Learning a Doubly-Exponential Number of Concepts From Few Examples

Ilia Sucholutsky (is3060@nyu.edu)
Center for Data Science
New York University

Bonan Zhao
School of Informatics
University of Edinburgh

Hee Seung Hwang
Dept. of Computer Science
Princeton University

Allison Chen
Dept. of Computer Science
Princeton University

Olga Russakovsky
Dept. of Computer Science
Princeton University

Thomas L. Griffiths
Dept. of Psychology
Dept. of Computer Science
Princeton University

Abstract

Recent research has shown that people can learn more new concepts than the number of examples they are presented with. However, these results relied on strong assumptions about what skills and prior knowledge are required to perform this kind of less-than-one-shot learning. This has included having participants disentangle soft labels that fuzzily map stimuli to multiple concepts, interpret continuous feature weights, and parse complex compositional statements. We propose a novel minimal paradigm that strips away these assumptions to explore how efficiently people can simultaneously learn visual and symbolic concepts. We show theoretically that it should be possible to learn up to 2^{k-1} binary features from k examples, and to learn up to $2^{2^{k-1}}$ unique combinations of those features. We validate this empirically, showing that people may be able to learn as many as 8 novel binary features and up to 256 concepts corresponding to unique compositions of those features from just 4 examples.

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

Introduction

Significant research has gone into determining how to efficiently teach people new concepts (Engelmann & Carnine, 1982), but what are the theoretical and practical limits on how many concepts people can learn from few examples? One-shot learning, or the ability to learn a new concept from a single example of it, has been studied both in human and machine learning settings (Fei-Fei et al., 2006; Lake et al., 2011, 2015; Tiedemann et al., 2022), but recent research has proposed that machines (Sucholutsky & Schonlau, 2021b,a; Sucholutskv et al., 2021) and humans (Malaviya et al., 2022; Sucholutsky, Zhao, & Griffiths, 2024) may be capable of *less-than-one-shot (LO-shot)* learning, where there are more novel concepts or categories than the number of training examples.

Initial studies on LO-shot learning leveraged "soft labels" (e.g., "this image of a handwritten digit looks 20% like the number 3, 30% like the number 8, and 50% like the number 9") as a way to associate information about multiple categories with each training example that learners could disentangle to learn about multiple categories at once. For example, these studies showed that people could learn 3 categories from 2 soft-labeled examples (Malaviya et al., 2022). However, while soft labels are highly informative (Sucholutsky, Battleday, et al., 2023), they are often quite counter-intuitive (Collins et al., 2023).

Recently, Sucholutsky, Zhao, & Griffiths (2024) showed that by leveraging cognitive mechanisms such as decomposition, contrasting, and compositional generalization, people could be taught 22 concepts from 4 training examples. However, this study relied on several strong assumptions including that a teacher can map training stimuli to continuous feature weights (e.g., "this creature is 20% tall and has 40% quantity of limbs"), that learners have strong numerical priors and can interpret those weights, and that learners have strong linguistic priors and can interpret complex compositional statements (e.g., "object B is similar to object A but with 80% feature F and much less feature G"). While this study identified mechanisms that are *sufficient* for LO-shot learning, it remains unclear what mechanisms or priors are *neccessary*.

Our aim in the current study is to find a minimal set of *naturalistic* assumptions that can enable people to learn more concepts than the number of available training examples. Engelmann & Carnine (1982) argue that people can learn almost any binary feature from simple contrasting examples (e.g., "this line is horizontal, this line is not horizontal"). We note that a set of k stimuli has 2^{k-1} possible unique contrast combinations (e.g., Object 1 has feature A, but Objects 2 and 3 do not; Objects 1 and 3 have feature B, but Object 2 does not) and propose that people can learn an exponential number of binary features (one per possible unique contrast) from a few examples. Furthermore, people can generalize to novel compositions of features they have learned (e.g., Murphy, 2004). A set of c binary features can have up to 2^c unique combinations, so we propose that people can learn as many as $2^{2^{k-1}}$ combinations from k stimuli, a doubly-exponential number of new concepts.

Bohn et al. (2021) suggest that three assumptions must be met for people to learn object-label associations: 1) the student must believe the teacher is both "cooperative and informative", 2) the student and the teacher must have "shared common ground" about what is being discussed, and 3) students should retain "semantic knowledge about previously learned word-object mappings". We leverage these principles to design an experiment for testing the limits of human learning from few examples with binary features by telling participants that they are on an alien world where they must learn the local language and use it to describe the aliens they encounter. Participants are shown four example aliens ac-

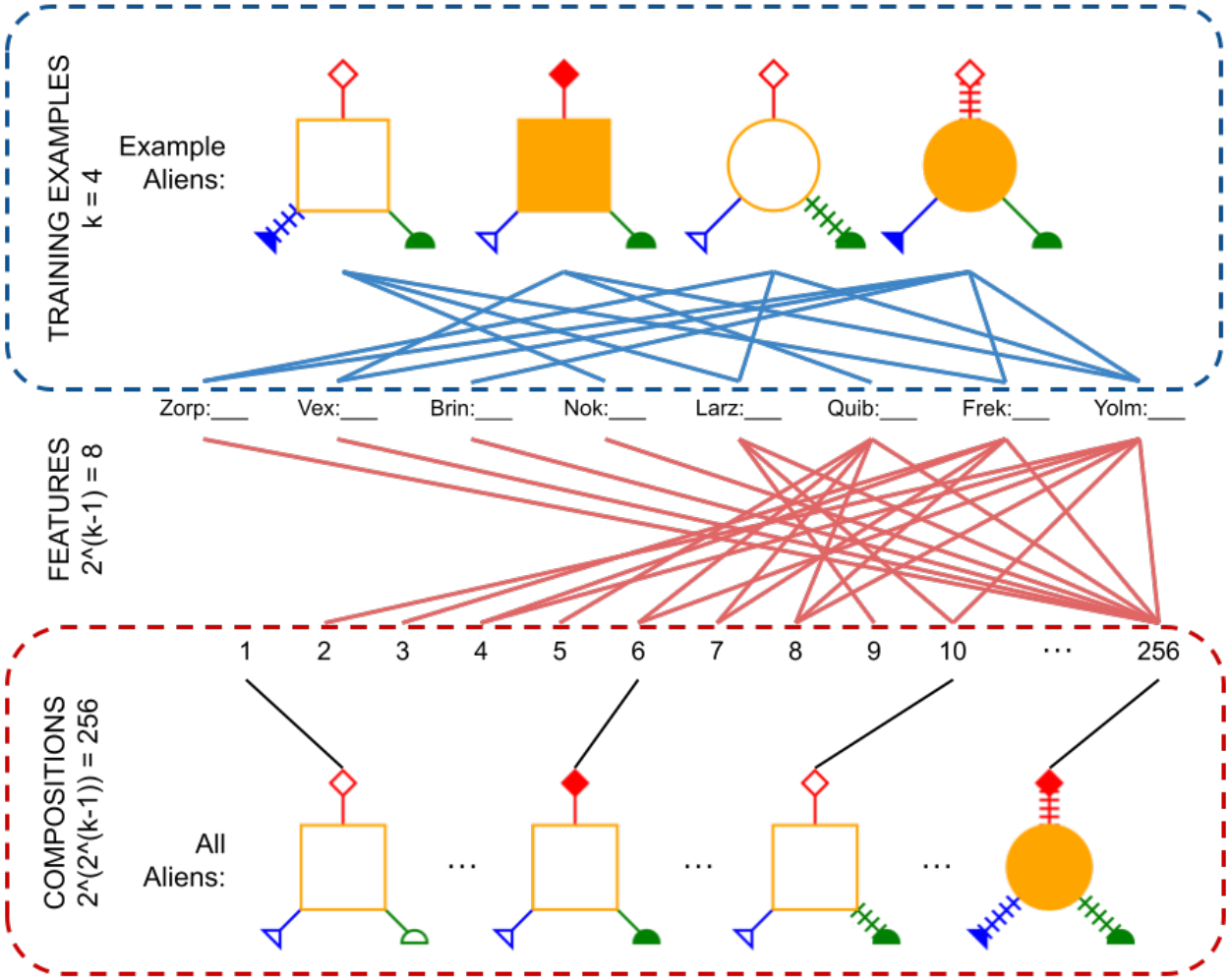

Figure 1: Visual summary of our experiment (adapted with permission from Sucholutsky, Zhao, & Griffiths, 2024). Everything within the dotted lines is provided to the participants and everything between the dotted lines is what participants need to learn. Everything within the dotted blue line is part of Phase 1 where participants compare-and-contrast $k = 4$ labeled alien examples to learn $2^{k-1} = 8$ binary features. Everything within the dotted red line is part of Phase 2 where participants label $2^{2^{k-1}} = 256$ aliens composed of unique combinations of the features.

companied by descriptions in the alien language, where the language consists of eight words, each corresponding to a binary feature (see Fig. 1). Participants are told that a local expert will teach them the language (assumption 1 – cooperative/informative teacher) by pointing at subsets of the example images and saying the alien word that corresponds to them (assumption 2 – the images are shared common ground) and participants take notes that they always have access to on what they think each feature means (assumption 3 – retained semantic knowledge of previous mappings).

We find that participants who receive this scaffolding and successfully learn the eight features can correctly label novel aliens sampled from the 256 possible combinations of features (i.e., 252 previously unseen combinations). In other words, people can learn as many as 8 feature concepts and 256 compositional concepts from just 4 training examples.

Background

The mind processes information compositionally (Fodor, 1983; Lake et al., 2017). We understand novel concepts from their components (Murphy, 2004; Chomsky, 2002), create new technologies by combining existing things (Allen et al., 2020; Arthur, 2010; Fleming, 2001), and can envision situations that never happen before from creatively composing other familiar elements (Bar, 2007). From language (Montague, 1970) to data structures (Sucholutsky, Zhao, & Griffiths, 2024), compositionality has been a central theme in understanding how a computational agent can efficiently learn from few data (Lake et al., 2015; Goodman et al., 2008), and generalize robustly to new situations (Gershman et al., 2015; Zhao et al., 2022).

We often use binary features to describe objects when communicating in the real world, often encoded as adjectives

(e.g., a big, fuzzy, red dog). People can learn binary features from a set of positive and negative examples of that feature (Engelmann & Carnine, 1982) – e.g., "this is a big dog", "this is a small dog". If the feature activations are not perfectly correlated across all of the examples, then people can learn multiple unique features from the same set of examples (e.g., "this is a big, red dog", "this is a small, red dog", "this is a big, blue dog").

A set of k stimuli can support 2^k unique binary activation patterns, or 'feature codes'. Formally, we define a feature code for feature ϕ_i and stimuli $s_1,...,s_k$ as the sequence $\phi_i(S^k) := \phi_i(s_1),...,\phi_i(s_k)$ where $\phi_i(s_j) = 1$ if ϕ_i is activated for s_j, and $\phi_i(s_j) = 0$ otherwise. For example, a set of 3 stimuli can have feature codes $\{000,001,010,100,011,101,110,111\}$. However, without a visual prior for what it means for a feature to be activated, learners cannot disentangle features encoded with reciprocal codes like 100 and 011 (we denote the reciprocal of a feature code $\phi_i(S^k)$ by $\bar{\phi}_i(S^k)$). Thus, a set of k stimuli can be used to teach at most 2^{k-1} binary features. The choice of the k stimuli, or the 'curriculum', is then crucial in supporting this kind of exponential feature learning. The stimuli must be carefully selected such that no features are perfectly correlated across the stimuli (i.e., no features have identical or reciprocal codes; $\phi_i(S^k) \neq \phi_j(S^k)$ and $\phi_i(S^k) \neq \bar{\phi}_j(S^k)$ for all $i \neq j$).

People are also capable of generalizing to novel compositions of previously learned features or concepts (Kemp et al., 2010; Zhao et al., 2024) and these compositions can themselves be considered (complex) concepts(Murphy, 1988). For c binary features, there are 2^c possible unique combinations. Formally, we define a combination code for stimulus s_j and features $\phi_1,...,\phi_c$ as the sequence $\Phi^c(s_j) := \phi_1(s_j),...,\phi_c(s_j)$. Thus from k training example stimuli, people may be able to learn 2^{k-1} binary features (i.e., learning from a set of example stimuli defined by $\Phi^{2^{k-1}}(S^k)$) and in turn recognize $2^{2^{k-1}}$ unique combinations of those features (i.e., generalize to a much larger set of stimuli defined by $\Phi^{2^{k-1}}(S^{2^{2^{k-1}}})$).

This suggests that people, and perhaps machines, may be able to learn a doubly-exponential number of compositional concepts from a small number of examples (see 1 for a visual summary). However, we note two caveats that may limit this ability. First, it is unclear whether binary features can be learned exclusively from positive examples, or if at least some negative examples are required (i.e., can a feature code consisting only of 1s be used). Second, any communication channel is likely to be noisy (e.g., a learner may misinterpret some examples or labels) and redundant bits may be necessary in the feature codes in practice for error-correction (Hamming, 1950) to make the learning and generalization robust.

Testing Binary LO-Shot Learning

To validate these theoretical results, we conduct an experiment where participants learn from $k = 4$ stimuli encoding 8 binary features that support 256 unique combinations.

Methods

Participants We recruited 100 participants through Prolific (47 females, $M_{\text{age}} = 40 \pm 11$). Two participants were excluded from analysis who failed to correctly respond to two attention checks, leaving us 98 participants in total. The task took 29 ± 15 minutes and participants were compensated $6 for their time. All participants gave informed consent prior to the study, in accordance with an approved Princeton University IRB protocol (#10859).

Material We generated images of an alien creature specified by eight binary features assigned to made-up monosyllabic words:

Zorp - body shape is circle (or not),
Vex - body is solid (or not),
Brin - top antenna is dashed (or not),
Quib - top tip is solid (or not),
Nok - left antenna is dashed (or not),
Frek - left tip is solid (or not),
Larz - right antenna is dashed (or not),
Yolm - right tip is solid (or not).

Training examples are visualized in the Phase 1 interface in Figure 3 and summarized by the following matrix, where rows correspond to feature codes, and columns correspond to training example combinations.

$$\Phi_1^8(S_1^4) = \begin{array}{c} \\ Zorp \\ Vex \\ Brin \\ Nok \\ Larz \\ Quib \\ Frek \\ Yolm \end{array} \begin{array}{c} \begin{array}{cccc} Alien_1 & Alien_2 & Alien_3 & Alien_4 \end{array} \\ \begin{pmatrix} 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 1 & 1 & 1 \end{pmatrix} \end{array}$$

Design and procedure The experiment consisted of 2 phases.

Phase 1, Language/Feature Learning: Each participant was first pre-trained on two simple practice trials (see Fig. 2) so that they could familiarize themselves with the Phase 1 interface and the compare-and-contrast learning strategy that they would need to use throughout it. Participants were then shown the 4 training aliens with labels only corresponding to one of the features (see Fig. 3) and told that a 'local expert points at [some] of them and says [feature name]' before being asked to write a free text response of what they think that feature name means. This was repeated for each of the 8 features. Each response was saved into the participant's 'notes' which were available to them throughout the entirety of both Phases 1 and 2. Phase 1 satisfies all 3 assumptions discussed above that Bohn et al. (2021) propose are required for learning object-label associations, and provides scaffold-

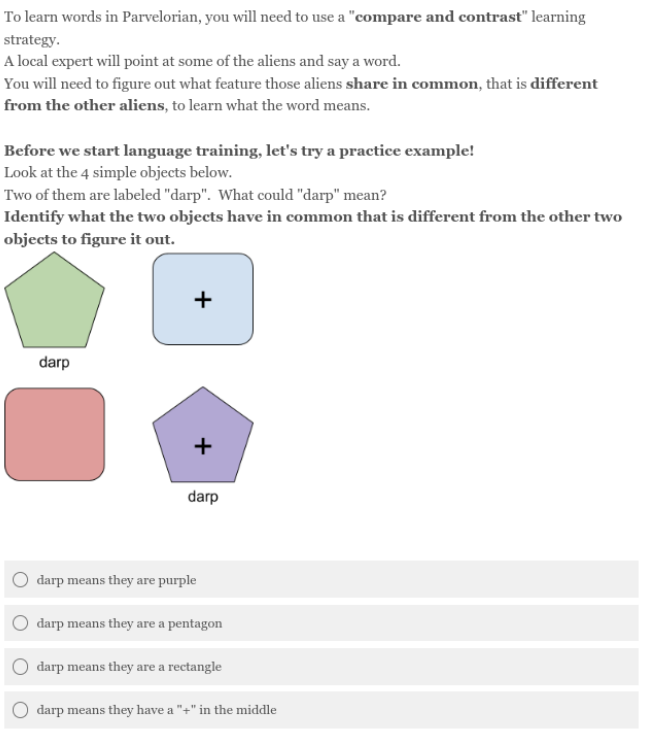

Figure 2: Pre-training and scaffolding for Phase 1, Language/Feature Learning. Before working with the Phase 1 interface, participants were first pre-trained on two simple practice trials (first trial pictured here) to familiarize them with the interface and the compare-and-contrast learning strategy.

ing for participants before they move on to Phase 2.

Phase 2, Generalization: Each participant was presented with 30 aliens (one at a time) with 5 five aliens fixed for all participants (those with combination codes {11111011, 11111100, 11111101, 11111110, 11111111}) and 25 aliens randomly sampled for each participant from the remaining 251 possible unique combinations of the binary features. Participants were asked to select which of the 8 features they believed applied to the presented alien (see Fig. 4).

Participants were evenly split into two groups. The experimental group went through both phases while the control group only went through Phase 2.

AI models We also repeat the experiment with two types of AI systems: multimodal vision-language models (VLMs), and unimodal vision-only models. To understand the limits of current AI capabilities, we use the state-of-the-art OpenAI o1 model (Jaech et al., 2024) from the Azure OpenAI API as our VLM, and to analyze the contribution of having strong language priors and cross-modal alignment to the AI system's ability to perform the task (Chen et al., 2024), we compare o1's performance to logistic classifiers trained on dimensionality-reduced embeddings (top 2 principal compo-

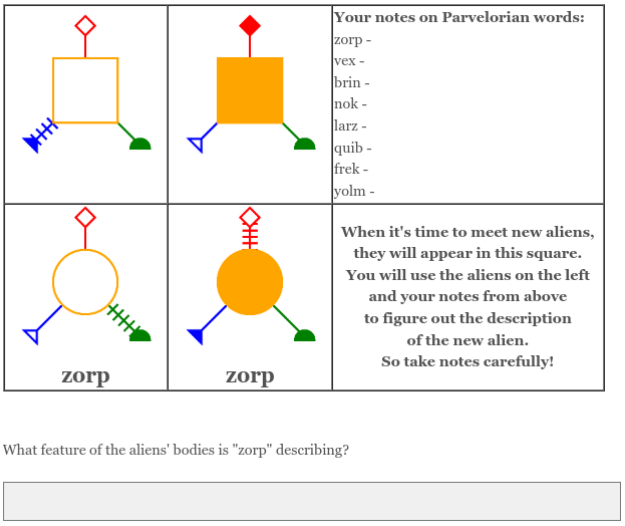

Figure 3: Interface for Phase 1, Language/Feature Learning. Participants were shown the 4 training example images with only one feature word at a time and are asked to write down what aspect of the aliens' bodies they believe this feature is describing. As participants progress through all 8 features in Phase 1, their responses to already-seen features appear in the top-right cell.

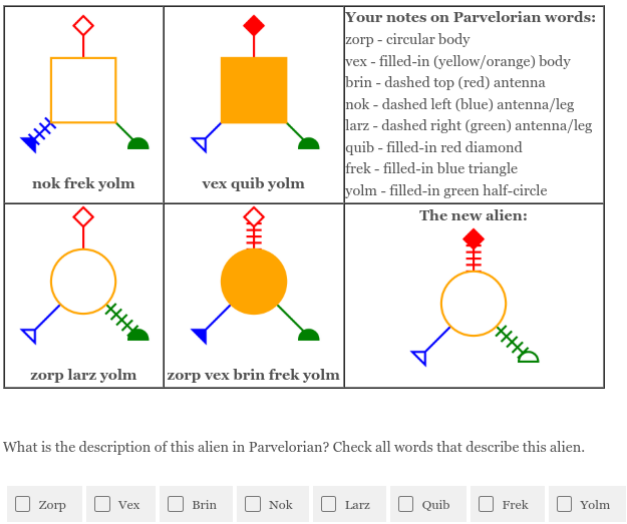

Figure 4: Interface for Phase 2, Generalization. Participants were shown the 4 training example images and their full descriptions, the responses the participant entered during phase 1, and a new unlabeled alien image. Participants are asked to pick which (if any) of the features apply to the unlabeled alien.

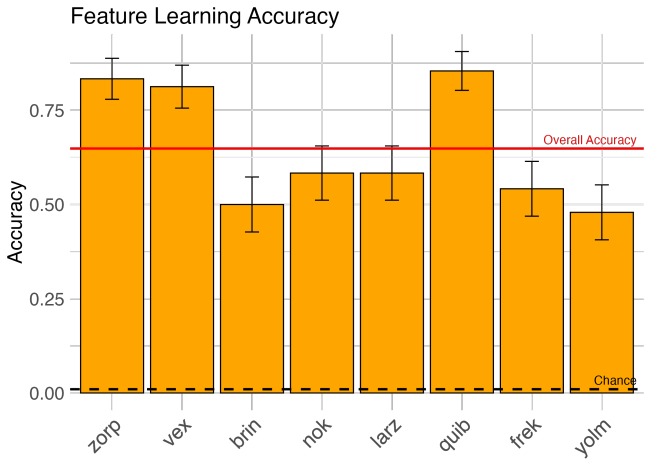

Figure 5: Human participant feature learning accuracy during Phase 1. The red line denotes average accuracy across all 8 features and the dotted line corresponds to chance-level performance. Error bars correspond to standard errors.

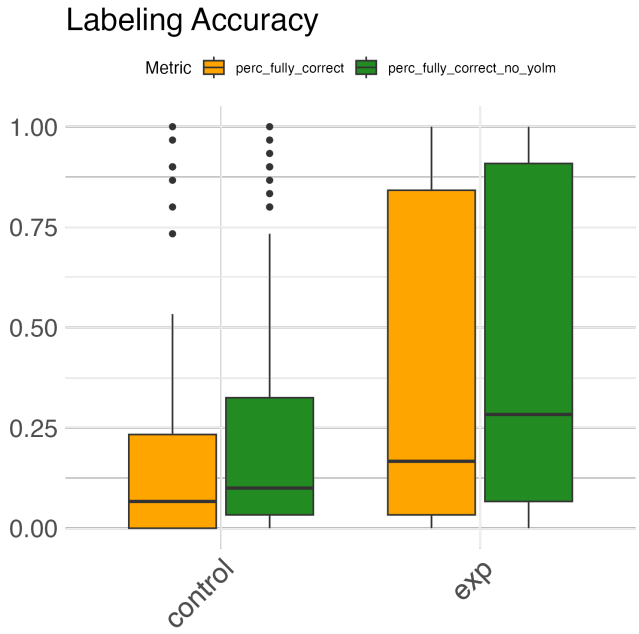

Figure 6: Participants in the experimental condition (participated in both Phase 1 and 2) had higher test accuracy (i.e., fully correct labels where all eight features were correctly turned on/off) during Phase 2 than the control group (participated in Phase 2 only). We show accuracy both with (orange) and without (green) considering the 'yolm' feature which had only positive examples.

nents outputted by PCA; Wold et al., 1987) from two state-of-the-art vision models, OpenSelfSup[1] implementations of BYOL (Grill et al., 2020) and MOCO (He et al., 2020).

[1]https://github.com/Berkeley-Data/OpenSelfSup

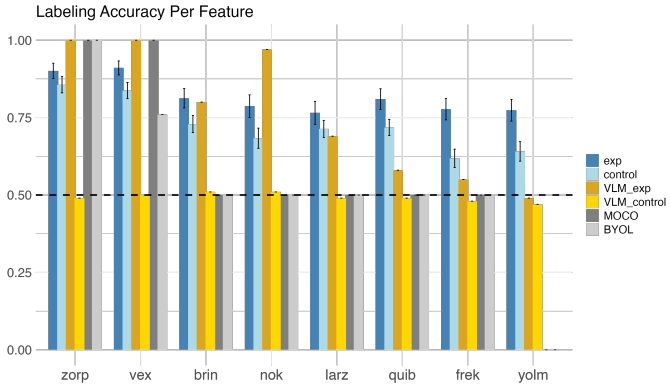

Figure 7: Per-feature labeling accuracy during Phase 2 for human participants in control and experiment conditions, a VLM in control and experiment conditions, and two vision-only AI models. The dotted line corresponds to chance-level performance. Error bars are standard errors.

Results

People can learn exponentially many binary features from a few examples Participants successfully learned the mapping between features and labels (see Figure 5). A one-sample t-test confirmed that participants' learning accuracy ($M = 0.65 \pm 0.3$) was significantly above chance, $t(47) = 14.95$, $p < 2.2 \times 10^{-16}$. As shown in Figure 5, learning accuracy among features were not perfectly even. However, a chi-square goodness-of-fit test did not reveal strong evidence that learning accuracy across features deviates from a uniform distribution ($\chi^2(7) = 12.88$, $p = 0.075$).

People may be able to learn some binary features from only positive examples One particularly intriguing finding is that 23 participants out of the total 48 were able to correctly deduce the meaning of 'yolm' (Figure 5), despite this feature being activated for all 4 training aliens (i.e., no negative examples). An exact binomial test revealed a significant deviation from chance, $p < 2.2 \times 10^{-16}$, 95% CI = [0.33, 0.63]. It is possible that some participants were able to do this by inferring that the yolm feature should a) have similar complexity or qualitative properties to the other features, and b) that it should be learnable and non-trivial (e.g., that it should not just generically refer to all aliens). This suggests that pragmatics may play an important role in supporting generalization when learning in such data-sparse settings.

Scaffolding and satisfying the three assumptions from Bohn et al. (2021) improves LO-shot feature learning During the test phase, participants in both the experiment and control groups successfully labeled the novel aliens significantly above chance ($t(97) = 7.97, p = 1.588 \times 10^{-12}$). Participants in the experiment group who went through Phase 1 and systematically learned each feature one-by-one using a compare-and-

contrast learning strategy, performed significantly better in Phase 2 ($M = 0.4 \pm 0.4$) than participants from the control group who did not go through Phase 1 ($M = 0.2 \pm 0.3$), $t(90.89) = 2.32, p = 0.02, 95\%$ CI: $[0.024, 0.311]$ (see Figure 6; we note that the accuracy gap between control and experiment conditions may be even greater when accuracy is measured after excluding the 'yolm' feature that had only positive examples).

Similarly, the VLM in the experimental condition ($M = 0.76 \pm 0.21$) significantly outperformed the VLM when it was in the control condition ($M = 0.5 \pm 0.01$) where its performance dropped to approximately chance levels. This suggests that scaffolding learning by focusing on each feature individually while satisfying the three assumptions proposed by Bohn et al. (2021) may be an important component (for both humans and machines) for learning successfully in settings like ours where there are few examples but with high information density.

People can learn a doubly-exponential number of concepts from a few examples Unlike during feature learning, during the Phase 2 test labeling for the experiment group, accuracy across features was significantly different from being uniform, $\chi^2(7) = 45.65, p = 1.02 \times 10^7$ (Figure 7). Participants who did better during learning also achieved higher performance in the test phase (Figure 8). A linear regression analysis revealed that feature learning accuracy in Phase 1 was a significant predictor of test accuracy in Phase 2, $\beta = 1.08, F(1,46) = 91.46, p < 1.65 \times 10^{12}, R^2 = 0.67$.

VLMs and human participants outperform vision-only AI systems Repeating the experiment with two vision-only AI models we find that people and VLMs both greatly outperform the two vision-only models which perform at chance levels for all features other than Zorp and Vex (Figure 7). These results suggest that strong language priors and cross-modal alignment may be important contributors to learning new concepts from few examples.

Conclusion

Humans have an incredible ability to learn from very little data and in this study we aimed to probe the limits of just how many new concepts people could learn from a few examples. We empirically showed that people can learn up to 8 binary features and 256 compositional concepts from just 4 examples, which validated our theoretical prediction that it should be possible to learn 2^{k-1} binary features and $2^{2^{k-1}}$ combinations of those features from k carefully chosen and labeled examples.

In practice, people are able to re-evaluate and rapidly update their beliefs about feature mappings. We saw this in the post-experiment free-text feedback responses of some of the participants who mentioned having changed what they think certain feature names corresponded to during the course

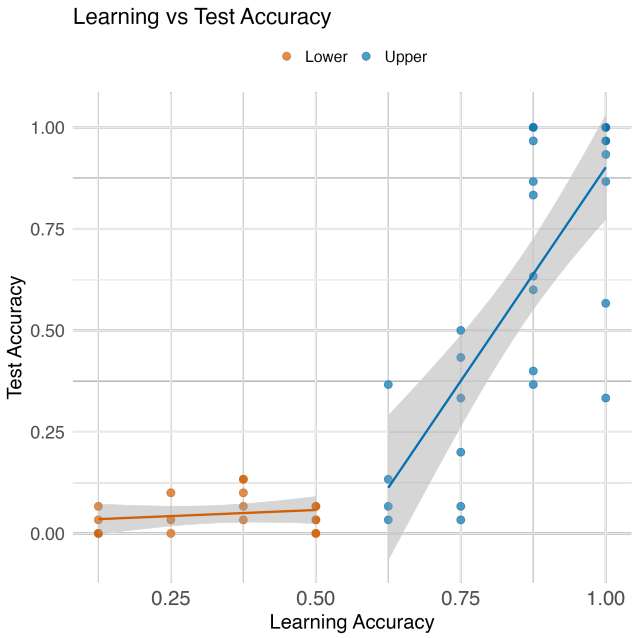

Figure 8: Relationship between feature learning accuracy during Phase 1 and test performance during Phase 2 for participants in the experiment condition. Participants are in two groups based on their feature learning accuracy being either above (upper; blue dots and line) or below (lower; orange dots and line) a threshold of 0.5, with a line of best fit plotted for each group.

of the experiment. This can still be considered LO-shot learning, as even though participants see additional alien images, they are unlabeled and participants receive no feedback/supervision on the labels they propose for these images. Future work should explore this backtracking ability both theoretically and empirically as a possible additional contributor to human less-than-one-shot learning abilities.

We further note that people's ability to learn this way may be hampered by factors like working memory limitations, noisy feature interpretations, and misaligned priors or representations (Sucholutsky, Muttenthaler, et al., 2023; Sucholutsky, Collins, et al., 2024). Our study examines the theoretical extreme limits of learning many concepts in the binary feature learning setting. Overall, our results suggest a more effective and naturalistic set of mechanisms that may enable humans to perform less-than-one-shot learning, acquiring an exponential number of features and a doubly-exponential number of new concepts from very few training examples.

Acknowledgments This work was supported by the NOMIS Foundation and a Microsoft AFMR grant. The authors thank Juju and Dan Nicolau Jr. for their help in inspiring the compositional binary feature learning setting used in this work.

References

Allen, K. R., Smith, K. A., & Tenenbaum, J. B. (2020). Rapid trial-and-error learning with simulation supports flexible tool use and physical reasoning. *Proceedings of the National Academy of Sciences*, *117*(47), 29302–29310.

Arthur, W. B. (2010). *The nature of technology: What it is and how it evolves*. Penguin.

Bar, M. (2007). The proactive brain: using analogies and associations to generate predictions. *Trends in Cognitive Sciences*, *11*(7), 280–289.

Bohn, M., Tessler, M. H., Merrick, M., & Frank, M. C. (2021). How young children integrate information sources to infer the meaning of words. *Nature Human Behaviour*, *5*(8), 1046–1054.

Chen, A., Sucholutsky, I., Russakovsky, O., & Griffiths, T. (2024). Analyzing the roles of language and vision in learning from limited data. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol. 46).

Chomsky, N. (2002). *Syntactic structures*. Mouton de Gruyter.

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).

Engelmann, S., & Carnine, D. (1982). *Theory of instruction: Principles and applications*. Irvington Publishers.

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.

Fleming, L. (2001). Recombinant uncertainty in technological search. *Management Science*, *47*(1), 117–132.

Fodor, J. A. (1983). *The modularity of mind*. MIT press.

Gershman, S. J., Horvitz, E. J., & Tenenbaum, J. B. (2015). Computational rationality: A converging paradigm for intelligence in brains, minds, and machines. *Science*, *349*(6245), 273–278.

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.

Grill, J.-B., Strub, F., Altché, F., Tallec, C., Richemond, P., Buchatskaya, E., ... others (2020). Bootstrap your own latent-a new approach to self-supervised learning. *Advances in Neural Information Processing Systems*, *33*, 21271–21284.

Hamming, R. W. (1950). Error detecting and error correcting codes. *The Bell System Technical Journal*, *29*(2), 147–160.

He, K., Fan, H., Wu, Y., Xie, S., & Girshick, R. (2020). Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 9729–9738).

Jaech, A., Kalai, A., Lerer, A., Richardson, A., El-Kishky, A., Low, A., ... others (2024). OpenAI o1 system card. *arXiv preprint arXiv:2412.16720*.

Kemp, C., Goodman, N. D., & Tenenbaum, J. B. (2010). Learning to learn causal models. *Cognitive Science*, *34*(7), 1185–1243.

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.

Lake, B., Ullman, T., Tenenbaum, J., & Gershman, S. (2017). Building machines that learn and think like people. *Behavioral and Brain Sciences*, *40*, e253.

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

Montague, R. (1970). Universal grammar. *Theoria*, *36*(3).

Murphy, G. (1988). Comprehending complex concepts. *Cognitive science*, *12*(4), 529–562.

Murphy, G. (2004). *The big book of concepts*. MIT press.

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., Collins, K. M., Malaviya, M., Jacoby, N., Liu, W., Sumers, T. R., ... others (2024). Representational alignment supports effective machine teaching. *arXiv preprint arXiv:2406.04302*.

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).

Sucholutsky, I., Zhao, B., & Griffiths, T. (2024). Using compositionality to learn many categories from few examples. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol. 46).

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

Wold, S., Esbensen, K., & Geladi, P. (1987). Principal component analysis. *Chemometrics and Intelligent Laboratory Systems*, *2*(1-3), 37–52.

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. (2024). A model of conceptual bootstrapping in human cognition. *Nature Human Behaviour*, *8*(1), 125–136.