Learning word-referent mappings and concepts from raw inputs

Wai Keen Vong (waikeen.vong@nyu.edu) and Brenden M. Lake (brenden@nyu.edu)

Center for Data Science, 60 5th Ave
New York, NY, 10011, USA

Abstract

How do children learn correspondences between the language and the world from noisy, ambiguous, naturalistic input? One hypothesis is via cross-situational learning: tracking words and their possible referents across multiple situations allows learners to disambiguate correct word-referent mappings (Yu & Smith, 2007). However, previous models of cross-situational word learning operate on highly simplified representations, side-stepping two important aspects of the actual learning problem. First, how can word-referent mappings be learned from raw inputs such as images? Second, how can these learned mappings generalize to novel instances of a known word? In this paper, we present a neural network model trained from scratch via self-supervision that takes in raw images and words as inputs, and show that it can learn word-referent mappings from fully ambiguous scenes and utterances through cross-situational learning. In addition, the model generalizes to novel word instances, locates referents of words in a scene, and shows a preference for mutual exclusivity.

Keywords: cross-situational word learning; word learning; deep learning; self-supervised learning; multi-modal learning

Introduction

Children must learn the meaning of words from noisy, sparse, and ambiguous information distributed across multiple modalities. Despite the computational challenges, children learn words at an impressive rate, estimated at upwards of ten per day on average between when they start speaking until the end of high school (Bloom, 2002). A key factor in understanding this efficiency is cross-situational learning: by tracking the co-occurrences between words and their referents across many individually ambiguous situations, learners can rapidly hone the meanings of words. Considerable evidence for cross-situational word learning has been found in laboratory studies of both adults (Yu & Smith, 2007; Kachergis, 2018; Yurovsky & Frank, 2015) and in infants (Smith & Yu, 2008).

There exist a variety of computational models of cross-situational word learning that provide theoretical accounts for the large body of empirical phenomena. Accounts based on "associative learning" track the observed statistics between words and referents across situations to determine the most plausible links (Kachergis, Yu, & Shiffrin, 2012; Fazly, Alishahi, & Stevenson, 2010). In contrast, accounts based on "hypothesis testing" consider only a limited number of hypotheses between words and possible referents (Trueswell, Medina, Hafri, & Gleitman, 2013). A third set of models use Bayesian approaches to infer lexicons with high posterior probability, assuming that words are intentionally selected based on the objects in a given situation (Frank, Goodman, & Tenenbaum, 2009; Yurovsky & Frank, 2015).

Despite their successes, each of these models is limited by the simplicity of the assumed input representation. These models observe objects and words that are parsed from their raw forms and encoded into simplified symbolic representations that can be directly manipulated, side-stepping the question of how cross-situational learning can proceed from raw sensory inputs. Additionally, because these simplified representations are discretized, it becomes challenging to explain how learners can generalize to novel instances of words they have learned (Lewis & Frank, 2013; Taxitari, Twomey, Westermann, & Mani, 2019). A complete computational account of cross-situational word learning should explain generalization to novel instances of words, even when learning from the raw inputs of individually ambiguous scenes.

In this paper, we present a neural network model that learns word-referent mappings from ambiguous scenes presented as pixel-level images. We leverage recent ideas from self-supervised learning to train a model on a proxy supervised task and show that as a byproduct, the model learns representations that can detect the correct correspondences between objects and words. Our model is intended to be a computational-level account (Marr, 1982) of cross-situational word learning—demonstrating a means of solving the computational problems outlined above—rather than providing a step-by-step process-level account of learning. We show that our model can capture four kinds of phenomena related to word and concept learning: (1) learning correct word-referent mappings from fully ambiguous images and words; (2) generalizing to novel instances of words; (3) determining the particular location of a referent in a scene; and (4) generalizing with a preference for mutual exclusivity.

Model

Our model takes two inputs: an image of a **scene** containing some number of objects, and a **caption** containing an array with the same number of words. Example scenes and captions are illustrated in the top of Figure 1. The caption associated with each scene can be either *matching*, where all of the words match with the objects in the scene, or *mismatching*, where at least one of the words does not match with one of the objects in the scene.

During learning, the model is trained to predict whether or not a scene and caption match. Our aim is that training on this discrimination task will produce representations that properly disambiguate the correct word-referent mappings from the training data. However, this set-up raises an important question: where do mismatching scenes and captions come from, if a child only experiences positive examples (matching scenes and captions)? One possibility is that the learner can implicitly construct mismatching scenes by re-

Figure 1: **Scenes and captions used for training (top), evaluation (middle) and testing for mutual exclusivity (bottom).** During training, the model is presented with a set of matching scenes and captions (illustrated with the same color), containing MNIST digits arranged in a 2x2 grid. Mismatching captions are created by permutation, and the model learns to discriminate between matching vs. mismatching scenes and captions. During evaluation, the trained model sees a novel scene and a single target word, and selects the location of the attention map with the highest value as its response for determining the target referent. To test for mutual exclusivity, we created a separate training set of scenes and captions that excluded a single digit, and then added either a single matching scene and caption containing the novel digit (0), or provided five additional mismatches, and examined the model's preference for the novel digit after training.

playing past scenes and combining them with captions from other scenes, as illustrated in Figure 1.[1] As the process of generating mismatching trials involves recombining the input data in a novel manner, and turning this from an unsupervised learning task to a supervised learning task, this is a form of *self-supervised learning* (Goyal, Mahajan, Gupta, & Misra, 2019).

The model architecture is inspired by the Audio-Visual Object Localization model (AVOL-net) from Arandjelovic and Zisserman (2018). In their work, the architecture is used to train a model that detects correspondences between images and audio sounds (not necessarily language), and can deter-

---

[1]This method of generating mismatching scenes assumes that the only words that match a given scene are the ones it was originally paired with, and not any other caption when performing this mismatching procedure, and provides a slight inductive bias towards mutual exclusivity.

Figure 2: **Model Architecture**. The network takes a scene (splitting it into four quadrants) and a caption as inputs, and embeds the information from each of these modalities using a CNN and Embedding layer respectively. It then applies a pairwise scalar product operation to detect correspondences, combining this information to produce attention maps that can be used to visualize word-referent mappings. As output, the network predicts whether a given scene and caption are matching or mismatching.

mine which parts of a particular image the sound may have come from. Using this as a starting point, we modified the architecture to match the designs used to study cross-situational word learning. A figure depicting the key elements of the architecture are shown in Figure 2.

Image and word embeddings. The scene and caption are first processed separately using an *image embedding subnetwork* (Figure 2; red shading) and *word embedding subnetwork* (green shading), respectively. As a pre-processing step, each scene x is broken up into four quadrants of equal size, and each 28x28px image x_i is passed into the image embedding subnetwork, although segmentation is not compulsory in this framework.[2] This subnetwork is a convolutional neural network, that outputs four separate image embedding vectors $u_i \in \mathbf{R}^d$, where d is the dimensionality of the embedding space. Because the same convolutional neural network is used to process each part of the image separately that contains the various objects, we call this model the "Object-CNN Model". Concurrently, each of the words w_j in the caption w is passed into the word embedding subnetwork (consisting of a single Embedding layer denoted as f_e) such that each word is represented by a word embedding vector $v_j \in \mathbf{R}^d$. The dimensionality of the word embedding vectors are designed to have the same dimensionality as the image embedding vectors. These operations are notated as follows,

$$u_i = \text{CNN}(x_i) \text{ and } v_j = f_e(w_j). \quad (1)$$

Attention maps. After computing the embeddings for each modality, the model computes a correspondence score $s_{ij} \in \mathbf{R}$ between each image embedding u_i and word embedding v_j

---

[2]Segmentation was found to reduce the sample complexity of cross-situational learning. One of our analyses compares this model to one without this pre-processing step.

via a scalar product operation. We divide the correspondence score by the square root of the size of the embedding dimensionality (Vaswani et al., 2017) and then apply a sigmoid operation to produce a bounded scalar attention score $a_{ij} \in [0,1]$,

$$s_{ij} = u_i \cdot v_j \quad \text{and} \quad a_{ij} = \sigma\left(\frac{s_{ij}}{\sqrt{d}}\right). \quad (2)$$

Using these attention scores, we can produce an **attention map** by concatenating all of the attention scores for a given word w_j, that depicts where the model believes the referents for each word are located in the scene (Figure 2; heatmaps for each label).

Output. Finally, by applying a max operator over the attention map for each word results in the sub-output $o_j \in [0,1]$, which represents the probability that the word w_j was detected in the scene. The final output of the model $o \in [0,1]$ is simply a product of all of the sub-outputs, reflecting the fact that for a match response, every word needs to be matched with a corresponding object in the scene,

$$o_j = \max_i a_{ij} \quad \text{and} \quad o = \prod_{j=1}^{k} o_j. \quad (3)$$

The model receives the correct response (match or mismatch) as binary feedback. More importantly, it does not receive any additional feedback for how the attention maps between words and objects should be organized, and must learn to find good representations to achieve this. Our goal is that by training the model on this discrimination task, we can investigate whether the learned representations can isolate the referents for each word in a manner that demonstrates cross-situational learning.

Simulations

We report extensive cross-situational learning simulations that vary key difficulty factors including scene complexity, generalization to new exemplars, and amount of training. All of the simulations were based on a synthetic dataset with generated scenes and captions, providing us with experimental control over all aspects of the evaluation. Scenes were 56x56px in size, and contained objects in some of the quadrants of the scene. The objects used in the scenes were digits from MNIST (LeCun, Bottou, Bengio, & Haffner, 1998), a database of thousands of handwritten digits and a commonly used dataset in machine learning. For each scene, we also generated captions that were **matching**, where the words in the caption matched the digits that appeared in the scene. We then generated an equivalent number of **mismatching** scenes and captions by switching the captions from its paired matching scene to a different scene, as illustrated in Figure 1.[3] We tested the model along three conditions:

---

[3]This method of generating mismatching scenes assumes that the only words that match a given scene are the ones it was originally paired with, and not any other caption when performing this mismatching procedure.

- *Scene complexity*: The referential ambiguity on each trial was manipulated, ranging from 2 digits per scene and 2 words per caption (TWO-OBJECTS), 3 digits and 3 words (THREE-OBJECTS), or 4 digits and 4 words per trial (FOUR-OBJECTS).

- *Generalization type*: In the FIXED condition, the same fixed instance of each digit was used in training and evaluation, requiring the network to generalize only to novel scenes that combine known digit instances in new ways.[4] On the other hand, in the VARYING condition, the model was presented with varying instances of the same digit sampled from the training set of MNIST, requiring the model to handle both new scenes and new instances. During evaluation, digit instances were chosen from the MNIST test set to ensure novelty.

- *Training set size*: The amount of matching word-object pairs presented to the network was also varied. In the FIXED example types, the model was presented with 36 to 720 matching word-object pairs, although in some cases we were limited by the number of possible unique combinations of captions that could be generated for some of the difficulty conditions. In the VARYING simulations, the model was presented with 36 to 3600 matching word-object pairs. An equal number of mismatching word-object pairs were also generated using the procedure described above, although in principle, many more mismatching scene and caption pairs could potentially be generated. Additionally, because we controlled for the number of matching word-object pairs, the exact number of scenes and captions for the same training set size varied across scene complexity conditions.

Training details. A few additional details are required to describe how we trained the model. The embedding dimensionality for both the image and word embedding subnetworks was set to be 64, the input size of the word embedding subnetwork was 10 (matching the number of possible digits) and the weights initialized as an identity matrix. The convolutional neural network consisted of two convolutional layers (with 16 and 32 feature maps) with 2x2 max pooling layers, followed by two fully connected layers. All of the activations in the convolutional layers and the first fully connected layer were ReLU activations, with a dropout layer (set to 50% dropout) in between the two fully connected layers. The model was trained for 1000 epochs, with 5 independent runs for each condition and the results averaged. The learning rate was set to 3e-4 and trained end-to-end with the AdamW optimizer (Loshchilov & Hutter, 2017), using a binary cross-entropy loss along with weight decay of 1e-4. The batch size used for training was 12 for all simulations, except for the varying example conditions with more than 360 matching word-object pairs, where the batch size was increased to 120.

---

[4]Arbitrarily chosen as the first instance of each digit in the MNIST training set.

Figure 3: Examples of attention maps produced by the model. Each row shows the scene (left), along with the three associated attention maps for each word in the caption for each scene (right). Lower attention scores are in purple, while higher scores are in yellow. The first two rows show that sufficient data produces highly peaked attention maps for each word in the correct locations of the scene, while the bottom two rows show that limited data results in some incorrect correspondences and fuzzier attention scores.

The model is trained from scratch, rather than using existing pre-trained representations that were trained in a supervised setting, to demonstrate that our model can indeed perform cross-situational word learning from ambiguous scenes and captions only. Additionally, the model is trained in batch, rather than on-line, as we are aiming at a computational-level rather than process-level account. Thus the network does not automatically make trial-by-trial predictions regarding behavior, similar to the model presented in Frank et al. (2009), but instead is compared on evaluation performance after training. Alternatively, behavior for varying amounts of experience can be modeled as varying the training set size, as described above.

Results

To begin, we examine whether our model learns the proxy discrimination task it is trained on. The model's training accuracy in the final epoch of training was 97.5%, when averaged across the different conditions, suggesting that it effectively learns to discriminate between matching and mismatching trials observed during training. However, and more importantly, does this result in representations that demonstrate the model isolates the correct word-referent mappings? We address this with two analyses: a qualitative analysis by presenting some attention maps generated by the model, and

a more thorough quantitative analysis to evaluate the performance of the model across the various simulation conditions using these attention maps. We also present results from two follow-up simulations investigating whether our model displays mutual exclusivity, and compare our model to a variant that does not require the pre-processing step that segments scenes.

Attention maps. Although the model discriminates between matching and mismatching training trials, does it do so by learning representations that isolate the correct underlying word-referent mappings, or by some other strategy? To investigate this, we can start by looking at the intermediate computations that produce the attention maps for each word. As shown in Figure 3, the attention map for each word visualizes the degree of correspondence the model thinks exists between each word embedding and each image embedding from the four quadrants of each scene. This provides us with a qualitative sense of the model's behavior, and where it looks at a given scene when presented with each word.

From Figure 3, we see that when the model is presented with a sufficient number of scenes and captions for training, the model learns the correct word-referent mappings where the attention is only active for the part of the scene containing the word, and zero otherwise. However, with a limited number of scenes and captions presented to the model during training, the model is both not confident about particular correspondences (even though they may be in the correct direction), and also fails to rule out incorrect associations between some words and referents.

Mapping evaluation. We also performed a more rigorous quantitative evaluation of the model's ability to learn word-referent mappings. Mimicking the evaluation procedure for behavioral experiments with children and adults (Yu & Smith, 2007; Smith & Yu, 2008), we generated novel evaluation scenes that consisted of a target digit, along with three foil digits chosen at random, and paired it with a caption containing a single word corresponding to the target digit, as illustrated in the bottom part of Figure 1. This four alternative forced-choice procedure was used regardless of the scene complexity seen at training. Each evaluation scene and caption was passed into a trained model, and used the location of the maximum value in the attention map produced to determine the model's response. If the position of the maximum attention value was the same as the location of the target digit, this indicates that the model had indeed learned the correct word-referent mapping.

We performed 100 evaluations for each trained model (consisting of ten separate evaluations per target word with different foils) across all conditions. Results are shown in Figure 4, where each plot shows the evaluation accuracy (averaged across the five training runs of each condition with different random seeds). Overall, we see that performance improves as more word-object pairs are observed for both the

Figure 4: **Evaluation performance of the Object-CNN model**. Accuracy as a function of the number of matching word-object pairs, with color representing the difficulty condition (number of objects in each scene). Error bars are 95% confidence intervals, and the dashed line indicates chance performance.

fixed and varying example types, with evaluation accuracy scores greater than 90% in each condition. Furthermore, the high evaluation accuracy achieved in the varying condition highlights our model's ability to generalize, as it suggests that for a learned word, the model can determine the correct referent using novel examples of that word that were not seen during training. However, learning to generalize concepts to novel instances requires an order of magnitude more data (3600 matching word-object pairs) than merely learning the correct mappings in novel scenes.

We also find that evaluation performance consistently decreases with increasing scene complexity (increasing ambiguity per scene), matching empirical studies of crosssituational word learning with adults (Yu & Smith, 2007). Furthermore, although the experimental design of Yu and Smith (2007) was slightly different to our simulations (18 different categories with 54 word-object pairs), they observed evaluation accuracy scores of 89%, 72% and 56% for the Two-Object, Three-Object and Four-Object conditions respectively in their task. (fixed examples only). We observed a qualitatively similar pattern of results in the limited data case with fixed examples, with evaluation performance of 93%, 72% and 58% respectively with 72 matching word-object pairs. It is surprising that the neural network and human participants achieve similar levels of accuracy for a limited number of word-object pairs, given that neural networks are notoriously data hungry (Geman, Bienenstock, & Doursat, 1992).

Mutual exclusivity. One of the hallmarks from both children's early word learning and empirical findings of crosssituational word learning is mutual exclusivity, the preference to map a novel word onto a novel object (Markman & Wachtel, 1988; M. Lewis, Cristiano, Lake, Kwan, & Frank, 2020). Many other models of cross-situational word learning can account for mutual exclusivity (Kachergis et al., 2012; Frank et al., 2009; McMurray, Horst, & Samuelson, 2012) although standard deep neural networks struggle with this type of reasoning (Gandhi & Lake, 2019), and thus we were interested in examining whether our model captures this important phenomenon. We conducted a separate set of simulations to determine whether or not our model displays a preference for mutual exclusivity.

We generated a separate set of training trials consisting of 72 matching word-object pairs, with two objects per scene using fixed examples, and excluded one digit from both scenes and captions from this training set.[5] To evaluate mutual exclusivity, the model was also provided with an additional training trial (or trials) involving the novel digit paired with a familiar foil digit, simulating the developmental paradigm where children are asked to "Show me the dax" when given a novel and familiar object (Markman & Wachtel, 1988) (see Figure 1 bottom). In the *Match Only* condition, we provided the model with this single additional mutual exclusivity trial, consisting of a novel digit and a foil digit along with the novel word as the caption, treating this as a matching trial. In the *Match plus Mismatch* condition, in addition to the single matching trial with the novel digit and word, we also paired the caption containing the novel word with five of the other training scenes (that only contained other digits) to create additional mismatched trials. The models were trained for 500 epochs, rather than 1000 epochs, but was otherwise trained in exactly the same manner as described earlier. For each excluded digit, we performed 10 independent runs.

To determine whether or not a trained model displays a preference for mutual exclusivity, we first examined whether the model produced the correct match output response for the mutual exclusivity trial, and from this, calculated the proportion of simulations where the model's attention for the novel word was higher for the novel digit than for the foil digit. In the *Match Only* simulations, the model's preference for the novel digit was 51%, suggesting that providing this single additional matching trial did not result in any preference for mutual exclusivity. On the other hand, the *Match plus Mismatch* simulations showed a greater preference for mutual exclusivity, with 73% of runs that favored the novel digit. This suggests that augmenting the model with a few additional mismatched trials with the novel word was sufficient to induce mutual exclusivity in our model, without requiring any additional changes.[6]

Cross-situational learning without segmentation preprocessing. A stated goal of this work is to learn wordreferent mappings from raw images, but the networks so far rely upon pre-processing to segment raw scenes into a set of candidate referent images. Children are not provided with such signals in realistic learning environments, and for the fi-

---

[5]This condition was chosen since this amount of training data was sufficient for very high evaluation performance in our main simulations.

[6]In the simulations for the *Match plus Mismatch* condition, we observed some instabilities during training that led to around 5% of runs showing a preference for the blank quadrants, rather than the novel or foil digit, which were excluded when calculating the preference for ME.

Figure 5: **Evaluation accuracy comparing the Object-CNN and Scene-CNN architectures**, with error bars showing 95% confidence intervals. Results shown are from the 3600 matching word-object pairs with varying examples (dashed line indicates chance performance), showing that while the Scene-CNN model can identify word-referent pairs after training without pre-processing the objects in the scene, its performance decreases with increasing ambiguity.

nal analysis, we examine whether or not this pre-processing is essential to our approach. Instead of the Object-CNN model, we consider a Scene-CNN model that takes in the full 56x56px scene as a single input, and applies a different convolutional neural network that outputs four image embeddings u_i as before. The Scene-CNN model was trained in exactly the same manner as the Object-CNN model described earlier on the 3600 matching word-object pair condition with varying examples, and then performed the same evaluation as described earlier. The results are shown in Figure 5, and show that while the performance of the Scene-CNN model decreases more than the Object-CNN model with the increasing scene complexity, performance is still far above chance suggesting that it can also learn correct word-referent mappings without additional segmentation.[7] Crucially, it shows that the approach can perform cross-situational learning from raw, unsegmented images of the scene. An interesting open question is whether an architecture that first performs object detection, or one that operates over the entire scene would scale to more realistic kinds of naturalistic data a child may encounter.

Discussion

We present a computational-level account of cross-situational word learning from images and words using a self-supervised learning approach. Our model provides a computational account for learning word-referent mappings from ambiguous raw inputs (images and words), and shows generalization to novel scenes and novel exemplars of these learned words, feats not been achieved by other models of cross-situational word learning. In addition, we can localize the intended referent from a given scene through the attention maps, and show

---

[7]The Scene-CNN model was also tested with the smaller training set sizes but evaluation performance was much lower than the Object-CNN model, despite achieving similar discrimination performance during training. This suggests that the inductive bias from extracting the objects in the Object-CNN model greatly helps in learning representations that lead to cross-situational word learning.

that the model displays a slight preference for mutual exclusivity.

While our work provides a proof-of-concept that cross-situational word learning can be achieved from raw inputs, there are a number of limitations of the current model due to the idealized set-up of our training procedure, but might be interesting directions for future research. First, the model requires the number of objects in a scene to be the same as the number of words, and relaxing this assumption may require other methods of detecting correspondences across modalities that are more graded than a binary match or mismatch response. Second, in our data generation process, we did not consider the effect of noise, where matching scenes and captions may have errors (Fazly et al., 2010). Finally, our model currently takes in words as text while cross-situational word learning experiments often provide participants the words as audio. One possible method for capturing this additional detail would be to replace the word embedding layer with a second convolutional neural network that takes in an audio spectrogram as input, as demonstrated by Arandjelovic and Zisserman (2018).

This work demonstrates how simultaneous cross-situational word learning and concept learning is possible with raw inputs from scratch, yet more work is needed before models of word learning in the lab generalize to word learning in the wild. Unlike our model, by the time children start learning words they also have access to object representations and the ability to segment objects and words, and these richer representations and abilities may be advantageous for word learning (Lake, Ullman, Tenenbaum, & Gershman, 2017). One avenue for future work is to try and scale up our model to naturalistic, longitudinal video headcam datasets (Sullivan, Mei, Perfors, Wojcik, & Frank, 2020). These datasets provide rich and detailed access to the kinds of environmental statistics that children receive from a first-person perspective, and such datasets may help determine the necessary computational machinery required for word learning at scale.

References

Arandjelovic, R., & Zisserman, A. (2018). Objects that sound. In *Proceedings of the European Conference on Computer Vision (ECCV)* (pp. 435–451).

Bloom, P. (2002). *How children learn the meanings of words*. MIT Press.

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.

Gandhi, K., & Lake, B. M. (2019). Mutual exclusivity as a challenge for neural networks. *arXiv preprint arXiv:1906.10197*.

Geman, S., Bienenstock, E., & Doursat, R. (1992). Neural networks and the bias/variance dilemma. *Neural Computation*, *4*(1), 1–58.

Goyal, P., Mahajan, D., Gupta, A., & Misra, I. (2019). Scaling and benchmarking self-supervised visual representation learning. *arXiv preprint arXiv:1905.01235*.

Kachergis, G. (2018). Word learning: Associations or hypothesis testing? *Current Biology*, *28*(9), R555–R557.

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

Lake, B. M., Ullman, T. D., Tenenbaum, J. B., & Gershman, S. J. (2017). Building machines that learn and think like people. *Behavioral and Brain Sciences*, *40*.

LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, *86*(11), 2278–2324.

Lewis, & Frank, M. (2013). An integrated model of concept learning and word-concept mapping. In *Proceedings of the 35th Annual Meeting of the Cognitive Science Society* (Vol. 35).

Lewis, M., Cristiano, V., Lake, B. M., Kwan, T., & Frank, M. C. (2020). The role of developmental change and linguistic experience in the mutual exclusivity effect. *Cognition*, *198*, 104191.

Loshchilov, I., & Hutter, F. (2017). Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Markman, E. M., & Wachtel, G. F. (1988). Children's use of mutual exclusivity to constrain the meanings of words. *Cognitive Psychology*, *20*(2), 121–157.

Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information*. W. H. Freeman.

McMurray, B., Horst, J. S., & Samuelson, L. K. (2012). Word learning emerges from the interaction of online referent selection and slow associative learning. *Psychological Review*, *119*(4), 831.

Smith, L., & Yu, C. (2008). Infants rapidly learn word-referent mappings via cross-situational statistics. *Cognition*, *106*(3), 1558–1568.

Sullivan, J., Mei, M., Perfors, A., Wojcik, E. H., & Frank, M. C. (2020). SAYCam: A large, longitudinal audiovisual dataset recorded from the infants perspective. Retrieved from https://psyarxiv.com/fy8zx/

Taxitari, L., Twomey, K. E., Westermann, G., & Mani, N. (2019). The limits of infants early word learning. *Language Learning and Development*, 1-21.

Trueswell, J. C., Medina, T. N., Hafri, A., & Gleitman, L. R. (2013). Propose but verify: Fast mapping meets cross-situational word learning. *Cognitive Psychology*, *66*(1), 126–156.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... Polosukhin, I. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems* (pp. 5998–6008).

Yu, C., & Smith, L. B. (2007). Rapid word learning under uncertainty via cross-situational statistics. *Psychological Science*, *18*(5), 414–420.

Yurovsky, D., & Frank, M. C. (2015). An integrative account of constraints on cross-situational learning. *Cognition*, *145*, 53–62.