Learning to Associate Spoken Words and Visual Objects from Egocentric Video of Parent-infant Social Interaction

Satoshi Tsutsui[1], Arjun Chandrasekaran[2], Md. Alimoor Reza[1], David Crandall[1], and Chen Yu[1]
[1] Indiana University, Bloomington, Indiana, United States
[2] Max Planck Institute for Intelligent Systems, Germany

Abstract

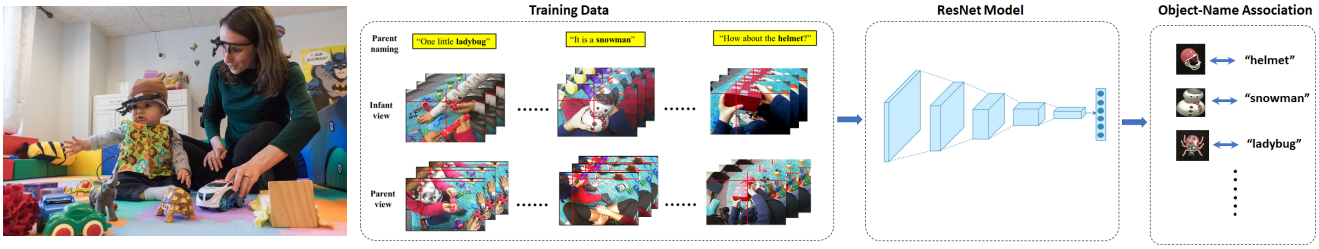
Human infants have the remarkable ability to learn the associations between object names and visual objects from inherently ambiguous experiences. Researchers in cognitive science and developmental psychology have built formal models that implement in-principle learning algorithms, and then used pre-selected and pre-cleaned datasets to test the abilities of the models to find statistical regularities in the input data. In contrast to previous modeling approaches, the present study used egocentric video and gaze data collected from infant learners during natural toy play with their parents. This allowed us to capture the learning environment from the perspective of the learner's own point of view. We then used a Convolutional Neural Network (CNN) model to process sensory data from the infant's point of view and learn name-object associations from scratch. As the first model that takes raw egocentric video to simulate infant word learning, the present study provides a proof of principle that the problem of early word learning can be solved, using actual visual data perceived by infant learners. Moreover, we conducted simulation experiments to systematically determine how attentional properties of infants' sensory experiences may affect word learning.

1. Introduction

Infants show knowledge of their first words as early as 6 months old and produce their first words at around a year. Learning object names — a major component of their early vocabularies — in everyday contexts requires young learners to not only find and recognize visual objects in view but also to map them with heard names. In such a context, infants seem to be able to learn from a sea of data relevant to object names and their referents because parents interact with and talk to their infants in various occasions — from toy play, to picture book reading, to family meal time [15].

However, if we take the young learner's point of view, we see that the task of word learning is quite challenging. Imagine an infant and parent playing with several toys jumbled together as shown in Figure 1a. When the parent names a particular toy at a particular moment, the infant perceives 2-dimensional images on the retina from a first-person point of view, as shown in Figure 1b. These images usually contain multiple objects in view. Since the learner does not yet know the name of the toy, how do they recognize all the toys in view and then infer the target to which the parent is referring? This *referential uncertainty* [10] is the classic puzzle of early word learning: because real-life learning situations are replete with objects and events, a challenge for young word learners is to recognize and identify the correct referent from many possible candidates at a given naming moment. Despite many experimental studies on infants [6] and much computational work on simulating early word learning [14], how young children solve this problem remains an open question.

Decades of research in developmental psychology and cognitive science have attempted to resolve this mystery. Researchers have designed human laboratory experiments by creating experimental training datasets and testing the abilities of human learners to learn from them [6]. In computational studies, researchers have built models that implement in-principle learning algorithms, and created training sets to test the abilities of the models to find statistical regularities in the input data. Some earlier work in modeling word learning has used sensory data collected from adult learners or robots [11, 14], while more recent models take symbolic data or simplified inputs [14]. Little is known about whether these models can scale up to address the same problems faced by infants in real-world learning. As recently pointed out in [5], the research field of cognitive modeling needs to move toward using realistic data as input because all the learning processes in human cognitive systems are sensitive to the input signals [13]. If our ultimate goal is to understand how infants learn language in the real world — not in laboratories or in simulated environment — we should model internal learning processes with natural statistics of the learning environment. This paper takes a step towards this goal and uses data collected by infants as they naturally play with toys and interact with parents.



(a) Environment

(b) Overview of our approach.

Figure 1. (a) An infant and parent play with a set of toys in a free-flowing joint play session. Both participants wore head-mounted cameras and eye trackers to record egocentric video and gaze data from their own perspectives. (b) The training data were created by extracting egocentric image frames around the moments when parents named objects in free-flowing interaction. The data was fed into ResNet models to find and associate visual objects in view with names in parent speech. As a result, the models built the associations between heard labels and visual presentations of target objects.

Recent advances in computational and sensing techniques (deep learning, wearable sensors, etc.) could revolutionize the study of cognitive modeling. In the field of machine learning, Convolutional Neural Networks (CNNs) have achieved impressive learning results and even outperform humans on some tasks [7]. In the field of computer vision, wearable miniaturized cameras have been used to capture an approximation of the visual field of their human wearer. Video from this egocentric point of view provides a unique perspective of the visual world that is inherently human-centric, giving a level of detail and ubiquity that may well exceed what is possible from environmental cameras in a third-person point-of-view [4]. Recently, head-mounted cameras and eye trackers have been used in developmental psychology to collect fine-grained information about what infants are seeing and doing in real time [3]. These new technologies make it feasible, for the first time, to build computational models using inputs that are very close to infants' actual sensory experiences, in order to understand the rich complexity of infants' sensory experiences available for word learning.

In the present study, we collect egocentric video and gaze data from infant learners as they and their parents naturally play with a set of toys. This allows us to capture the learning environment from the perspective of the learner's own point of view. We then build a computational system that processes this infant sensory data to learn name-object associations from scratch. As the first model taking raw egocentric video to simulate infant word learning, the present study has two primary goals. The first aim is to provide a proof of principle that the problem of early word learning can be solved using raw data. The second aim is to systematically determine the computational roles of attentional strategies that may influence word learning. This examination allows us to generate quantitative predictions which can be further tested in future experimental studies.

2. Data Collection and Word Learning Model

To closely approximate the input perceived by infants, we collected visual and audio data from everyday toy play — a context in which infants naturally learn about objects (24 toys) and their names. Following a study on infant object recognition [1], we developed and used an experimental setup in which we placed a camera on the infant's head to collect egocentric video of their field of view, as shown in Figure 1a. We also used a head-mounted eye gaze tracker to record their visual attention, which is later used to approximate acuity effects [9]. Thirty-four child-parent dyads participated in our study. Each dyad was brought into a room with toys scattered on the floor. Children and parents were told to play with the toys, without more specific directions.

Parents' speech during toy play was fully transcribed and divided into spoken utterances, each defined as a string of speech between two periods of silence lasting at least 400ms [15]. Spoken utterances containing the name of one of the objects were marked as "naming utterances" (e.g. "that's a helmet"). For each naming utterance, trained coders annotated the intended referent object. On average, parents produced 15.51 utterances per minute (σ=4.56), 4.82 of which were referential (σ=2.09). In total, the entire training dataset contains 1,459 naming utterances.

Recent studies on infant word learning show that the moments during and after hearing a word are critical for young learners to associate seen objects with heard words [15]. In light of this, we temporally aligned speech data with video data, and used a 3-sec temporal window starting from the onset of each naming utterance. Given that each naming utterance lasted about 1.5 to 2 seconds, a 3-sec window captured both the moments that infants heard the target name in parent speech and also the moments after hearing the name. For each temporal window, a total of 90 image frames (30 frames per second) were extracted. To summarize, the final training dataset consists of all the naming instances in parent-child joint play, with each instance containing a tar-

get name and a set of 90 image frames from the child's first-person camera that co-occur with the naming utterance. As shown in Figure 1b, each image typically contains multiple visual objects and the named object may or may not be in view.

To evaluate the result of word learning, we prepared a separate set of clean canonical images for each of the 24 objects varying in camera view and object size and orientation in a similar manner to previous work [2]. In particular, we took pictures of each toy from eight different points of view (45 degree rotations around the vertical axis), which allowed us to examine whether the models generalized the learned names to novel visual instances from a substantially different data distribution. During test, we presented one image at a time to a trained model and measured whether the model could generate the correct label for the test image.

As a model of child word learning mechanism, we use a state-of-the-art CNN model, ResNet50 [8], and trained with stochastic gradient descent (SGD). Because training was stochastic, there is natural variation across training runs; we thus ran each of our experiments 10 times and report means and standard deviations

3. Experiments and Results

3.1. Study 1: Learning object names from raw egocentric video

The aim of Study 1 is to demonstrate that a state-of-the-art machine learning model can be trained to associate object names with visual objects by using egocentric data closely approximating sensory experiences of infant learners. We also evaluated models learned with parent view data in order to compare the informativeness of these different views. Moreover, to examine the impact of properties of the training data, we created several simulation conditions by sub-sampling the whole set into seven subsets with different numbers of naming events (50, 100, 200, 400, 600, 800, 1459). While we expected that more naming instances would lead to better learning, we sought to quantify this relationship.

Figure 2 reveals two noticeable patterns in the models trained on the infant data and the model trained on the parent data. First, when there are 200 or more naming events, models trained with infant data consistently outperformed the same models trained on parent data. Second, as the quantity of training data increased, the models trained on infant data obtained better performance while the models trained on the parent data saturated. Taken together, the results here provide convincing evidence that the model can solve the name-object mapping problem from raw video, and that the infant data contain certain properties leading to better word learning.

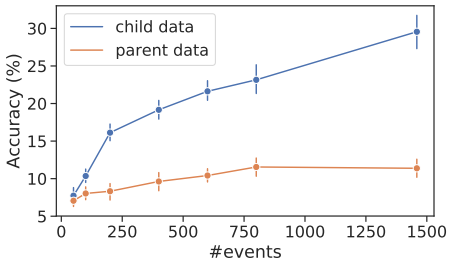


Figure 2. Results from models trained with infant data improve with more naming instances, while the models trained with the parent data show no improvement.

3.2. Study 2: Examining the effects of different attentional strategies

Humans perform an average of approximately three eye movements per second because our visual system actively selects visual information which is then fed into internal cognitive and learning processes. Thus during the 3-second window during and after hearing a naming utterance, an infant learner may generate multiple looks on different objects in view, or, alternatively, they may sustain their attention on one object during the whole time. The aim of Study 2 is to investigate whether different attention strategies during naming events influence word learning, and if so, in which ways.

To answer these questions, we first assigned each naming event into one of two categories: *sustained attention* if the infant attended to a single object for more than 60% of the frames in the naming event, and otherwise *distributed attention*. This split resulted in 750 sustained attention (SA) and 709 distributed attention (DA) events. In either case, the infant may or may not attend to the named object because the definition is based on the distribution of infant attention, *not* on which objects were attended in a naming event. We trained two identical models, one on SA instances and one on DA instances. The results in Figure 3 reveal that the model trained with sustained attention events outperformed the model trained with distributed attention events, suggesting that sustained attention on a single object while hearing a name leads to better learning.

Of course, infants may or may not show sustained attention on the object actually named in parent speech. In total, infants attended to the target in 452 out of 750 SA events, and attended to a non-target object in the other 298 SA events. Attending to the target object with sustained attention should help learning while sustained attention on a non-target object should hinder learning. To test this prediction, we sub-sampled 298 on-target events from 452 SA events, and compared them with the remaining 298 on-non-target events. As shown in Figure 3, the model trained with the on-target events achieved significantly higher accuracy

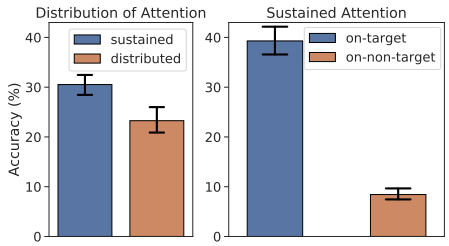


Figure 3. The model trained with sustained attention events outperformed the model trained with distributed attention events. Within the sustained attention events, the model trained with on-target instances outperformed the model trained with on-non-target instances.

than the model trained on on-non-target events.

In everyday learning contexts such as toy play, young learners do not passively perceive information from the environment; instead, the visual input to internal learning processes is highly selective moment-to-moment. The ability to sustain attention in such contexts is critical for early development and has been linked to healthy developmental outcomes [12]. The results from the present study suggests a pathway through which sustained attention during parent naming moments creates sensory experiences that facilitate word learning.

4. Conclusion

Despite the fact that the referential uncertainty problem in word learning was originally proposed as a philosophical puzzle, infant learners need to solve this problem at the sensory level. From the infant's point of view, learning object names begins with hearing an object label while perceiving a visual scene having multiple objects in view. However, many computational models on language learning use simple data pre-selected and/or pre-cleaned to evaluate the theoretical ideas of learning mechanisms instantiated by the models. We argue that to obtain a complete understanding of learning mechanisms, we need to examine not only the mechanisms themselves but also the data on which those mechanisms operate. For infant learners, the data input to their internal processes are those that make contact with their sensory systems. Using egocentric video and head-mounted eye tracking, the present study is the first, to our knowledge, to use actual visual data from the infant's point of view to reconstruct infants' sensory experiences and to show how a computational model can solve the referential uncertainty problem with the information available to infant learners. Our findings show that the available information from the infant's point of view is sufficient for a machine learning model to successfully associate object names with visual objects. Moreover, our findings here provide a sensory account of the role of sustained attention in early word learning. Previous research showed that infant sustained attention at naming moments during joint play is a strong predictor of later vocabulary [16]. The results here offer a mechanistic explanation that the moments of sustained attention during parent naming provide better visual input for early word learning compared with the moments when infants show more distributed attention.

References

[1] Sven Bambach, David Crandall, Linda Smith, and Chen Yu. Toddler-inspired visual object learning. In *NeurIPS*, 2018.

[2] Sven Bambach, David J Crandall, Linda B Smith, and Chen Yu. Active viewing in toddlers facilitates visual object learning: An egocentric vision approach. In *CogSci*, 2016.

[3] Sven Bambach, David J. Crandall, Linda B. Smith, and Chen Yu. An egocentric perspective on active vision and visual object learning in toddlers. In *ICDL*, 2017.

[4] A. Betancourt, P. Morerio, C. S. Regazzoni, and M. Rauterberg. The evolution of first person vision methods: A survey. *IEEE Trans. Circuits Syst. Video Technol.*, 25(5):744–760, 2015.

[5] Emmanuel Dupoux. Cognitive science in the era of artificial intelligence: A roadmap for reverse-engineering the infant language-learner. *Cognition*, 173:43–59, 2018.

[6] Roberta Michnick Golinkoff, Kathryn Hirsh-Pasek, Lois Bloom, Linda B Smith, Amanda L Woodward, Nameera Akhtar, Michael Tomasello, and George Hollich. *Becoming a word learner: A debate on lexical acquisition*. Oxford University Press, 2000.

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *CVPR*, 2015.

[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

[9] Jeffrey S. Perry and Wilson S. Geisler. Gaze-contingent real-time simulation of arbitrary visual fields. In *Human Vision and Electronic Imaging*, 2002.

[10] Willard Van Orman Quine. *Word and Object*. MIT press, 1960.

[11] Deb K. Roy and Alex P. Pentland. Learning words from sights and sounds: a computational model. *Cognitive Science*, 26(1):113–146, Jan 2002.

[12] Holly Alliger Ruff and Mary Klevjord Rothbart. *Attention in early development: Themes and variations*. Oxford University Press, 2001.

[13] Linda B Smith, Swapnaa Jayaraman, Elizabeth Clerkin, and Chen Yu. The developing infant creates a curriculum for statistical learning. *Trends in cognitive sciences*, 22(4):325–336, 2018.

[14] Chen Yu and Dana H Ballard. A unified model of early word learning: Integrating statistical and social cues. *Neurocomputing*, 70(13-15):2149–2165, 2007.

[15] Chen Yu and Linda B Smith. Embodied attention and word learning by toddlers. *Cognition*, 125(2):244–262, 2012.

[16] Chen Yu, Sumarga H Suanda, and Linda B Smith. Infant sustained attention but not joint attention to objects at 9 months predicts vocabulary at 12 and 15 months. *Developmental science*, 22(1):e12735, 2019.