Big Generalizations with *Small* Data: Exploring the Role of Training Samples in Learning Adjectives of Size

Sandro Pezzelle and **Raquel Fernández**
Institute for Logic, Language, and Computation
University of Amsterdam
{s.pezzelle|raquel.fernandez}@uva.nl

Abstract

In this paper, we experiment with a recently proposed visual reasoning task dealing with quantities – modeling the multimodal, contextually-dependent meaning of size adjectives ('big', 'small') – and explore the impact of varying the training data on the learning behavior of a state-of-art system. In previous work, models have been shown to fail in generalizing to *unseen* adjective-noun combinations. Here, we investigate whether, and to what extent, seeing some of these cases during training helps a model understand the rule subtending the task, i.e., that being *big* implies being *not small*, and vice versa. We show that relatively few examples are enough to understand this relationship, and that developing a specific, mutually exclusive representation of size adjectives is beneficial to the task.

1 Introduction

A recently proposed visual reasoning task challenges models to learn the meaning of *size* adjectives ('big', 'small') from visually-grounded contexts (MALeViC; Pezzelle and Fernández, 2019). Differently from standard approaches in language and vision treating size as a *fixed* attribute of objects (Johnson et al., 2017), in MALeViC what counts as 'big' or 'small' is defined *contextually*, based on a cognitively-motivated threshold function evaluating the size of all the *relevant* objects in a scene (Schmidt et al., 2009). In the most challenging version of the task, SET+POS, the subset of relevant objects (i.e., the reference set) comprises all the objects belonging to the same category as the queried one. Given a scene depicting a number of colored shapes (e.g., the leftmost image in Figure 1) and a sentence about one object's size (e.g., 'The white rectangle is a *big* rectangle'), models have to assess whether the sentence is *true* or *false* in that context; i.e., whether the white rectangle is *big* given the other rectangles in the scene.

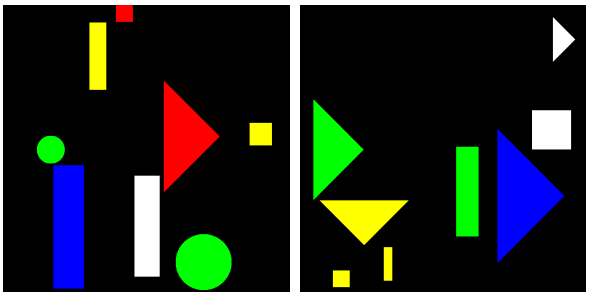

Figure 1: SET+POS. Two original (ORIG) examples. **Left (ORIG):** The white rectangle is a *big* rectangle, True. **Right (ORIG):** The blue triangle is a *small* triangle, False. To test model abilities in handling *unexpected* cases, an increasing number of ORIG training samples is modified by *swapping* both the size adjective and its ground-truth (SWAP). **Left (SWAP):** *small*, False. **Right (SWAP):** *big*, True. Best viewed in color.

Among the tested models, FiLM (Perez et al., 2018) turned out to be the best overall architecture for the task. However, when tested with adjective-noun combinations that were never *seen* in training (i.e., the model has been taught what means to be *big* for circles and rectangles or *small* for triangles and squares, but not, e.g., what means to be *small* for a circle), FiLM was shown to use a default strategy which ignores the adjective rather than applying it *compositionally*. This finding is in line with previous evidence showing the lack of compositionality in neural networks (Baroni, 2019), either in multimodal tasks like visual question answering (Agrawal et al., 2017) and visual reasoning (Johnson et al., 2017), or when coping with language data (Lake and Baroni, 2018; Loula et al., 2018). To solve this well-known issue, several attempts have been made to develop new models and techniques (Agrawal et al., 2018; Ramakrishnan et al., 2018; Korrel et al., 2019), and several datasets have been proposed to test compositional abilities of systems (Agrawal et al., 2017, 2018).

Proceedings of the Beyond Vision and LANguage: inTEgrating Real-world kNowledge (LANTERN), pages 18–23
Hong Kong, China, November 3, 2019.

In this work, we focus on a tightly related problem, that is, generalization with little data. Since models do not learn an *abstract* representation of 'big' and 'small' that can be applied compositionally to *unseen* examples, we test whether, and to what extent, this problem can be alleviated by seeing some of these 'unseen' cases during training. We refer to these cases as *unexpected* (due to their low frequency compared to the more frequent, *expected* ones), and test whether injecting an increasing proportion of these examples in the training data helps models understand the rule subtending the task, i.e., that being *big* implies being *not small*, and vice versa. Intuitively, the model could (1) stick to the default strategy and correctly predict only the *expected* examples, or (2) learn a more general rule that also accounts for the *unexpected* cases. Here, we are interested in checking how much data is required to start adopting the latter strategy, and aim to understand what information the model exploits while performing the task.

To explore these issues, we focus on the SET+POS task and the best-performing FiLM model, and build 7 new training settings with an increasing proportion of *unexpected* cases.[1] Such examples are obtained by simply swapping the original size adjective and its ground-truth answer, as described in Figure 1. By training the model on each of these settings, we show that very little *unexpected* data is needed to obtain high generalizations in testing, and that seeing these examples is beneficial to learn the rule subtending the task.

2 Generalizing to *Unexpected* Data

Method We explore whether injecting some *unexpected* cases in training data helps the model understand the relation that holds between the adjectives 'big' and 'small'. We use the 10K-datapoint (8K training, 1K val, 1K test) SET+POS dataset (hence, A) used by Pezzelle and Fernández (2019) in their *compositional* experiment, and build 7 new training settings containing an increasing percentage of *unexpected* examples. We refer to these settings using capital letters from B to H. They contain 0.8%, 1.6%, 3.2%, 6.4%, 12.8%, 25.6%, and 50.0% *unexpected* cases, respectively. To generate the new training settings, we sample a given percentage of datapoints (e.g., 0.8% for B) from the original training/validation files and simply *swap* the original adjective and ground-truth answer (see Figure 1). While doing so, we ensure that a balanced number of cases is modified for each <adjective-noun, ground truth> tuple. To illustrate, out of the 8 modified cases in the validation split of B, 2 involve circles; out of these, one is originally a <big-circle, true> case, the other a <big-circle, false>. This makes all 7 settings perfectly balanced with respect to shape, size, and ground truth.[2] This prevents biases in the data, e.g., that circles are more likely to be *big* than squares. It is worth mentioning that, compared to A, only (some) sentences and answers are modified. As for the visual data, all settings employ the exact same 10K images and visual features precomputed using ResNet-101 (He et al., 2016).

Model We experiment with FiLM (Perez et al., 2018) using the best configuration of hyperparameters and the same experimental pipeline reported in Pezzelle and Fernández (2019). In each setting, the model is trained for 40 epochs with 3 random initializations. For each of these 3 runs, the best model epoch based on accuracy on the validation split is selected and then tested on 3 different test sets: (a) *seen* (1K datapoints), where all the examples are *expected*, (b) *unseen* (1K), where all the examples are *unexpected*, and (c) *balanced* (2K), where a balanced number of *expected* and *unexpected* cases is present. All test sets are taken from Pezzelle and Fernández (2019).

Results In Table 1 we report, for each setting, average model accuracy and standard deviation (sd) over 3 runs (the same results are visualized in Figure 2). Starting from the *unseen* test set, we notice that injecting an extremely low percentage of *unexpected* cases in B (0.8%, i.e., 64/8000 cases in training) has already some impact on the accuracy, with a 12-point increase (27%) compared to A (15%). This pattern is observed in the subsequent settings, with accuracy increasing to 44% in C and to 45% in D. The most striking result is observed in setting E, where model accuracy gets well above chance level (65%) with a percentage of just 6.4% *unexpected* cases seen in training (see also Figure 2, where the blue line exceeds chance level in E). This clearly indicates that the model, instead of just trying to correctly predict all the *expected* cases, which would potentially lead to a

[1] Data, code, and trained models are available at: https://github.com/sandropezzelle/malevic.

[2] Note that we do not balance with respect to color since this would increase by 5 the number of modified examples.

test set	average accuracy ± sd								
	A [0.0]*	**B [0.8]**	**C [1.6]**	**D [3.2]**	**E [6.4]**	**F [12.8]**	**G [25.6]**	**H [50.0]**	**16K [50.0]***
seen	**0.85 ± 0.01**	0.84 ± 0.02	0.83 ± 0.04	0.74 ± 0.01	0.75 ± 0.04	0.81 ± 0.01	0.74 ± 0.03	0.75 ± 0.01	0.91 ± 0.02
unseen	0.15 ± 0.02	0.27 ± 0.03	0.44 ± 0.00	0.45 ± 0.03	0.65 ± 0.04	0.74 ± 0.03	0.72 ± 0.02	**0.75 ± 0.03**	0.90 ± 0.02
balanced	0.50 ± 0.00	0.54 ± 0.03	0.65 ± 0.04	0.60 ± 0.01	0.71 ± 0.05	**0.79 ± 0.03**	0.73 ± 0.03	0.77 ± 0.03	0.88 ± 0.02

Table 1: Average accuracy ± standard deviation by FiLM on 3 test sets in settings A-H (in brackets, proportion of *unexpected* cases seen in training). For comparison, performance by best model trained with 16K datapoints is reported (16K). * refers to models trained in Pezzelle and Fernández (2019). In **bold**, highest number in the row.

93.6% accuracy, employs a learning strategy that is valuable also for *unexpected* examples.

It is interesting to note, in this regard, that on the *seen* test set the model experiences a performance drop from A (85%) to H (75%), which shows how an increasing proportion of *unexpected* cases makes guessing the *expected* ones a bit harder (this is, to some extent, intuitive since in A there are only 4 *seen* adjective-noun combinations); indeed, the overall best accuracy in *seen* is obtained with A, while the best accuracy in *unseen* is obtained with H, where the highest proportion of *unexpected* examples is given in training. As for the *balanced* test set, we observe that an increasing proportion of *unexpected* cases in training boosts model generalization, though F turns out to slightly outperform H (79% vs 77%) due to its better performance on the *expected* (*seen*) instances. Finally, it should be noted that training with twice as many samples (16K) leads to a significantly higher accuracy in all test sets (+11-16 points compared to H), which shows a 'the bigger, the better' effect of training set size on model performance in the task.

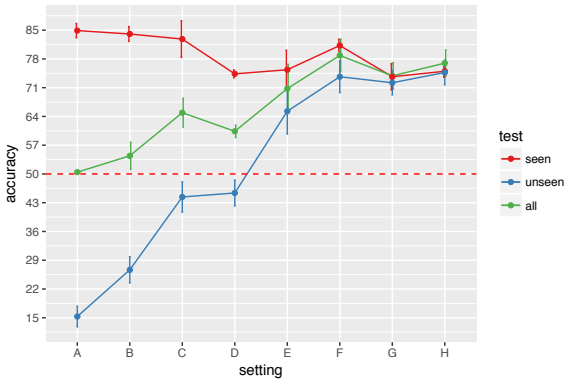

Figure 2: FiLM performance on 3 test sets across settings A-H. Average accuracy over 3 runs (dots) and standard deviation (bars) are reported. The red dashed line indicates chance level. Best viewed in color.

3 Analysis

Linguistic representations In FiLM, the representation of the sentence obtained via the Gated Recurrent Unit (GRU; Chung et al., 2014) influences the CNN computation to focus on image features that are important to solve the task. Thus, examining it could shed light on the type of linguistic information exploited by the model. Here, we are interested in checking how much *size* information is encoded by the GRU in each setting. We run each setting's best trained model on the *balanced* test set and, for each sentence, we extract the final 4096-d GRU hidden state. We then perform a 2-dimensional PCA analysis on these 2K embeddings: if the model pays attention to size adjectives, embeddings containing 'big' ('small') should be overall similar/close to each other, but different/far from those containing 'small' ('big').

In Figure 3, we plot the results of the PCA analysis for settings A, B, C, and H (from left to right). In A, where each shape type is always either 'big' or 'small', embeddings are clearly grouped in 4 clusters corresponding to each shape (labels not reported for clarity of presentation), while no pattern regarding size is observed (i.e., red and blue dots are mixed together). This shows that, in A, the GRU does not learn a specific representation for 'big' and 'small', in line with the hypothesis that the model just ignores these words (Pezzelle and Fernández, 2019). This is confirmed by the results of an additional analysis where we tested the models trained in A on sentences (either from the *seen* or *unseen* test set) from which the size adjective is removed (e.g., 'The white rectangle is a rectangle'). As conjectured, no differences in accuracy compared to the standard setting were observed (i.e., 0.85 in *seen*; 0.15 in *unseen*). In B, in contrast, some information about size is encoded (embeddings containing a 'big' shape are 'South-East' to those containing the same shape 'small'), with this pattern becoming clearer in C, where 'big' and 'small' are neatly separated by

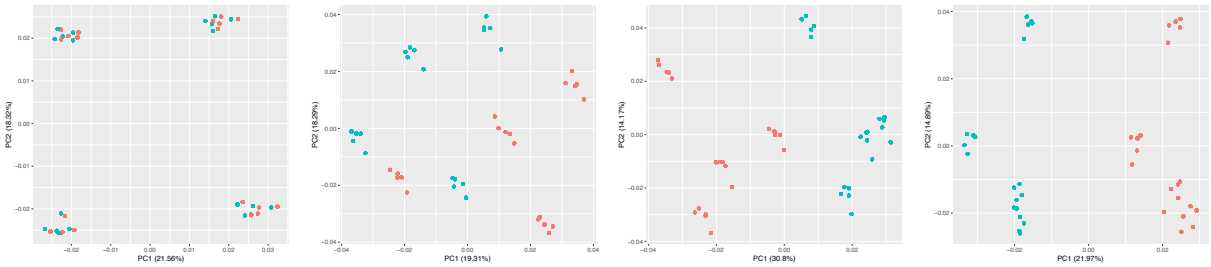

Figure 3: PCA analysis on 2K GRU sentence embeddings by each best model on the *balanced* test set in settings A, B, C, and H (from left to right). Red dots correspond to sentences embedding 'big', blue to 'small'. 'Big', 'small' become progressively separated as the proportion of *unexpected* samples increases. Best viewed in color.

PC1. This distance increases in the subsequent settings (not reported), and becomes extreme in H, where the size adjective is the most discriminative linguistic feature. By testing the models trained in H on the 'without adjective' test sentences, indeed, we obtain an accuracy that is close to chance level (i.e., 0.49 in *seen*; 0.53 in *unseen*), which clearly indicates that the model is unable to perform the task without the size adjective. In sum, seeing more and more *unexpected* cases helps the model develop an increasingly specific, mutually exclusive representation of size adjectives, which goes hand in hand with a better performance.

To quantitatively assess this pattern, we evaluate the similarity between 'big'/'small' embeddings by (1) averaging all the embeddings containing the same adjective, (2) computing the cosine similarity between the two centroids. If the model progressively develops a mutually exclusive representation for 'big' and 'small', the similarity should decrease across settings; in contrast, such a pattern should not be found for shape (the meaning of, e.g., *square* is not supposed to change).[3] The expected pattern is shown in Figure 4, with similarity starting very high in A and rapidly decreasing with an increasing proportion of *unexpected* cases. Note that, in A, there is almost no difference between 'big' and 'small'. This is somehow intuitive since, in the *balanced* test set, the sentences in the 'big' centroid are exactly the same as those in the 'small' one, except for the size adjective. As for shape, a rather 'flat' pattern is observed.

Mutual exclusivity of predictions An insightful way to test whether FiLM has learned a mutually exclusive representation for 'big' and 'small' is to consider its predictions for the original (ORIG) and swapped (SWAP) test samples. If the model has learned that being *big* implies being *not small*, and vice versa, we should expect it not to output the same answer (e.g., *true*) to both questions. To explore this issue, we first obtain model predictions on both the *seen* and *unseen* test set. We then take either test set and, for each sample, we swap the size adjective and the ground truth (see Figure 1). This way, we obtain two SWAP test sets where ground truths are systematically reversed compared to ORIG. We obtain model predictions on each SWAP test set, compare them to those on the corresponding ORIG, and count the number of non-overlapping (i.e., mutually exclusive) predictions. As shown in Figure 5, mutual exclusivity is close to 0 in A, where FiLM outputs (almost) always the same answer to both ORIG and SWAP samples, and progressively increases across settings, which boosts FiLM's generalization ability. This pattern of results is in line with what is reported by Gandhi and Lake (2019), i.e., that standard neural networks lack the ability to reason with mutual exclusivity. Until there is a balanced enough number of 'big' and 'small' ex-

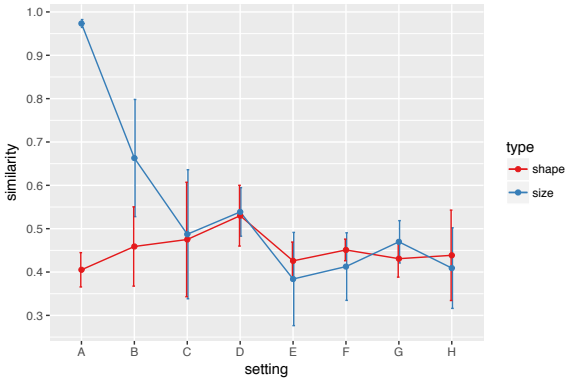

Figure 4: Similarity between sentence embeddings grouped by *shape* or *size*. Best viewed in color.

[3]For shape, we obtain an average representation for each shape (*circle*, *square*, etc.), compute all pairwise similarities between the 4 centroids, and compute the average similarity.

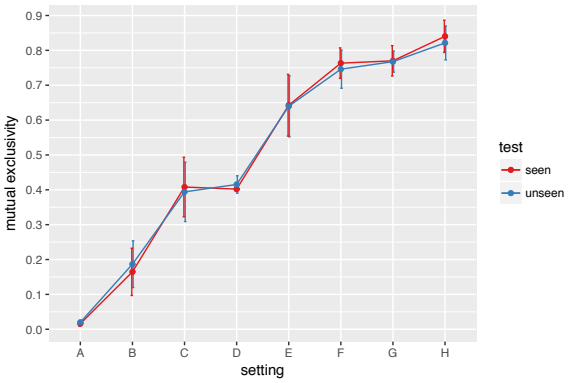

Figure 5: Mutual exclusivity (ME) of predictions between ORIG and SWAP *seen/unseen*. The less overlapping predictions, the higher ME. Best viewed in color.

amples in training, indeed, the model does not fully *understand* the mutually exclusive relation tying the two adjectives; rather, it makes predictions that are biased toward the most frequent, *expected* instances.

4 Generalization vs Compositionality

The results described above show the ability of FiLM to make powerful generalizations with little data. However, this is not informative of its *compositional* skills since, in settings B-H, the same proportion of *unexpected* cases is seen by the model for each shape type. As a consequence, the model is not required to apply the 'big'/'small' relation learned for, say, circles to, say, squares. Here, we test whether learning the rule for some shape types makes the model able to apply it to other shapes. Crucially, this is different from the compositional experiment in Pezzelle and Fernández (2019) (here referred to as setting A) where the 'big'/'small' relation had to be learned across shapes.

We train the model with perfectly balanced data (as in H) for triangles and circles, and perfectly unbalanced data (as in A) for squares and rectangles. More in detail, the model is trained with sentences containing the following queries:[4] *big triangle* (1K datapoints), *small triangle* (1K), *big circle* (1K), *small circle* (1K), *small square* (2K), *big rectangle* (2K), and is then tested with the usual *seen* and *unseen* test sets. If the model learns the abstract, mutually exclusive relation between 'big' and 'small' by being exposed to examples of these two adjectives combined with two different shape types, it should then be able to *compositionally* apply the rule to the other two types of shape. Otherwise, a similar pattern as the one observed in setting A should be found for squares and rectangles.

On the *seen* test set, where all the adjective-noun combinations are seen in training, the model obtains an average accuracy (over 3 runs) of 0.81. On the *unseen* one, in contrast, it stops at 0.64. As expected, this worse performance is due to the extremely low accuracy on *big square* (0.22) and *small rectangle* (0.23), i.e., the cases that were never seen in training. This opposite pattern of results (triangle and circle vs square and rectangle) suggests that the model learns a 'big'/'small' rule that is shape-dependent and cannot be *compositionally* applied to other shapes. This is confirmed by the results obtained when testing the model on the 'without adjective' test sentences: in the best model run, e.g., chance-level accuracy is observed for triangles and circles in either test set (i.e., the model 'needs' the adjective to perform the task), while the same numbers as those obtained with the default sentences are observed for squares and rectangles (i.e., the adjective is 'ignored').

5 Conclusion

Previous work has reported the inability of FiLM to apply 'big', 'small' to *unseen* adjective-noun combinations (Pezzelle and Fernández, 2019). Here, we show that seeing some of these cases in training mitigates the problem, leading to high generalizations (in line with Lake and Baroni, 2018) and helping the model understand the mutually exclusive status of size adjectives. Although the model can learn the 'big'/'small' rule, this rule is shown to be shape-dependent; i.e., it cannot be learned for some nouns and *compositionally* applied to others for which direct evidence was not observed during training. Taken together, these findings indicate that models fail to apply rules *compositionally*, but are extremely good at generalizing to even rarely *seen* examples.

Acknowledgments

We are grateful to the anonymous reviewers of the LANTERN workshop for their valuable feedback. This work was funded by the Netherlands Organisation for Scientific Research (NWO) under VIDI grant no. 276-89-008, *Asymmetry in Conversation*.

[4] We employ the same 8K training datapoints and images used in the previous experiments.

References

Aishwarya Agrawal, Dhruv Batra, Devi Parikh, and Aniruddha Kembhavi. 2018. Don't just assume; look and answer: Overcoming priors for visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4971–4980.

Aishwarya Agrawal, Aniruddha Kembhavi, Dhruv Batra, and Devi Parikh. 2017. C-VQA: A compositional split of the Visual Question Answering (VQA) v1.0 dataset. *arXiv preprint arXiv:1704.08243*.

Marco Baroni. 2019. Linguistic generalization and compositionality in modern artificial neural networks. ArXiv preprint arXiv:1904.00157, to appear in the Philosophical Transactions of the Royal Society B.

Junyoung Chung, Caglar Gulcehre, Kyunghyun Cho, and Yoshua Bengio. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. In *Workshop on Deep Learning at NIPS-2014*.

Kanishk Gandhi and Brenden M Lake. 2019. Mutual exclusivity as a challenge for neural networks. *arXiv preprint arXiv:1906.10197*.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. 2017. CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2901–2910.

Kris Korrel, Dieuwke Hupkes, Verna Dankers, and Elia Bruni. 2019. Transcoding compositionally: Using attention to find more generalizable solutions. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 1–11, Florence, Italy. Association for Computational Linguistics.

Brenden Lake and Marco Baroni. 2018. Generalization without systematicity: On the compositional skills of sequence-to-sequence recurrent networks. In *International Conference on Machine Learning*, pages 2879–2888.

Joao Loula, Marco Baroni, and Brenden Lake. 2018. Rearranging the familiar: Testing compositional generalization in recurrent networks. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 108–114.

Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville. 2018. Film: Visual reasoning with a general conditioning layer. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Sandro Pezzelle and Raquel Fernández. 2019. Is the Red Square Big? MALeViC: Modeling Adjectives Leveraging Visual Contexts. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing*. Forthcoming.

Sainandan Ramakrishnan, Aishwarya Agrawal, and Stefan Lee. 2018. Overcoming language priors in visual question answering with adversarial regularization. In *Advances in Neural Information Processing Systems*, pages 1541–1551.

Lauren A Schmidt, Noah D Goodman, David Barner, and Joshua B Tenenbaum. 2009. How tall is tall? Compositionality, statistics, and gradable adjectives. In *Proceedings of the 31st annual Conference of the Cognitive Science Society*, pages 2759–2764.