Using Compositionality to Learn Many Categories from Few Examples

Ilia Sucholutsky (is2961@princeton.edu)
Department of Computer Science
Princeton University

Bonan Zhao (bnz@princeton.edu)
Department of Computer Science
Princeton University

Thomas L. Griffiths (tomg@princeton.edu)
Department of Computer Science
Department of Psychology
Princeton University

Abstract

Humans have the remarkable ability to learn new categories from few examples, but how few examples can we actually learn from? Recent studies suggest it may be possible to learn more novel concepts than the number of examples. Previous approaches to such less-than-one-shot (LO-shot) learning used soft labels to provide weighted mappings from each example to multiple categories. Unfortunately, people find soft labels unintuitive and this approach did not provide plausible, cognitively-grounded mechanisms for LO-shot learning at scale. We propose a new paradigm that leverages well-established learning strategies: reducing complex stimuli to primitives, learning by discrimination, and generalizing to novel compositions of features. We show that participants can learn 22 categories from just 4 examples, shedding light on the mechanisms involved in LO-shot learning. Our results provide valuable insights into the human ability to learn many categories from limited examples, and the strategies people employ to achieve this impressive feat.

Keywords: Categorization, few-shot learning; soft labels; compositional generalization; machine-learning

Introduction

It is well-established that humans can learn new concepts from a few examples, but just how few can it be? Previous research in both human and machine learning has treated one-shot learning, where the participant must learn a new concept from a single example, as the limit on sample-efficiency in supervised learning settings (Tiedemann et al., 2022; Fei-Fei et al., 2006). Recent research in machine learning has shown that it is theoretically possible to learn more novel concepts than the number of presented examples, so-called less-than-one-shot (LO-shot) learning (Sucholutsky & Schonlau, 2021a; Sucholutskv et al., 2021), by associating examples with "soft labels" that describe their closeness to each concept as opposed to traditionally-used "hard labels" which associate each example with a single concept. LO-shot learning has recently been replicated in humans in a study that showed that participants presented with two examples of novel stimuli paired with soft labels relating them to three categories could infer the structure of those three categories (Malaviya et al., 2022). However, it remains unclear what the limit of LO-shot learning is in humans and what mechanism enables us to learn in this unexpected way.

The use of soft labels poses a barrier to studying the limits of LO-shot learning in humans, as well as the amount of insight it provides into how humans learn in naturalistic settings. Under this paradigm, participants are presented with data that effectively corresponds to statements of the form "this image is 60% dog, 35% wolf, 5% white fox, and 0% horse" when shown something like a photograph of a husky. While such statements are in theory highly informative (Sucholutsky, Battleday, et al., 2023), they are also highly unintuitive – it is difficult for people to interpret and produce them (Collins, Bhatt, et al., 2023). As a result, previous studies have only been able to show human generalization from a couple of soft-labeled examples to a few categories (Malaviya et al., 2022) and have been unable to pinpoint the mechanism underlying this kind of learning.

We approach this challenge from a different direction. Rather than trying to recreate this phenomenon by replicating machine-learning studies as human studies, we instead examine the extensive literature on human cognition to identify potential mechanisms that could enable LO-shot learning. In particular, research in cognitive science has shown that people are capable of reducing complex stimuli to primitives (Tversky, 1977; Goldstone, 1998; Austerweil & Griffiths, 2013), learning by discrimination Tversky (1977); Kruschke (1992), aligning concepts across multiple domains (Aho et al., 2022), and generalizing to novel compositions of features (Lake et al., 2011; Zhao et al., 2023). We combine these insights into a novel LO-shot learning paradigm (visualized in Figure 1, where participants first learn new visual features by discriminating between a small set of examples labeled with feature weights, and then learn about new categories only from textual descriptions of their compositional structure in terms of the new features.

With this new paradigm, we can elicit much more extreme forms of LO-shot learning in participants, with our results showing that people can learn 22 categories from 4 examples. Furthermore, we can probe the mechanisms that enable people to perform this kind of LO-shot learning. In particular, we show that people perform two different types of LO-shot learning in sequence to achieve this level of generalization. Participants are first able to learn 8 distinct features from the 4 examples, and they are then able to learn 22 categories from descriptions of compositions of those 8 features.

Background

Machine learning researchers have long sought to develop models that could learn new classes from a few examples

In L. K. Samuelson, S. L. Frank, M. Toneva, A. Mackey, & E. Hazeltine (Eds.), *Proceedings of the 46th Annual Conference of the Cognitive Science Society.*

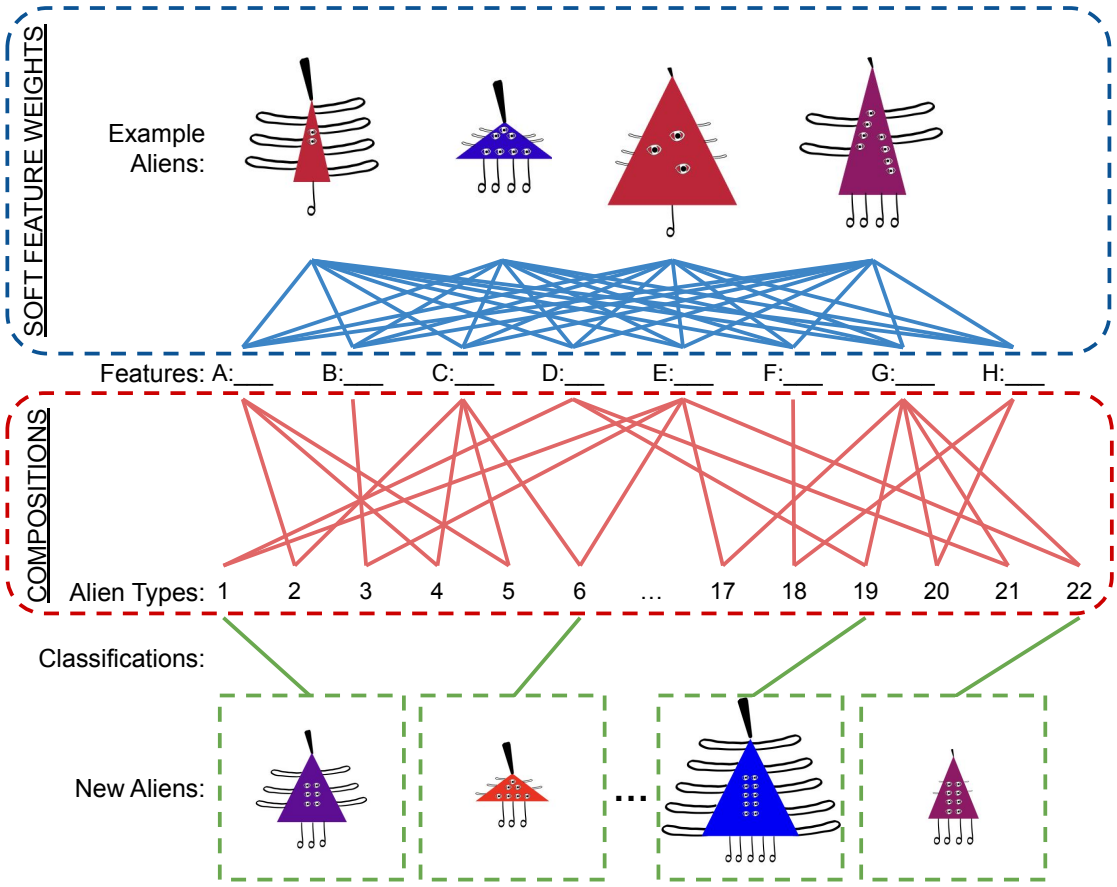

Figure 1: Schematic of our approach. Components within dashed lines are provided to participants. **Top (blue)**: Participants must first use 4 visual examples paired with dense feature weights to learn what the 8 unknown features (A-H) correspond to. **Middle (red)**: Participants are also presented with sparse and dense compositional information (as text) that describes each category of aliens. **Bottom (green)**: Participants are presented with alien images and must use all the previous information (examples, soft weights, feature inferences, and compositional information) to classify them into the correct categories.

per class (Wang et al., 2020) often turning to the rich history of cognitive science research on how humans can learn so efficiently for inspiration in designing AI systems that generalize more effectively. One-shot learning, where models learn new classes from just a single example of each, has long been considered the limit for traditional supervised learning in both machines (Fei-Fei et al., 2006) and humans (Lake et al., 2011). Recent research suggests that it may theoretically be possible to learn more new classes than the number of presented examples, so-called less-than-one-shot (LO-shot) learning (Sucholutsky & Schonlau, 2021a).

There are some limited pieces of empirical evidence that machines (Sucholutsky & Schonlau, 2021b; Sucholutskv et al., 2021) and humans (Malaviya et al., 2022) can perform LO-shot learning by leveraging soft labels that encode the relationship between each example and every known class. While people can somewhat understand and produce soft labels, and can even use them to communicate beliefs to AI systems (Collins, Barker, et al., 2023), they often find these labels unintuitive and difficult to interpret even in simple visual settings (Collins, Bhatt, et al., 2023). As a result, most soft label studies instead rely on populations of participants probability matching to their uncertainty levels when making hard classifications (Peterson et al., 2019).

Theoretical investigations of soft labels have found that they enable LO-shot learning because they contain significantly more representational information per label than hard labels do (Sucholutsky, Battleday, et al., 2023), meaning fewer soft labels are required to align the learner's internal representations of the classes and stimuli with the teacher's representations. This type of *representational alignment* (Sucholutsky, Muttenthaler, et al., 2023) between a teacher and a student is closely linked with few-shot learning performance (Sucholutsky & Griffiths, 2023) and efficient value learning (Wynn et al., 2023). Analogous claims have been made about *concept alignment* (Rane et al., 2024) as a way to leverage prior knowledge from multiple domains to speed up learning (Aho et al., 2022) or for a student to efficiently learn a teacher's values (Rane et al., 2023).

Meanwhile, people can combine primitive components

in principled ways to aid rapid generalization to new situations (Tenenbaum et al., 2011). People tend to identify sub-components from complex structures (Wong et al., 2022), and make use of base concepts to form rich categories (Piantadosi et al., 2016), learn complex new concepts effectively from few examples (Lake et al., 2015), and bootstrap to a diverse collection of novel concepts (Zhao et al., 2023). In particular, many studies have pinpointed compositionality as a key mechanism in human few-shot (Lake & Baroni, 2023; Dehaene et al., 2022) and one-shot (Lake et al., 2011; Zhao et al., 2022) learning.

We propose that efficient human learning relies on all three of these mechanisms: soft labels, representational/concept alignment, and compositional generalization. We show that people can a) decompose a small number of complex stimuli into a greater number of simple features and use associated soft labels (or weights) to align those features with familiar concepts; and b) learn to recognize an even greater number of novel compositions of those features without seeing additional examples.

Testing Compositional LO-shot Learning

We summarize our approach to investigating compositional less-than-one-shot learning in people in Figure 1. Participants are first shown a small number of example aliens with associated soft feature weights without revealing what characteristics those features actually correspond to (we just call them features A-H). We use this to test whether people can decompose the stimuli into simple features and then use the soft weights to align these with familiar concepts (like "number of arms", and "height"). Participants are then shown textual information describing how novel compositions of these features correspond to categories, before being asked to classify new alien images into the categories. We use this to test whether people can learn new categories from compositional information without ever having seen examples from those categories. The remainder of this section provides a more detailed explanation of our experimental setup and lays out the results.

Methods

We tested whether people can make use of soft featural and compositional information of T training examples to identify K novel classes, where $T < K$. Specifically, we challenged participants to learn 22 new classes from just 4 examples.

Participants 105 participants were recruited through Prolific Academic (45 females, $M_{\text{age}} = 40 \pm 10$). We excluded 17 participants from analysis who failed to correctly identify at least 1 out of the 4 learning examples, leaving us 88 participants in total. Participants were paid six dollars for their time. The task took 36 ± 31 minutes. The experiment was performed with IRB approval. All participants gave informed consent before undertaking the experiment. An anonymous pre-registration is available here.

Material We designed 22 alien creatures in total (Table 1). Each alien creature is defined by eight features: antenna length (Feature A, ranging from 0 to 30mm), body width (Feature B, 20–70mm), body height (Feature C, 20–70mm), color (Feature D; i.e., proportion of blue pigment of the skin, RGB(0,0,255) to RGB(255,0,0)), number of eyes (Feature E, 0–10), number of arms (Feature F, 0–10), number of legs (Feature G, 0–5), and arm length (Feature H, 0–30mm). The 4 training examples are shown in the top panel in Figure 1. Aliens 5 to 10 are homogeneous across features—"all features are X"—where aliens 5 and 6 test extrapolation of extreme feature values (lowest/highest possible value of each feature), and aliens 7–10 test interpolation; aliens 11–14 are like training aliens 1–4 but each with 2 features modified to be in the extrapolation region (the modified features are underlined in Table 1); aliens 15–18 are like training aliens 1–4 but each with 2 features modified to be in the interpolation region (the modified features are underlined in Table 1); aliens 19–22 are like 11–18 but instead of having both modified features be extra/interpolation, they have one feature in extrapolation and one in interpolation (the modified features are underlined in Table 1).

To convert proportions to exact feature values, let x, y be the minimal and maximal feature values and p the proportion, the values used in the stimuli are computed as $x + p \times (y - x)$. To translate the color feature to RGB, we kept G to be 0 throughout, set R to $255 - 255 \times p$, and B to $255 \times p$. Note that we only used feature labels A to H in the experiment, and participants had to infer the mappings between these uninformative labels with the visualized features.

There are two ways we communicated soft labels to participants. One is the compact description that compares a new alien to a known one, for example, "Alien 11 is very similar to Alien 1 but with high Feature E and low Feature F". Another is the detailed description that lists a percentage (0 to 100) indicating how strongly that alien exhibits that feature, in tabular form (i.e., a subset of rows from Table 1.

Design and procedure We manipulated whether participants had the chance to learn soft featural and compositional information (yes for the experimental group, and no for the control group), and the order in which participants receive compact or detailed descriptions first in the test phase. These manipulations led to $2 \times 2 = 4$ between-subject conditions. After giving informed consent, participants were told that they were explorers to the wonderful world of Parvelor, and we needed their help to classify the many quirky creatures on the planet. Participants were then shown a field guide with three sections.

In Section 1, participants in the experimental group saw the four example aliens (Figure 1 top) with corresponding soft feature information (Table 1, first four rows) summarized in a table. Participants in the control group saw the same

Table 1: Soft information for the aliens participants categorized in the experiment. Aliens 1-4 are examples. Underlined values are mentioned in the compact descriptions.

Alien	A	B	C	D	E	F	G	H
1	0.8	0	0.5	0.2	0.2	0.7	0.2	0.8
2	0.2	1	1	0.2	0.3	0.6	0.2	0.2
3	0.8	0.66	0	0.8	0.7	0.5	0.8	0.2
4	0.2	0.33	1	0.4	0.8	0.4	0.8	0.8
5	0	0	0	0	0	0	0	0
6	1	1	1	1	1	1	1	1
7	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2
8	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4
9	0.6	0.6	0.6	0.6	0.6	0.6	0.6	0.6
10	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.8
11	0.8	0	0.5	0.2	0.9	0.1	0.2	0.8
12	1	1	0.25	0.2	0.3	0.6	0.2	0.2
13	0.8	0.66	0	0	0.7	1	0.8	0.2
14	0	0.33	1	1	0.8	0.4	0.8	0.8
15	0.8	0.4	0.5	0.2	0.2	0.5	0.2	0.8
16	0.2	1	0.5	0.5	0.3	0.6	0.2	0.2
17	0.8	0.66	0	0.8	0.4	0.5	0.8	0.6
18	0.5	0.33	1	0.4	0.8	0.4	0.4	0.8
19	0.8	0.5	0.5	0.2	0.2	0	0.2	0.8
20	1	1	1	0.2	0	0.6	0.2	0.2
21	0.8	0.66	0	0	0.7	0.5	0.6	0.2
22	0.2	0.33	0.5	0.4	0.8	0.4	0.8	0.1

four example aliens, but without soft feature information. In Section 2, all participants had access to text descriptions of aliens 5–22. Aliens 5–10 were always presented in the compact format. Participants in the compact-first group saw aliens 11–16 in the compact format and aliens 17–22 in the detailed format, and *vice versa* for participants in the detailed-first condition. In Section 3, participants were asked to pick a short description of what they thought each feature corresponded to, and selecting answers from a drop-down menu with the eight features and one distractor (position of the eyes). After that, participants were shown all 22 aliens (4 examples and 18 tests) in random order and categorized them by dragging an alien image to the corresponding label (Figure 2). Participants had access to their field journal (the training aliens, descriptions of the test aliens, and the feature mapping they created themselves) at all times. The experiment ended with a debriefing and feedback form.

Results

We analyze how well people categorized the new aliens as a function of the two factors we manipulated: learning soft labels (*exp* group) or not (*control* group), and seeing test alien descriptions in *compact*-first or *detailed*-first orders. Test accuracy is defined as the number of correct categorizations by a participant divided by the total number of categorizations. We also examine how performance in the feature learning phase predicts test accuracy.

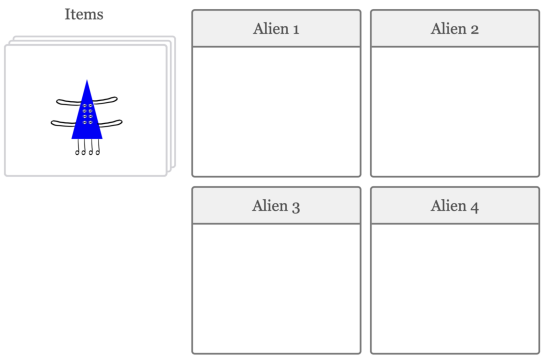

Figure 2: Interface for categorizing new aliens. Participants can drag images in the stack on the left to corresponding labels on the right. There were 22 alien images and 22 labels.

Participants use soft information to identify more novel categories. As illustrated in Figure 3a, participants with soft information achieved overall test accuracy of 0.7 ± 0.3, significantly higher than the average accuracy of 0.4 ± 0.3 in the control group ($t(83.2) = 5.4$, $p < .001$, 90% CI $= [0.2, 0.4]$, Cohen's $d = 1.28$). Breaking down the test aliens by different types, participants in both the experimental and control groups achieved similar accuracy when categorizing the four learning examples during test phase ($M_{\text{exp}} = 0.8 \pm 0.2$, $M_{\text{control}} = 0.8 \pm 0.2$, $t(84.7) = 0.1$, $p = .89$), serving as a sanity check that the control group paid attention to the task and were not making purely random choices (Figure 3b).

Participants with soft feature information identified more aliens correctly for each other type of the classification task (Figures 3c-e): test aliens with homogeneous features ($M_{\text{exp}} = 0.7 \pm 0.4$, $M_{\text{control}} = 0.4 \pm 0.4$, $t(84.9) = 3.9$, $p < .001$, 90% CI $= [0.2, 0.5]$, Cohen's $d = 0.8$), test aliens described in compact formats ($M_{\text{exp}} = 0.6 \pm 0.4$, $M_{\text{control}} = 0.1 \pm 0.3$, $t(78.6) = 6.6$, $p < .001$, 90% CI $= [0.3, 0.6]$, Cohen's $d = 1.7$), and test aliens described using detailed feature tables ($M_{\text{exp}} = 0.6 \pm 0.4$, $M_{\text{control}} = 0.2 \pm 0.3$, $t(82.5) = 5.4$, $p < .001$, 90% CI $= [0.3, 0.5]$, Cohen's $d = 1.3$). Overall, this is strong evidence that with the help of soft information people can identify more novel categories than the number of learning examples when given compositional instructions.

Participants learn equally well with compact and detailed descriptions. We did not find the compact-first or detailed-first manipulation to be a significant predictor of test accuracy ($M_{\text{compact-first}} = 0.5 \pm 0.3$, $M_{\text{detailed-first}} = 0.5 \pm 0.3$, $t(84.2) = 0.1$, $p = .88$). Focusing on the groups of aliens that were described in compact or detailed formats, there are no significant differences between test accuracy

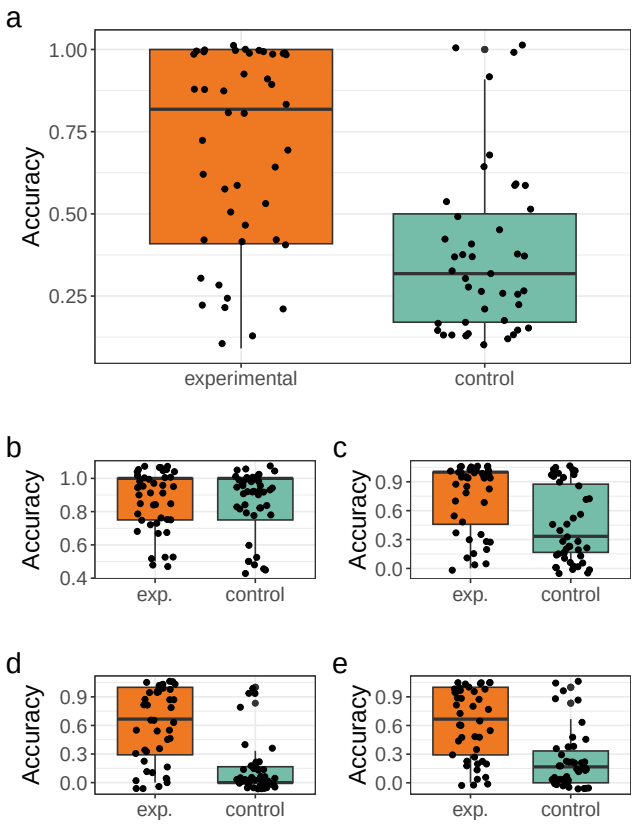

Figure 3: Categorization accuracy. (a) For all 22 alien creatures. (b) For the learned examples. (c) For new aliens with homogeneous features. (d) For new aliens with compact descriptions. (e) For new aliens with detailed descriptions.

($M_{\text{compact}} = 0.4 \pm 0.4$, $M_{\text{detailed}} = 0.4 \pm 0.4$, $t(173.6) = -0.9$, $p = .36$). This holds for both participants in the experimental group ($M_{\text{compact}} = 0.6 \pm 0.4$, $M_{\text{detailed}} = 0.6 \pm 0.4$, $t(41.2) = 1.2$, $p = .2$), and participants in the control ($M_{\text{compact}} = 0.1 \pm 0.3$, $M_{\text{detailed}} = 0.2 \pm 0.3$, $t(85.9) = -0.2$, $p = .8$). These results suggest that how soft information is presented has little impact on how well people use this information in categorizing new observations.

Implicit soft information for the control group Although participants in the control group achieved lower test accuracy for almost all the test aliens, their overall test accuracy is significantly above chance ($\chi^2(17) = 62.6$, $p < .001$). In particular, test accuracy for aliens 1–6 (with homogeneous features) in the control group is significantly higher than for aliens 7–18 ($M_{\text{homog.}} = 0.4 \pm 0.4$, $M_{\text{control}} = 0.2 \pm 0.3$, $t(76.4) = 3.5$, $p < .001$, 90% CI $= [0.1, 0.4]$, Cohen's $d = 0.9$). This implies that people may be sensitive to the feature space and feature strengths, and can make efficient use of such information to navigate through hard categorization tasks.

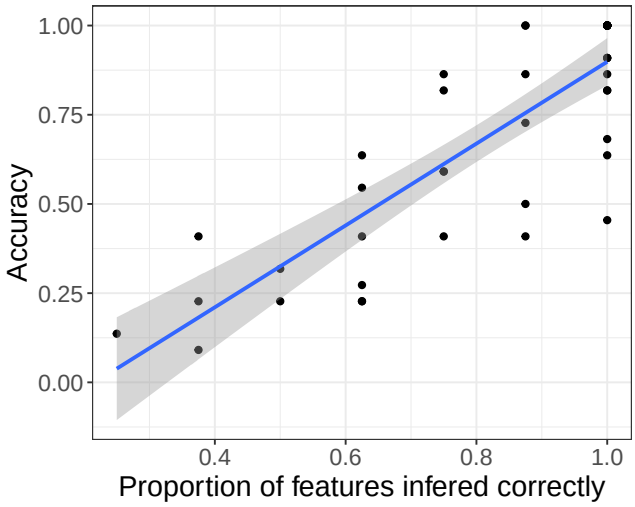

Figure 4: Performance on learning features from soft labels predicts test classification accuracy.

Mastering soft labels improves performance in categorization. Participants in the experimental group took a longer time completing the task ($M_{\text{exp}} = 45 \pm 19$ min, $M_{\text{control}} = 34 \pm 42$ min, $t(43) = -5.3$, $p < .001$, 95% CI $= [-46, -20]$, Cohen's $d > 10$). On average, participants in the experimental group achieved 0.8 ± 0.2 accuracy in identifying the feature mappings, and this performance is a significant predictor of test accuracy ($F(1,42) = 93.2$, $\beta = 1.27$, $p < .001$), in addition to explaining 68.9% of the variance (Figure 4). This further confirms that people are making effective use of soft information in compositional generalizations.

Conclusion

People use features compositionally to represent categories, and this helps us to learn novel categories even without having encountered their category members. Extending previous work in learning with soft labels with compositional generalization, we showed that people can indeed learn 22 new categories from just 4 examples by using soft featural information compositionally. In fact, people are good at using such information with either compact or detailed formats, and exhibit a strong positive correlation between how well they learn the soft featural information and how good they are at identifying new categories compositionally. Taken together, these results provide a novel perspective on how people learn so much despite their limited experience in the world: being able to extract features and combine them compositionally to create new concepts is a rapid way to learn from limited data.

Acknowledgments This research project and related results were made possible with the support of the NOMIS Foundation, and an NSERC fellowship (567554-2022) to IS.

References

Aho, K., Roads, B. D., & Love, B. C. (2022). System alignment supports cross-domain learning and zero-shot generalisation. *Cognition*, *227*, 105200. doi: https://doi.org/10.1016/j.cognition.2022.105200

Austerweil, J. L., & Griffiths, T. L. (2013). A nonparametric Bayesian framework for constructing flexible feature representations. *Psychological Review*, *120*(4), 817-851.

Collins, K. M., Barker, M., Espinosa Zarlenga, M., Raman, N., Bhatt, U., Jamnik, M., ... Dvijotham, K. (2023). Human uncertainty in concept-based AI systems. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society* (pp. 869–889).

Collins, K. M., Bhatt, U., Liu, W., Piratla, V., Sucholutsky, I., Love, B., & Weller, A. (2023). Human-in-the-loop mixup. In *Uncertainty in Artificial Intelligence* (pp. 454–464).

Dehaene, S., Al Roumi, F., Lakretz, Y., Planton, S., & Sablé-Meyer, M. (2022). Symbols and mental programs: a hypothesis about human singularity. *Trends in Cognitive Sciences*, 751–766.

Fei-Fei, L., Fergus, R., & Perona, P. (2006). One-shot learning of object categories. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *28*(4), 594–611.

Goldstone, R. L. (1998). Perceptual learning. *Annual Review of Psychology*, *49*(1), 585–612.

Kruschke, J. K. (1992). ALCOVE: An exemplar-based connectionist model of category learning. *Psychological Review*, *99*(1), 22–44.

Lake, B., & Baroni, M. (2023). Human-like systematic generalization through a meta-learning neural network. *Nature*, *623*(7985), 115–121.

Lake, B., Salakhutdinov, R., Gross, J., & Tenenbaum, J. (2011). One shot learning of simple visual concepts. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol. 33).

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

Malaviya, M., Sucholutsky, I., Oktar, K., & Griffiths, T. (2022). Can humans do less-than-one-shot learning? In *Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol. 44).

Peterson, J. C., Battleday, R. M., Griffiths, T. L., & Russakovsky, O. (2019). Human uncertainty makes classification more robust. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 9617–9626).

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.

Rane, S., Bruna, P. J., Sucholutsky, I., Kello, C., & Griffiths, T. L. (2024). Concept alignment. *arXiv preprint arXiv:2401.08672*.

Rane, S., Ho, M., Sucholutsky, I., & Griffiths, T. L. (2023). Concept alignment as a prerequisite for value alignment. *arXiv preprint arXiv:2310.20059*.

Sucholutskv, I., Kim, N.-H., Browne, R. P., & Schonlau, M. (2021). One line to rule them all: Generating lo-shot soft-label prototypes. In *2021 International Joint Conference on Neural Networks (IJCNN)* (pp. 1–8).

Sucholutsky, I., Battleday, R. M., Collins, K. M., Marjieh, R., Peterson, J., Singh, P., ... Griffiths, T. L. (2023). On the informativeness of supervision signals. In *Uncertainty in Artificial Intelligence* (pp. 2036–2046).

Sucholutsky, I., & Griffiths, T. L. (2023). Alignment with human representations supports robust few-shot learning. In *Thirty-seventh Conference on Neural Information Processing Systems.*

Sucholutsky, I., Muttenthaler, L., Weller, A., Peng, A., Bobu, A., Kim, B., ... others (2023). Getting aligned on representational alignment. *arXiv preprint arXiv:2310.13018*.

Sucholutsky, I., & Schonlau, M. (2021a). 'Less than one'-shot learning: Learning N classes from $M < N$ samples. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 35, pp. 9739–9746).

Sucholutsky, I., & Schonlau, M. (2021b). Soft-label dataset distillation and text dataset distillation. In *2021 International Joint Conference on Neural Networks (IJCNN)* (pp. 1–8).

Tenenbaum, J. B., Kemp, C., Griffiths, T. L., & Goodman, N. D. (2011). How to grow a mind: Statistics, structure, and abstraction. *Science*, *331*(6022), 1279–1285.

Tiedemann, H., Morgenstern, Y., Schmidt, F., & Fleming, R. W. (2022). One-shot generalization in humans revealed through a drawing task. *eLife*, *11*, e75485. doi: 10.7554/eLife.75485

Tversky, A. (1977). Features of similarity. *Psychological Review*, *84*(4), 327.

Wang, Y., Yao, Q., Kwok, J. T., & Ni, L. M. (2020). Generalizing from a few examples: A survey on few-shot learning. *ACM Computing Surveys (csur)*, *53*(3), 1–34.

Wong, C., McCarthy, W. P., Grand, G., Friedman, Y., Tenenbaum, J. B., Andreas, J., ... Fan, J. E. (2022). Identifying concept libraries from language about object structure. In *Proceedings of the 44th Annual Meeting of the Cognitive Science Society* (p. 2701—2708).

Wynn, A., Sucholutsky, I., & Griffiths, T. L. (2023). Learning human-like representations to enable learning human values. *arXiv preprint arXiv:2312.14106*.

Zhao, B., Lucas, C. G., & Bramley, N. R. (2022). How do people generalize causal relations over objects? a non-parametric bayesian account. *Computational Brain & Behavior*, *5*(1), 22–44.

Zhao, B., Lucas, C. G., & Bramley, N. R. (2023). A model of conceptual bootstrapping in human cognition. *Nature Human Behaviour*, 1–12.