COGNITIVE SCIENCE
A Multidisciplinary Journal

Cognitive Science 46 (2022) e13122
© 2022 Cognitive Science Society LLC
ISSN: 1551-6709 online
DOI: 10.1111/cogs.13122

Cross-Situational Word Learning With Multimodal Neural Networks

Wai Keen Vong,[a] Brenden M. Lake[a,b]

[a]Center for Data Science, New York University
[b]Department of Psychology, New York University

Received 17 June 2021; received in revised form 2 December 2021; accepted 21 January 2022

Abstract

In order to learn the mappings from words to referents, children must integrate co-occurrence information across individually ambiguous pairs of scenes and utterances, a challenge known as cross-situational word learning. In machine learning, recent multimodal neural networks have been shown to learn meaningful visual-linguistic mappings from cross-situational data, as needed to solve problems such as image captioning and visual question answering. These networks are potentially appealing as cognitive models because they can learn from raw visual and linguistic stimuli, something previous cognitive models have not addressed. In this paper, we examine whether recent machine learning approaches can help explain various behavioral phenomena from the psychological literature on cross-situational word learning. We consider two variants of a multimodal neural network architecture and look at seven different phenomena associated with cross-situational word learning and word learning more generally. Our results show that these networks can learn word-referent mappings from a single epoch of training, mimicking the amount of training commonly found in cross-situational word learning experiments. Additionally, these networks capture some, but not all of the phenomena we studied, with all of the failures related to reasoning via mutual exclusivity. These results provide insight into the kinds of phenomena that arise naturally from relatively generic neural network learning algorithms, and which word learning phenomena require additional inductive biases.

Keywords: Cross-situational word learning; Word learning; Concept learning; Multimodal neural networks; Mutual exclusivity

1. Introduction

Children effortlessly acquire the meaning of words from sparse and ambiguous sights and sounds, estimated at a rate of around 10 words per day between when they start speaking until

Correspondence should be sent to Wai Keen Vong, Center for Data Science, New York University, New York, NY, 10011, USA. E-mail: waikeenvong@gmail.com

the end of high school (Bloom, 2002). How do children pull off this seemingly incredible, yet ordinary feat? One candidate explanation that has received considerable attention in the literature is cross-situational learning: the mapping of words to their intended referents can be determined by tracking the co-occurrences between words and their referents across multiple individually ambiguous situations. Considerable evidence for cross-situational word learning has been found in laboratory studies of both adults (Medina, Snedeker, Trueswell, & Gleitman, 2011; Trueswell, Medina, Hafri, & Gleitman, 2013; Yu & Smith, 2007) and children (Halberda, 2003; Smith & Yu, 2008). In addition, research on cross-situational word learning (and word learning more broadly) has led to a wide array of empirical phenomena and inductive biases associated with this kind of learning, examining the circumstances under which learners find it easier or more difficult to determine the underlying word-referent mappings in ambiguous contexts.

Within cognitive science, different types of computational models have been proposed to explain the mechanisms behind cross-situational word learning and to capture various empirical phenomena. Computational models based on *associative learning* track the co-occurrence statistics between words and referents across situations, typically taking the form of pairwise counts or associative strengths (Fazly, Alishahi, & Stevenson, 2010; Kachergis, Yu, & Shiffrin, 2012; McMurray, Horst, & Samuelson, 2012). A second class of models instead takes a *hypothesis testing*-based approach, where models only consider a single word-referent mapping at a time, and staying or switching hypotheses depending on if the observed data are consistent with the hypothesis or not (Stevens, Gleitman, Trueswell, & Yang, 2017; Trueswell et al., 2013). A third class of models uses Bayesian approaches to infer lexicons (the full set of word-referent mappings) with high posterior probability, trading off between a prior that favors a simple lexicon versus a lexicon that properly captures the observed data (Frank, Goodman, & Tenenbaum, 2009; Yurovsky & Frank, 2015). These three model classes successfully account for various behavioral phenomena, but they all share a common limitation: They operate by encoding visual referents as discrete symbols rather than their raw perceptual inputs, side-stepping a crucial aspect of how cross-situational word learning is possible in the wild.

In this paper, we look to machine learning for potential solutions to this problem. Recent advances have led to the development of multimodal neural networks that combine language and vision information and can be trained from raw data such as images and text. These networks are capable of learning a variety of vision and language tasks ranging from image captioning (Xu et al., 2015), visual question answering (Antol et al., 2015; Johnson et al., 2017), and grounded language learning (Hill, Clark, Hermann, & Blunsom, 2020).[1] One consequence of these successes is the intriguing possibility that multimodal neural networks are effectively performing large-scale cross-situational word learning and are capable of doing so from naturalistic data. Additionally, their ability to generalize to new exemplars suggests that they may address some of the shortcomings of symbolic, count-based approaches.

Despite the application-driven successes of multimodal neural networks, it is unclear how these approaches would fare as accounts of psychological processes. Which empirical phenomena from the cross-situational word learning literature can they explain? Although previous researchers have explored similar questions, they have typically focused on just one or two phenomena. For example, Chrupała, Kádár, and Alishahi (2015) evaluate their model on measures of word similarity. Other work explores individual phenomena such as fast mapping (Hill et al., 2020; Lazaridou, Bruni, & Baroni, 2014) or mutual exclusivity (Gulordava, Brochhagen, & Boleda, 2020). Moreover, each of these studies differs in the architectures and training procedures used, suggesting the need for a more comprehensive and standardized account of what these models are capable of.

We simulated two different kinds of multimodal neural networks and their ability to capture a wide range of key phenomena in cross-situational learning. We base our modeling efforts on existing, successful architectures in machine learning and natural language processing, examining the extent to which they capture empirical phenomena out of the box. Since these methods were developed for machine learning applications rather than cognitive modeling, it would be entirely unexpected if these models were to provide a complete account of the behavioral phenomena we consider (indeed, they do not). Instead, our goal is to better understand which kinds of word learning phenomena naturally emerge from this powerful model class, and which ones require additional mechanisms or inductive biases.

Overall, our results show that the two multimodal neural networks presented in this work can be trained in an online fashion and reach similar levels of accuracy as humans do from only a single epoch of training. Their apparent sample efficiency is quite surprising, considering that neural networks are notoriously data hungry (Geman, Bienenstock, & Doursat, 1992; Lake, Ullman, Tenenbaum, & Gershman, 2017) and that other associative models require far more training for successful learning (McMurray et al., 2012). We also find that these networks successfully capture a diverse set of phenomena from the literature, and yet they fail to capture a number of phenomena linked to mutual exclusivity. Although our results pertain most directly to the two networks used in our simulations, we believe these results are a representative, though not exhaustive, account of how simple multimodal networks fare as models of cross-situational word learning.

2. Model

2.1. Experiment

We start by introducing the standard design of a cross-situational learning experiment, laying out the kind of inputs to be passed into a multimodal neural network to train it, as well as how it will be evaluated. An example cross-situational word learning experiment is shown in Fig. 1a. During the training phase as shown in the top-left panel, participants are presented with multiple referents alongside multiple words on a single trial (either as text or heard through speakers), where there is ambiguity between which words map onto which referents on each trial. However, as seen in the two training trials displayed, the word “Toma” occurs twice in conjunction with the same referent (a set of colored balls) in both trials, suggesting that this is what the word “Toma” refers to. Participants then take part in an evaluation phase as shown in the top-right panel of Fig. 1a, where their knowledge of word-referent mappings is tested. Each of the words presented during training is tested individually

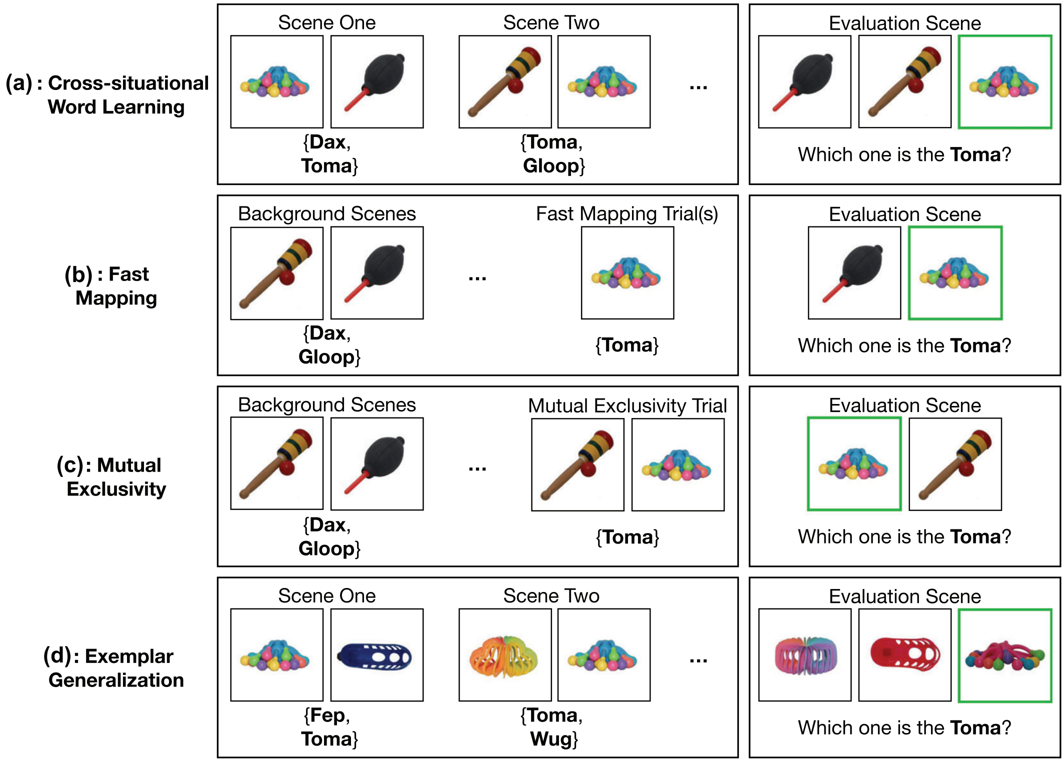

Fig. 1. The training and evaluation setup for various cross-situational word learning experiments. Starting at the top (a), the standard cross-situational word learning experiment involves presenting participants with a set of words and a set of objects per trial during training (with ambiguous alignment) and testing them on their knowledge of each word-referent mapping at evaluation time. In the fast mapping setup (b; Experiment 2), participants are presented with one or a few trials of a novel word-referent pair in an unambiguous fashion and evaluated on their ability to retain this mapping based on a limited number of presentations. In the mutual exclusivity setup (c; Experiment 3), participants are presented with a single ambiguous trial of a novel word-referent pair alongside an existing referent and evaluated on their ability to infer that the novel referent is associated with the novel word. Finally, in the exemplar generalization setup (d; Experiments 6 and 7), participants are trained exactly like the standard cross-situational learning experiment, but during evaluation time they are tested on visually similar but distinct exemplars to the learned word-referent mappings. The correct referent for each case is highlighted with a green border in each experiment type.

(and perhaps multiple times) by presenting a single target word, such as the word "Toma," alongside an array containing the target referent and a number of foil referents. Accuracy for the evaluation phase is calculated by averaging the number of correct selections of the target referent during the evaluation phase and is used as a measure of the number of word-referent mappings learned.

This particular experimental paradigm has been the dominant approach to the study of cross-situational word learning in both children and adults (Smith & Yu, 2008; Yu & Smith, 2007), as it attempts to isolate the problem of cross-situational learning to its core of

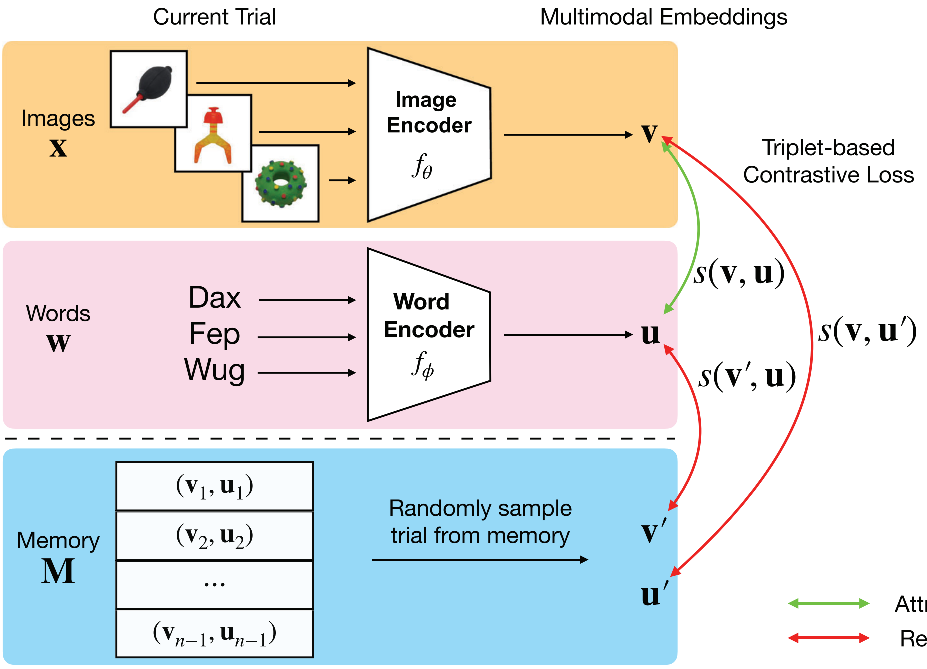
Current Trial
Multimodal Embeddings
Images
$\mathbf{x}$
Image Encoder
f_θ
$\mathbf{v}$
Triplet-based Contrastive Loss
Words
$\mathbf{w}$
Dax
Fep
Wug
Word Encoder
f_ϕ
$\mathbf{u}$
$s(\mathbf{v}, \mathbf{u})$
$s(\mathbf{v}, \mathbf{u}')$
$s(\mathbf{v}', \mathbf{u})$
Memory
$\mathbf{M}$
$(\mathbf{v}_1, \mathbf{u}_1)$
$(\mathbf{v}_2, \mathbf{u}_2)$
...
$(\mathbf{v}_{n-1}, \mathbf{u}_{n-1})$
Randomly sample trial from memory
$\mathbf{v}'$
$\mathbf{u}'$
Attract
Repel

Fig. 2. The scene-caption multimodal neural network. On each trial, the set of images $\mathbf{x}$ and the set of words $\mathbf{w}$ are passed into the image and word encoders, respectively, producing a set of image and word embeddings $(\mathbf{v}, \mathbf{u})$. Separately, we randomly sample from memory a previous set of image and word embeddings $(\mathbf{v}', \mathbf{u}')$. A triplet-based contrastive loss function is employed to bring the image and word embeddings from the current trial closer together (green arrows), while separating the image and word embeddings from separate trials (red arrows). Over the course of training, the network learns to encode images and words to produce multimodal embeddings in a manner that can disambiguate the underlying word-referent mappings. The object-word network differs by performing the contrastive loss over individual object-word pairs, rather than across the full scene.

determining the correct mappings of words to a set of referents. In particular, it simplifies some other aspects of the word learning problem: the various referents are presented as distinct objects, and the spoken language is simply a list of labels. Therefore, participants do not need to perform object detection to determine the available referents in a given scene nor do they need to extract object names from naturalistic speech.

2.2. *Architecture*

This section outlines the details of the multimodal neural network architectures used in this paper, describing how they can discover word-referent mappings from ambiguous presentations of multiple images and words. An overview of our method is shown in Fig. 2.

On each trial, the network receives as input a set of images $\mathbf{x} = [\mathbf{x}_1, \ldots, \mathbf{x}_N]$ and a set of words $\mathbf{w} = [\mathbf{w}_1, \ldots, \mathbf{w}_M]$, where N is the number of images and M is the number of words. The network encodes images using an *image encoder* f_θ and words using a *word encoder* f_ϕ, mapping images and words, respectively, into a shared multimodal embedding space consisting of d-dimensional vectors.[2] Over the course of training, images and words will be encoded into this shared representational space that disambiguates word-referent pairs, despite the inherent ambiguity present in each trial.

The *image encoder* consists of a VGG-16 convolutional neural network (CNN) pre-trained on ImageNet (Simonyan & Zisserman, 2014), with the classifier head removed and replaced with a non-linear projection head consisting of two fully connected layers (with a Rectified Linear Unit (ReLU) non-linearity in between) to map images as d-dimensional vectors.[3] The convolutional head of the image encoder is frozen, and only the projection head is learned in our network. Since the classifier head is removed and replaced with a learned projection head, this means that the output of the image encoder does not output a discrete category classification, but rather a distributed embedding representation. The image encoder is applied to each image on a given trial as follows:

$$\mathbf{v}_i = f_\theta(\mathbf{x}_i), \tag{1}$$

where $\mathbf{v}_i \in \mathbb{R}^d$. Similarly, the *word encoder* consists of a single word embedding layer, such that each word is mapped to a d-dimensional vector as follows:

$$\mathbf{u}_j = f_\phi(\mathbf{w}_j), \tag{2}$$

where $\mathbf{u}_j \in \mathbb{R}^d$. The set of image embeddings is denoted as $\mathbf{v} = [\mathbf{v}_1, \ldots, \mathbf{v}_N]$ and the set of word embeddings as $\mathbf{u} = [\mathbf{u}_1, \ldots, \mathbf{u}_M]$.

Given these two sets of embedding vectors, how does the network determine which words map onto which referents in a given trial? Since the word embeddings are randomly initialized, there is no relationship between a given word embedding to the corresponding image embedding of its matching referent at the beginning of training. One popular choice is the use of a *contrastive loss function*, which has been employed in a number of other recent multimodal architectures (Gulordava et al., 2020; Harwath et al., 2018; Lazaridou, Chrupała, Fernández, & Baroni, 2016). Although our multimodal neural networks share many similarities with previous approaches, the specific networks we use in the current paper have been simplified to allow these networks to be trained in an online fashion, mimicking the training process humans undergo in cross-situational word learning experiments. In the remainder of this section, we first provide a high-level description of how the contrastive loss function works and how it can learn word-referent mappings, followed by the additional technical details.

In supervised learning, the loss function reflects a network's ability to correctly predict the output but requires unambiguous labels to do so. On the other hand, contrastive loss functions can work with weakly labeled data by telling the network which pairs of points should be more similar to each other, and which pairs of points should be more dissimilar. In our case, there is some connection between the words and referents that appear together on the same training trial, and a contrastive loss can use this as a learning signal to embed these entities to be more similar to each other. On the other hand, the set of words from one trial typically bears no relationship to the set of referents on a separate trial, and the contrastive loss can likewise use this as another learning signal to embed these entities to be more dissimilar to each other. Even though the similarity calculation in the contrastive loss occurs across all words and referents for a given trial, the network is able to correctly discern the underlying word-referent mappings, the similarity calculations favor learning correct word-referent mappings as they result in a lower overall loss compared to when incorrect word-referent mappings are learned.

To fully specify how the contrastive loss works, the next few sections cover the remaining technical details: (1) How the network samples contrastive items to compare against, (2) how similarity is computed between words and referents, and (3) how the contrastive loss is computed.

Memory: In order to sample the contrastive items to compare against, the network has a memory $\mathbf{M} = \{(\mathbf{v}^1, \mathbf{u}^1), (\mathbf{v}^2, \mathbf{u}^2), \ldots\}$ that stores the set of observed word and referent embeddings from previous trials. On a new training trial, the network requires a set of contrasting words and referents from a previous trial $(\mathbf{v}', \mathbf{u}')$ to compare against the current embedded words and referents $(\mathbf{v}, \mathbf{u})$. This is achieved by sampling a previous trial's words and referents randomly from the memory of previous trials, for example, $(\mathbf{v}', \mathbf{u}') \sim \mathbf{M}$.[4] Our model performs a check to ensure that the set of words and referents from the sampled trial is not exactly the same as the set of words and referents from the current trial, and if so, will resample from memory until a proper mismatch is found. After each training trial, the network updates its memory by adding the current set of words and referents to a new slot in memory.[5]

Similarity: The similarity score determines how similar a given set of words is to a given set of referents. For each word $\mathbf{u}_i$, we calculate the dot product between the word embedding to all of the image embeddings. This dot product provides a scalar correspondence score for any given word and any given image, where a higher score represents a higher correspondence between a word and an image. We then take the maximum dot product for a given word for all of the possible referents (capturing the idea that each word maps to a single referent), and then apply this process across all of the other remaining words. The similarity score is calculated by taking the mean across these maximal dot products per word,[6] as shown in the equation below:

$$s(\mathbf{v}, \mathbf{u}) = \frac{1}{M} \sum_{i=1}^{M} \max_{\mathbf{v}_j \in \mathbf{v}} (\mathbf{v}_j \cdot \mathbf{u}_i). \tag{3}$$

Contrastive loss: These two components are combined in the contrastive loss function, which is comprised of three different similarity computations. First, the similarity between the current set of *matching* words and referents in the current trial is computed: $s(\mathbf{v}, \mathbf{u})$. Second, a previous trial's embeddings are sampled from memory $(\mathbf{v}', \mathbf{u}') \sim \mathbf{M}$. We then compute two additional *mismatching* similarity scores, by pairing either the current set of words with the previously observed set of referents $s(\mathbf{v}', \mathbf{u})$, or the current set of referents with the previously sampled set of words $s(\mathbf{v}, \mathbf{u}')$. Then, the contrastive loss function can be specified via the following equation:

$$\begin{aligned} L(\theta, \phi) \quad = \quad & \max(0, s(\mathbf{v}', \mathbf{u}) - s(\mathbf{v}, \mathbf{u}) + \eta) \\ & + \max(0, s(\mathbf{v}, \mathbf{u}') - s(\mathbf{v}, \mathbf{u}) + \eta). \end{aligned} \tag{4}$$

The loss function contains one hyperparameter η corresponding to a margin variable, which we set to be 1 in all of our simulations. This margin hyperparameter means that for a given matching pair $s(\mathbf{v}, \mathbf{u})$ and a mismatching pair $s(\mathbf{v}', \mathbf{u})$, the network will adjust their embeddings such that the matching similarity score is at least η larger than the mismatching similarity, $s(\mathbf{v}, \mathbf{u}) > s(\mathbf{v}', \mathbf{u}) + \eta$. Any further separation does not further decrease the loss.

Initially, as the network has not learned to associate any words with its referents, the similarities scores for the matching words and referents and the mismatching ones will be random. Over the course of learning, as the network updates its representations of words and referents on the basis of this contrastive loss function, it begins to correctly output higher similarity scores for sets of words and referents that match and lower similarity scores for sets of words and referents that mismatch. Thus, through this training process, it can acquire the correct underlying word-referent mappings.

Response function: Once the network has been trained, we can evaluate the trained network in the same fashion as a cross-situational word learning experiment. First, we present the network with a single word $\mathbf{w}^*$, and then an array containing the target referent and a number of other randomly selected foil referents $(\mathbf{x}_1, \ldots, \mathbf{x}_N)$. Second, the network separately embeds the target word $\mathbf{u}^* = f_w(\mathbf{w}^*)$, and each of the referents $\mathbf{v} = [\mathbf{v}_1, \ldots, \mathbf{v}_N] = [f_\theta(\mathbf{x}_1), \ldots, f_\theta(\mathbf{x}_N)]$. Finally, the network calculates the dot product for the target word embedding against each of the referent embeddings and selects the corresponding referent y with the highest dot product as follows:

$$y = \arg\max_{i \in \mathbf{v}} (\mathbf{u}^* \cdot \mathbf{v}_i). \tag{5}$$

2.3. Scene-caption network

We consider two slightly different variants of how the network combines the word and referent embeddings on a given trial into this triplet-based contrastive loss function. The network described in the above equations represents the *scene-caption network*, as the network combines all of the available words (caption) and the available referents (scene) in a given trial as input to the similarity function. This similarity function is closely related to the MISA (max-image sum-audio) similarity function from Harwath et al. (2018), although the max operation in our setup is performed over the set of possible referents rather than across different patches within a single image.

2.4. Object-word network

We also consider a variant of this architecture, which we call the *object-word network*, which computes all pairwise similarities between each word and potential referents instead of aggregating across a scene as in Equation 3. That is, for *each* possible word-referent pair $(\mathbf{v}_i, \mathbf{u}_j)$ on the current trial, its dot-product similarity is calculated as follows:

$$s(\mathbf{v}_i, \mathbf{u}_j) = \mathbf{v}_i \cdot \mathbf{u}_j. \tag{6}$$

Likewise, the network then will randomly sample a previously observed word-referent pair (rather than a scene-caption pair) from memory $(\mathbf{v}'_i, \mathbf{u}'_j) \sim \mathbf{M}$, where $\mathbf{M}$ now stores previously observed object-word pairs rather than scene-caption pairs. Finally, the network uses a modified version of the previous contrastive loss:

$$L(\theta, \phi) = \sum_{\mathbf{v}_i, \mathbf{u}_j \in (\mathbf{v}, \mathbf{u})} \max(0, s(\mathbf{v}'_i, \mathbf{u}_j) - s(\mathbf{v}_i, \mathbf{u}_j) + \eta) + \max(0, s(\mathbf{v}_i, \mathbf{u}'_j) - s(\mathbf{v}_i, \mathbf{u}_j) + \eta), \tag{7}$$

summing over this contrastive loss for each possible for each potential object-word pair in a given trial before a gradient update is performed. Thus, in a scene with two words and two referents, there are four possible word-referent pairs and the network will calculate the combined loss, each time sampling a new contrasting word-referent pair from memory to compare to the current word-referent pair. Again, the model checks that the sampled mismatching word-referent pair differs from the current word-referent pair, resampling if necessary. The evaluation procedure for the object-word network is the same as the scene-caption network.

2.5. *Dataset*

The images we used for these simulations were chosen from the NOUN (novel object and unusual name) database, consisting of 60 images depicting unusual objects that are commonly used in word learning experiments (Horst & Hout, 2016). Each image was resized to 224 × 224 pixels to match the required input size to the image encoder, and the output after passing an image through this encoder was a 64-dimensional embedding vector. A subset of images from the NOUN database are depicted in Fig. 2. The inputs to the word encoder are random indices for each unique word, resulting in a 64-dimensional vector representing each word from the word encoder, the same dimensionality as the visual embedding.

2.6. *Training*

For the majority of our simulations, we report results of our networks trained in an online manner. On each trial, a single set of matching words and referents is presented to the network, a mismatching set of words and referents is sampled from memory, and the network's parameters are updated via a contrastive loss function, as shown in Fig. 2. Additionally, the network was only trained for a single epoch, updating its parameters only a single time for each trial. This training procedure mimics the trial-by-trial learning found in cross-situational learning experiments, in contrast with standard, epoch-based training where the network cycles over the data many times. Despite the very limited nature of the training data and parameter updates—compared to previous associative models of cross-situational word learning (McMurray et al., 2012) and neural networks more generally (Geman et al., 1992)—we observe that our networks can indeed uncover the underlying the word-referent mappings.

All of our networks were trained using stochastic gradient descent, with a learning rate of 0.01. The results of each condition within each simulation were averaged across 20 independent runs, randomly selecting the subset of images from the NOUN database presented to the model in each run. This ensured that the resulting number of word-referent mappings learned by the network was not a by-product from any specific set of images from the database.

3. Experiments

In this section of the paper, we catalog the range of experiments we conducted with these two multimodal neural networks. We selected a broad range of empirical phenomena related to cross-situational learning and word learning. Although we covered a large range of phenomena, this is not meant to be an exhaustive list, and it would be valuable for future work to examine how well other multimodal neural networks with additional inductive biases or learning mechanisms could capture other aspects of cross-situational learning.

The seven simulations we investigated were (1) referential ambiguity, (2) fast mapping, (3) mutual exclusivity, (4) relaxation of mutual exclusivity, (5) learning from Zipfian distributions, (6) exemplar generalization, and (7) learning visual representations from scratch. For each simulation, we first describe the key empirical phenomena. We then present simulation results from both network types, examining whether they can reproduce the critical behavioral findings.

As mentioned, we had no expectation that these multimodal neural networks would capture all of the phenomena under consideration. Indeed, our aim is to catalog which findings are captured and which are not, given straightforward machine learning approaches that work at scale and address practical applications. To foreshadow our results, we find that the networks capture four out of the seven phenomena. Additionally, the three remaining phenomena the networks are unable to capture are all linked to mutual exclusivity. We address the implications of our findings in the respective sections covering these simulations and in the general discussion.

3.1. Experiment 1: Referential ambiguity

In the first set of simulations, we investigated whether or not our two multimodal neural networks could capture the referential ambiguity effect. This phenomenon refers to the degree of uncertainty for which words map onto which referents in a given scene. Increasing the number of words and referents in a given scene increases the number of potential mappings between words and referents to consider. Therefore, the increased uncertainty of determining which words map onto which referents reduces the likelihood of learning any given word-referent mapping (Yu & Smith, 2007). The training setup from the top-left panel of Fig. 1 illustrates learning with two objects and two referents per scene for example.

This phenomenon was empirically demonstrated in Experiment 1 from Yu and Smith (2007), where adult participants were presented with 18 different word-referent mappings in a standard cross-situational word learning experiment, as illustrated in the top-left panel of Fig. 1. Each word-referent mapping was presented six times over the course of training, where the degree of referential ambiguity was controlled by showing participants either two words and two referents (**2** × **2**), three words and three referents (**3** × **3**), or four words and four

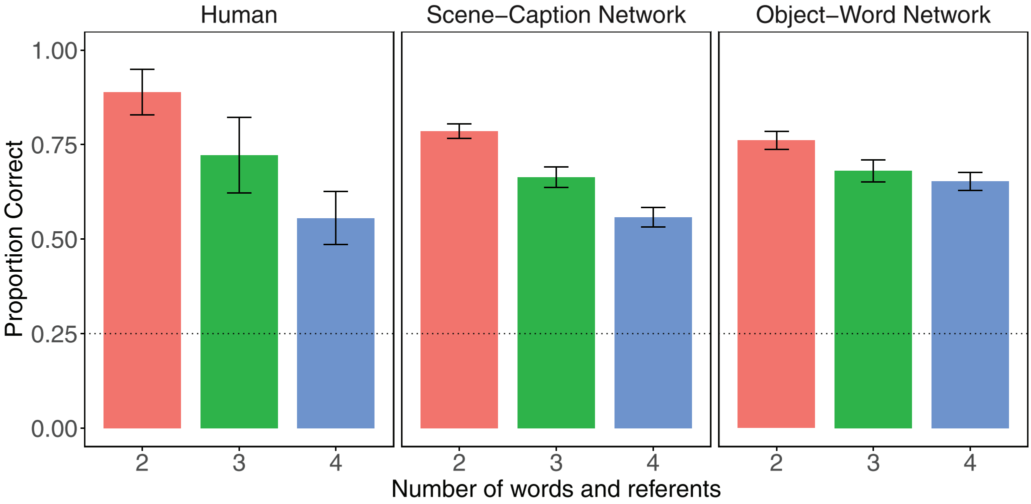

Fig. 3. Referential ambiguity simulation results. Our results show that both neural networks exhibit decreased accuracy with increasing referential ambiguity in scenes. Additionally, we find that accuracy across the referential ambiguity conditions is comparable to humans even from a single epoch of training. The dotted line represents chance accuracy, and error bars depict 95% confidence intervals.

referents (**4** × **4**) per trial.[7] After training, participants' knowledge of word-referent mappings was evaluated by presenting them with a target word along with the target referent and three other randomly selected foil referents and asking them to select the referent that matched the given word, as depicted in the top-right panel of Fig. 1. The critical finding from this study showed that the average number of word-referent mappings participants were able to learn decreased with additional referential ambiguity. As shown in Fig. 3 (left), participants learned 16 out of 18 words on average when trained on two-words two-referents. In contrast, they learned 13 out of 18 when trained on three-words three-referents, and 10 out of 18 when trained on four-words four-referents.

Simulation: The networks were trained in a manner that matched the experimental designs, in terms of the number of presentations for each word-referent mapping (six presentations each) as well as the number of words and referents per trial (two, three, or four depending on the condition). During the evaluation phase, the referent selected by each model was determined by the one whose dot-product similarity was highest for the target word on each trial.

Results: In Fig. 3, we show results from the simulations with the scene-caption and object-word networks alongside the behavioral findings from Yu and Smith (2007). As mentioned in the introduction, these results show that both multimodal neural networks are able to learn a comparable number of word-referent mappings as humans from a single epoch of training, despite receiving the same amount of trial-by-trial experience. We observed that accuracy in both models in the $\mathbf{2} \times \mathbf{2}$ condition was slightly lower than human performance but was indeed comparable for the $\mathbf{3} \times \mathbf{3}$ and $\mathbf{4} \times \mathbf{4}$ conditions.

How do both our networks obtain such high accuracy even with such limited experience? First, the contrastive loss function aims to adjust the image and word embeddings closer for words and referents presented together on the same trial and further apart for words and referents paired from different trials. Since the embeddings are high dimensional, there are many feasible ways of adjusting the embeddings to align with the observed data. Furthermore, because there is a consistent relationship between the true word-referent mappings, only a handful of gradient updates that align these pairs may be sufficient to disambiguate between correct and incorrect mappings.

In addition to achieving human-level accuracy, both networks also captured the referential ambiguity effect, showing an increased difficulty in acquiring word-referent mappings when additional words and referents at present on each trial. This is also consistent with the above explanation. In situations where there is less referential ambiguity within trials, the embeddings will update to quickly disambiguate the true word-referent pairs. However, in situations with higher referential ambiguity, the embeddings will stay consistent with multiple mappings, requiring more examples and more gradient updates to resolve the ambiguities present from the observed data. A closer look at the attention maps as shown in Fig. 4 aligns with this explanation, where low referential ambiguity situations like the $\mathbf{2} \times \mathbf{2}$ condition allow the model to easily resolve almost all of the word-referent mappings in the experiment with a single epoch of training. However, higher referential ambiguity means that the correspondence scores for each word are more diffuse to the set of referents, highlighting the increased uncertainty of determining the correct word-referent mappings. Furthermore, we also see that training the model for additional epochs allows the model to resolve almost all of the word-referent pairs regardless of the degree of referential ambiguity.

3.2. Experiment 2: Fast mapping (retention)

The second set of simulations we conducted looked at whether a multimodal neural network can retain the knowledge of word-referent mappings from a small number of exposures, as captured in studies from fast mapping (Carey, 1978; Carey & Bartlett, 1978). In a typical fast mapping experiment, participants are first presented with either one or multiple instances of a novel word and asked to choose from an array of referents which one they think the novel word refers to (*referent selection*). At some point after the initial presentation, participants are then tested on their knowledge of this novel word-referent pair and checking whether they remembered this association (*retention*). It is this latter aspect of fast mapping that we examine in this simulation, as referent selection is the focus of subsequent experiments.

Previous studies involving children show that children can flexibly map novel words to novel objects via referent selection (Carey & Bartlett, 1978; Golinkoff, Hirsh-Pasek, Bailey, & Wenger, 1992), with the ability emerging around the age of 24 months (Horst & Samuelson, 2008). However, there remains an active debate whether children can retain these fast mapped words after this initial referent selection trial. In the original studies by Carey and Bartlett (1978), children were able to retain a novel color word ("chromium") when tested

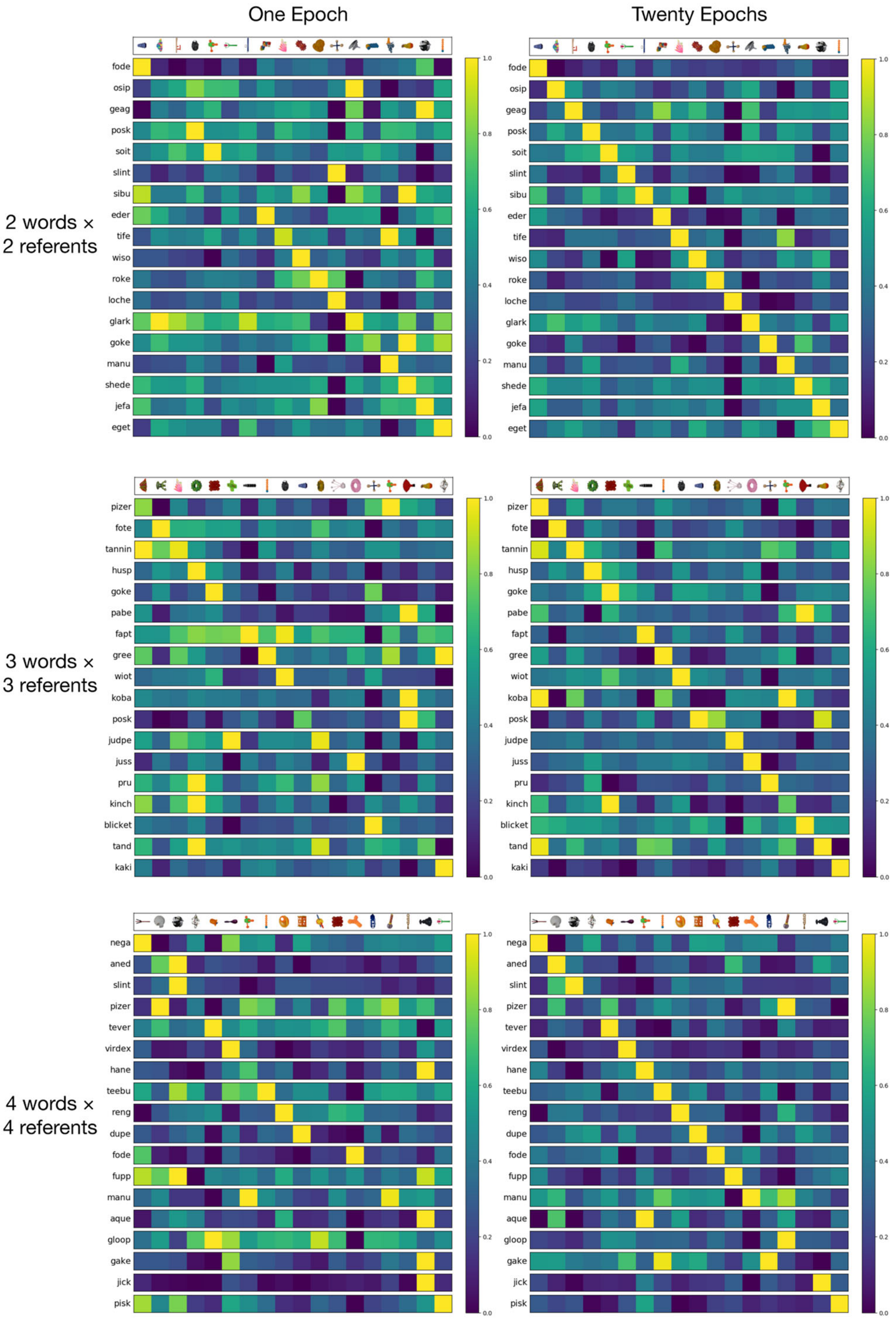
One Epoch
Twenty Epochs
2 words ×
2 referents
fode
osip
geag
posk
soit
slint
sibu
eder
tife
wiso
roke
loche
glark
goke
manu
shede
jefa
eget
3 words ×
3 referents
pizer
fote
tannin
husp
goke
pabe
fapt
gree
wiot
koba
posk
judpe
juss
pru
kinch
blicket
tand
kaki
4 words ×
4 referents
nega
aned
slint
pizer
tever
virdex
hane
teebu
reng
dupe
fode
fupp
manu
aque
gloop
gake
jick
pisk
1.0
0.8
0.6
0.4
0.2
0.0

Fig. 4. Attention map visualizations for the scene-caption network from Experiment 1. Here, we show the resulting attention maps after a single epoch of training, or after 20 epochs across the three referential ambiguity conditions. Each row indicates the degree to which each word is associated with each referent based on the dot product (with lighter colors indicating a higher correspondence and darker lowers showing a lower correspondence, and attention values scaled uniformly to lie between 0 and 1). Our results show that after a single epoch, the network learns more word-referent pairs when there is less referential ambiguity, but with sufficient training almost all pairs are resolved.

again many weeks later after the initial training session. On the other hand, Horst & Samuelson (2008) showed that even after a 5-min interval children were unable to retain any of the fast mapped words they had previously learned, except under conditions with ostensive naming events, where the experimenter provided the child with additional explicit instruction by directly holding the target referent after each naming trial in a clear and unambiguous fashion, to highlight that the novel word was linked with the novel referent, and not the other referents in the scene. Due to some of the experimental differences across these studies, we were motivated to test our models ability to perform fast mapping under the simplest conditions, focusing on a minimal number of unambiguous presentations.

Simulation: In our simulations, we examine whether multimodal neural networks can retain the knowledge of word-referent mappings by testing the network's ability to remember the correct referent for a novel word with minimal experience. The setup for our fast mapping simulations is illustrated in the second row in Fig. 1B. The first aspect of training involved presenting a single novel word-referent pair ("Toma") in an *unambiguous* manner at some point during training, such that the network only saw the novel word and the novel referent together without any other referents. One point of departure from this setup relative to empirical work studying fast mapping is that the presentations of the novel word-referent pair to the model are unambiguous. This is closest to the ostensive naming procedure from Experiment 2 of Horst and Samuelson (2008), and we chose to simulate retention via this design (rather than including additional familiar referents), to first check that these networks could indeed learn and retain word-referent mappings with minimal exposure.

We varied the number of times the network saw the novel word-referent mapping (**1**, **3**, or **5** times), as well as the timing of the presentation of these unambiguous trials (either at the *start*, *middle*, or *end* of training). As additional background training to the fast mapping trials, the networks were also trained on 10 regular word-referent pairs, with individually ambiguous trials consisting of two words and two referents. Each of these word-referent pairs was shown six times each, matching the previous simulation, where our results showed that six presentations were sufficient to reach a high degree of accuracy for learning word-referent mappings. These regular word-referent mappings served as the familiar referents upon which we tested whether our networks could demonstrate evidence of fast mapping. After the training phase, we evaluated whether the networks had retained the novel word-referent mappings (Fig. 1B, right panel) by presenting the single novel word along with the novel referent and a second foil referent (randomly selected from the set of 10 familiar referents observed during

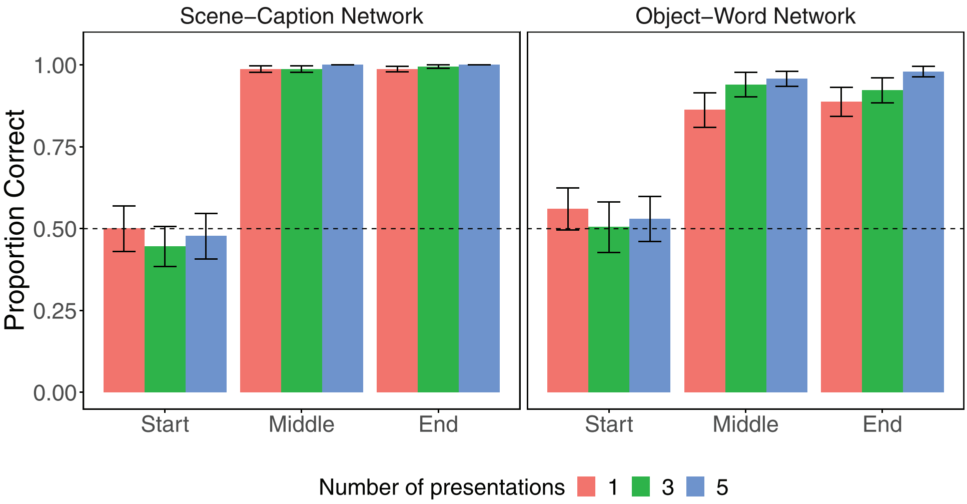

Fig. 5. Fast mapping simulation results. Both the scene-caption and object-word networks displayed evidence for fast mapping, even from a single example, but only if presented at the middle or the end of training. A slight benefit was observed for additional presentations for the object-word network too. However, in both cases, the presentation of the novel word at the beginning of training before seeing any other scenes, resulted in a failure to learn. The dotted line represents chance accuracy, and error bars depict 95% confidence intervals.

training), performing 20 separate evaluations per run. Accuracy was scored as the average proportion the network selected the novel referent instead of the foil referent.

Results: Results for Experiment 2 are shown in Fig. 5. Our results show that both networks demonstrate retention in fast mapping, if the unambiguous novel word-referent pair is presented at the middle or end of the training. Additionally, we see that even with a single example, the network correctly selects the novel referent, matching some of the empirical evidence that a single novel word used in a naturalistic context is sufficient for fast mapping (Carey & Bartlett, 1978). Higher accuracy scores were observed by providing the object-word network with additional unambiguous examples of the novel word-referent pair. These results extend the findings from Experiment 1, highlighting that multimodal neural networks can pick up word-referent mappings with minimal experience. One reason this may be accentuated in this particular experiment is that the network is provided with unambiguous information for a single word-referent pair, allowing the contrastive loss to adjust the network in a manner that makes the novel word-referent pair distinct from what was observed in the background training where other words and referents were presented together in an ambiguous fashion.

A failure was observed for both of these networks if the fast mapping trials were presented at the very beginning of training, and neither network was able to perform above chance during the evaluation phase. One explanation for this failure is a form of catastrophic forgetting (French, 1999), where the additional word-referent mappings presented after these initial presentations interfered with the existing knowledge of the novel word-referent pair. A second potential explanation is that since all of these unambiguous trials are presented at the beginning of training, the networks cannot apply contrastive learning to learn the novel word-referent pair because it does not have other examples to properly contrast against. Regardless, this issue could be alleviated with some kind of additional experience replay mechanism (McClelland, McNaughton, & O'Reilly, 1995), allowing the network to also sample previously observed trials as matching examples (rather than the current memory mechanism which only samples mismatches).

This ability to retain words from fast mapping was also recently demonstrated in Hill et al. (2020) in a more complex simulated three-dimensional environment where the agent was first presented with a novel label (when fixating on a particular novel object), and in a room with multiple other novel objects, and then afterwards asked to pick up the indicated novel object. In contrast to our setup, the additional complexity of their environment required a more specialized multimodal architecture with multiple kinds of loss functions, as well as extensive training via reinforcement learning before it could consistently demonstrate evidence of retention from fast mapping.

3.3. Experiment 3: Mutual exclusivity

For the third experiment, we explored whether multimodal neural networks capture *mutual exclusivity*, the assumption that each object has a single label associated with it (Halberda, 2003; Markman & Wachtel, 1988). The setup for our mutual exclusivity simulations is illustrated in the third row in Fig. 1C. In a typical experiment, children are presented with one familiar object and one novel object and asked to "Show me the Toma" (evaluation scene), where "Toma" is a novel word. If children select the novel object over the familiar object, they are applying the principle of mutual exclusivity: since the familiar object already has a label, the child infers that the novel word must refer to the novel object.[8] Reasoning by mutual exclusivity has been widely observed in children, with experiments showing success in this task as early as 17-month olds, with an increasing preference for mutual exclusivity as children get older (Halberda, 2003).

Simulation: Our setup for examining mutual exclusivity is illustrated in the third row of Fig. 1. Similar to the fast mapping simulations, both the scene-caption and object-word networks were trained with a set of 10 background word-referent mappings with six presentations for each word-referent pair, with two objects and two referents per trial. This set of 10 word-referent mappings served as the basis for the set of familiar objects in this experiment.

At the end of training, the networks were presented with a single mutual exclusivity trial consisting of a single novel word, along with a novel referent and another randomly selected foil referent (from one of the 10 familiar objects the network was already trained on). This provides the network with ambiguous information about which one of the two referents the novel word should be mapped to. This setup contrasts with the fast mapping simulation that unambiguously introduced a novel word attached to a single novel referent, making the mutual exclusivity task more difficult, where the model needs to recognize that the novel word should be mapped to just the novel referent. For both of the networks, we treat the presentation of

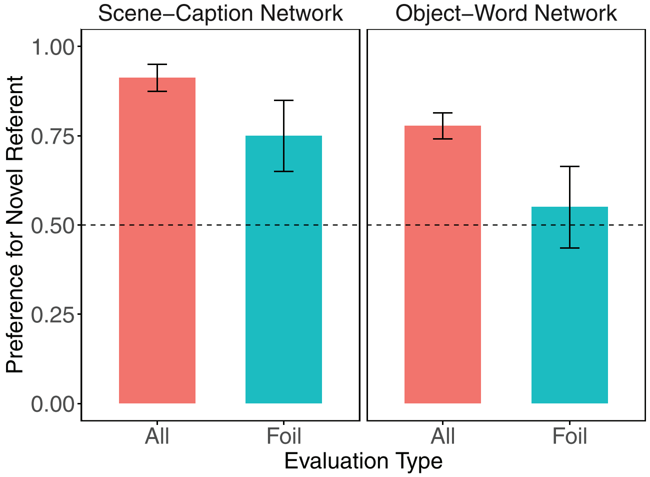

Fig. 6. Mutual exclusivity simulation results. We examined whether networks could learn a word-referent mapping from a novel word presented ambiguously with a novel and familiar referent. Both networks showed a preference for the novel referent when placed against other referents observed for training in the *All* condition, but for the tougher *Foil* condition, only the scene-caption network showed a slight preference for mutual exclusivity. The dotted line represents chance accuracy, and error bars depict 95% confidence intervals.

the novel word in an ambiguous setting as a training trial (and allow the network to perform a gradient update), in order to reflect the typical "Show me the Toma" wording of the evaluation prompt.[9] After this single mutual exclusivity trial, the network was evaluated for its preference of mutual exclusivity in two different ways. In the *All* condition, the novel word was paired with the novel referent and a randomly selected referent from training. In the more challenging *Foil* condition, the foil referent was the familiar referent that appeared with the network during the mutual exclusivity trial, and therefore success on these trials would require the network to correctly infer that the novel word mapped to the novel referent and not the foil referent that co-occurred with it. These two conditions were designed as different measures for capturing the extent to which the two multimodal neural networks could display a mutual exclusivity bias.

Results: The results of the mutual exclusivity simulations are shown in Fig. 6. Examining the results from the *All* condition, we find that both networks reliably selected the novel referent compared to a randomly selected referent from the set of other learned word-referent mappings. In the more challenging evaluation *Foil* condition, we do not observe a consistent pattern of success. The scene-caption network demonstrates evidence of mutual exclusivity on these trials, selecting the novel referent 75% of the time over the foil referent, while the object-word network shows no preference to either the novel or the foil referent. However, despite the success observed from the scene-caption network and the failure of the object-word network shown here, under a wider range of configurations varying the learning rate and degree of gradient clipping our results showed that mutual exclusivity could not be consistently demonstrated in either network, in contrast to the robust demonstrations of mutual exclusivity in children and adults (Halberda, 2006). The additional results are presented in Appendix A. This lack of mutual exclusivity has been observed in other deep neural network architectures looking at more traditional tasks such as classification (Gandhi & Lake, 2020), and it suggests that standard multimodal architectures do not capture mutual exclusivity reliably without additional mechanisms.

A recent paper by Gulordava et al. (2020) was able to reproduce the mutual exclusivity bias using a model similar to the object-word network, but with slight differences in their training procedure. In their setup, the model was allowed to sample the novel word or novel referent as negative items in the contrastive loss during the background training process, providing the model with implicit negative evidence that these novel items should not be associated with any of the background words and referents. In contrast, in our setup, the gradient update performed on the final ambiguous trial is the first time our networks obtain the relevant information about the novel word and referent. However, it is not clear whether sampling the novel word and referent pairs as negative items counts as true evidence for mutual exclusivity, as they are no longer truly "novel" when they are presented on the final mutual exclusivity trial like our setup. Another approach they explored to induce mutual exclusivity was adding an extra pragmatic reasoning step into the referent selection mechanism, by comparing the novel referent to all of the words in the vocabulary, in a similar manner to earlier work (Alishahi, Fazly, & Stevenson, 2008). However, this step may have also been influenced by observing novel items as negatives during the training process, suggesting that further work may be needed to tease out when and where multimodal neural networks can consistently display mutual exclusivity. In the next two simulations, we explore the need to capture mechanisms such as mutual exclusivity during the learning process too.

3.4. Experiment 4: Relaxation of mutual exclusivity

The fourth simulation, we examined the relaxation of mutual exclusivity effect observed in Kachergis et al. (2012), demonstrating that not only do people employ the principle of mutual exclusivity to learn word-referent mappings but that they can also relax this principle to endorse multiple mappings for a given word or referent when provided with sufficient evidence to do so. A summary of the cross-situational learning task used in this experiment is shown in Fig. 7.

First, during the *early training* phase, participants were trained on a trial drawn from a set of six word-referent pairs ($w_1 - x_1, \dots, w_6 - x_6$), with two words and two objects per trial. Second, during the *late training* phase, participants were trained on trials drawn from an additional six word-referent pairs ($w_7 - x_7, \dots, w_{12} - x_{12}$). In this second phase, some word-referent pairs (e.g., $w_7 - x_7$) always co-occurred with another word-referent pair ($w_1 - x_1$), suggesting that w_1 and x_7 are also paired in a secondary sense (likewise for w_7 and x_1). During a subsequent evaluation phase, participants were tested on each word twice. The first test asked participants to select the target referent for the word (w_1), with its early referent as the target (x_1 shown, x_7 not shown) and 10 other objects as distractors. The second test presented the same word, but swapped the target object, so participants saw the same word (w_1) but

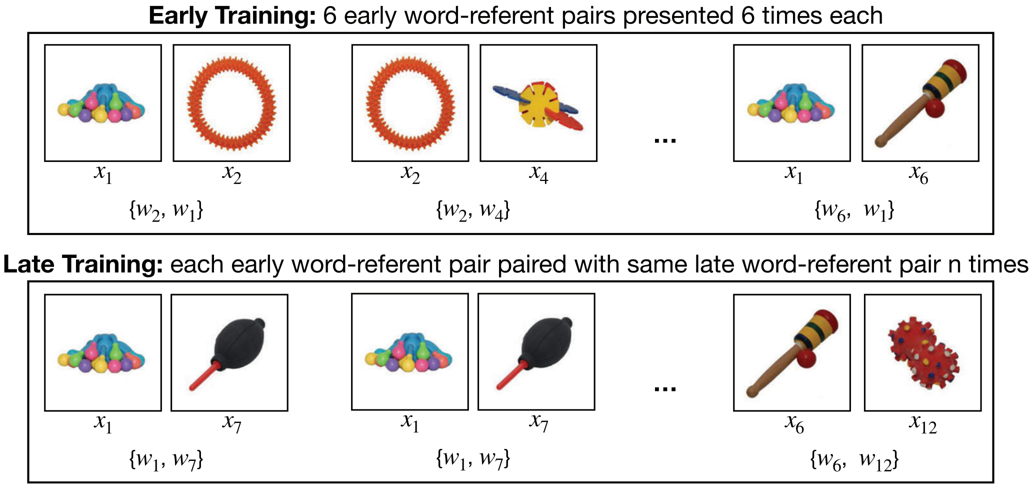
Early Training: 6 early word-referent pairs presented 6 times each
x_1
x_2
$\{w_2, w_1\}$
x_2
x_4
$\{w_2, w_4\}$
...
x_1
x_6
$\{w_6, w_1\}$
Late Training: each early word-referent pair paired with same late word-referent pair n times
x_1
x_7
$\{w_1, w_7\}$
x_1
x_7
$\{w_1, w_7\}$
...
x_6
x_{12}
$\{w_6, w_{12}\}$

Fig. 7. Experimental design for the relaxation of mutual exclusivity simulations. Here, we simulated a subset of the conditions in the original study from Kachergis et al. (2012). In the early training phase, participants learned six early word-referent pairs, with two objects and two referents per trial, and six presentations of each word-referent pair. During the late training phase, participants learned six late word-referent pairs, where each trial consisted of an early word-referent pair always appearing with the same late-referent pair, and this was repeated either three, six, or nine times. In the evaluation phase, participants were evaluated on the four possible word-referent pairings possible from the set of matched early and late word-referent pairs.

now with the late referent as the target (x_1 not shown, x_7 shown) and 10 other distractors. This process was mirrored for all of the late words. This evaluation design was chosen to examine which of the four possible mappings between early and late word-referent pairs participants would endorse ($w_1 - x_1$, $w_1 - x_7$, $w_7 - x_1$, $w_7 - x_7$). The number of presentations of both the early pairs was varied between subjects (0, 3, 6, or 9), and late pairs (3, 6, or 9) were varied within subjects.

Two critical findings emerged from this work. First, even after only three presentations of the late word-referent pairs (always in conjunction with the same early word-referent pair), participants showed high accuracy in selecting this pair ($w_7 - x_7$). Kachergis et al. (2012) argued that this result can be explained as an inference using mutual exclusivity. As this pair was always shown with $w_1 - x_1$, and participants would have learned this particular word-referent mapping from the first phase of training, participants should be able to infer that the new word (w_7) should map onto the new referent (x_7) through mutual exclusivity, rather than endorsing the two possible mappings between the first and second phases of training ($w_1 - x_7$ or $w_7 - x_1$) despite their patterns of co-occurrence. This provides evidence that people can employ mutual exclusivity not just during evaluation with novel words (as shown in Experiment 3) but also during the learning process itself as a means of quickly acquiring new word-referent mappings. The second major finding was that as the number of presentations increased in the *late training* phase, participants began to display a relaxation of mutual exclusivity. That is, in addition to endorsing $w_1 - x_1$ and $w_7 - x_7$ during the evaluation phase,

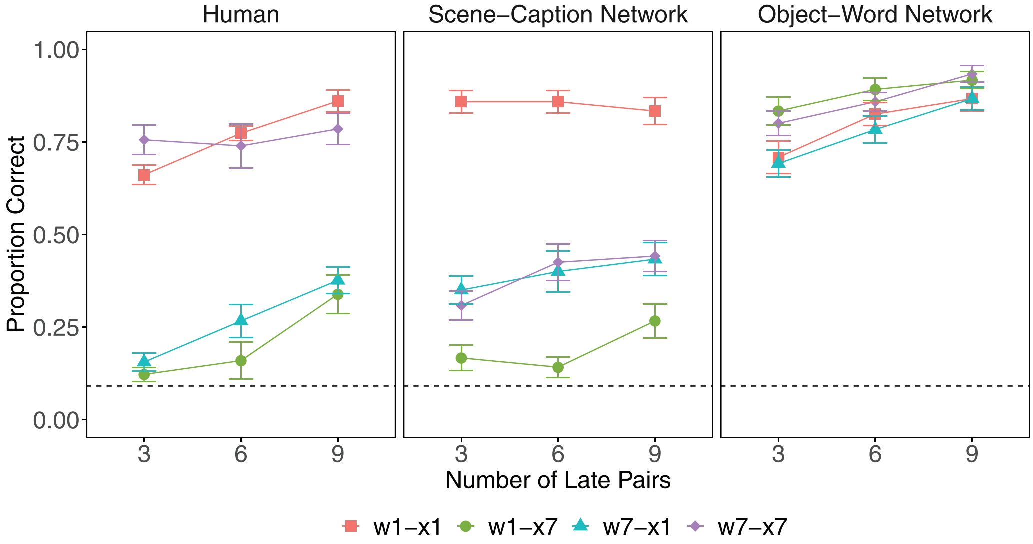

Fig. 8. Relaxation of mutual exclusivity simulation results. The left panel shows the human results from Kachergis et al. (2012), indicating that participants initially employed mutual exclusivity to exclude certain word-referent mappings, and then gradually relaxed mutual exclusivity given a sufficient number of late pairs. However, the two panels on the right show that neither of the two networks captures the same qualitative phenomena. The scene-caption network shows an effect of mutual exclusivity for the early word-referent pair $w_1 - x_1$ but not for the late pair $w_7 - x_7$. On the other hand, the object-word network learns all four possible mappings strongly, ignoring any kind of mutual exclusivity. The dotted line represents chance accuracy, and error bars depict 95% confidence intervals.

participants also increased their endorsements of $w_1 - x_7$ or $w_7 - x_1$, mapping each word to multiple referents (rather than a single referent as the mutual exclusivity bias would entail), although the proportion these cross mappings were selected were lower on average than the true mappings. This pattern of results is displayed in the left panel of Fig. 8.

Simulation: We simulated a subset of the conditions from Kachergis et al. (2012) to examine whether our two multimodal neural networks capture these findings; in particular, whether they can use mutual exclusivity to aid cross-situational word learning, as well as relax mutual exclusivity when provided with sufficient evidence during the late training phase. More specifically, since there were no substantive differences between varying the number of times each early pair was shown in the original experiment, we only considered the condition where the six early word-referent pairs were shown six times each (and excluding the 0, 3, or 9 repetition conditions). However, we varied the number of late word-referent pairs to be either **3**, **6**, or **9** repetitions, and pairing each late pair with an early pair in the same manner as the original experiment. Finally, the evaluation trials matched the original experiment, where each word was presented twice alongside 11 possible referents to select from: once with the early referent, but not the late referent, and then a second time with the late referent, but not the early referent, and both cases alongside the 10 remaining referents. This allowed us to determine which of the four possible pairings between early and late word-referent mappings the networks would endorse.

Results: The simulation results are summarized and compared with human behavior in Fig. 8. Each plot shows the four possible word-referent mappings that are tested, with results averaged across the six different early-late word-referent pair combinations. We find that both networks capture some, but not all of the qualitative patterns of learning as humans in this particular study, with both networks showing distinct preferences for each of the four types of word-referent mappings that were evaluated. Both networks, but the scene-caption network in particular broadly accounts for the same level of accuracy as humans in this experiment.

In the scene-caption network, we note a couple of patterns starting with the case with three presentations of the late word-referent pairs. First, the accuracy of $w_1 - x_1$ is highest as this was the least ambiguous word-referent pairing. On the other hand, the corresponding accuracy for $w_1 - x_7$ is lowest indicating that the network was hesitant to form an additional mapping to a second referent. With three presentations of the late word mappings, the network learns both possible mappings ($w_7 - x_1$ and $w_7 - x_7$) equally well. The lack of a preference for either $w_7 - x_7$ or $w_7 - x_1$ suggests that this network was *not* employing mutual exclusivity during training by limiting mappings from one word to one referent, and where the greatest divergence between human behavior appears. Finally, with additional presentations of the late word-referent pairs, we find that the network's preference for each of these four mappings increases in line with human behavior.

In the object-word-network, a very different pattern was observed for both the results from human participants and the scene-caption network. Even after three presentations of the late word-referent pairs, the object-word network selected the intended referent for all of the four word-referent mappings we evaluated, and this increased with additional presentations. Here the differences between the two networks are more pronounced than in some of the earlier simulations. We hypothesize that the similarity calculation in the scene-caption network which aggregates across all of the words and referents, through the use of the max operator, leads to different patterns of update where only the word embedding for the maximally active referent is updated. On the other hand, in the object-word network, each potential word–object pair in a given scene is considered a matching pair, and since the same early and late word-referent pairs are matched for multiple trials, the network treats all of these pairs as equally valid. This causes the network to update the embeddings to be consistent for all possible pairings, regardless of whether they were observed early or late during training.

Overall, neither of the two networks captures the full set of trends found in the behavioral results in Kachergis et al. (2012). However, both networks are capable of endorsing multiple mappings without issue, indicating that their inductive bias towards mutual exclusivity may be less strict or non-existent relative to humans. Leveraging mutual exclusivity during the learning process may enable faster learning, but this study also demonstrates how models of cross-situational word learning also need to accommodate cases when a single word maps to multiple referents and vice versa, and our results suggest that neither network captures both of these tendencies.

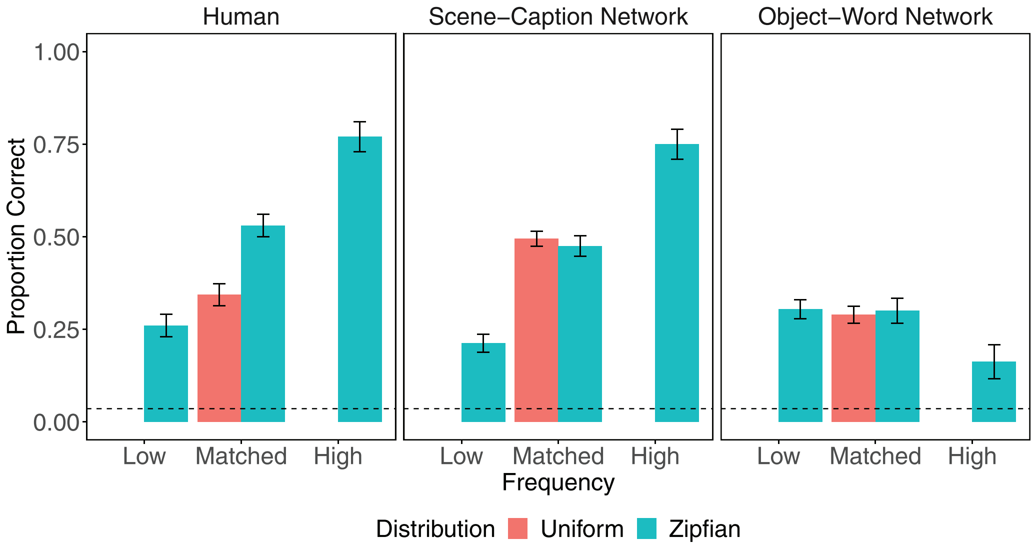

Fig. 9. Learning from Zipfian distributions simulation results. The left panel shows the data collected from Experiment Two of Hendrickson and Perfors (2019), showing that the Zipfian condition led participants to learn more words for the subset of words whose frequency matched the *uniform* condition. Although the accuracy for the Zipfian condition in the scene-caption network is comparable to human-level accuracy, its performance in the *uniform* condition is also equivalent, and thus does not qualitatively match the behavior of human participants. On the other hand, the object-word network exhibited low overall performance, from the large amount of referential ambiguity and the absolute number of words to be learned. The dotted line represents chance accuracy, and error bars depict 95% confidence intervals.

3.5. Experiment 5: Learning from Zipfian distributions

The fifth simulation is based on a recent study by Hendrickson and Perfors (2019), comparing cross-situational learning between word-referent pairs that were distributed either uniformly or as a Zipfian distribution, showing a benefit for cross-situational word learning in the Zipfian case to a uniform distribution. In the *uniform* case, each of the possible word-referent mappings is presented the same number of times throughout training, akin to many of the previous simulations reported earlier in this paper. In the *Zipfian* case, a few word-referent pairs are presented many times over the course of training, while the remainder of the word-referent pairs is only presented a few times. This skewed distribution is intended to be more representative of real-world language learning, where children may hear some words many times and others very infrequently. As shown in the left panel of Fig. 9, Experiment Two from Hendrickson and Perfors (2019) showed that participants in the Zipfian condition learned roughly twice as many words as participants in the Uniform condition, when comparing words that were matched in the frequency of presentation.[10] Hendrickson and Perfors (2019) argued that under a Zipfian distribution, the very frequent word-referent mappings would be learned first. Then, on subsequent trials containing these easily learned words and referents, participants could reason using mutual exclusivity to exclude these known words/referents when reasoning about unknown words/referents, effectively reducing the degree of referential ambiguity based on existing knowledge. On the other hand, under a uniform distribution, participants would be more limited in their ability to reduce the space of potential space of word-referent mappings. Therefore, learners would have higher referential ambiguity throughout the course of the experiment, leading to a decrease in word-referent pairs learned relative to the Zipfian condition, which is what was empirically observed.

Simulation: In their experiment, participants learned 28 word-referent mappings, with four words and four referents per trial.[11] In the *uniform* condition, participants were shown each word-referent mapping 10 times throughout the course of training. In the *Zipfian* condition, the most frequent word-referent pair appeared 62 times throughout the course of training, the next most frequent word-referent pair appeared 33 times, and the 12 least frequent word-referent pairs occurred five or fewer times. To generate trials for the Zipfian condition, we randomly generated training trials consisting of four words and four referents that matched the frequency counts from the Zipfian condition in the original paper for the 28 word-referent pairs. Due to the skewed distribution in the Zipfian condition, mismatches are sampled from a memory that is also distributed in the same Zipfian fashion. The evaluation phase consisted of a challenging 28-way classification, testing each target word against all the potential referents observed during training.

Results: A summary of the simulation results is shown in Fig. 9. We find that for the scene-caption network, the accuracy in the *Zipfian* condition matches the qualitative and quantitative aspects found in the human data, which is quite remarkable given the network was only trained for a single epoch and the evaluation involved identifying the correct referent among 28 options. As expected, the network is more accurate on higher frequency items. Moreover, the level of accuracy matches what we find in humans at the three levels of frequency. This high level of accuracy arises because the process for sampling mismatched scenes and captions will often involve these high-frequency items, providing a strong training signal for these particular word-referent pairs. Despite these successes, this network does not capture the critical finding showing a benefit for humans in the *Zipfian* condition compared to the *uniform* condition for the words whose frequency of presentations were matched. Rather, the performance in the *uniform* condition was equal (and better than human performance). Instead of relying on mutual exclusivity to learn word-referent mappings, the performance of the scene-caption network may be explained by the fact that it observed a sufficient number of presentations to learn these word-referent pairs in both the Zipfian and *uniform* conditions, regardless of the kind of referential ambiguity from other words and referents present on each trial.

For the object-word network, we see a reversal in performance compared to Experiment 4. Here, we find that performance of the network in both distribution conditions is low across all frequency groupings, and surprisingly the most frequent items result in the lowest accuracy scores. Due to these discrepancies, the network fails to capture any of the qualitative patterns from the human data in this experiment. Because the network considers all 16 possible pairings per trial as viable, one possible explanation for this network's failures is that it learns multiple incorrect word-referent mappings due to the high level of referential ambiguity. This is further enhanced as the process for sampling mismatches will be skewed towards the high-frequency words and referents, but incorrectly paired with other referents or words that co-occurred on previous trials.

3.6. Experiment 6: Exemplar generalization

In the sixth simulation, we examined whether multimodal neural networks could generalize to visually similar referents. Although it is quite common in developmental studies to examine generalization to novel exemplars of a word (Carey & Bartlett, 1978; Samuelson & Horst, 2007; Taxitari, Twomey, Westermann, & Mani, 2020; Wojcik, 2017), previous models of cross-situational word learning do not capture this form of generalization as referents are symbolically encoded, bypassing the problem of generalization completely.[12] This leads us to question whether multimodal neural networks can naturally capture this kind of generalization to novel exemplars. Due to the limited number of words and referents used in all of the previous simulations, one hypothesis for how multimodal neural networks is successful is that they converge to something like a discrete code that matches words to referents, akin to a symbolic encoding. Such a representation, however, would fail to generalize for novel referents of a concept, as these may not necessarily map to the same discrete encoding for the learned referent. Another hypothesis is that these networks instead learn a distributed representation in a manner that satisfies the constraints of the contrastive loss function. In this scenario, generalization to novel referents from an existing word-referent pair occurs as visually similar referents are mapped close together in embedding space, in a manner that preserves the strength of the correspondence score with the corresponding original word embedding. In this simulation, we find that both multimodal neural networks we use to capture this kind of generalization via their use of distributed representations, in line with the latter hypothesis.

Simulation: The setup for the exemplar generalization simulation is depicted in the bottom row in Fig. 1D. For this simulation, training is the same as the basic cross-situational learning procedure (Fig. 1A), where the model observes multiple words and referents per trial, but multiple presentations of each word-referent pair across trials. However, during evaluation, in addition to testing the word-referent pairs seen during training, the network is also asked to classify novel exemplars of a concept (Fig. 1D). Concretely, we used a different subset of the NOUN database consisting of 30 images grouped into 10 categories with three exemplars each, where each of the exemplars was similar in shape but varied by color or texture. Despite the simplicity of the visual generalization demonstrated here, we were primarily motivated to use the same kinds of stimuli that are often found in developmental studies of exemplar generalization (Twomey, Ranson, & Horst, 2014; Wojcik, 2017), rather than more complex, naturalistic images from other work (Chrupała et al., 2017). The network was trained on 10 word-referent pairs, using only one out of the three exemplars per category. Two words and two referents were presented on each trial, with a total of six presentations for each word-referent pair, matching the training setup for many of the previous simulations. During the evaluation, the network was tested by pairing each target words with either the set of

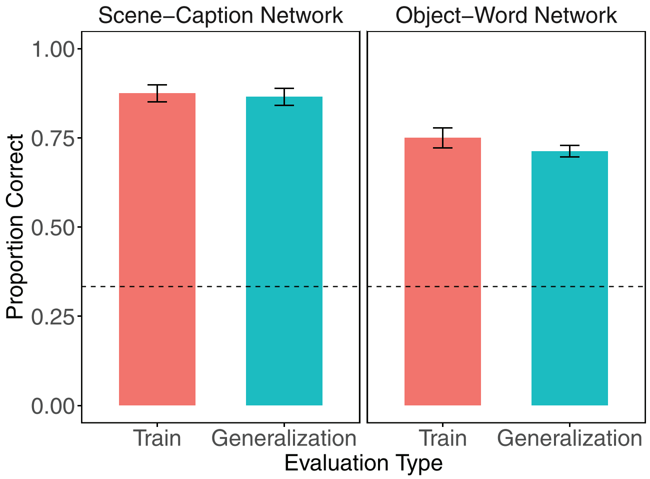

Fig. 10. Exemplar generalization results. The left bars show performance on the evaluation trials for the scene-caption network, while the right bars show performance for the object-word network. Performance is separated into the trials that evaluated the word with familiar examples of a concept (*train*) versus novel examples (*generalization*). Both networks are easily generalized to novel examples with an accuracy comparable to the training examples, although a slight performance advantage was observed in the scene-caption network relative to the object-word network. The dotted line represents chance accuracy, and error bars depict 95% confidence intervals.

referents observed during training (*train accuracy*) presented alongside two other foil referents observed during training. Second, the networks were also evaluated using the held out set of the two other referents for each category (*generalization accuracy*), presented alongside two other novel foil referents from other categories, which controlled for the familiarity of observed referents.

Results: A summary of the results is shown in Fig. 10. Both the scene-caption and the object-word networks exhibit high accuracy on the evaluation trials, regardless of whether familiar or novel exemplars of a category were evaluated, with slightly higher accuracy in the scene-caption network. The networks were able to generalize existing learned word-referent mappings to novel exemplars not seen during the training process, utilizing the fact that the image encoder can map novel (but similar looking) visual referents in the embedding space close to the embedding of the original referent, retaining a high similarity score to the corresponding word embedding that was paired with the original referent. This result highlights a strength of distributed representations compared to purely symbolic models of cross-situational word learning, and that it automatically emerges as a by-product after training.

3.7. *Experiment 7: Learning multimodal representations without pre-training*

In all of the previous simulations, we relied on a pre-trained CNN for the image encoder. The visual representations were learned from another dataset and fixed throughout training, allowing the networks to focus on acquiring the word-referent mappings. In this final simulation, we train convolutional networks from scratch to explore whether useful visual representations can be acquired solely through the process of cross-situational learning. Answering this question in the affirmative would provide evidence that the methods here need not rely on pre-trained features obtained from fully supervised training (like all of the previous simulations); instead, it would provide a proof of concept for how these networks could account for learning at larger scale, more naturalistic settings such as those experienced by children during development. We also study, as in Experiment 6, how these networks generalize to novel exemplars of a category. The next two parts of this section describe the dataset used to train this network, followed by how the network architecture and training process was adapted for these simulations.

Dataset: A much larger dataset than the NOUN database is needed to train a CNN from scratch. For these simulations, we used MNIST (LeCun, 1998), a standard machine learning dataset, consisting of 60,000 training examples and 10,000 test examples of 28 × 28 px handwritten digits from 0 to 9. Based on these images, we created our own *Multi-MNIST* training and evaluation sets. The training sets were generated by randomly sampling a certain number of digits per scene from the original MNIST training set. The number of digits per scene was either 2, 3, or 4, similar to the referential ambiguity simulations from Experiment 1. We also generated a matching caption specifying the digit labels (in a permuted order). We varied the total number of exemplars presented during training—using either 480, 1,920, 4,800, 19,200, or 48,000 digit exemplars—thus, the total number of exemplars presented was controlled regardless of how many digits were shown per scene.

The evaluation dataset was generated in a manner similar to previous simulations. We generated 100 distinct evaluation trials for each digit (so 1,000 in total). On each trial, the network was presented with a target digit word alongside an array of digits from 0 to 9, each of which was randomly sampled from the original MNIST test set. Similar to Experiment 6, the network is tested on its ability to generalize to novel exemplars—as all the test images are new to the network—providing a stronger test of the network's generalization capabilities when trained from scratch.

Network: The results in this simulation rely on a variant of the scene-caption network used throughout this work, but with three notable changes.[13] First, rather than using a pre-trained VGG-16 CNN as our image encoder, a separate CNN was constructed and randomly initialized. This CNN consisted of two convolutional layers (with 32 and 64 feature maps) with ReLU non-linearities in between, and then followed by a single 2 × 2 max pooling layer. This was followed by a layer of dropout, and a single fully connected layer, resulting in an image embedding of size 128. For the word encoder, the digit labels were embedded using a single embedding layer as before, but mapping the digit labels to word embeddings of size 128 to match the dimensionality of the image embeddings.

As the trial-by-trial performance was not of interest in this simulation, the network was trained instead using minibatches (of size 256), as is standard practice. Due to this change, mismatching examples for the contrastive loss were sampled from other scene–caption pairs within the same batch, as is common in other papers employing a contrastive loss function

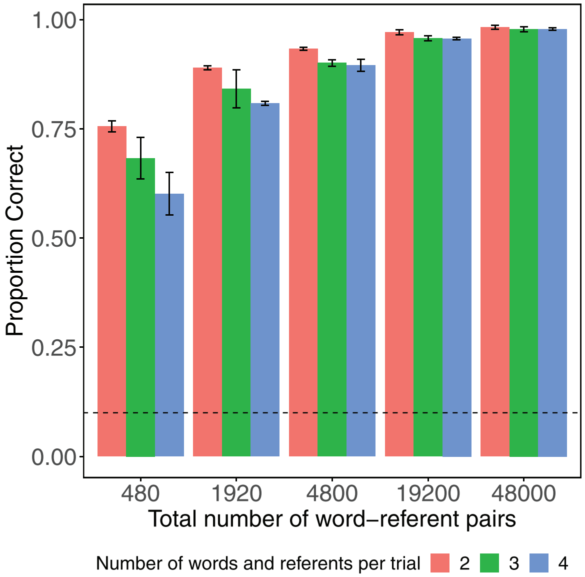

Fig. 11. Learning representations from scratch. Results show the evaluation performance of a multimodal neural network from scratch. We find that performance increases as the network is provided with more examples during training, while the same qualitative decreases occur with greater referential ambiguity. However, networks trained with the maximum number of scenes achieved close to perfect performance, showing that these networks can be trained from scratch given enough data. The dotted line represents chance accuracy, and error bars depict 95% confidence intervals.

(Harwath et al., 2018), and removing the need for a memory mechanism as used in previous simulations. Finally, all of the networks were trained for 10 epochs using the Adam optimizer with a learning rate of 1e-4.

Results: The results from this experiment are shown in Fig. 11. Overall, the network is successful at solving the cross-situational word learning problem via a randomly initialized CNN but also reveals a number of other interesting findings. First, it can acquire visual representations from scratch that generalize well to novel exemplars in different categories, even in the condition with the fewest number of examples. Second, regardless of the degree of referential ambiguity, evaluation accuracy approaches perfect performance as the number of training examples increases. This provides strong evidence that these multimodal neural networks can indeed be trained from scratch entirely from trials that are individually ambiguous, learning to resolve cross-situational mappings and generalize to novel exemplars (Lewis & Frank, 2013). Finally, the degree of referential ambiguity affected evaluation accuracy in the same manner as observed in behavioral findings (Yu & Smith, 2007) and simulated in Experiment 1. Notably, this referential ambiguity effect attenuates when more examples are provided during training and model accuracy approaches ceiling.

Other work has also demonstrated the ability of multimodal neural networks to learn visual representations from scratch with different kinds of training. For example, Harwath et al. (2018) showed that a similar contrastive learning method between paired images of visual scenes and speech descriptions could be used to train a CNN from scratch. Another paper from Desai and Johnson (2020) showed that performing an image captioning task (predicting a language description for a given image) was also effective for training a CNN from scratch and demonstrated strong performance on a number of downstream tasks after this pre-training procedure. Our results and these other findings suggest a distinct advantage for multimodal neural networks over other classes of models for cross-situational word learning. The scalability of this approach suggests that these models can not only capture the learning of novel word-referent pairs in the lab by using a pre-trained CNN but also as a model for cross-situational word learning during development, starting from a randomly initialized CNN and jointly learning and aligning visual and language embeddings without any explicit prior knowledge.

4. General discussion

In this work, we evaluated two different multimodal neural networks on a variety of cross-situational word learning experiments, examining their ability to explain key empirical phenomena from the psychological literature. Our primary motivation was to understand the kinds of phenomena that emerge from training relatively generic neural networks on multimodal data. Just as importantly, we want to understand which kinds of phenomena do not naturally emerge from such an account, suggesting additional learning mechanisms or inductive biases may be responsible for these behaviors in humans. Our approach is shared by a number of other recent works using state-of-the-art architectures from machine learning to provide insights into human cognition, such as using pre-trained models in categorization (Lake, Zaremba, Fergus, & Gureckis, 2015; Peterson, Abbott, & Griffiths, 2018) and in the language (Arehalli & Linzen, 2020; Manning, Clark, Hewitt, Khandelwal, & Levy, 2020).

Our investigation was instructive in understanding the capabilities of relatively generic multimodal neural networks; they were able to capture four diverse empirical phenomena, out of the seven phenomena we studied. Some of these phenomena have previously been shown to be accounted for by other symbolic models (Fazly et al., 2010; Frank et al., 2009; Kachergis et al., 2012; Yu & Smith, 2012). Moreover, the successes we observed in the last two simulations involving exemplar generalization are something that traditional symbolic accounts have had difficulty capturing. Our results show that exemplar generalization occurs naturally as a by-product of using CNNs to encode visual referents, by embedding visually similar images close together in representational space. Another advantage of these networks is their ability to learn from a single epoch of stimuli presented in an online manner, matching the amount of experience presented to human participants. This was a surprising finding to us, but it appears to be in line with other recent findings using multimodal architectures that employ a contrastive learning procedure for word learning; a single epoch of training can be sufficient to disambiguate ambiguous word-referent mappings (Lazaridou et al., 2016).

The failures we observed in some simulations do not imply that all *other* multimodal neural network architectures will perform identically. The two models we considered were, in some sense, minimal instantiations of a contrastive learning account using multimodal neural networks, and therefore our results imply a floor of performance for this class of models, not the ceiling. Other work has explored ways this base model could be modified, such as adding extra forms of attention or accounting for social interactions (Lazaridou et al., 2016), using other kinds of similarity or loss functions, such as finding the optimal mapping between words and referents on a single trial (Gulordava et al., 2020; Harwath et al., 2018), or increasing the number of negative mismatches to contrast against (Oord, Li, & Vinyals, 2018; Radford et al., 2021).

4.1. Failures relating to mutual exclusivity

We observed a striking qualitative discrepancy between human behavior and the models in three of the seven simulations: in Experiment 3 (mutual exclusivity, except in limited cases), Experiment 4 (relaxation of mutual exclusivity), and Experiment 5 (learning from Zipfian distributions). Despite the differences between simulations, there was a *correlated* failure mode: neither network seemed to utilize mutual exclusivity in a manner similar to humans. Both networks showed evidence of cross-situational learning in these tasks, but key behavioral findings were not evident in either of the networks. The failures of the object-word network appeared to be more pronounced than the scene-caption network, displaying less human-like responses for all three of these simulations.

In each of these three simulations, it has been argued that the principle of *mutual exclusivity* is employed by humans to solve these tasks (Halberda, 2003; Hendrickson & Perfors, 2019; Kachergis et al., 2012). In Experiment 4, participants leveraged knowledge about prior learning of the word-referent mapping $w_1 - x_1$ to quickly form a new mapping for the late word-referent pair $w_7 - x_7$, even though the raw co-occurrence counts were equally consistent with other hypotheses. In Experiment 5, participants used the knowledge obtained from frequent word-referent pairs in the Zipfian condition to reduce the referential ambiguity on later trials, again applying the assumption of mutual exclusivity to do so. This commonality across these three experiments suggests that solving the singular problem of building an inductive bias for mutual exclusivity into the learning process for these networks may be sufficient for capturing all three of these results, rather than requiring distinct architectural changes for each of these three phenomena.

Reasoning by mutual exclusivity remains a challenge for standard deep learning architectures, despite the promised benefits of incorporating this inductive bias. Earlier computational accounts of cross-situational word learning proposed multiple mechanisms to handle mutual exclusivity, via inductive biases for novelty (Kachergis et al., 2012) or by applying Bayesian inference over a prior that favors simpler lexicons (Frank et al., 2009). However, these earlier computational accounts rely on the symbolic encodings of referents that implicitly encode novel referents as distinct entities from familiar referents, a procedure that cannot be easily translated to how multimodal neural networks embed words and referents into a continuous multidimensional embedding space. In particular, having a learnable word embedding for each novel word that is distinct from other word embeddings would likely be insufficient to produce mutual exclusivity, as the novel word embedding would also need to produce a corresponding similarity score for any possible novel referent and not any of the observed referents.

In recent work, Gandhi and Lake (2020) showed that many architectures actually have a bias *against* mutual exclusivity. Given a novel input, networks tend to respond with a familiar output response, rather than a novel output response. However, there have been some successes in getting neural networks to demonstrate mutual exclusivity by training models via memory-augmented meta-learning, allowing the networks to learn this inductive bias (Lake, 2019; Santoro, Bartunov, Botvinick, Wierstra, & Lillicrap, 2016). Gulordava et al. (2020) proposed more sophisticated referent selection mechanisms that incorporated pragmatic reasoning, although their approach required that the novel word and novel referent to be sampled as negative contrasting items. Furthermore, while each of these approaches can utilize mutual exclusivity at evaluation time, the behavioral phenomena studied in Experiments 4 and 5 suggest that people also apply mutual exclusivity during training too. One potential avenue for future research would be to investigate how mutual exclusivity could be incorporated as another learning mechanism during the training process, which could both speed up word learning and enable more human-like patterns of cross-situational learning.

4.2. Associations versus hypothesis testing

One longstanding debate in cross-situational word learning is whether people learn word-referent mappings in an associative manner or through hypothesis testing (Khoe, Perfors, & Hendrickson, 2019; Yu & Smith, 2012). The manner in which our multimodal neural networks gradually update their representations of words and referents over time via contrastive learning aligns more closely with associative accounts of cross-situational word learning rather than hypothesis testing accounts (Fazly et al., 2010; Kachergis et al., 2012; McMurray et al., 2012).

Yet, there is evidence that humans engage in forms of explicit hypothesis testing during word learning (Berens, Horst, & Bird, 2018; Medina et al., 2011; Stevens et al., 2017; Trueswell et al., 2013), suggesting that a fuller explanatory account of word learning would require the addition of such mechanisms. One difficulty, however, is translating the kinds of hypothesis testing mechanisms from these computational accounts to the framework presented in this paper. Existing hypothesis testing models are restricted to generating hypotheses in a format that requires referents to be symbolically encoded (Stevens et al., 2017; Trueswell et al., 2013), which means that existing models cannot be used with raw images straightforwardly. A plausible hypothesis testing account that could be integrated within a multimodal neural network would need some other representational format that does not symbolically encode referents, but would instead need to represent or generate hypotheses for words that can flexibly deal with high-dimensional embeddings, like the outputs of our vision encoder.

4.3. Incorporating prior linguistic experience

Both of the multimodal neural networks we explored consisted of a pre-trained CNN as our visual encoder (in all simulations besides the final one), and combined this with a randomly initialized embedding layer for the text encoder. Would there be any advantages to also using a pre-trained text encoder in our simulations, such as word2vec (Mikolov, Chen, Corrado, & Dean, 2013) or BERT (Devlin, Chang, Lee, & Toutanova, 2018)? We argue this is unlikely to play a major role, for two reasons. First, cross-situational word learning experiments generally present participants with novel words rather than familiar words, and pre-trained models would initially treat these novel words as random embeddings, exactly like our current setup. Second, cross-situational word learning experiments also typically present the words in a given trial as a disjoint set of words, rather than embedded in a natural language sentence, removing any syntactic benefits a pre-trained model might be able to exploit to learn novel word-referent mappings.

Nevertheless, there could be cases where using pre-trained linguistic or pre-trained multimodal models might confer some kind of benefit, especially as one considers the problem of cross-situational word learning in more naturalistic contexts. For example, pre-trained models could be used to more closely model the exact natural language instructions provided to children in these experiments, for example, "bring me the chromium tray, not the blue one, the chromium one" (Carey, 1978). Models with pre-trained linguistic representations could also take advantage of their knowledge of syntax to restrict the set of possible referents under consideration, via processes such as *syntactic bootstrapping* (Gleitman, 1990).

4.4. Conclusion

In this work, we explored which word learning phenomena arise from relatively generic multimodal neural networks trained on multimodal data, focusing on capturing the kinds of phenomena found in human experiments with limited training of novel word-referent pairs. Our main contribution has been to increase the number of simultaneous phenomena studied and to perform a more comprehensive evaluation of the capabilities of multimodal networks using the same training setup across multiple simulations. Nevertheless, some of the phenomena we examined overlap with past work, including fast mapping (Hill et al., 2020; Lazaridou et al., 2014; Lazaridou, Marelli, & Baroni, 2017) and mutual exclusivity (Gulordava et al., 2020). Other multimodal architectures have also been used to study word learning phenomena we did not consider such as the shape bias (Hill et al., 2020) and learning from child-directed input (Lazaridou et al., 2016). Lastly, although we did not test the use of raw speech for this work, other research has shown that neural network architectures can be applied to cross-situational learning with raw speech instead of text using similar contrastive-based methods (Chrupała et al., 2017; Harwath et al., 2018), suggesting that contrastive learning is a powerful general-purpose tool to align cross-modal representations from raw sensory input.

One limitation of experimental approaches to cross-situational learning is the degree to which the learning problem is simplified. In contrast, a child learning language is embedded in naturalistic contexts that present multiple additional learning challenges. Other work has explored this problem of scalability in a variety of ways, from early multimodal approaches (Roy & Pentland, 2002), to more recent work using large-scale naturalistic headcam data (Orhan, Gupta, & Lake, 2020; Tsutsui, Chandrasekaran, Reza, Crandall, & Yu, 2020) and studying the ways in which children or machines play an active role in word learning

(Gelderloos, Kamelabad, & Alishahi, 2020; Zettersten & Saffran, 2019). The fact multimodal neural networks can be trained from scratch, as demonstrated in Experiment 7 and other works (Harwath et al., 2018; Radford et al., 2021), suggests that these kinds of networks could be further developed to provide a unifying account of artificial word learning in the lab and naturalistic word learning in the wild (Meylan & Bergelson, 2021). Finally, while we attempted to test a broad range of phenomena, our list was by no means exhaustive. Future work should aim to examine other aspects of word learning not considered here, such as visual grounding of referents from other kinds of lexical classes such as verbs and adjectives (Ebert & Pavlick, 2020; Nikolaus & Fourtassi, 2021).

Acknowledgments

We would like to thank the anonymous reviewers as well as the Lake and Gureckis labs for helpful discussions and feedback on this work. The authors have no conflicts to disclose. All code and data are available at https://osf.io/nh4jk/.

Open Research Badges

This article has earned Open Materials badges. Data are available at https://osf.io/nh4jk/?view_only=e3355708e7104e6f9d4fb1f9c0a4cef3.

Notes

[1] Other recent work has shown that these neural network approaches can also learn not just from raw images and text but also raw images and raw audio (Chrupała, Gelderloos, & Alishahi, 2017; Harwath et al., 2018), highlighting the flexibility of these networks to form cross-modal representations with different combinations of modalities.

[2] For all our simulations, we set $d = 64$, except the final simulation where we set $d = 128$.

[3] Since the purpose of cross-situational word learning experiments is to focus on learning the mappings between words and their referents, the use of a pre-trained visual backbone is intended as a rough proxy for the prior visual experience of participants performing these tasks. However, Experiment 7 demonstrates how to jointly train a CNN from scratch in this task, showing that using pre-trained representations are not required with sufficient training data.

[4] In all our simulations, we only consider the case where the model samples a single mismatch. We explored the effect of sampling multiple mismatches in a subset of the simulations, but we found no qualitative differences in results.

[5] In general, contrastive approaches sample the contrastive items from the same minibatch rather than from an explicit memory, but because our networks are trained in an online fashion with a single trial at a time, this necessitated the use of an explicit memory mechanism to sample contrasting items.

[6] Although it is common in cross-situational word learning experiments to match the number of words and referents per trial, this form of similarity allows some additional flexibility in handling situations where the number of words differs from the number of referents, like when the set of words is a sentence in natural language and not every word can be mapped onto a visually grounded referent.

[7] Because the experiment controlled for the number of presentations of each word-referent pair, the three conditions each had a different total number of training trials (54, 36, and 27 for the **2** × **2**, **3** × **3**, and **4** × **4** conditions, respectively).

[8] Note that the ability to perform referent selection in fast mapping experiments can be explained via the principle of mutual exclusivity, so the results of this section can also be interpreted as whether multimodal neural networks can perform referent selection in fast mapping (Carey & Bartlett, 1978; Horst & Samuelson, 2008), in addition to their ability to display retention as observed in Experiment 2.

[9] The prompt, as provided to children, implies that one of the two objects corresponds to the new word "Toma." Thus, the network is allowed a gradient update to incorporate this information; otherwise, it would have no way of knowing whether or not "Toma" refers to any object in this scene.

[10] For our simulations, we only considered the ZipfianFrequency condition where the frequency of words mattered, compared to the ZipfianLength condition where the word lengths also varied according to a Zipfian distribution, as the difference between the two conditions in the original work were quite minor and this would not have affected either network's predictions.

[11] The original experiment had 32 word-referent mappings, but four of these were displayed once as check trials for human participants, which we excluded in our simulations, resulting in 70 total trials rather than 71.

[12] One such attempt to model this kind of generalization can be found in Lewis and Frank (2013), where they used a Bayesian cross-situational word learning model (Frank et al., 2009) combined with a model of Boolean concept learning (Goodman, Tenenbaum, Feldman, & Griffiths, 2008), although in their model the features were still symbolically encoded (but at the level of features rather than referents).

[13] We did not explore the use of the object-word network for this simulation, as the previous simulations generally showed better performance with the scene-caption network.

References

Alishahi, A., Fazly, A., & Stevenson, S. (2008). Fast mapping in word learning: What probabilities tell us. In A. Clark & K. Toutanova (Eds.), *CoNLL 2008: Proceedings of the Twelfth Conference on Computational Natural Language Learning* (pp. 57–64). New York: Association for Computational Linguistics.

Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Lawrence Zitnick, C., & Parikh, D. (2015). Vqa: Visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision* (pp. 2425–2433). Piscataway, NJ: IEEE Press.

Arehalli, S., & Linzen, T. (2020). Neural language models capture some, but not all, agreement attraction effects. In S. Denison, M. Mack, Y. Xu, & B. C. Armstrong (Eds.), *Proceedings of the 42nd Annual Meeting of the Cognitive Science Society* (pp. 370–376). Austin, TX: Cognitive Science Society.

Berens, S. C., Horst, J. S., & Bird, C. M. (2018). Cross-situational learning is supported by propose-but-verify hypothesis testing. *Current Biology*, *28*(7), 1132–1136.

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

Carey, S. (1978). The child as word learner. In M. Halle, J. Bresnan, & G. Miller (Eds.), *Linguistic theory and psychological reality* (pp. 264–293).

Carey, S., & Bartlett, E. (1978). Acquiring a single new word. *Papers and Reports on Child Language Development*, *15*, 17–29.

Chrupała, G., Gelderloos, L., & Alishahi, A. (2017). Representations of language in a model of visually grounded speech signal. [Preprint]. arXiv:1702.01991.

Chrupała, G., Kádár, A., & Alishahi, A. (2015). Learning language through pictures [Preprint]. arXiv:1506.03694.

Desai, K., & Johnson, J. (2020). Virtex: Learning visual representations from textual annotations [Preprint]. arXiv:2006.06666.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of deep bidirectional transformers for language understanding [Preprint]. arXiv:1810.04805.

Ebert, D., & Pavlick, E. (2020). A visuospatial dataset for naturalistic verb learning [Preprint]. arXiv:2010.15225.

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.

French, R. M. (1999). Catastrophic forgetting in connectionist networks. *Trends in Cognitive Sciences*, *3*(4), 128–135.

Gandhi, K., & Lake, B. M. (2020). Mutual exclusivity as a challenge for deep neural networks. *Advances in Neural Information Processing Systems*, *33*.

Gelderloos, L., Kamelabad, A. M., & Alishahi, A. (2020). Active word learning through self-supervision. In S. Denison., M. Mack, Y. Xu, & B. C. Armstrong (Eds.), *Proceedings of the 42nd Annual Meeting of the Cognitive Science Society* (pp. 1050–1056). Austin, TX: Cognitive Science Society.

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

Gleitman, L. (1990). The structural sources of verb meanings. *Language Acquisition*, *1*(1), 3–55.

Golinkoff, R. M., Hirsh-Pasek, K., Bailey, L. M., & Wenger, N. R. (1992). Young children and adults use lexical principles to learn new nouns. *Developmental Psychology*, *28*(1), 99.

Goodman, N. D., Tenenbaum, J. B., Feldman, J., & Griffiths, T. L. (2008). A rational analysis of rule-based concept learning. *Cognitive Science*, *32*(1), 108–154.

Gulordava, K., Brochhagen, T., & Boleda, G. (2020). Deep daxes: Mutual exclusivity arises through both learning biases and pragmatic strategies in neural networks. In S. Denison, M. Mack, Y. Xu, & B. C. Armstrong (Eds.), *Proceedings of the 42nd Annual Meeting of the Cognitive Science Society* (pp. 2089–2095). Austin, TX: Cognitive Science Society.

Halberda, J. (2003). The development of a word-learning strategy. *Cognition*, *87*(1), B23–B34.

Halberda, J. (2006). Is this a dax which i see before me? Use of the logical argument disjunctive syllogism supports word-learning in children and adults. *Cognitive Psychology*, *53*(4), 310–344.

Harwath, D., Recasens, A., Surís, D., Chuang, G., Torralba, A., & Glass, J. (2018). Jointly discovering visual objects and spoken words from raw sensory input. In V. Ferrari, M. Hebert, C. Sminchisescu, & Y. Weiss (Eds.), *Proceedings of the European Conference on Computer Vision (ECCV)* (pp. 649–665). Cham, Switzerland: Springer International Publishing.

Hendrickson, A. T., & Perfors, A. (2019). Cross-situational learning in a Zipfian environment. *Cognition*, *189*, 11–22.

Hill, F., Clark, S., Hermann, K. M., & Blunsom, P. (2020). Simulating early word learning in situated connectionist agents. In S. Denison, M. Mack, Y. Xu, & B. C. Armstrong (Eds.), *Proceedings of the 42nd Annual Meeting of the Cognitive Science Society* (pp. 875–881). Austin, TX: Cognitive Science Society.

Hill, F., Tieleman, O., von Glehn, T., Wong, N., Merzic, H., & Clark, S. (2020). Grounded language learning fast and slow [Preprint]. arXiv:2009.01719.

Horst, J. S., & Hout, M. C. (2016). The novel object and unusual name (NOUN) database: A collection of novel images for use in experimental research. *Behavior Research Methods*, *48*(4), 1393–1409.

Horst, J. S., & Samuelson, L. K. (2008). Fast mapping but poor retention by 24-month-old infants. *Infancy*, *13*(2), 128–157.

Johnson, J., Hariharan, B., van der Maaten, L., Fei-Fei, L., Lawrence Zitnick, C., & Girshick, R. (2017). CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 2901–2910). Piscataway, NJ: IEEE Press.

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.

Khoe, Y. H., Perfors, A., & Hendrickson, A. T. (2019). Modeling individual performance in cross-situational word learning. In A. K. Goel, C. M. Seifert, & C. Freksa (Eds.), *Proceedings of the 41st Annual Meeting of the Cognitive Science Society* (pp. 560–567). Montreal, Canada: Cognitive Science Society.

Lake, B. M. (2019). Compositional generalization through meta sequence-to-sequence learning. In *Advances in Neural Information Processing Systems* (pp. 9791–9801).

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*, E253.

Lake, B. M., Zaremba, W., Fergus, R., & Gureckis, T. M. (2015). Deep neural networks predict category typicality ratings for images. In D. C. Noelle, R. Dale, A. S. Warlaumont, J. Yoshimi, T. Matlock, C. D. Jennings, & P. P. Maglio (Eds.), *Proceedings of the 37th Annual Meeting of the Cognitive Science Society* (pp. 1243–1249). Austin, TX: Cognitive Science Society.

Lazaridou, A., Bruni, E., & Baroni, M. (2014). Is this a wampimuk? cross-modal mapping between distributional semantics and the visual world. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 1403–1414). Stroudsburg, PA: Association for Computational Linguistics.

Lazaridou, A., Chrupała, G., Fernández, R., & Baroni, M. (2016). Multimodal semantic learning from child-directed input. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 387–392). Stroudsburg, PA: Association for Computational Linguistics.

Lazaridou, A., Marelli, M., & Baroni, M. (2017). Multimodal word meaning induction from minimal exposure to natural text. *Cognitive Science*, *41*, 677–705.

LeCun, Y. (1998). The MNIST database of handwritten digits. http://yann.lecun.com/exdb/mnist/.

Lewis, M., & Frank, M. (2013). An integrated model of concept learning and word-concept mapping. In M. Knauff, M. Pauen, N. Sebanz, & I. Wachsmuth (Eds.), *Proceedings of the 35th Annual Meeting of the Cognitive Science Society* (pp. 882–887). Austin, TX: Cognitive Science Society.

Manning, C. D., Clark, K., Hewitt, J., Khandelwal, U., & Levy, O. (2020). Emergent linguistic structure in artificial neural networks trained by self-supervision. *Proceedings of the National Academy of Sciences*, *117*, 30046–30054.

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.

McClelland, J. L., McNaughton, B. L., & O'Reilly, R. C. (1995). Why there are complementary learning systems in the hippocampus and neocortex: Insights from the successes and failures of connectionist models of learning and memory. *Psychological Review*, *102*(3), 419.

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.

Medina, T. N., Snedeker, J., Trueswell, J. C., & Gleitman, L. R. (2011). How words can and cannot be learned by observation. *Proceedings of the National Academy of Sciences*, *108*(22), 9014–9019.

Meylan, S. C., & Bergelson, E. (2021). Learning through processing: Toward an integrated approach to early word learning. *Annual Review of Linguistics*, *8*, 77–99.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space [Preprint]. arXiv:1301.3781.

Nikolaus, M., & Fourtassi, A. (2021). Evaluating the acquisition of semantic knowledge from cross-situational learning in artificial neural networks. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics* (pp. 200–210). Stroudsburg, PA: Association for Computational Linguistics.

Oord, A. v. d., Li, Y., & Vinyals, O. (2018). Representation learning with contrastive predictive coding. [Preprint]. arXiv:1807.03748.

Orhan, E., Gupta, V., & Lake, B. M. (2020). Self-supervised learning through the eyes of a child. In H. Larochelle, M'A. Ranzato, R. Hadsell, M.-F. Balcan, & H.-T. Lin (Eds.), *Advances in Neural Information Processing Systems* (Vol. 33, pp. 9960–9971).

Peterson, J. C., Abbott, J. T., & Griffiths, T. L. (2018). Evaluating (and improving) the correspondence between deep neural networks and human representations. *Cognitive Science*, *42*(8), 2648–2669.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. (2021). Learning transferable visual models from natural language supervision [Preprint]. arXiv:2103.00020.

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

Samuelson, L. K., & Horst, J. S. (2007). Dynamic noun generalization: moment-to-moment interactions shape children's naming biases. *Infancy*, *11*(1), 97–110.

Santoro, A., Bartunov, S., Botvinick, M., Wierstra, D., & Lillicrap, T. (2016). Meta-learning with memory-augmented neural networks. In *International Conference on Machine Learning* (pp. 1842–1850). MLR Press.

Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. [Preprint]. arXiv:1409.1556.

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

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

Taxitari, L., Twomey, K. E., Westermann, G., & Mani, N. (2020). The limits of infants' early word learning. *Language Learning and Development*, *16*(1), 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.

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 [Preprint]. arXiv:2006.02802.

Twomey, K. E., Ranson, S. L., & Horst, J. S. (2014). That's more like it: Multiple exemplars facilitate word learning. *Infant and Child Development*, *23*(2), 105–122.

Wojcik, E. H. (2017). 2.5-year-olds' retention and generalization of novel words across short and long delays. *Language Learning and Development*, *13*(3), 300–316.

Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., Zemel, R., & Bengio, Y. (2015). Show, attend and tell: Neural image caption generation with visual attention. In *International Conference on Machine Learning* (pp. 2048–2057). MLR Press.

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

Yu, C., & Smith, L. B. (2012). Modeling cross-situational word–referent learning: Prior questions. *Psychological Review*, *119*(1), 21.

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

Zettersten, M., & Saffran, J. R. (2019). Sampling to learn words: Adults and children sample words that reduce referential ambiguity. In A. K. Goel, C. M. Seifert, & C. Freksa (Eds.), *Proceedings of the 41st Annual Conference of the Cognitive Science Society* (pp. 1261–1267). Montreal, Canada: Cognitive Science Society.

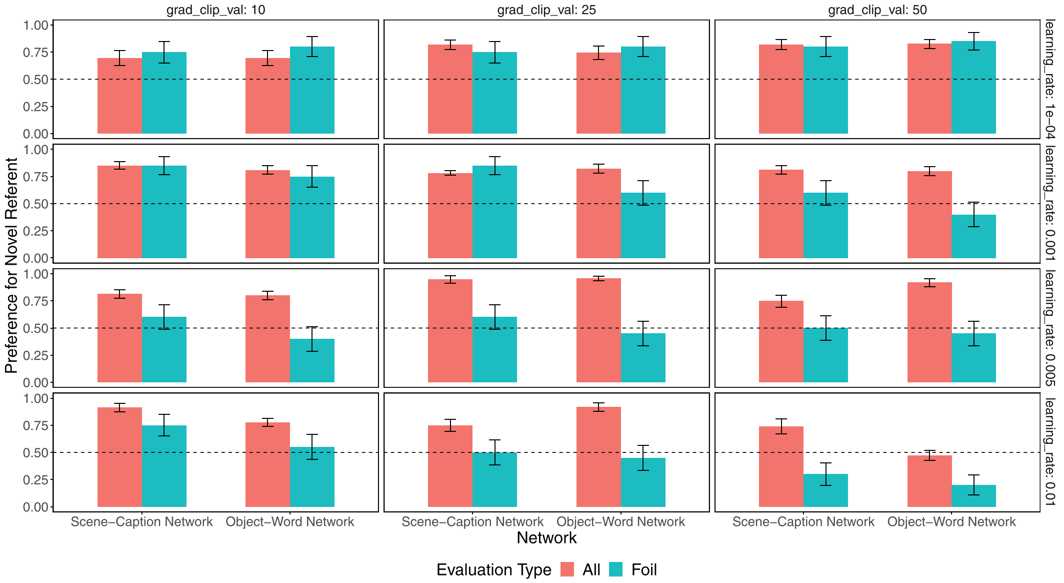

Fig. A1. Mutual exclusivity results. We further examined the effect of mutual exclusivity under different training configurations for both networks, varying the learning rate (rows) or the degree of gradient clipping employed (columns). Both models were also evaluated in the *All* and *Foil* conditions.

Appendix A: Mutual exclusivity

As shown in Fig. A1, we ran additional simulations looking at some other factors that influenced whether a model would display evidence for mutual exclusivity. We varied both the learning rate the models were trained on, as well as the amount of gradient clipping to apply. In the *All* condition, where the model is presented with the novel referent and a randomly selected familiar referent, both the scene-caption and object-word networks show a strong preference for the novel referent. However, in the more challenging *Foil* condition where the familiar referent was the one presented alongside the novel referent, we can observe that there is a lot of variation in which referent the networks favor, and that results are not as consistent across the different training configurations. We did not observe this kind of qualitative shift in any of the other simulation results.