Flexible Compositional Learning of Structured Visual Concepts

Yanli Zhou (yanlizhou@nyu.edu)
Center for Data Science
New York University

Brenden M. Lake (brenden@nyu.edu)
Department of Psychology and Center for Data Science
New York University

Abstract

Humans are highly efficient learners, with the ability to grasp the meaning of a new concept from just a few examples. Unlike popular computer vision systems, humans can flexibly leverage the compositional structure of the visual world, understanding new concepts as combinations of existing concepts. In the current paper, we study how people learn different types of visual compositions, using abstract visual forms with rich relational structure. We find that people can make meaningful compositional generalizations from just a few examples in a variety of scenarios, and we develop a Bayesian program induction model that provides a close fit to the behavioral data. Unlike past work examining special cases of compositionality, our work shows how a single computational approach can account for many distinct types of compositional generalization.

Keywords: concept learning; Bayesian inference; few-shot learning; visual learning; compositionality

Introduction

Humans have a remarkable capacity to learn new concepts from limited data. Early in development, children can make meaningful generalizations from just one or few positive examples of a new word (Smith et al., 2002; F. Xu & Tenenbaum, 2007), an ability known as few-shot learning. Critical to few-shot learning is compositional generalization, the reuse and manipulation of preexisting knowledge of parts and relations to understand novel combinations (e.g., Biederman, 1987). For example, people who are familiar with *coffee maker*, *toaster oven* and *griddle* can effortlessly grasp the concept of *breakfast machine* upon seeing it for the first time (Fig. 1A).[1] On the other hand, computer vision models, while highly successful in many applications, are far more limited in their abilities to form compositional generalizations (Lake et al., 2017). For instance, a pre-trained ResNet-50 (He et al., 2016) classifies the new concept in Fig. 1A as a *"waffle iron,"* whereas a strong image captioning system (K. Xu et al., 2015) describes it as *"a close up of a toaster oven with some muffins in it."*

There are qualitatively different types of composition present in real-world visual concepts, posing a challenging learning problem that demands manipulating parts and relations at various levels of abstraction (see examples in Fig. 1B). A concept like *bicycle* stipulates a fixed configuration of parts and relations (e.g. bikes have handlebars, a seat, and two wheels in a consistent configuration), whereas a concept like *vehicle* allows category members to have freer combinations of parts and relations (varying numbers of wheels, motors, etc. are acceptable). A concept like *sun shield* requires selectivity of object orientation, in order to fulfill a given conceptual constraint. Finally, a concept like *clothing that comes in pairs* requires an additional degree of compositional abstraction, allowing a variety of parts to fill a role as long as they are duplicated.

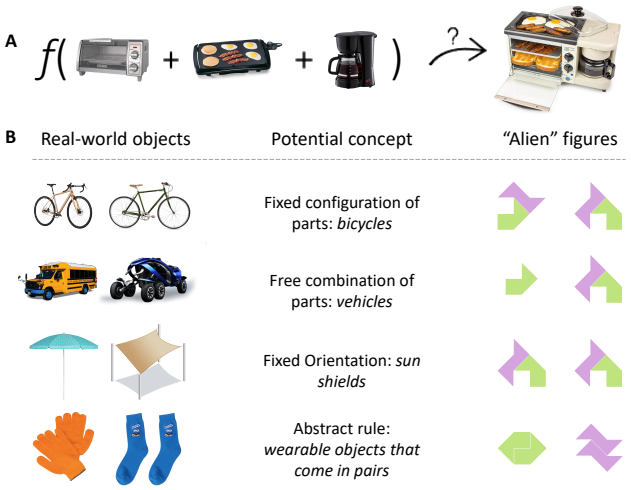

Figure 1: Visual concept learning requires flexible notions of compositional structure. (A) Humans can learn the concept of *breakfast machine* with a single example by recognizing familiar components and reasoning about their relations. Leading computer vision models tend to struggle with this concept. (B) Real-world visual concepts are defined by different types of compositions: 1. A *bicycle* is a well-defined collection of parts in a consistent configuration; 2. *vehicles* allow a set of stereotyped parts to be combined more freely; 3. To be a *sun shield*, an upright orientation is required; 4. *Wearable objects that come in pairs* stipulate a repetition of *wearable* elements. The rightmost column contains examples of experimental stimuli that are analogous to these concepts.

Although previous work on few-shot learning has examined special cases of compositionality, we are still far from understanding the full variety of compositions present in real-world visual concepts (Fig. 1B). In a seminal study, F. Xu and Tenenbaum (2007) examined word learning as Bayesian inference over tree-structured hypothesis spaces. Their model explains how children can make meaningful inferences from just a few examples, but compositional concepts were not considered. Lake, Salakhutdinov, and Tenenbaum (2015) developed compositional models of learning handwritten characters, although individual characters are highly constrained in how their parts and configuration are allowed to vary (as in the 1st row of Fig. 1B). Other studies have considered sequential patterns (Overlan, Jacobs, & Piantadosi, 2017; Lake, Linzen, & Baroni, 2019) and recursive structures (Stuhlmuller, Tenenbaum, & Goodman, 2010; Lake & Piantadosi, 2020) more akin to the 4th row of Fig. 1B, or free combinations of parts akin to the 2nd row of Fig. 1B ar-

[1] Example from Vicarious Research Blog.

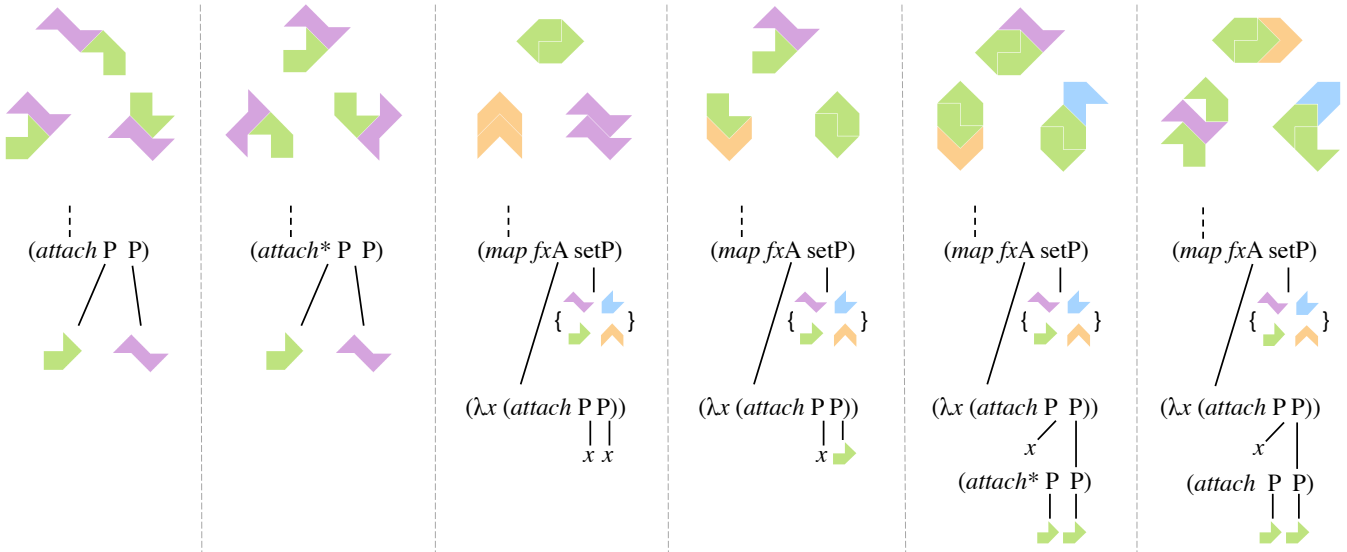

Figure 2: Examples of trial types tested in the experiments. Top: example alien figures given to participants to study on a given trial. Bottom: simplified parse tree of the most likely concept inferred by the Bayesian program induction model for each trial. The grammar over programs specifies primitive shapes and operations including the *attach* function, which returns the set of all possible configurations of two parts, and the *attach** function which returns a set containing the single specified configuration. We can see that the spatial arrangement among components cannot be described with simple relations such as *left*, *right*, *above* or *under*.

ranged in grid-like scenes (Orbán, Fiser, Aslin, & Lengyel, 2008), although each of these concept types considered relatively simplistic spatial relations.

Our goal here is to study these various types of visual composition in a single experimental paradigm, and evaluate the success of Bayesian program induction in accounting for the inferences people make. To do so, we consider a domain of visual concepts that is richly hierarchical, compositional, and relational. Using "alien figures" as our stimuli, we conducted two experiments on few-shot concept learning, asking participants to make generalization judgements on test items after observing only a small number of positive examples. Following previous modeling work on Bayesian program induction in the visual domains (Stuhlmuller et al., 2010; Lake et al., 2015; Overlan et al., 2017; Lake & Piantadosi, 2020), we formalize learning in the alien categorization game as a search for the best programs for explaining the examples under a Bayesian score. The space of possible programs is constructed using a probabilistic language of thought (PLoT) (Goodman, Tenenbaum, Feldman, & Griffiths, 2008; Piantadosi, 2011; Piantadosi & Jacobs, 2016), allowing for a wide range of compositions and abstractions. We found that our Bayesian program induction model provides an excellent account of experimental data, outperforming alternative models that lack key capacities to represent relations and compositionality. In addition, the fitted model parameters are psychologically meaningful, providing insight into people's inductive biases for these few-shot learning tasks.

Behavioral Experiments

Our experiments aimed to evaluate the flexibility of human compositional learning across a range of concept types. We adapted the few-shot learning paradigm of F. Xu and Tenenbaum (2007) for our purposes, as described below.

Stimuli. The stimuli were described to participants as "alien figures," which were programmatically generated by composing one to three shape primitives (see examples in Fig 2). A composition of two parts is considered valid when they are non-overlapping and connected via two sides of identical length. Participants saw black-and-white outlines of each primitive[2]. We left these primitives uncolored to motivate closer observations of stimulus shapes. As a visual aid in the experiment, rolling one's mouse over a primitive led all identical primitives in the display to become highlighted. The primitives were constructed through an additional level of compositionality, as they were composed of four isosceles right triangles. To form each set of training examples, we varied (1) which primitives can appear, (2) how many primitives appear in each exemplar (3) how the parts are composed and (4) if the configuration has a fixed orientation.

Task. Participants took part in an "alien figure categorization game" in which they are the assistant to a professor who collected samples of alien figures on a newly discovered planet. Their job in the game is to help the professor categorize a series of unnamed alien figures based on a small set of named examples.

During each trial, participants were first familiarized with four different shape primitives. They were also informed that all relevant figures within the trial were built from these four primitives and no other primitives were possible. Next, par-

[2]In this paper, all alien figures are shown with color-coded shape primitives for clarity.

ticipants were given a small set of example figures that shared a common name (see Fig. 4 for example trials). To minimize the effect of memory demands on learning, a display of the examples and primitives remained on screen throughout the trial. After an untimed observation period, participants entered a test stage in which they categorized a series of 9-13 unnamed alien figures. Specifically, participants chose 'yes' or 'no' for each test image to indicate whether it belongs to the same named category as the example images. We constructed each test set to cover a wide range of both possible and impossible extensions of potential concepts.

We conducted two separate experiments with identical task procedures. The two experiments differed only in terms of the training and test sets in each trial. In Experiment 1, for every participant we tested 11 trials with each trial containing one to three training examples, followed by judgments on the test examples. Experiment 2 consisted of 10 trials and considered concept types that were more complex compared to those used in Experiment 1. We also used Experiment 2 to evaluate out-of-sample model predictions, since all model parameters were fit on the basis of Experiment 1. To study the effect of the exemplar set size on learning, participants in Experiment 2 were randomly separated into two conditions, based on whether they saw three or six exemplars of each concept. Trial orders were randomized for each participant; the set of allowable primitives were also randomized per participant per trial.

Participants. We used Amazon Mechanical Turk to recruit participants for both online experiments. Forty participants took part in Experiment 1, and 30 for each condition of Experiment 2. Responses from participants that failed one or more attention checks during either task were excluded. In the end, generalization judgements from 32, 25 and 20 participants were used in our reported analyses of Experiment 1, the 3-example condition of Experiment 2, and the 6-example condition of Experiment 2, respectively. All participants finished the task within an hour and were paid $5.00 at the completion of the experiment.

Computational models

We explored several types of computational models, with the aim of characterizing human generalizations in computational terms.

Bayesian program induction

To provide a unifying computational account of the wide range of generalization behavior elicited by various composition types, we developed a Bayesian program induction model that considers explicit, structural hypotheses as explanations for novel visual concepts. The model updates its beliefs over these hypotheses using a Bayesian framework (Goodman et al., 2008; Piantadosi & Jacobs, 2016) which generates human-like graded predictions with very limited data. In particular, the alien concepts were represented as probabilistic programs, which are structured generative models that produce distributions of examples. The goal of the learner is to infer programs consistent with the observed examples and the prior beliefs over programs. Inspired by Piantadosi (2011) and Piantadosi, Tenenbaum, and Goodman (2016), we formed a compositional hypothesis space using a probabilistic grammar based on λ-*calculus*. The grammar defines a set of primitive parts and operations which can be combined to build up programs of various levels of complexity (see Fig. 2 for examples of programs and output). Each sample from the grammar corresponds to a visual concept, and the production rules of the grammar specify the infinite space of possible concepts.

START		
Actions		
ATTACH	→ (*attach* P P)	*Returns the set of all allowable configurations of two parts*
ATTACH*	→ (*attach** P P)	*Returns a specific configuration of two parts*
ROTATE	→ (*rotate* P *d*)	*Returns a rotated copy of input at d degrees*
Parts		
PART	→ ATTACH*	*A fixed configuration*
PART	→ p_1, …, p_4	*A shape primitive*
PART	→ *x*	*A part variable*
Mapping & λ-expressions		
MAP	→ (*map fx*A SET)	*Maps an expression onto a set*
*fx*A	→ (λ *x* FUNC)	*Action expression with part variable*
Part-based functions		
PARTIAL	→ (*has* PART)	*Returns the set of all possible figures that contain a particular part*
PARTIAL	→ (*only* PART)	*Returns the set of all possible figures that consist only of a particular part*
Sets		
SET	→ (*diff* SET, SET)	*Removes the second set from first set*
SET	→ (*union* SET, SET)	*Returns a combined set of two sets*
SET	→ {p_1, …, p_4}	*The set of all shape primitives*

Figure 3: Core grammatical rules used to generate concept programs. The hypothesis space used in the study consisted of valid compositions of these primitives. Full grammar and code will be available online: https://github.com/yanlizhou/AlienFigures.

Prior over programs. To generate a concept, our grammar begins with expanding the START symbol into downstream nodes according to applicable rewrite rules. These nodes are subsequently rewritten until no further expansions are possible. Fig 3 shows the core set of rules used to generate the programs (concepts) considered in our study. The output of each program is the set of all possible alien figures under such concept. In the example `(rotate (attach`^{1} `p1 p2), 180)`, the inner most expression is first evaluated and returns the 1^{st} allowable configuration of primitives p_1 and p_2, which gets passed on to the outside expression that generates a rotated copy at 180°. This program has only a single element in its output set, as it corresponds to a generative process that fully specifies the types of parts, their configuration, and overall rotation. Figure orientation is based on four discrete possibilities, and two identical configurations at different rotations are considered distinct alien figures.

The grammar also supports λ-expressions; together with mapping and set operations, the grammar can produce abstract concepts like `(map (lambda x (attach x x)) S)`

which outputs the set of all possible configurations of two identical components sampled from the set S. Other function primitives in the grammar support hypotheses that do not fully specify a composition process. For example, `(has p)` returns the set of all possible alien figures with p as a part.

Likelihood and inference. In Bayesian concept learning, the learner aims to compute the probability of a hypothesis h given a set of examples $X = \{x_1, \ldots, x_k\}$, or the posterior probability $P(h|X)$, which can be calculated by applying Bayes' rule: $P(h|X) \propto P(X|h)P(h)$. The first component, the likelihood of X assuming hypothesis h is true is defined as

$$P(X|h) = \prod_i^k P(x_i|h) = \frac{1}{|h|^k},$$

where $|h|$ is the size of the concept. A likelihood function that is inversely proportional to the concept size, in our case the number of all unique outputs of a program, reflects the *size principle* which assigns more weight to smaller hypotheses (Tenenbaum, 1999). The second component $P(h)$, the prior probability of a concept, can be naturally derived from the grammar (Goodman et al., 2008). Since each production of the grammar is a sequence of expansions of non-terminals, the probability of the production is the product of the probabilities associated with each expansion. This formulation operationalizes an important psychological preference for simplicity (Chater & Vitányi, 2003) as shorter programs require fewer multiplications of expansion probabilities. To generate a model prediction for each test item y after making a set of observations, we calculate the probability that the label $l_y \in \{0,1\}$ of y is consistent with the set of observed examples X as

$$P(l_y = 1|X) = \sum_{h \in \mathcal{H}} P(l_y = 1|h)P(h|X)$$

where $\mathcal{H}$ is the hypothesis space considered in our study. Approximate posterior inference was implemented in the LOTlib3 software package (Piantadosi, 2014). For each trial, we ran three Monte Carlo chains for 100,000 steps of a tree-regeneration Markov chain Monte Carlo (MCMC) procedure (Goodman et al., 2008).

Parameter fitting. Given behavioral data collected in our experiments, we are interested in finding the set of grammar parameters that most likely generated people's generalization patterns. Formally, we would like to infer the probability of the set of parameters of interest, given human response data: $\arg\max_{\vec{\theta},\alpha,\beta} P(\vec{\theta},\alpha,\beta|R,Y)$, where $\vec{\theta}$, α and β are parameters of the learning model and R is the set of human responses to the set of test items Y. To account for possible response noise in our collected generalization judgements, we fit a lapse rate α, which determines the probability that a response was made at random. In the case of a lapse trial, we also represented a baseline preference for answering *Yes* with parameter β. $\vec{\theta}$ is the set of grammar parameters, which are the probabilities associated with the distribution of expansions for each non-terminal. We only considered two grammar parameters that are psychologically meaningful, and we fixed the rest of expansions to have uniform probabilities. These two grammar parameters encode participants' preferences for *orientation invariance* and *configuration invariance*, respectively. We discuss the implications of the fitted values of these parameters in the Results section. The model-fitting procedure closely followed the one implemented by Piantadosi et al. (2016), in which we performed stochastic search for the best fitting parameters via MCMC. The prior over the parameters used beta distributions with uninformative, uniform priors for $\vec{\theta}$, α and β.

Alternative models

We compare the Bayesian program induction model with two versions of an exemplar model known as the Generalized Context Model (GCM) (Nosofsky, 1986). In a GCM, the probability of extending a category label l_y to a new stimulus y is based on its similarity to the training examples X:

$$P(l_y = 1|X) \propto \frac{1}{k}\sum_i^k \exp(-w \cdot d(y, x_i))$$

where d is a distance function and w a scaling parameter. We evaluated two variants of the GCM with different distance measures.

Pixel-GCM. We used a deep convolutional neural net (CNN) to extract features of our visual stimuli from raw pixel data. A pre-trained 50-layer ResNet (He et al., 2015) was used to encode all images into vectorized representations. Cosine distance between two feature vectors was calculated as a measure of their similarity.

String-GCM. We also used a weighted Levenshtein distance to measure the distance between the string representations of two alien figures. For every image, its string format is a concatenation of 3 substrings that separately encode shape primitive types, primitive configurations and orientation. For example, an alien figure consisted of two primitives p_1 and p_2 connected according to their 1^{st} allowable configuration and rotated to 180° can represented in the string format as "$(p_1 p_2) + 1 + 180$". We fit a weight parameter for each type of substring and the overall distance is a weighted average of the distances between each pair of corresponding substrings.

Results

The scatter plots in Fig. 5 summarize the correlations between human responses and model predictions for every trial type and model. Fig. 4 shows examples of model predictions alongside human data. Overall, we found that the Bayesian model provides an excellent account of human behavior with an average correlation of $r = 0.955$ across all trial types studied in both experiments. We observe that the Bayesian model consistently assigns high probabilities to the test item that human participants found most likely, and produced graded predictions that tracked people's willingness to extend the

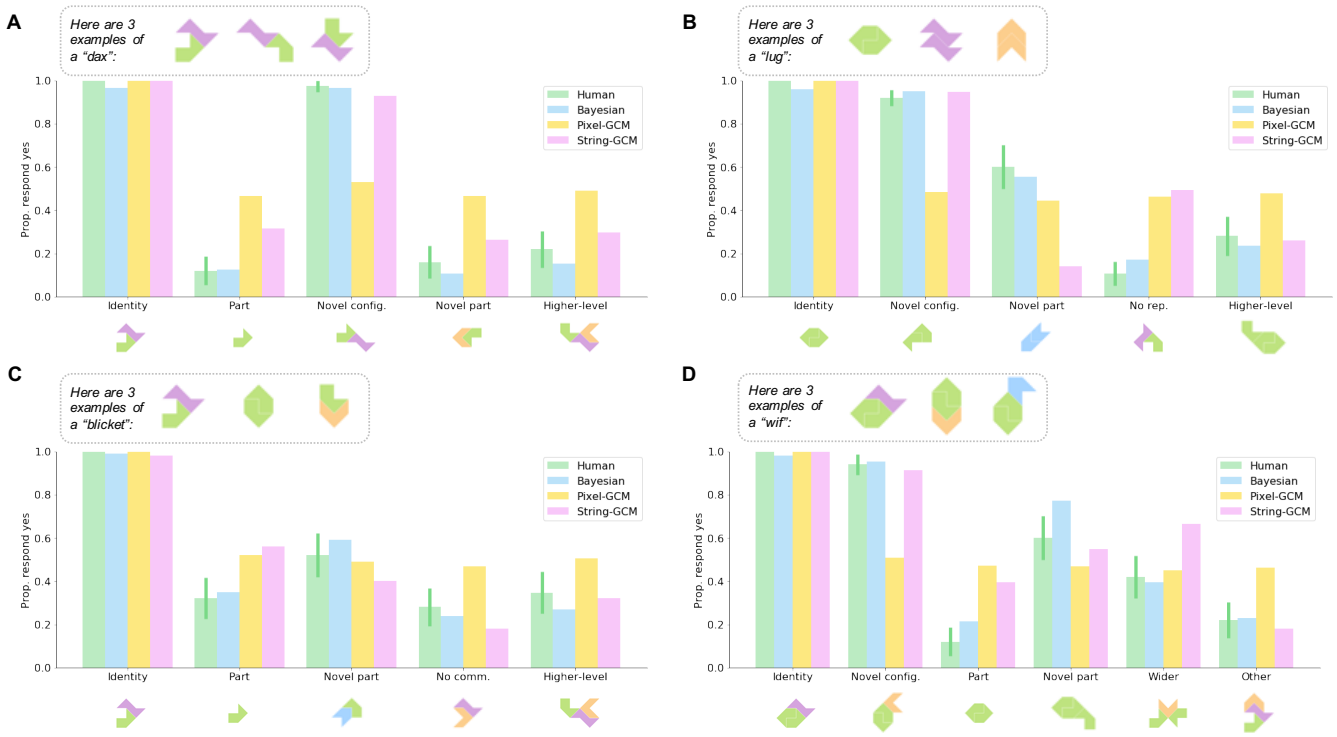

Figure 4: Model predictions on four of the trial types used in Experiment 2. The set of training examples is shown on the top of each panel; examples of test items were shown at the bottom. *Identity* test items are identical to one of the examples; *Part* test items are parts that appeared in one of examples; *Novel configurations* items were new configurations of parts in examples; *Novel part* items were conceptually consistent with examples but contained unseen parts; *Higher-level* items were configurations with one of examples as subpart; *No repetition*, *No common* and *Other* items in B,C and D were conceptually inconsistent with examples; *Wider* items in D are samples from a wider concept for which the set of possible extensions is a superset of the concept of interest.

Table 1: Fitted parameters values.

Type	Probability
Orientation invariance	0.999
Configuration invariance	0.725
α (1-lapse rate)	0.839
β (base rate for responding 'yes')	0.714

concept. Importantly, the Bayesian model makes robust out-of-sample predictions. Using maximum-a-posteriori (MAP) parameters fit based on Experiment 1, the model was able to make good predictions for the more complex concepts in Experiment 2 ($r = 0.952$ for Experiment 1, $r = 0.948$ for 3-exemplar condition of Experiment 2 and $r = 0.947$ for 6-exemplar condition of Experiment 2). All Bayesian model predictions regarding Experiment 2 trials reported were generated in this manner.

On the other hand, the two GCM variants fit the human data less closely, with average correlations of $r = 0.604$ for the pixel-GCM and $r = 0.874$ for the string-GCM. In the case of the pixel-GCM, the model responds strongly to the identity match, but unlike people, it does not clearly distinguish between the other types of generalization (Fig. 4). The pre-trained CNN seemingly fails to perceive the stimuli in terms of their underlying parts and relations, at least without further fine-tuning. The string-GCM is a reasonably good account of the trial types with example figures sharing common parts, with no additional configuration constraints (e.g Fig. 4A). This is unsurprising since the string format precisely encodes which shape primitives are present in each alien figure. The string-GCM struggles with more abstract rules that extends to unseen primitives (e.g. Fig. 4B) or contain configurations of primitives not previously observed (e.g. Fig. 4C). It also has a hard time grasping partial configuration constraints in a concept (e.g. Fig. 4D).

The MAP values of fitted free parameters are reported in Table 1. Values of the two grammar parameters reveal two inductive biases people brought to bear when performing this visual concept learning task. The first parameter is the probability that a given concept contains images of the same configuration at different orientations. This probability is found to be very high, suggesting that our participants had a strong preference for orientation invariance when judging unnamed alien figures. People may have been influenced by their experience with named objects in the real world, which are usually orientation invariant. People were also biased towards concepts that do not require fixed configurations of parts. This is exemplified by their willingness to generalize to novel configurations, even when all examples shared the same configu-

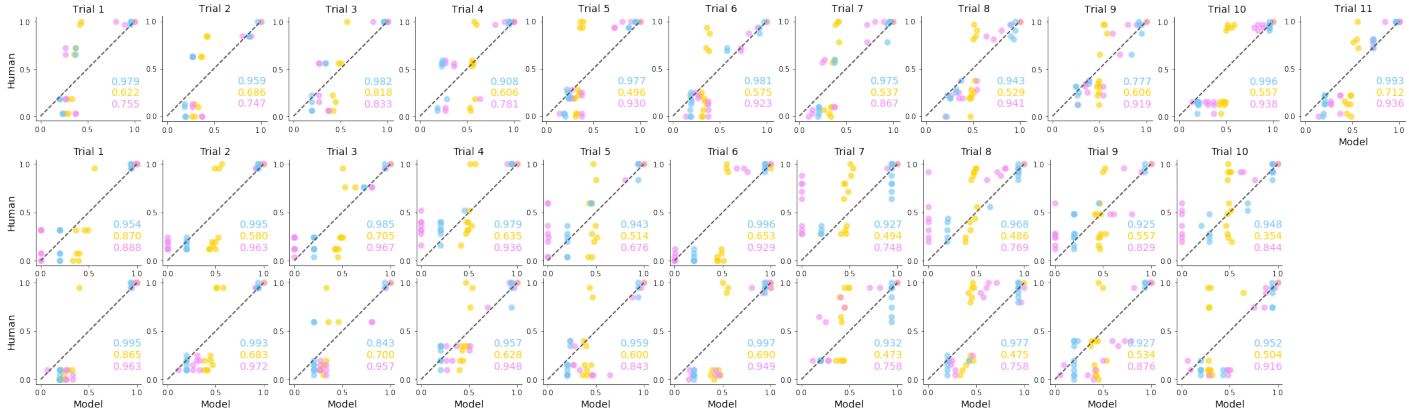

Figure 5: Comparison between human responses and model predictions for each trial type of Experiment 1 (top row), 3-exemplar condition of Experiment 2 (middle row), and 6-exemplar condition of Experiment 2 (bottom row). Each dot in a scatter plot indicates the probability of responding 'Yes' for each test item. The color of dots corresponds to the model type: blue for the Bayesian model, yellow for pixel-GCM and purple for string-GCM. Human-model correlations are also shown for each trial.

ration.

Discussion

We carried out an investigation of human few-shot visual concept learning, with an emphasis on concepts that compose primitives together in different ways. We studied "alien figures" that are richly structured, defined in terms of visual shapes connected in different systems of relations and at various levels of abstraction. Extending previous work on few-shot learning, we provided new empirical results on a set of concepts that better reflect the variety of ways parts combine in real wold visual concepts.

Our Bayesian program induction model provided predictions that closely matched human generalization patterns. Although the model is formulated exclusively to describe the class of alien figures, the model is flexible enough to be further extended by incorporating more or different primitives, or by adding grammatical rules to represent other types of visual concepts. Alternatively, we can formulate a set of different grammars and perform model comparison to distinguish between different language of thought theories within our existing framework.

Importantly, our paradigm is readily applicable to other learning approaches such as neural network (NN) models. The probabilistic grammar used in our studies can be used to sample many more concepts, as is needed for training NN models capable of few-shot learning through meta-learning (Vinyals et al., 2016). By training NN models on this distribution of concepts, we can examine their ability to make compositional generalizations. We can also further refine NN models by fine-tuning them on a subset of the behavioral data, with the aim of better capturing more complex types of inductive bias. Direct comparisons between human and model behavior may further inform how to build machines with more compositional forms of learning (Lake et al., 2019), and help identify potential ingredients that can endow NN models with more human-like capabilities.

In addition, various architectures and algorithms have been developed for problems such as Raven's Progressive Matrices (Zhang et al., 2019) and Bongard problems (Nie et al., 2020). In these datasets, simple visual forms are used to compose problems that test for compositional and relational reasoning abilities. Our task is related in some ways, but with a greater focus on understanding a variety of different types of composition. It's not obvious that models developed for these other domains will generalize to our tasks, but it's an important path for future work to consider.

Using our framework, we also plan to compare humans and computational models on generative tasks. Our Bayesian program induction model can generate new examples; however, generative tasks can provide a particularly direct window into human inductive biases, and it's likely that some modification will be needed to bring the prior closer to human expectations. We hope that generative tasks, building on the findings presented here, will further inform efforts to develop models of flexible, human-like compositional learning.

Acknowledgements

This work was supported by NSF Award 1922658 NRT-HDR: FUTURE Foundations, Translation, and Responsibility for Data Science. We thank Wai Keen Vong for helpful discussions of this manuscript.

References

Biederman, I. (1987). Recognition-by-components: A theory of human image understanding. *Psychological Review*, *94*(2), 115–147.

Chater, N., & Vitányi, P. (2003). Simplicity: a unifying principle in cognitive science? *Trends in Cognitive Sciences*, *7*(1), 19–22.

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.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 770–778). Las Vegas, NV, USA: IEEE.

Lake, B. M., Linzen, T., & Baroni, M. (2019). Human few-shot learning of compositional instructions. In *Proceedings of the 41st annual conference of the cognitive science society.*

Lake, B. M., & Piantadosi, S. T. (2020). People Infer Recursive Visual Concepts from Just a Few Examples. *Computational Brain & Behavior*, *3*(1), 54–65.

Lake, B. M., Salakhutdinov, R., & Tenenbaum, J. B. (2015). Human-level concept learning through probabilistic program induction. *Science*, *350*(6266), 1332–1338.

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.

Nie, W., Yu, Z., Mao, L., Patel, A. B., Zhu, Y., & Anandkumar, A. (2020). Bongard-logo: A new benchmark for human-level concept learning and reasoning. In *Advances in neural information processing systems (neurips).*

Nosofsky, R. M. (1986). Attention, similarity, and the identification–categorization relationship. *Journal of Experimental Psychology: General*, *115*(1), 39–57.

Orbán, G., Fiser, J., Aslin, R. N., & Lengyel, M. (2008). Bayesian learning of visual chunks by human observers. *Proceedings of the National Academy of Sciences*, *105*(7), 2745–2750. doi: 10.1073/pnas.0708424105

Overlan, M. C., Jacobs, R. A., & Piantadosi, S. T. (2017). Learning abstract visual concepts via probabilistic program induction in a Language of Thought. *Cognition*, *168*, 320–334.

Piantadosi, S. T. (2011). *Learning and the language of thought*. Thesis, Massachusetts Institute of Technology.

Piantadosi, S. T. (2014). *LOTlib: Learning and Inference in the Language of Thought.* available from https://github.com/piantado/LOTlib.

Piantadosi, S. T., & Jacobs, R. A. (2016). Four Problems Solved by the Probabilistic Language of Thought. *Current Directions in Psychological Science*, *25*(1), 54–59.

Piantadosi, S. T., Tenenbaum, J. B., & Goodman, N. D. (2016). The logical primitives of thought: Empirical foundations for compositional cognitive models. *Psychological Review*, *123*(4), 392–424.

Smith, L. B., Jones, S. S., Landau, B., Gershkoff-Stowe, L., & Samuelson, L. (2002). Object name Learning Provides On-the-Job Training for Attention. *Psychological Science*, *13*(1), 13–19.

Stuhlmuller, A., Tenenbaum, J. B., & Goodman, N. D. (2010). Learning Structured Generative Concepts. In *Proceedings of the thirty-second annual conference of the cognitive science society.*

Tenenbaum, J. B. (1999). *A Bayesian framework for concept learning*. Thesis, Massachusetts Institute of Technology.

Vinyals, O., Blundell, C., Lillicrap, T., kavukcuoglu, k., & Wierstra, D. (2016). Matching networks for one shot learning. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, & R. Garnett (Eds.), *Advances in neural information processing systems* (Vol. 29). Curran Associates, Inc.

Xu, F., & Tenenbaum, J. B. (2007). Word learning as Bayesian inference. *Psychological Review*, *114*(2), 245–272.

Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhutdinov, R., ... Bengio, Y. (2015). Show, attend and tell: Neural image caption generation with visual attention. *arXiv preprint arXiv:1502.03044*.

Zhang, C., Gao, F., Jia, B., Zhu, Y., & Zhu, S.-C. (2019). RAVEN: A Dataset for Relational and Analogical Visual REasoNing. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 5312–5322). Long Beach, CA, USA: IEEE.