In-the-Moment Visual Information from the Infant's Egocentric View Determines the Success of Infant Word Learning: A Computational Study

Andrei Amatuni[1], Sara Schroer[1], Yayun Zhang[1], Ryan Peters[1], Md. Alimoor Reza[2], David Crandall[2], Chen Yu[1]

{andreiamatuni, saraschroer, yayunzhang}@utexas.edu

{ryan.peters, chen.yu}@austin.utexas.edu

mdreza@iu.edu, djcran@indiana.edu

Department of Psychology, University of Texas at Austin[1]

Luddy School of Informatics, Computing, and Engineering, Indiana University[2]

Abstract

Infants learn the meaning of words from accumulated experiences of real-time interactions with their caregivers. To study the effects of visual sensory input on word learning, we recorded infant's view of the world using head-mounted eye trackers during free-flowing play with a caregiver. While playing, infants were exposed to novel label-object mappings and later learning outcomes for these items were tested after the play session. In this study we use a classification based approach to link properties of infants' visual scenes during naturalistic labeling moments to their word learning outcomes. We find that a model which integrates both highly informative and ambiguous sensory evidence is a better fit to infants' individual learning outcomes than models where either type of evidence is taken alone, and that raw labeling frequency is unable to account for the word learning differences we observe. Here we demonstrate how a computational model, using only raw pixels taken from the egocentric scene image, can derive insights on human language learning.

Keywords: word learning; egocentric vision; sensory grounding; deep neural networks; computational modeling

Introduction

Infants learn the meanings of their first words through their everyday experiences. Linking spoken words to their correct visual referents requires young learners to successfully integrate what they see with what they hear. This is a challenging task, because the visual-audio sensory input during learning contains a high degree of uncertainty with many candidate words co-occurring with many candidate objects across space and time. To examine the underlying mechanisms employed to solve this difficult problem, researchers have long turned to computational modeling. These computational studies have advanced our understanding by specifying computational principles inherent to this learning task. Additionally, they allow us to simulate potential cognitive mechanisms that might support word learning. Early modeling work has heavily relied on simplified and artificial stimuli due to the lack of datasets capturing infants' learning environments in the real world as well as limited computational power to handle high-density input collected from real world settings. However, recent advances in sensing and computational technologies have revolutionized the cognitive modeling field. These new technologies allow us to collect high-density behavioral data and simulate infant learning in naturalistic settings to test precise theories of how word learning might unfold in the real world.

Computational Models of Infant Word Learning

Early modeling work used a connectionist system to explain several word learning phenomena previously documented in young children, such as over/under-extension effects, vocabulary spurts and prototype effects (Plunkett, Sinha, Møller, & Strandsby, 1992). To simulate word learning, the model learned to associate images to their corresponding labels. This work demonstrated the power of applying computational models in researching infant word learning, by explicitly linking the model's internal mechanisms to various features of early language development. While these simulations offered insights on potential mechanisms that might support learning, the stimuli in this study were highly abstract random dot patterns paired with discrete binary valued word labels. This limits the study's generality when considering real world learning data which involves highly complex visual and auditory sensory input. Later work would use more advanced neural network models to learn word-referent mappings from more naturalistic stimuli (Roy & Pentland, 2002). However, this later model was only able to learn using highly reduced forms of the raw audio and visual sense data, which first went through considerable pre-processing before being input to a word learning module. While both of these early models demonstrated powerful learning capabilities, neither of them were tied to actual human infant word learning performance, though in the case of Plunkett et al. (1992) they were able to simulate a few general features observed during early language acquisition.

Going beyond these early word learning simulations, there has been considerable modeling work which relates more directly to human word learning performance. These studies have largely come from statistical learning paradigms which model human word learning performance by linking distributional features of training stimuli to human learning outcomes. In these experiments, researchers will typically encode different statistical regularities in their training stimuli and then test subjects' learning outcomes after they have been exposed to these data during a series of training trials. Computational models have linked statistical word learning performance to associative memory mechanisms (Kachergis, Yu, & Shiffrin, 2012a, 2012b), as well as proposed local learning rules which test specific word-meaning hypotheses (Stevens, Gleitman, Trueswell, & Yang, 2017). Other work

Figure 1: Infant head camera frames taken from multiple different infants at the moment at which "moose" was uttered by their caregiver. One group of frames (A or B) is from infants who learned the object name by the end of the free-play session, and the other is from infants who did not learn the name. Can you tell which group of infants learned "moose" by the end of the session?

has used computational models to study how social information, such as prosodic and attentional cues (Yu & Ballard, 2007) or speakers' referential intentions (Frank, Goodman, & Tenenbaum, 2009), might be integrated with statistical information in order to learn word meanings. To model iterative word learning across referentially ambiguous labeling episodes, some authors have proposed probabilistic associative models of learning (Fazly, Alishahi, & Stevenson, 2010). Prior work has also used modeling to study the relative timecourse of word learning. Some studies have modeled acquisition rates as a function of the inherent uncertainty for learning an entire lexicon (Blythe, Smith, & Smith, 2010), while other work, which hoped to explain certain rapid word learning phenomena, has suggested general principles of Bayesian inference may be at play (Xu & Tenenbaum, 2007). These studies have all used formal models to specify the cognitive mechanisms and computational principles that may govern word learning. While some of this prior work has attempted to explain general patterns of word acquisition in human subjects, none of these studies has attempted to explain individual infants' word learning performance for individual words.

Simulating Individual Word Learning In a Naturalistic Setting

During statistical word learning, subjects will accumulate different statistics as a product of their unique selective attention dynamics. This leads to individual differences in both the statistical evidence they collect as well as the specific words they end up learning. Prior work has shown that fine grained patterns of selective attention can be used to derive individual word learning outcomes as well as to decode subjects' internal states of knowledge (Amatuni & Yu, 2020). Here we present work that aims to simulate individual infants' word learning performance. In contrast to previous modeling work using idealized stimuli and subjects, we use sensory data collected from infants' first person views while they and their caregiver play with a set of toys in a naturalistic environment.

With recent advances in mobile sensing technologies such as small head-mountable cameras, there have been numerous studies recording infants' naturalistic first-person visual and auditory experiences (Yu & Smith, 2012; A. F. Pereira, Smith, & Yu, 2014; Tsutsui, Chandrasekaran, Reza, Crandall, & Yu, 2020; Bergelson & Aslin, 2017; Sullivan, Mei, Perfors, Wojcik, & Frank, 2021). These recordings help us characterize the learning data young learners have access to, and allow us to draw insights on the cognitive processes which make use of this sensory input to support word learning.

Here we present a study which uses a computational model to simulate real-time word learning outcomes for a set of 24 infants as they play freely with a caregiver in a naturalistic home-like environment. At the end of this free-play session infants perform a word-learning test to measure which of the specific toys they successfully learned the names for. While infants and their caregivers are allowed to behave freely as they normally would, the lab based environment offers a controlled setting in which to extract natural behaviors from the dyads and to test immediate word learning outcomes. See Figure 1 for example egocentric frames taken from two different groups of infants at the moment that their caregiver labeled the object "moose". One of these groups of infants would ultimately learn that the label "moose" was associated with the toy moose.

Our goal in this paper is to test whether our model, trained by egocentric vision recorded during naturalistic infant-parent toy play, can successfully simulate individual infant

learning performance for individual words. Specifically, in Study 1 we used our model to predict infants' word learning performance using only images taken from the child's point of view during labeling instances. In Study 2, we used the model to simulate different types of statistical evidence that learners may aggregate. Here we used the strength of association between visual features in the egocentric scene and their associated learning outcomes as a measure of the quality of visual sensory input delivered to each infant. Using this measure we computed each subject's accumulated sensory evidence for each individual object in order to test whether the *quality* of visual sensory input may explain individual word learning performance.

Behavioral Data Collection

Twenty-four infants (mean age: 17.5mo, range: 12.6-25.8) and their parent were recruited to play in a home-like lab as part of a larger experiment. Parents were not told we were interested in word learning. The parent-infant dyads were given 10 novel toys to play with for a 10 minute session where parents were asked to use specific labels for each toy. The objects and their associated labels were chosen so that there was a low probability that children in this age group already knew these specific words. Parents were told to simply play as they would at home, without further instructions. While playing, both parent and infant wore wireless head-mounted eye trackers (Pupil Labs). The parent's eye tracker was worn like a pair of glasses and the infant's eye tracker was modified to be attached to a soft hat. The eye trackers were connected to a smart phone (Google). Participants wore a custom-made jacket with a pocket in the back to hold the phone while they played. The wireless eye tracking allowed infants and parents to move freely, capturing more naturalistic interactions than most lab studies on word learning.

Following the play session, infants' knowledge of the 10 label-object mappings was tested using a screen-based task. Two objects were presented on the screen at a time and the infant was prompted to look at a labeled object. A trial was considered correct if the infant looked at the target object longer than the distractor. Infants were tested on each object twice. A word was considered "learned" if both trials were correct and "not learned" if both trials were incorrect. On average, 2.3 objects were coded as "learned" and 1.9 objects as "not learned" per infant. The remaining items could not be conclusively scored as the infant did not attend to the screen for the majority of a trial and/or the infant only got one trial correct. After the experiment, the eye tracking videos were calibrated to generate an estimate of the (x, y) coordinates corresponding to subjects' gaze. Parent speech was also transcribed to find instances when they labeled any of the 10 objects (both those that the infants did and did not learn).

Study 1: Simulating Infant Word Learning Outcomes Using Egocentric Video Frames

As an initial step towards linking first-person visual scenes to early word learning, we demonstrate that a computational model can predict infants' learning outcomes above chance using only images from their field of view (FOV) during a labeling event. We train a deep convolutional neural network (CNN) to discriminate frames from subjects who ultimately learned the meanings of these words at the end of the play session vs. those who did not. By solving this classification task, the model must learn to associate visual features in the egocentric scene with the relative success of those labeling moments, and in so doing it becomes sensitive to the visual features that constitute an ideal learning moment.

Data and Computational Models

We collect all the video frames centered around a labeling moment (3 seconds surrounding the utterance), collapsing across subjects and grouping together all the frames associated with a common label. Due to differences in relative rates of naming and word learning for different objects, we omit half of the items from further analysis due to lack of data. These specific objects had large class imbalances ($> 75\%$ of frames are either all "learned" or all "not learned"), preventing us from training and testing models on these particular items. This yielded 5 viable objects which had a balance of both learned and not-learned instances associated with them. For each of these 5 objects, we train a separate set of binary classifiers on all the frames associated with labeling events for that specific object. The goal of these models is to discriminate whether or not an image was taken from a subject who ultimately learned the word uttered by the caregiver vs. a subject who did not learn the meaning of the specific word uttered at that labeling moment.

Our networks are organized in a feedforward architecture, where input images are fed through a deep convolutional network followed by a fully connected layer with two outputs. We train the network end-to-end using a cross-entropy loss. Ground truth labels for each frame (i.e. learned vs. not learned) are determined by each specific infants' word learning results at the end of the play session. For example, if an infant learned "koala" by the end of the experiment, the ground truth for all the frames associated with their own "koala" labeling events would be marked as "learned;" if they did not learn "koala," these frames would be marked "not-learned" (Figure 2).

We report analyses using a pretrained ResNet-50 as our deep CNN architecture (He, Zhang, Ren, & Sun, 2016), although we found similar results using other backbone architectures. This model was pretrained on the ImageNet dataset (Deng et al., 2009) and fine-tuned on our set of infants' egocentric scene images. To robustly estimate accuracy of our models, we test using 5-fold cross-validation with 3 separate training trials on each of the folds (each initialized with different random seeds). When splitting data for training and

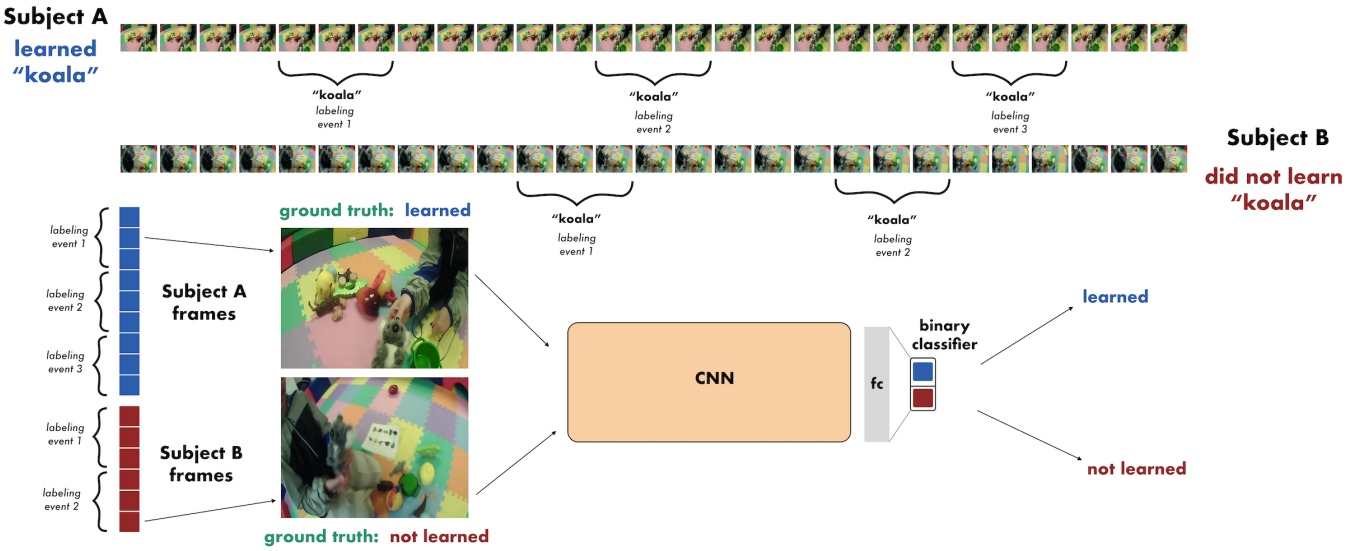

Figure 2: Using a classification based approach to model visual features associated with successful word learning outcomes. We train independent binary classifiers for each of the 5 different objects

testing, we partition by event rather than by frames, so that all scene images associated with a given labeling event will either be in the training or test set. Otherwise, due to the close temporal contiguity for frames within a labeling event, the model could memorize the visual similarities specific to a particular event rather than learning features that generalize across labeling events. See Figure 2 for a schematic of our analyses.

Testing and Cross-Validation Procedure During training we use a frequency weighted loss to mitigate the model's exposure to class imbalances inherent to the training set. During testing, for any given testing fold, we subsample from the larger class to produce a balanced dataset (with a 50% random baseline). The classification accuracies we report here are averages across multiple random subsamplings (n=100).

Results

Figure 3 presents the accuracies of our learned versus not-learned classifier across the five objects in the dataset. Our models successfully classify word learning outcomes above chance (50% random baseline) for all 5 objects using the held out images taken from never before seen labeling events.

These results demonstrate how infants' embodied interaction with their environment leaves a unique signature on the visual scenes that they experience. Our models are able to extract these signatures from the first-person visual signal and successfully link patterns in these signals to infants' own word learning, offering a proof-of-concept that this is possible, in principle, using a sensory grounded computational model. While these results demonstrate how word learning outcomes can be derived from information taken in-the-moment, we know that word learning is a process that's extended across multiple episodes, as numerous pieces of evi-

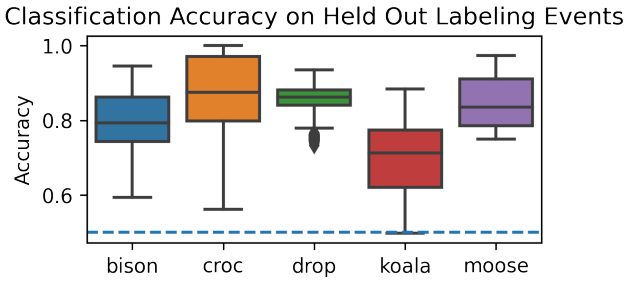

Figure 3: Classification accuracy on held out test frames, where independent models are trained for each object. Our models are able to predict learning outcomes for labeling events that were never seen by the model during training, demonstrating an ability to generalize to novel labeling events. Plots reflect aggregate classification accuracies of balanced test samples across 5 folds.

dence regarding the meanings of words are collected and integrated over time. This statistical aspect of word learning will be the focus of Study 2 where we address how different forms of evidence are integrated over time during word learning.

Study 2: Aggregating FOV frames to predict learning of individual words from individual infants

Here we use our model to quantify the quality of visual information present in an individual infant's FOV during labeling moments, so that we may study the integration of visual sensory information during statistical word learning. To do this, we make use of the model's own internal sense of uncertainty in its classifications for each input frame. For every input im-

age the network will assign a final confidence score which it will use to determine learned vs. not-learned classifications (see the red and blue units in the "binary classifier" section in Figure 2). When the score for "learned" classification is greater than the score for "not-learned" (i.e. > 0.5), the model classifies the input image as "learned", and vice versa in the case of "not-learned".

While the accuracy results reported in Study 1 reflect the model's general performance characteristics in predicting learning, they do not reflect the model's internal sense of how strongly a frame is associated with "learned" vs. "not-learned" outcomes. Even in the frames correctly classified as "learned," some frames may be more strongly associated with "learned" outcomes compared to others, reflecting a greater association between the visual features in that frame and successful word learning. When a frame has a large confidence score associated to it, we interpret this to mean the sensory evidence contained within the frame is more strongly associated with learning. This is because the target of the referent uttered at that specific labeling moment ended up being learned given the specific sensory information contained within that frame.

Information Integration During Word Learning When these confidence scores are taken in aggregate across all the frames associated with a subject-object pair, they serve as a weighted measure of the accumulated sensory information present in the visual input from interactions associated with each subject and object pair. In contrast to analyses in Study 1 which only looked at word learning in-the-moment, classifying single frames as learned vs. not learned, the analyses in Study 2 allow us to study information integration over time, incorporating both labelling frequency and quality effects in our modeling.

We use our model to study three different forms of evidence accumulation, determining the type of a labeling instance (informative vs. misleading) on the basis of our model's classification accuracy and using the confidence values associated with each frame to model the degree of informativeness for any specific interaction. Here different subjects will have different number of frames associated with each object, reflecting differences in the *quantity* of labeling input across subject-object pairs. To model differences in the *quality* of their visual input, we use these confidence values to score frames associated with each subject-object pair for their degree of association with learning outcomes. Here we modeling 3 different types of evidence accumulation during statistical word learning.

Correct Classifications To model the accumulation of sensory evidence, both highly informative and ambiguous, we accumulate the scores for all the frames that were *correctly* classified by the model, reflecting the total degree of sensory evidence that was positively associated to learning at the individual subject-object level.

Incorrect Classifications To model the accumulation of misleading sensory evidence, we accumulate the scores for all the frames that were *incorrectly* classified by the model. This reflects the total amount of misleading evidence, that is, when frames were visually similar to successful word learning moments but did not actually lead to successful learning outcomes, or alternatively frames that were similar to not-learned instances but where these subjects in fact ended up learning these words given this specific input.

Integrated Evidence To model the integration of both informative as well as misleading sensory evidence, we accumulate the confidence scores for all the frames which were *correctly* classified by the model and then subtract the accumulated scores for all the *incorrectly* classified frames. This is meant to reflect the interaction of both positive and negative evidence during statistical word learning as misleading information is integrated with high quality visual evidence over time.

Results

We plot the accumulated sensory evidence for the 3 different integration types in Figure 4, along with the associated ground-truth learning outcomes for each subject-object pair. We also include model comparisons of 3 different logistic fits predicting individual subject-object learning outcomes using confidence values from the 3 different evidence accumulation types. We find that a model that integrates both negative as well as positive evidence is a better fit to the individual subjects' learning outcomes than models that use either positive or negative evidence alone. Moreover, a two-sided Mann-Whitney U test indicated that, at the per-subject/item level, the raw number of frames (which reflects the frequency of an individual object's naming) was not significantly different for learned vs. not learned items ($U = 993.0$, $p = 0.43$). This suggests that the specific *quality* of the sensory information, rather than the mere labeling frequency, may better explain the learning effects we observe here.

Discussion

We used computational modeling to quantifying infants' visual sensory experiences and track learning progress during free-flowing, naturalistic parent-infant interactions – a type of analysis that would be impossible with conventional behavioral studies. In Study 1, we trained a computational model to classify word learning outcomes using images taken from infants' FOV and showed that it can generalize beyond its training set to predict infant word learning on never-before-seen labeling events. In Study 2, we quantified the real-time sensory evidence in free-flowing interactions by using the model's internal degrees of uncertainty as a measure of the quality of information in each naming event. This allowed us to study statistical word learning at the sensorimotor level, by tracking how infants may accumulate and integrate informative vs. misleading evidence over time. We found that a model which incorporates both positive and mislead-

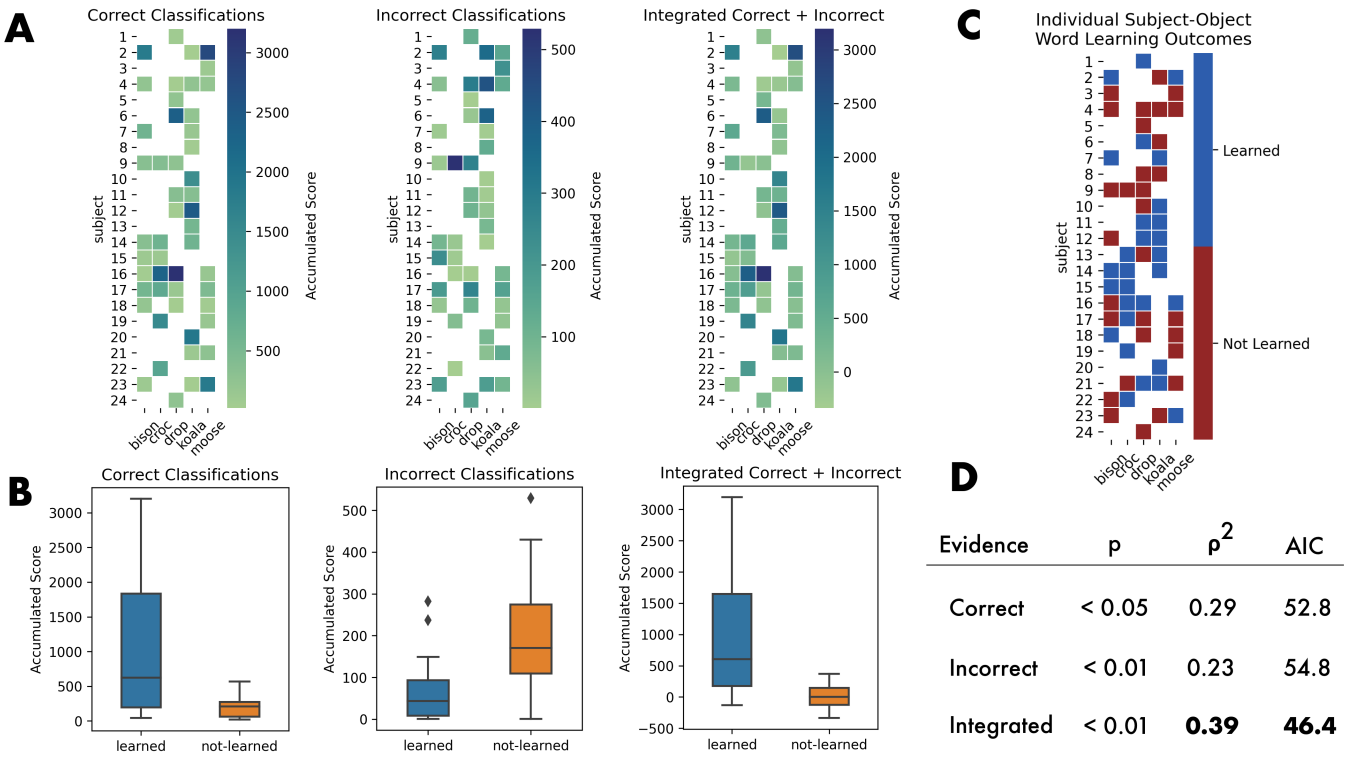

Figure 4: Using the model to quantify the relative quality of sensory evidence at the individual subject/object level. A) Individual differences in the relative rates of evidence accumulation for 3 different forms of accumulation scores - correct, incorrect, and integrated. B) Accumulated evidence for the 3 types of accumulation at the group level, collapsing across subjects and objects. C) Ground truth learning outcomes for all subject-object pairs. White squares represent missing data for a specific pair. D) Model comparisons of logistic fits across the three types of evidence accumulation. The model which integrates both correct and incorrect scores together is a better fit to the individual word learning outcomes than models that accumulate scores from either type of classification alone.

ing evidence accumulation is better at predicting individual subject-object word learning outcomes than models which only include positive or only negative evidence accumulation alone, consistent with prior work studying infant word learning (Zhang & Yu, 2017). Moreover, we find that mere labeling frequency is unable to account for this learning effect, and that we are only able to successfully simulate infants' word learning outcomes by considering the specific *quality* of the sensory information available to individual infants.

We used a novel classification-based approach to begin studying the visual properties associated with informative as well as misleading word learning moments. To our knowledge, this is the first computational model to successfully associate infants' raw, first-person visual input with their own word learning outcomes. While these machine learning methods have regularly been used in functional neuroimaging work to model complex patterns of neural activity (F. Pereira, Mitchell, & Botvinick, 2009), few studies have leveraged this class of techniques to study real-world language learning. With steady improvements in technology, we are able to collect high density behavioral and sensory data which require similar improvements in our analytical approaches. Rather than reduce the complexity of natural data for the sake of simple statistical models, classifiers like the ones we have presented here have started to find use in discovering meaningful patterns in rich multimodal data associated with language learning (Piazza, Iordan, & Lew-Williams, 2017; Ludusan, Mazuka, & Dupoux, 2020; Amatuni & Yu, 2020).

Nonetheless, there are limitations to our approach. One limitation concerns the use of the pretrained CNN in our classifier. While pretraining allows us to bypass early visual sensory learning (i.e. learning about edges, textures and shapes, which is not the focus of the present work), the pretraining likely introduces biases which are specific to both the pretraining dataset (e.g. object-centered scenes in non-naturalistic poses in ImageNet) as well as the pretraining task (i.e. image classification). In future analyses we hope to mitigate this bias by training networks from scratch using infants' own first person experiences. Another limitation is that our current work only approximates the input for early language learning. In the course of real language learning, multimodal and social information beyond the visual modality plays a crucial role in learning — e.g., the content of parent speech or whether the infant is holding the labeled object. Our current

paper focuses purely on the visual domain, using a deep CNN as a model of complex sensory input in the visual modality. In future work we hope to incorporate multiple sources of sensory and social information in a unified computational model of multimodal word learning.

Here we used a computational model to successfully associate visual properties in infants' FOV with infants' word learning, allowing us to model ideal visual contexts for learning words. We accomplished this by training a deep CNN classifier to discriminate whether or not a frame taken from infants' FOV within a free-flowing parent-infant interaction would lead to word learning by the infant. Our model used only images collected from infants' egocentric scenes during naming instances, the same visual information an infant has access to. Further, we show that both positive and negative evidence may play a role in language learning at the sensory level, demonstrating how a model which integrates both forms of evidence is a better fit to individual learning outcomes than models that look at only positive or only negative sensory evidence alone. Our results suggest that complex sensory experience, rather than being a problem for language learning, may be a critical part of the solution, and that we may better understand it when we look at how it is grounded in infants' sensory input.

Acknowledgments

This work is supported by National Institute of Child Health and Human Development R01HD074601 and R01HD093792.

References

Amatuni, A., & Yu, C. (2020). Decoding eye movements in cross-situational word learning via tensor component analysis. In *Proceedings of the 42nd annual meeting of the cognitive science society.*

Bergelson, E., & Aslin, R. N. (2017). Nature and origins of the lexicon in 6-mo-olds. *Proceedings of the National Academy of Sciences*, *114*(49), 12916–12921.

Blythe, R. A., Smith, K., & Smith, A. D. (2010). Learning times for large lexicons through cross-situational learning. *Cognitive Science*, *34*(4), 620–642.

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* (pp. 248–255).

Fazly, A., Alishahi, A., & Stevenson, S. (2010). A probabilistic computational model of cross-situational word learning. *Cognitive Science*, *34*(6), 1017–1063.

Frank, M. C., Goodman, N. D., & Tenenbaum, J. B. (2009). Using speakers' referential intentions to model early cross-situational word learning. *Psychological science*, *20*(5), 578–585.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 770–778).

Kachergis, G., Yu, C., & Shiffrin, R. M. (2012a). An associative model of adaptive inference for learning word–referent mappings. *Psychonomic bulletin & review*, *19*(2), 317–324.

Kachergis, G., Yu, C., & Shiffrin, R. M. (2012b). Cross-situational word learning is better modeled by associations than hypotheses. In *2012 ieee international conference on development and learning and epigenetic robotics (icdl)* (pp. 1–6).

Ludusan, B., Mazuka, R., & Dupoux, E. (2020). Does infant-directed speech help phonetic learning? a machine learning investigation.

Pereira, A. F., Smith, L. B., & Yu, C. (2014). A bottom-up view of toddler word learning. *Psychonomic bulletin & review*, *21*(1), 178–185.

Pereira, F., Mitchell, T., & Botvinick, M. (2009). Machine learning classifiers and fmri: a tutorial overview. *Neuroimage*, *45*(1), S199–S209.

Piazza, E. A., Iordan, M. C., & Lew-Williams, C. (2017). Mothers consistently alter their unique vocal fingerprints when communicating with infants. *Current Biology*, *27*(20), 3162–3167.

Plunkett, K., Sinha, C., Møller, M. F., & Strandsby, O. (1992). Symbol grounding or the emergence of symbols? vocabulary growth in children and a connectionist net. *Connection Science*, *4*(3-4), 293–312.

Roy, D. K., & Pentland, A. P. (2002). Learning words from sights and sounds: A computational model. *Cognitive science*, *26*(1), 113–146.

Stevens, J. S., Gleitman, L. R., Trueswell, J. C., & Yang, C. (2017). The pursuit of word meanings. *Cognitive science*, *41*, 638–676.

Sullivan, J., Mei, M., Perfors, A., Wojcik, E., & Frank, M. C. (2021, 03). SAYCam: A Large, Longitudinal Audiovisual Dataset Recorded From The Infant's Perspective. *Open Mind*, 1-10. Retrieved from https://doi.org/10.1162/opmi_a_00039 doi: 10.1162/opmi$_{a0}$0039

Tsutsui, S., Chandrasekaran, A., Reza, M. A., Crandall, D., & Yu, C. (2020). A computational model of early word learning from the infant's point of view. In *Proceedings of the 42nd annual meeting of the cognitive science society.*

Xu, F., & Tenenbaum, J. B. (2007). Word learning as bayesian inference. *Psychological review*, *114*(2), 245.

Yu, C., & Ballard, D. H. (2007). A unified model of early word learning: Integrating statistical and social cues. *Neurocomputing*, *70*(13-15), 2149–2165.

Yu, C., & Smith, L. B. (2012). Embodied attention and word learning by toddlers. *Cognition*, *125*(2), 244–262.

Zhang, Y., & Yu, C. (2017). How misleading cues influence referential uncertainty in statistical cross-situational learning..