Inferring Structured Visual Concepts from Minimal Data

Peng Qian (pqian@mit.edu) Luke Hewitt (lbh@mit.edu)
Joshua B. Tenenbaum (jbt@mit.edu) Roger Levy (rplevy@mit.edu)
Department of Brain and Cognitive Sciences
Massachusetts Institute of Technology
43 Vassar Street, Cambridge, MA 02139 USA

Abstract

Humans can learn and reason about abstract concepts quickly, flexibly, and often from very little data. Here, we study how people learn novel concepts within a binary grid domain, and find that even this minimal task nonetheless necessitates the inference of highly structured parts as well as their compositional relationships. Furthermore, by changing the presentation condition of the learning examples, we reveal different approaches involved in learning such visual concepts: given the same images, human generalizations differ between rapid and static presentation conditions. We investigate this difference by developing several computational models that vary in their use of structured primitives and composition. We find that learning in the rapid presentation condition is best described as inference in simple models, while learning in the static presentation condition is best described as inference in a more structured space of graphics programs.

Keywords: Bayesian inference; concept learning; few-shot learning; program induction

Introduction

Human concept learning can involve remarkably fast and flexible abstraction. When we see a bridge or appreciate a sculpture, we not only perceive a set of objects, but also the underlying parts and their relationships. With such intuitive understanding of how the parts make the whole structure, human can productively compose learned primitives, generalize to new kinds of objects, and imagine new scenes.

We wish to study the compositional structure that underlies the richness of human visual concept learning by comparing computationally explicit models with human behavior. Prior work in cognitive psychology has built compositional models to describe human visual concept learning, typically by presupposing relevant, symbolically represented parts as inputs to the model, rather than operating directly on images (Shepard, Hovland, & Jenkins, 1961; Rehder & Hoffman, 2005; Goodman, Tenenbaum, Feldman, & Griffiths, 2008). These models are limited to a small stimulus space generated from the conjunction of the few predetermined features. In contrast, machine vision models successfully perform classification from arbitrary natural images (Krizhevsky, Sutskever, & Hinton, 2012), but recent work has found that these models lack the compositional structure necessary to recapitulate human visual concept learning in specific domains (Lake, Salakhutdinov, & Tenenbaum, 2015).

Here, we add to this literature by introducing a new minimal domain to incorporate both of these necessary ingredients: the inference of primitive parts directly from images, and the discovery of compositional structure that relates them. The domain we choose is 5×5 images with binary pixels. Despite the simplicity of this setup, Figure 1 shows that the visual concepts implied by these images can be complex and compositionally structured. In comparison to existing datasets that also occupy this space, our dataset focuses on occlusion and spatial juxtaposition that makes the basic parts particularly ambiguous, as well as concepts that lack prototypical images.

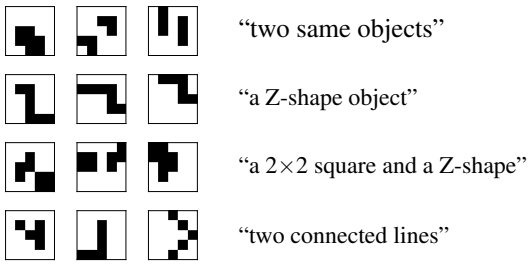
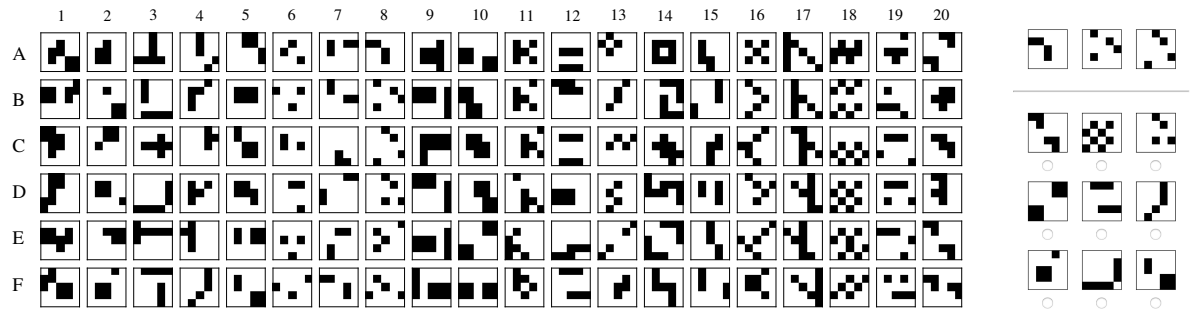


Figure 1: Abstract visual concepts represented by sets of images on a 5×5 binary grid.

Based on these images we develop a few-shot learning task to be presented under either static or rapid viewing conditions. Participants are asked to perform a 9-way classification, for which we compare several computational models that vary in their degree of compositionality and type of structured primitives present. We include a hierarchical Bayesian program learning model, and several additional Bayesian models with alternative primitives. We evaluate these models by quantitatively comparing how the model predictions match human judgments in few-shot generalization. Across the several Bayesian models tested, we find that the ability to jointly infer parts and compose them is critical to explain human generalizations in even this minimal domain, so long as participants are given sufficient time to view the stimulus. However, for rapid viewing conditions, participants' judgements are better explained as inference in a much simpler model with less rich compositional structure.



(a) Set of 20 visual concepts (b) Behavioral task

Figure 2: Visual stimuli and task paradigm used in the behavioral experiments. 60 images (rows A-C) are used for learning concepts and 60 (rows D-F) for testing generalization.

Learning grid concepts

We manually design 20 sets of binary images in the 5×5 grid, covering object occlusion, repetitive structures, and other interesting visual patterns. Each column in Figure 2a is a set of images representing a certain concept. For each concept, three examples (A-C) are designated for learning and a further three (D-F) for testing generalization behaviors. We have 60 different test trials in total.

We use a classification task to compare how humans and models generalize from three examples. The basic task is to learn the underlying concept from the 3 provided examples, and then to select from 9 novel query images the one that most likely displays the same concept. To create each trial, we sample one query image from the same visual concept as the three observed examples, and 8 from distinct other concepts which are drawn uniformly at random (See Fig 2b, 2a col. 8).

To collect human judgements, 216 participants were recruited via Amazon Mechanical Turk to participate in a few-shot classification task, each completing 20 trials: Participants were instructed to observe interesting objects on the visual scenes in the grid world. Subjects were presented with three example images, and then asked to choose one of the new query images that most likely displays the same concept, as is illustrated in Figure 2b.

Each participant was assigned either to the 'rapid' or 'static' viewing condition. In the 'static' condition, subjects could see all three of the example images simultaneously, for as long as required to make a judgement. However, in the 'rapid' condition, subjects instead watched only a video containing the stimuli in quick succession, with an interval between stimulus onsets of 72ms. At the end of the video, a 5×5 grey noise patch was displayed for backward masking.

Bayesian models

Concept learning, from the computational perspective, is fundamentally linked to the generalization problem $P(e'|e_1, e_2, \ldots, e_k)$. Consider a set of k observed examples $e_1, e_2, \ldots, e_k$, and a new observation e'. A concept c naturally plays a role when we factorize the conditional probability $P(e'|e_1, e_2, \ldots, e_k)$ as $\sum_{c \in \mathcal{C}} P(e'|c)P(c|e_1, \ldots, e_k)$.

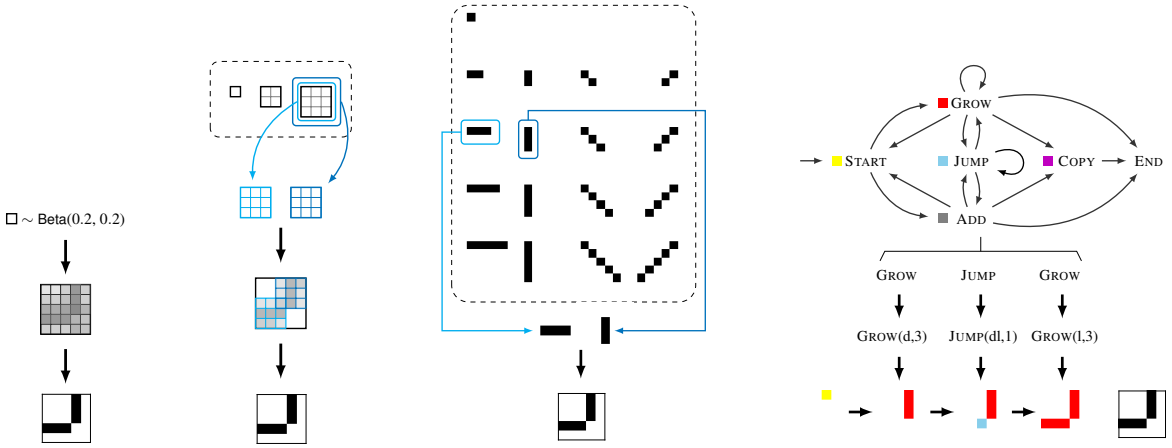
In the Bayesian framework of concept learning, we have the following according to Bayes rule and assuming conditional independence of observations given the concept c:

$$P(c|e_1, \ldots, e_k) = \frac{P(e_1, \ldots, e_k, c)}{\sum_{c \in \mathcal{C}} P(e_1, \ldots, e_k, c)} \propto \prod_{i=1}^{k} P(e_i|c)P(c) \tag{1}$$

The key component is about the structure of $P(e_1, \ldots, e_k, c)$, or more specifically $P(e|c)$ and $P(c)$. Here we construct four different models $P(e_1, \ldots, e_k, c)$ with various assumptions, levels of abstraction, and types of structured representation.

Independent Pixel Model This model assumes that the latent concept $c \in \mathcal{C}$ is a 25-element list of Bernoulli distribution parameters $[p_1, p_2, \ldots, p_{25}]$, each of which corresponds to one of the pixels in the grid and is sampled independently from a prior distribution $\text{Beta}(0.2, 0.2)$. An image instance e is generated from the concept c by sampling the binary state of each pixel in the grid according to its Bernoulli distribution parameter, as is shown in Figure 3a. This model lacks compositionality and complex structured representation, as the primitive available is just a single independent pixel.

Patch Model This model assumes that the latent concept $c \in \mathcal{C}$ is a list of patches drawn from a patch inventory of three different sizes (1×1, 2×2, 3×3), as is shown in Figure 3b. Specifically, c consists of the total number of patches as well as the size of each patches. To generate an image e from the concept c, the model first randomly localizes each patch on the grid and independently samples the Bernoulli distribution parameter for each pixel within the patches. Then an image instance is generated by sampling the binary state of each pixel within each localized patch according to the corresponding Bernoulli distribution parameter. The pixels out of the localized patches will always be turned off. This model has limited compositional structure, as it abstracts an image



(a) Pixel (b) Patch (c) Chunk (d) Program

Figure 3: Generative process of a concept and a image for different models.

as composition of several patches. However, the model lacks explicit structure within a patch, as the Bernoulli distribution parameters of the pixels within a certain patch are not shared at the concept level across multiple generated images.

Chunk Model This model assumes that the latent concept $c \in \mathcal{C}$ is a list of n chunks that are uniformly drawn from an inventory of line primitives (e.g. lines of various sizes and directions), as is shown in Figure 3c. An image e is generated by randomly placing on the grid the list of chunks from the concept c. The locations of the chunks are sampled during the image generation process and not shared at the concept level. While the built-in inventory of basic chunk primitives supports explicit structured representation, the model, however, lacks the mechanism to compose chunks into complex objects during the generative process.

Full Program Model Drawing inspiration from Lake et al. (2015), we design a model in which images are generated from a sequence of consecutive drawing actions. In the full program model, each concept $c \in \mathcal{C}$ takes the form of a probabilistic action program $\{a, \theta\}$, where a refers to the action type and θ refers to the parameters of each action. Once a concept c is generated, a binary image e is sampled from the concept by executing each action in the program step by step. Figure 3d illustrates how an example image is generated from the concept 'GROW(d,3) $\rightarrow$ JUMP(dl,1) $\rightarrow$ GROW(l,3)'.

To generate a concept, namely an action program in this model, the length of the program n is first sampled from an exponential distribution over all the possible program lengths ranging from 1 to 5 ($P(n) \propto \lambda^n$, where $\lambda = 0.9$), with preference to short programs. After that, a sequence of n actions, a, is sampled step by step from the plausible action primitives to construct the template of the program, under the constraints of the action transition grammar specified in Figure3d. For each action, the plausible transitions to other actions are uniformly distributed. The action primitives include GROW (adding pixels in a certain direction), JUMP (skipping over pixels in a certain direction), COPY (making copies of the current drawing trace and placing them randomly on the grid), ADD (generating a square patch of certain size and placing it randomly on the grid.), and START (placing currently generated trace on the grid and initializing a new trace).

After sampling the program template, the parameters θ_i of each action a_i in the program a (e.g. the direction and size of GROW) is uniformly sampled from the plausible values that a certain parameter type can take. There are eight basic values for the direction parameter, u (up), d (down), l (left), r (right), ul (upleft), ur (upright), dl (downleft), and dr (downright). The size parameter can take a number that is smaller or equal to the grid width size for GROW and JUMP actions, and a number less than 3 for COPY action. Both the direction and size parameter can also take a special parameter value '*any*', which refers to randomly sampling one of the basic directions or plausible size values during the image generation process.

Regarding the execution of an action program, the initial empty trace starts at the reference point $(0,0)$ on the temporary canvas. Following the action instructions, we draw pixels or move to other location on the canvas consecutively. The trace generated on the temporary canvas will be placed at a random place on the 5×5 grid once we encounter the end of the program or a START action. It is worth noticing here that the mechanism of composing action traces and starting new traces gives rise to the model's ability of utilizing more relational and object-like compositional structure. Therefore, Bayesian program learning model has more expressive compositionality and explicitly structured representation.

Few-Shot Classification and Generation

In order to evaluate each model against our collected human data, we must perform inference. However, this is computationally challenging to do exactly, and so we perform approx-

	H $_{\text{static}}$	H $_{\text{rapid}}$	M_1	M_2	M_3	M_4
Human $_{\text{static}}$	-	36	51	35	15	10
Human $_{\text{rapid}}$	36	-	36	25	11	10
Program [M_1]	51	36	-	35	15	9
Chunk [M_2]	35	25	35	-	14	6
Patch [M_3]	15	11	15	14	-	8
Pixel [M_4]	10	10	9	6	8	-

Table 1: Proportion of the same choices between model predictions and human judgements for 60 trials.

Model	Static presentation	Rapid presentation
Program	**0.39**	0.49
Chunk	0.43	0.47
Patch	0.52	**0.44**
Pixel	0.78	0.78
Uniform	0.56	0.46

Table 2: Hellinger distance averaged across 60 trials for each model compared to human data under each presentation condition (lower is better)

imate inference using a neural network trained for amortized few-shot classification in each model.

For each model, we train a separate network with a shared architecture, comprising a single convolutional layer and two fully connected layers with 200 hidden units. Each network was trained on model-generated data to produce a distribution of responses for 9-way classification of novel images. Specifically, we generate synthetic training data by sampling 9 concepts from each model's prior, drawing one image from each concept as the query examples, and a further 3 images from one concept as the observed examples. We optimise the network to classify the correct query example given the observed examples.

We then evaluate each of these trained networks on the same stimuli as presented to human subjects. Thus, regarding the behavioral task, each model's inference network is used to select the most likely query image from the 9 options.

For few-shot generation, we approximate the posterior $P(c|e_1,\ldots,e_k)$ using Markov Chain Monte Carlo (MCMC) implemented in WebPPL (Goodman & Stuhlmüller, 2014). Then we are able to produce novel instances from the inferred concepts.

Results

We are particularly interested in how human and the models proposed in this work make generalizations from few examples. We evaluate model predictions with respect to human judgments on 60 trials in the behavioral task, in each presentation condition.

Evaluation results of the models are listed in Table 1. We compute the proportion of choosing the same test images as the top choice for each pair between different models and human judgments. It is shown that the predictions of Bayesian program learning model largely matches the most popular (top 1) choice of human judgments in the static condition.

We compare the probability distribution of model's prediction to the distribution of human judgments over 9 test images for each trial in the experiment. We normalize human judgments to get a distribution of choice over the 9 test items, and similarly calculate $P(e'_i|e_1,e_2,e_3)$ over the 9 test items for each model. For each of 60 trials, we compute the Hellinger distance (Hellinger, 1909) between the distribution of model prediction and human judgement to quantify the distance between human and model responses. The average Hellinger distances are shown in Table 2, highlighting a difference between the two presentation conditions. For static presentations of the stimulus, the highly structured Bayesian program model is by far closest to human judgments in terms of the distribution of the choice in each trials. However, for rapid presentations of the stimulus, the program model suffers from overconfidence while the less structured 'chunk' model provides the best prediction of human judgements. Figure 4 visualizes the distribution of human judgments and models' predictions for several trials.

Regarding the question of what type of compositional structure supports human concept learning, the differences among the proportions of matched choices between human and four Bayesian models of different level of abstraction provide some interesting insights. As is discussed before, these Bayesian models can be summarized briefly with how much abstraction and what level of abstraction is built into the architecture: The pixel model does not have any compositional structures, while the patch model composes a scene by combining several patches. However, neither of these match human judgements well: the compositional ability of the patch model is largely limited due to the lack of explicitly structured primitives in its representation, as the patch only vaguely specify a pattern instead of clearly defining the structure of the pattern. With more structured primitives, the chunk model achieves significantly stronger results.

While the lack of structured representation makes it hard for the patch model to take the advantage of compositional structure in learning concepts, comparison between chunk and program models further suggest that hierarchical compositional structures are important in capturing human few-shot learning of simple visual concepts.

One final advantage of a Bayesian generative model is its generative process. Table 3 lists three of the inferred concepts by Bayesian program learning model, the approximate log posterior probabilities of these concepts, and the posterior samples for several sets of binary images used in the classification experiment. We can see that Bayesian program learning model successfully inferred the program and generated reasonable novel images of the same concept.

Discussion

Our work is an advanced investigation of similarity and generalization, along the line of research of classic Bayesian

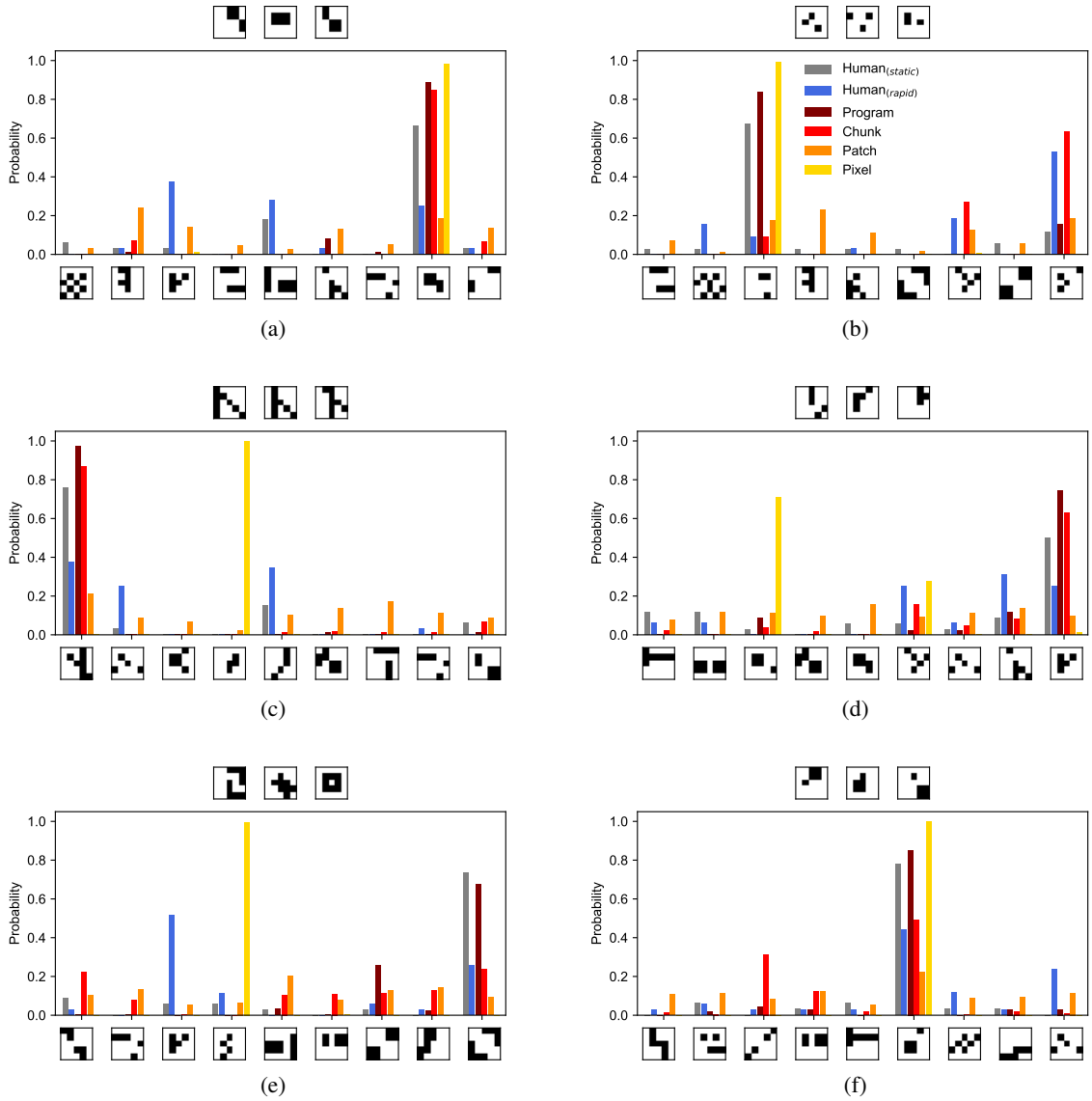


Figure 4: Fine-grained comparison of model responses to human responses.

concept learning (Tenenbaum, 2000; Tenenbaum & Griffiths, 2001; Kemp, Bernstein, & Tenenbaum, 2005; Goodman et al., 2008; Stuhlmuller, Tenenbaum, & Goodman, 2010) in computational cognitive science. We investigated visual concepts with more abstract, relational, compositional, hierarchical and object-like structure.

Compared to previous work (Orbán, Fiser, Aslin, & Lengyel, 2008) that studied learning visual scenes in a grid world composed of simple chunks (i.e. the statistical dependencies are simple associations between adjacent objects), this work explores more complex scenes that allow for more abstract (non-statistical) relations between objects in a scene. Further, objects in the visual scenes might occlude each other, which propose yet another challenge for learners, both model and humans, in identifying the latent structure.

Other important related works are Bayesian program learning of hand-written characters (Lake et al., 2015) and abstract visual concepts (Overlan, Jacobs, & Piantadosi, 2017). Our study introduces a richer grid concept domain, and develops computational account of different levels of abstraction. Although Lake et al. (2015) presents a Bayesian program learning model for few-shot learning of hand-written characters, which are images on a larger grid than what we use here, some interesting differences are worth mentioning here. Human might have a lot of practical experience with hand-written characters in daily life. There could be reasonably good prototype for hand-written characters as they are often standardized for communication purpose. People might rely on inferring a single visual prototype and generalize through similarity matching to the prototypical image. In our case, in contrast, it is hard to infer a single visual prototype for many of our concepts, even though there are only a small number

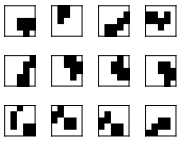
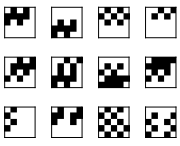
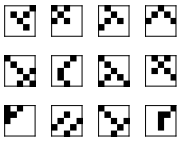
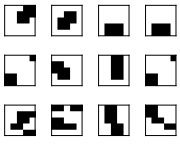
Examples	$\log(P)$	Concepts	Posterior samples
	-1.41	GROW(dl,2) → GROW(u,2) → GROW(ur,2) → START → ADD(2×2)	
	-1.41	GROW(dl,2) → JUMP(d,1) → GROW(ur, 2) → START → ADD(2×2)	
	-7.65	GROW(any,2) → JUMP(u,1) → GROW(ur,2)→ START → ADD(2×2)	
	-1.47	GROW(ur,2) → JUMP(l,2) → GROW(dl,2) → COPY(1)	
	-6.43	GROW(dl,2) → JUMP(r,3) → JUMP(u,1) → GROW(dl,2) → COPY(2)	
	-12.66	GROW(dl,2) → JUMP(r,any) → JUMP(u,1) → GROW(dl,2) → COPY(any)	
	-2.12	GROW(dr,3) → START → GROW(dl,3)	
	-6.28	GROW(dr,any) → START → GROW(ur,3)	
	-8.35	GROW(dl,3) → START → GROW(any,3)	
	-0.42	ADD(2×2) → START → ADD(2×2)	
	-3.07	ADD(2×2) → JUMP(any,any) → ADD(2×2)	
	-23.04	ADD(2×2) → START → GROW(r,2) → COPY(2)	

Table 3: Programs found by MCMC for several test concepts, with corresponding posterior-predictive samples of new images.

of observations to choose and generalize from.

This work also shows that compositionality is not the only important aspect behind human few-shot learning, in line with previous work (Schulz, Tenenbaum, Duvenaud, Speekenbrink, & Gershman, 2016) that demonstrates human's preferences of compositional pattern in function learning domain. The level of abstraction in the representation also plays an important role in building models that can better match the generalization behaviors observed in human concept learning.

We believe that our visual concept learning task contributes to an understanding of how humans learn and reason about novel visual concepts, addressing two questions: (1) what kinds of representation and architecture support flexible inference of underlying abstract structure, and the impressive generalizations that humans achieve from often minimal data? (2) Is this same architecture necessary, and is it sufficient, to explain the kind of rapid inferences humans are able to make given only a short glimpse of a concept? Comparisons among several Bayesian models with different degrees of abstraction demonstrate that, even in this minimal domain, humans can infer concepts with a rich compositional structure, but that the extent of this structure is dependent on the condition of presentation.

Acknowledgments

We would like to thank Maddie Cusimano and members of the MIT Computational Psycholinguistics Lab for their helpful comments on this project.

References

Goodman, N. D., & Stuhlmüller, A. (2014). *The Design and Implementation of Probabilistic Programming Languages.* `http://dippl.org`. (Accessed: 2017-12-17)

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.

Hellinger, E. (1909). Neue begründung der theorie quadratischer formen von unendlichvielen veränderlichen. *Journal für die reine und angewandte Mathematik*, *136*, 210–271.

Kemp, C., Bernstein, A., & Tenenbaum, J. B. (2005). A generative theory of similarity. In *Proceedings of the 27th annual conference of the cognitive science society* (pp. 1132–1137).

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems* (pp. 1097–1105).

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

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.

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.

Rehder, B., & Hoffman, A. B. (2005). Thirty-something categorization results explained: selective attention, eyetracking, and models of category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*(5), 811.

Schulz, E., Tenenbaum, J., Duvenaud, D. K., Speekenbrink, M., & Gershman, S. J. (2016). Probing the compositionality of intuitive functions. In *Advances in neural information processing systems* (pp. 3729–3737).

Shepard, R. N., Hovland, C. I., & Jenkins, H. M. (1961). Learning and memorization of classifications. *Psychological monographs: General and applied*, *75*(13), 1.

Stuhlmuller, A., Tenenbaum, J. B., & Goodman, N. D. (2010). Learning structured generative concepts. In *Proceedings of the cognitive science society* (Vol. 32).

Tenenbaum, J. B. (2000). Rules and similarity in concept learning. In *Advances in neural information processing systems* (pp. 59–65).

Tenenbaum, J. B., & Griffiths, T. L. (2001). Generalization, similarity, and bayesian inference. *Behavioral and brain sciences*, *24*(4), 629–640.