained on egocentric child video and then frozen; only its projection layer is trained. The language encoder is trained from scratch.

Relation to Cross-Situational Learning

Recall cross-situational learning from Lecture 14:

$$\text{word } \textit{ball} + \text{many different scenes} \implies \text{stable association with object BALL}$$

CVCL implements a modern version of this idea:

- it does not count discrete word–object co-occurrences;
- it learns distributed representations of whole images and utterances;
- repeated co-occurrence gradually aligns visual and word representations.

Association learning is combined with representation learning.

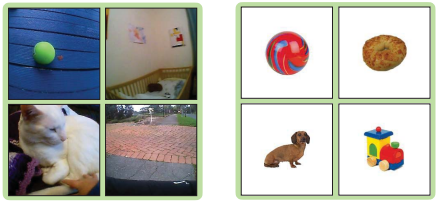
Evaluation

How Is Word Knowledge Tested?

The evaluation resembles a classic infant word-comprehension task:

1. give the model a target word, e.g., *ball;*
2. show four candidate images;
3. encode the word and all four images;
4. choose the image with highest cosine similarity to the word.

No category labels are used to train CVCL.



What Should CVCL Be Compared Against?

The paper includes ablations and upper bounds:

- CVCL-Shuffled: break the true image–utterance pairings.
- Random Features: remove useful learned visual representations.
- CLIP: web-scale contrastive image–text model trained on vastly more data.
- Linear Probe: directly supervised classifier trained on manually labeled images.

Ablations ask which ingredients are necessary; upper bounds show what is possible with more or cleaner supervision.

Time for a short quiz on WooClap!

Go to wooclap.com

Enter the event code in the top banner

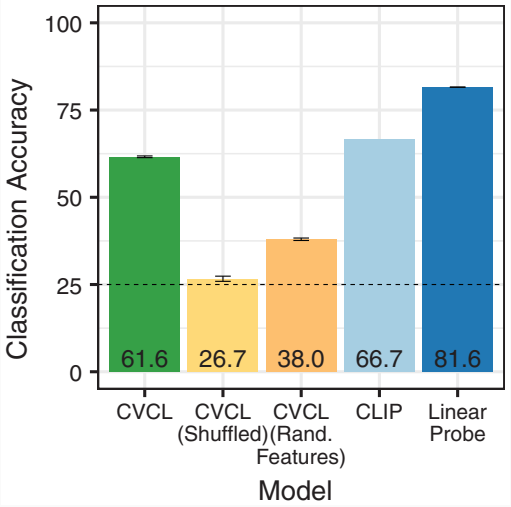
Event code

EIAPJOX



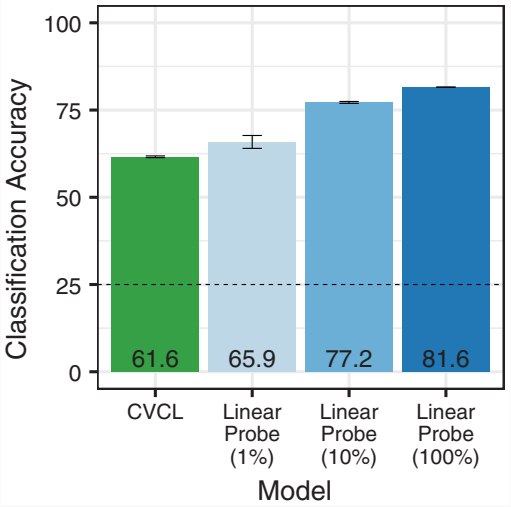
Results

Result 1: Acquiring Word–Referent Mappings



- Evaluate on Label-S: manually annotated dataset.
- Shuffling pairs destroys learning: co-occurrence matters.
- Random visual features hurt: representation learning matters.
- Chance performance is 25%.
- CLIP and Linear Probe as upper bounds.

How Efficient Is Natural Language Supervision?



- Category labels are a much cleaner learning signal than child-directed speech.
- one directly labeled example is worth at least seven natural-language examples (authors' estimate).

Result 2: Zero-Shot Generalization

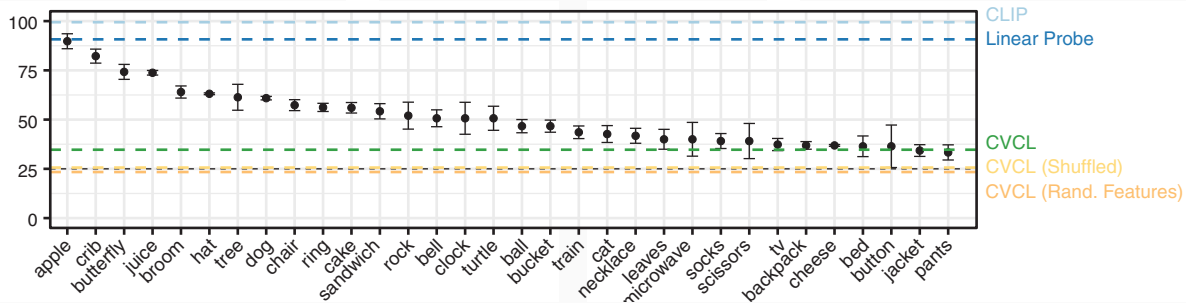
Learning a word should mean more than recognizing the exact scenes seen during training: Learners need to generalize.

The model is therefore tested on the Konkle Objects dataset:

- 64 object categories;
- isolated objects on a white background;
- visual exemplars are very different from the child's training images;
- the relevant words occurred in the training input.

Can the learned word meanings generalize to these new visual instances?

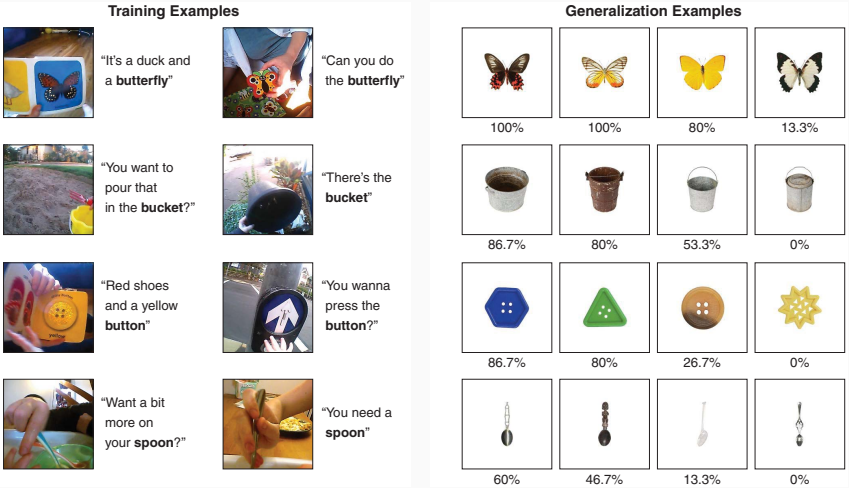
Zero-Shot Generalization: Results



Overall CVCL accuracy: 34.7% (chance: 25%).

- 16/64 categories exceed 50%;
- another 42/64 are above chance;
- performance is far below the web-scale upper bounds.

What Does Zero-Shot Generalization Look Like?



The model can sometimes extract a category that transfers well beyond the original scene, but generalization is uneven.

Result 3: Visual and Linguistic Concepts Align

If the model has learned grounded meanings, then its visual and word representations should have similar structure.

For 22 concepts, the authors compare:

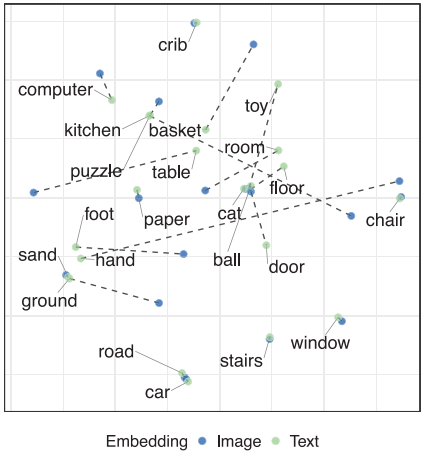
- each word embedding;
- the mean visual embedding for that concept.

This extends the idea from Lecture 10: rather than asking whether a visual embedding provides a useful psychological space, we now ask whether visual and linguistic spaces acquire a common conceptual organization.

Visual and linguistic spaces show significant conceptual alignment ($r = .37$, $p < .001$).

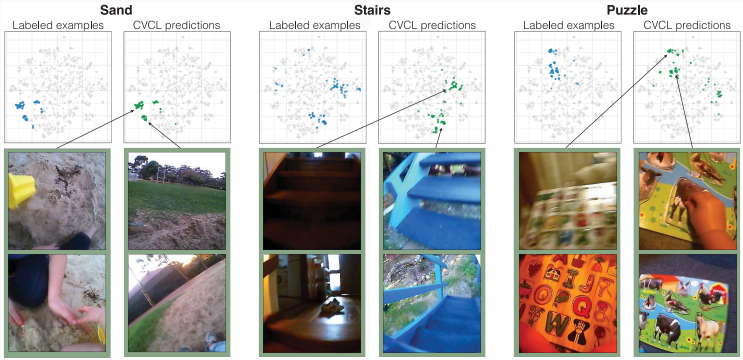
Better cross-modal alignment also predicts better classification performance.

Multimodal Conceptual Alignment



The learned word and visual systems are not merely associated pairwise; their broader conceptual organization becomes aligned.

Concepts Can Have Internal Structure

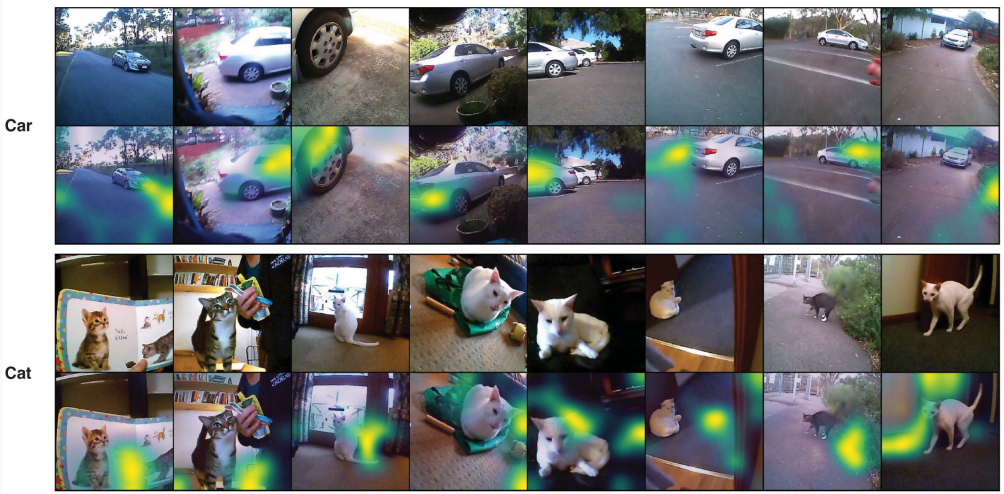


A single word embedding can activate multiple visual subclusters:

- indoor versus outdoor stairs;
- alphabet versus animal puzzles.

Structured concepts emerge without explicitly assuming prototypes or subcategories.

Does the Model Localize Referents?



Grad-CAM attention maps show that CVCL sometimes focuses on the correct visual referent. But localization is inconsistent across categories.

Cognitive Interpretation

Frank et al. (2009) vs. Vong et al. (2024)

	Frank et al. (2009)	Vong et al. (2024)
Input	words + symbolic objects	pixels + transcribed utterances
Representation	discrete word–object mappings	distributed visual and word embeddings
Learning	Bayesian inference	contrastive neural representation learning
Cross-situational?	yes	yes
Speaker intention	explicit latent variable	not explicitly represented
Main question	can intentions help learn a lexicon?	what is learnable from naturalistic sensory input?

What Does CVCL Establish?

CVCL demonstrates learnability:

- realistic first-person visual and linguistic data contain enough statistical structure to learn many early word–referent mappings;
- simple temporal co-occurrence can provide a useful supervisory signal;
- modern representation learning helps turn noisy perceptual input into useful concepts;
- learned meanings can generalize beyond the exact training scenes.

Some grounded lexical knowledge learnable without explicit word–object labels.

What Does CVCL Not Establish?

Success with a minimal model does not show that children use only these mechanisms.

The model does not explicitly include:

- mutual exclusivity or the principle of contrast;
- shape bias;
- syntactic cues;
- social or gestural reasoning;
- active information seeking;
- explicit hypothesis generation.

These mechanisms may still contribute to human word learning.

A Cognitive Modeling Perspective

Why is this interesting for computational cognitive science?

- It tests a hypothesis about the information available to the learner.
- It asks whether a relatively generic learning mechanism can recover psychologically meaningful structure.
- Ablations allow us to ask which model ingredients are necessary.
- Out-of-distribution tests ask whether the model has learned something more abstract than memorized scenes.

But matching children's learning outcomes is not yet the same as matching their learning process.

Summary

- CVCL learns from naturalistic head-camera video and child-directed language from a single child.
- Contrastive learning aligns visual and word embeddings by attracting temporally co-occurring pairs and repelling mismatches.
- CVCL learns many word–referent mappings and performs well above chance; shuffled pairings and weak visual features hurt performance.
- Learned meanings generalize to novel visual exemplars and produce partially aligned visual and linguistic conceptual spaces.
- Compared with Frank et al. (2009), CVCL uses more realistic input and learned representations, but doesn't model speaker intentions.
- The results show what is learnable from realistic sensory input; they do not show that children used these mechanisms.

References

Battleday, Ruairidh M., Joshua C. Peterson, and Thomas L. Griffiths. 2020. Capturing human categorization of natural images by combining deep networks and cognitive models. *Nature Communications* 11(1):5418.

Frank, Michael C., Noah D. Goodman, and Joshua B. Tenenbaum. 2009. Using speakers' referential intentions to model early cross-situational word learning. *Psychological Science* 20(5):578–585.

Radford, Alec, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*. PMLR, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763.

Vong, Wai Keen, Wentao Wang, A. Emin Orhan, and Brenden M. Lake. 2024. Grounded language acquisition through the eyes and ears of a single child. *Science* 383(6682):504–511.