# Few-shot image classification by generating natural language rules

**Wai Keen Vong** and **Brenden M. Lake**
Center for Data Science
New York University
New York, NY, United States

## Abstract

The ability to generate rules and hypotheses plays a key role in multiple aspects of human cognition including concept learning and explanation. Previous research has framed this ability as a form of inference via probabilistic program induction. However, this approach requires careful construction of the right grammar and hypothesis space for a particular task. In this work, we propose an alternative computational account of rule generation and concept learning that sidesteps some of these issues. By leveraging advances in multimodal learning and large language models, we extend the latent language framework from Andreas et al. (2017) to work in a zero-shot manner. Taking naturalistic images as input, our computational model is capable of generating candidate rules that are specified in natural language, and verifying them against the observed data. We show that our model can generate, in a zero-shot manner, plausible rules for visual concepts in two domains.

## 1 Introduction

Humans can generate a wide variety of rules and hypotheses to make sense of the world, ranging from perceptual rules that distinguish between safe and poisonous plants to complex scientific hypotheses (Schulz, 2012). Where do these hypotheses come from? Within the last decade, there has been a resurgence of interest in rule and hypothesis generation, led by accounts positing concept learning as a form of probabilistic program induction (Goodman et al., 2008; Piantadosi and Jacobs, 2016; Bramley et al., 2018). Such approaches have been successful at capturing human concept learning in domains such as hand-written characters (Lake et al., 2015) and logical concepts (Piantadosi et al., 2016).

However, these approaches have important limitations. First, these models are often instantiated with a minimal grammar to construct novel concepts, whereas humans can leverage their knowledge of pre-existing lexical concepts to guide the rule generation process. Second, the grammars used in these with are often *domain-specific*, requiring researcher effort to reverse-engineer from behavioral data or to transfer to different domains.

In this paper, we outline an alternative account of rule-based concept learning that attempts to overcome some of the above limitations. We combine two key ideas in our approach. First, rather than specifying a domain-specific grammar, we leverage natural language as a *domain-general* representational medium for rules, inspired by the latent language framework from Andreas et al. (2017). Second, we utilize recent advances from multimodal neural networks (Radford et al., 2021; Li et al., 2022) and large language models (Brown et al., 2020), to instantiate a few-shot image classification model via latent language, allowing it to classify images via generating sensible rules specified in natural language without *any* additional training. We demonstrate our model's ability to generate rules in a *zero-shot* manner across two kinds of visual concept learning tasks.

## 2 Related Work

Within cognitive science, our work builds off a long history of research into rule and hypothesis generation approaches to concept learning (Nosofsky et al., 1994; Goodman et al., 2008; Piantadosi et al., 2016; Bramley et al., 2018; Ellis et al., 2020). Many of these accounts treat concept learning as a form of probabilistic program induction, sampling hypotheses or rules from an grammar-based hypothesis space. However, these models are specified with a minimal starting representation that is in stark contrast to the vast library of concepts already available from natural language.

Within machine learning, our work is related to models of few-shot visual classification (Lake et al., 2015; Snell et al., 2017; Mu et al., 2020), but especially the work of Andreas et al. (2017), which

leverages natural language as a representational bottleneck for performing visual classification. Within few-shot learning, there also exist models that allow the use of auxiliary class information for zero-shot classification (Snell et al., 2017), but in this work our model generates the relevant semantic class label information itself from the support examples. We also take advantage of recent work in multimodal neural networks (Radford et al., 2021; Li et al., 2022) which have been shown to contain extensive knowledge of multimodal concepts (Goh et al., 2021), and can be used to perform tasks like image captioning and zero-shot image classification, and large language models (Brown et al., 2020) which offer complementary strengths such as reasoning in natural language.

## 3 Model

Our model extends the *Learning with Latent Language* (L3) approach from Andreas et al. (2017). The L3 model consists of a **proposal model** $g_\phi$ that generates candidate natural language rules via an image captioning system, and an **interpretation model** $h_\eta$ that provides a numerical similarity score between a natural language rule and an image, allowing one to check the goodness of a sampled rule. Training the L3 model first required a *language learning* phase where domain-specific language annotations were used to train each component separately, which could then be utilized at test time by generating candidate language descriptions for new concepts. In this work, we skip the language learning phase, and instead leverage and combine existing pre-trained multimodal and large language models to perform both the proposal and interpretation steps. Additionally, the use of natural language as a representational bottleneck in this architecture means that different multimodal or language models that take natural language as input, or generate natural language as output, can be stitched together despite not being trained jointly. [1]

To generate proposals, Andreas et al. (2017) used the mean embedding across multiple images as input to a trained image captioning model to propose natural language rules. However, pre-trained image captioning models are generally trained to predict captions from a single image, not across multiple images. Furthermore, these captions are typically only applicable to that particular image, and do not describe what might be common amongst multiple images, which is our intended goal.

To alleviate these issues, we split the proposal model into two separate steps in our model, as shown in Figure 1. First, in the **captioning** step, we use BLIP (Li et al., 2022), a recent pre-trained multimodal transformer, to generate captions for each image in the support set separately. Second, in the **abstraction** step, we leverage large language models like GPT-3 (Brown et al., 2020) to propose natural language rules that convert these instance-

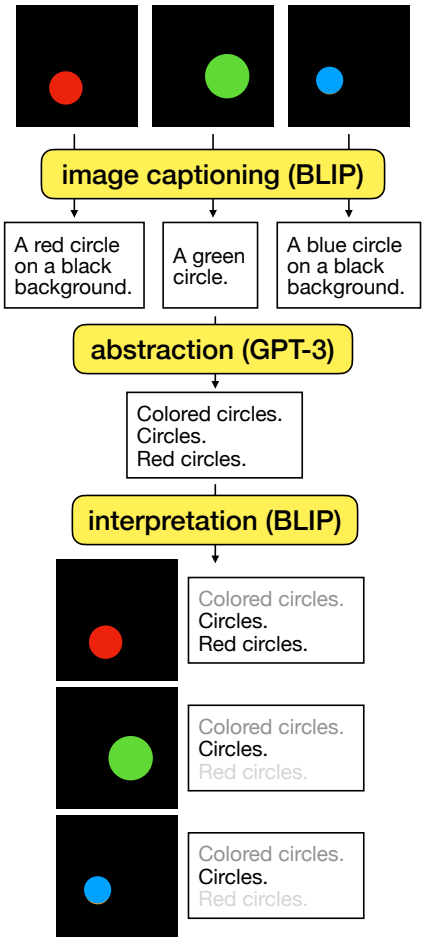


Figure 1: **Latent Language Model**. Given a set of support images, our model first lifts these images into natural language via a pre-trained image captioning model. Second, we prompt GPT-3 to convert these instance-level captions into an abstract natural language rule, sampling multiple candidate rules in the process. Finally, these proposed rules are interpreted against the support images via image-text matching to determine the best rule (rules in darker text are better).

[1] Zeng et al. (2022) concurrently proposed Socratic Models, a similar approach for combining together different pre-trained models via natural language as the medium for communicating between models to solve other kinds of multimodal reasoning tasks.

level captions to an abstract rule specified in natural language. Specifically, we provided GPT-3 with a prompt consisting of a task description followed by some examples of descriptions and a natural language rule for how these descriptions were related, and then appended each of the generated captions in a random order to the prompt and asked it to generate a corresponding rule (see Appendix A). Across the experiments reported, we run the captioning step once and sample multiple rules via prompting GPT-3.

Adapting the **interpretation** step for the zero-shot setting is relatively straightforward. The image-text matching procedure from L3 is very similar to how image-text matching is used to train contrastive image-text models (Radford et al., 2021; Li et al., 2022). For convenience purposes, we reuse the BLIP architecture from the captioning step, but instead use the separate image-text contrastive head as $h_\eta$. The interpretation model takes in an image and text as input, and outputs a cosine similarity score describing the match between the given image and text. The interpretation model is used to determine the best generated rule by averaging over the cosine similarities between each of the sampled rules against all of the support examples. Once a given rule is chosen as the best amongst the set of sampled rules, that same rule is then used with the interpretation model for prediction on new query examples.

## 4 Experiments

**ShapeWorld.** We adapted the ShapeWorld visual reasoning dataset which has been used to investigate few-shot visual concept learning (Kuhnle and Copestake, 2017; Andreas et al., 2017; Mu et al., 2020). While the kinds of images in this dataset are relatively simple abstract shapes of varying colors, one important difference is that our model has not been trained directly on any images from ShapeWorld, and needs to determine the relevant dimensions to generate valid rules from the observed exemplars without any prior experience in this domain.

Using 8 colors and 7 shapes, we generated three types of concepts consisting of either a single primitive **shape** concept (where color could vary), or a primitive **color** concept (where shape could vary), or a conjunction of **shape and color** primitives (e.g. a *blue triangle*, or a *red square*).

Since the concepts generated in this domain consist of a single class of positive examples, we opted to test the model via *set completion* (Andonian et al., 2020) rather than few-shot classification. Given with a set of support images that follow some rule, set completion involves selecting one image out of many from a separate query set that matches the concept from the support set. Given a set of four support images from a given concept, we used the captioning model to generate four captions, which were passed into GPT-3 to generate 10 candidate natural language rules. The best rule was determined by the rule with the highest average cosine similarity to the set of support examples using the interpretation model $h_\eta$. From a separate query set of another four examples, this rule was subsequently used to select the example from the query set that matched the concept from the support set. We sampled 300 concepts roughly split between the three different types for testing.

| Model | Overall | S | C | S+C |
|---|---|---|---|---|
| Prototype | 89.7 | 92.9 | 77.1 | 98.1 |
| L3 (1 rules) | 64.6 | 70.4 | 63.5 | 60.3 |
| **L3 (10 rules)** | 88.3 | 86.7 | 89.6 | 88.7 |
| L3 (oracle) | 91.0 | 90.8 | 91.7 | 90.6 |

Table 1: ShapeWorld set completion results (as percent correct). **S** refers to shape concepts, **C** refers to color concepts, and **S+C** refers to shape and color concepts

**Results.** Results are shown in Table 1. Overall, our model performed quite well at the set completion task in this domain, and allowing the model to generate multiple candidate rules led to a substantial increase in accuracy. Qualitatively, the kinds of rules generated were sensible and matched the underlying concepts (e.g. "*a purple triangle*", "*squares on a black background*"), whereas incorrect rules often specified the wrong target feature (such as color instead of shape).

Overall performance was similar to a baseline prototype model, that predicted the matching exemplar based on the cosine similarity of each query example to the mean embedding from the visual representations of the support examples from BLIP's vision encoder. Breaking down performance for the different rule types showed interesting differences. Although performance in our model was the same across the three types of concepts, the prototype-based model was more accurate at classifying conjunctive concepts, likely because both shape and color dimensions were cor-

related leading to examples being clustered more tightly in the visual embedding space. However, its performance on shape versus color was lower, with color concepts being substantially more difficult. This suggests that generating candidate natural language rules might be especially helpful with determining the relevant axis of generalization compared to a model that generalizes from fixed visual representations, and allowing for more flexible kinds of generalizations.

**Abstract Rules.** In the second experiment, we explored two kinds of more challenging rules using naturalistic images. We were interested to see whether our model could generate sensible rules, and whether using these natural language rules for classification would be more beneficial than image features alone. We filtered a subset of the COCO training dataset (Lin et al., 2014), for images with annotations with either one or two animals. We used this subset to create few-shot classification splits for determining whether images contained **one vs. two** animals, or determining whether an image contained two **same vs. different** animals, an analogue of the same-different task that has been extensively studied in cognitive science (Carstensen and Frank, 2021). Both of these kinds of rules are more challenging since they rely on picking up relational properties across objects in a single image, rather than more salient features like color or shape.

Out of the 10 animal classes in COCO, we randomly split half to be sampled as images for support examples, and the other half to be sampled as query examples.[2] For both types of rules, we sampled 6 exemplars of each class to form a set of support examples, and repeated this process to create 10 different instances of each concept split. To evaluate the models, we sampled 6 additional exemplars from each class as query examples. Due to the added complexity of the rules in this experiment, we allowed the model to sample up to 20 rules (sampling one rule for each class separately), and selected the best rule on the basis of classification on the support set. Initial testing showed that the previous prompt to GPT-3 used in the ShapeWorld domain was not sufficient generating sensible rules, and required an adaptation to the original prompt for this task with slightly modified examples (see Appendix A for details).

[2]Any images that contained one animal from the support animal classes, and one from the query animal classes was discarded during this sampling process.

| Model | One vs. Two | Same vs. Diff |
|---|---|---|
| Prototype | 63.3 | 52.5 |
| **L3 (20 rules)** | 59.1 | 70.0 |
| L3 (oracle) | 87.5 | 85.8 |

Table 2: COCO few-shot classification results (as percent correct).

**Results.** Results are shown in Table 2. Compared to the ShapeWorld domain, we observed more variability in the results for this simulation. Even with the ability to generate 20 candidate rules, performance on both tasks was quite variable, depending on the quality of the rules generated in each separate instance.[3] For the **one vs. two** condition, the model often generalized too strictly, e.g. "*two giraffes*" which led to incorrect classifications on the query set that involved other animals, and its overall performance was similar to the prototype model. Interestingly, for **same vs. different**, the model showed reasonable classification accuracy relative to the chance-level performance observed in the prototype model, and was able to generate plausible rules like "*animals in a pair*", or "*a couple of something*" for the **same** instances. Furthermore, providing the model with ground truth rules in the oracle model resulted in high classification performance in both tasks, suggesting that classification via natural language rules might be especially effective for more abstract visual properties. Of course, the model must reliably generate compelling rules to succeed, or learn these distinctions by receiving natural language rules from another person in a social context (Chopra et al., 2019; Acquaviva et al., 2021).

## 5 Discussion

In this paper, we extended the learning from latent language (L3) framework from Andreas et al. (2017) to enable generation of natural language rules in a zero-shot manner, and demonstrated its use in two few-shot image classification settings. By combining large-scale pre-trained neural networks to act as proposal and interpretation models, we constructed a system that could generate natural language rules across a number of concept learning

[3]We also noticed that even in the context of a single rule, classification responses demonstrated a graded nature, for example, an image with two animals but one partially occluded would be more confusable than an image with two distinct animals.

tasks by generating valid yet interpretable rules in natural language. Our results highlight the *flexibility* of using natural language for generalization: our model was better at generalizing to color concepts than a baseline prototype model that relied on visual features alone, as well as to more abstract relational rules in the second experiment (particularly when it was provided with the underlying rule to use), suggesting that natural language can modulate generalization and classification behavior.

One limitation of the current approach is that is reliant on the text captions to generate proposed rules, and thus the quality of the generated rules is dependent on what aspects or features the pretrained captioning model chooses to highlight and describe. Some ways to overcome this issue might be using a model trained to caption multiple images directly (Hernandez et al., 2022), or incorporating images directly into the abstraction prompt via multimodal few-shot learning (Tsimpoukelli et al., 2021). Second, the method we prompted GPT-3 to propose rules only used captions from one support class as a time, performing a comparison *within* captions of that class (and repeating this process independently for the other class). However, it is very likely that adapting the prompt to allow for comparison *across* captions from multiple classes would also be beneficial, as that could allow the language model to better reason about the relevant distinction for classification (Williams and Lombrozo, 2010; Edwards et al., 2019). Finally, given the generality of the proposed approach, in future work we hope to test this model more extensively against other kinds of concepts such as ones from Bongard Problems (Bongard, 1970; Nie et al., 2020), to determine whether natural language rules can explain other kinds of human-like generalization behavior.

## References


Samuel Acquaviva, Yewen Pu, Marta Kryven, Catherine Wong, Gabrielle E Ecanow, Maxwell Nye, Theodoros Sechopoulos, Michael Henry Tessler, and Joshua B Tenenbaum. 2021. Communicating natural programs to humans and machines. *arXiv preprint arXiv:2106.07824*.

Alex Andonian, Camilo Fosco, Mathew Monfort, Allen Lee, Rogerio Feris, Carl Vondrick, and Aude Oliva. 2020. We have so much in common: Modeling semantic relational set abstractions in videos. In *European Conference on Computer Vision*, pages 18–34. Springer.

Jacob Andreas, Dan Klein, and Sergey Levine. 2017. Learning with latent language. *arXiv preprint arXiv:1711.00482*.

M Bongard. 1970. Pattern recognition.

Neil Bramley, Anselm Rothe, Josh Tenenbaum, Fei Xu, and Todd Gureckis. 2018. Grounding compositional hypothesis generation in specific instances. In *Proceedings of the 40th annual conference of the cognitive science society*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Alexandra Carstensen and Michael C Frank. 2021. Do graded representations support abstract thought? *Current Opinion in Behavioral Sciences*, 37:90–97.

Sahil Chopra, Michael Henry Tessler, and Noah D Goodman. 2019. The first crank of the cultural ratchet: Learning and transmitting concepts through language. In *CogSci*, pages 226–232.

Brian J Edwards, Joseph J Williams, Dedre Gentner, and Tania Lombrozo. 2019. Explanation recruits comparison in a category-learning task. *Cognition*, 185:21–38.

Kevin Ellis, Catherine Wong, Maxwell Nye, Mathias Sable-Meyer, Luc Cary, Lucas Morales, Luke Hewitt, Armando Solar-Lezama, and Joshua B Tenenbaum. 2020. Dreamcoder: Growing generalizable, interpretable knowledge with wake-sleep bayesian program learning. *arXiv preprint arXiv:2006.08381*.

Gabriel Goh, Nick Cammarata, Chelsea Voss, Shan Carter, Michael Petrov, Ludwig Schubert, Alec Radford, and Chris Olah. 2021. Multimodal neurons in artificial neural networks. *Distill*, 6(3):e30.

Noah D Goodman, Joshua B Tenenbaum, Jacob Feldman, and Thomas L Griffiths. 2008. A rational analysis of rule-based concept learning. *Cognitive science*, 32(1):108–154.

Evan Hernandez, Sarah Schwettmann, David Bau, Teona Bagashvili, Antonio Torralba, and Jacob Andreas. 2022. Natural language descriptions of deep visual features. *arXiv preprint arXiv:2201.11114*.

Alexander Kuhnle and Ann Copestake. 2017. Shapeworld-a new test methodology for multimodal language understanding. *arXiv preprint arXiv:1704.04517*.

Brenden M Lake, Ruslan Salakhutdinov, and Joshua B Tenenbaum. 2015. Human-level concept learning through probabilistic program induction. *Science*, 350(6266):1332–1338.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. *arXiv preprint arXiv:2201.12086*.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer.

Jesse Mu, Percy Liang, and Noah Goodman. 2020. Shaping visual representations with language for few-shot classification. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4823–4830.

Weili Nie, Zhiding Yu, Lei Mao, Ankit B Patel, Yuke Zhu, and Anima Anandkumar. 2020. Bongard-logo: A new benchmark for human-level concept learning and reasoning. *Advances in Neural Information Processing Systems*, 33:16468–16480.

Robert M Nosofsky, Thomas J Palmeri, and Stephen C McKinley. 1994. Rule-plus-exception model of classification learning. *Psychological review*, 101(1):53.

Steven T Piantadosi and Robert A Jacobs. 2016. Four problems solved by the probabilistic language of thought. *Current Directions in Psychological Science*, 25(1):54–59.

Steven T Piantadosi, Joshua B Tenenbaum, and Noah D Goodman. 2016. The logical primitives of thought: Empirical foundations for compositional cognitive models. *Psychological review*, 123(4):392.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR.

Laura Schulz. 2012. Finding new facts; thinking new thoughts. *Advances in child development and behavior*, 43:269–294.

Jake Snell, Kevin Swersky, and Richard Zemel. 2017. Prototypical networks for few-shot learning. *Advances in neural information processing systems*, 30.

Maria Tsimpoukelli, Jacob Menick, Serkan Cabi, SM Eslami, Oriol Vinyals, and Felix Hill. 2021. Multimodal few-shot learning with frozen language models. *Advances in Neural Information Processing Systems*, 34.

Joseph J Williams and Tania Lombrozo. 2010. The role of explanation in discovery and generalization: Evidence from category learning. *Cognitive science*, 34(5):776–806.

Andy Zeng, Adrian Wong, Stefan Welker, Krzysztof Choromanski, Federico Tombari, Aveek Purohit, Michael Ryoo, Vikas Sindhwani, Johnny Lee, Vincent Vanhoucke, et al. 2022. Socratic models: Composing zero-shot multimodal reasoning with language. *arXiv preprint arXiv:2204.00598*.

# A Appendix

## A.1 ShapeWorld

The following prompt was used to pass the individual captions into GPT-3 to generate rules:

```
Task: given a list of
descriptions, predict what they
have in common
List: a bowl of apples | a
photograph of some cherries |
a person eating a banana |
Rule: fruit
List: people on a soccer field |
a cricket game | a basketball |
Rule: sports that involve a ball
List: {caption_1} | {caption_2}
| ... | {caption_n} |
Rule:
```

For all concepts we use the `text-davinci-001` version of GPT-3 with a temperature setting of 0.7.

## A.2 Abstract Rules

The adapted prompt passed into GPT-3 for both rules in the Abstract Rules simulations was as follows:

```
Task: given a list of captions,
predict what they have in common
List: a red apple in a bowl | a
child in a crib | coffee in a mug
|
Rule: things inside of another
object
List: a flock of seagulls | a
swarm of bees | a collection of
books
Rule: things in a collection
List: {caption_1} | {caption_2}
| ... | {caption_n} |
Rule:
```

For all concepts we use the `text-davinci-001` version of GPT-3

with a higher temperature setting of 1.0 to encourage different rules for each class.