Extended Abstract: Concept Acquisition Through Meta-Learning

Erin Grant[12], **Chelsea Finn**[12], **Joshua Peterson**[13], **Joshua Abbott**[13],
Sergey Levine[12], **Trevor Darrell**[12], **Thomas Griffiths**[13]

[1] Berkeley AI Research (BAIR), University of California, Berkeley
[2] Department of Electrical Engineering & Computer Sciences, University of California, Berkeley
[3] Department of Psychology, University of California, Berkeley
{eringrant,cbfinn,svlevine,trevor}@eecs.berkeley.edu
{jpeterson,joshua.abbott,tom_griffiths}@berkeley.edu

Abstract

We propose a novel method for few-shot learning of visual concepts from only a small number of positive examples. Our experiments on a large-scale visual concept dataset confirm that our gradient-based meta-learning method can learn new visual concepts strictly from positive examples, akin to how humans learn new concepts. Furthermore, we compare our method with human performance on a classic concept-learning task, showing that both are similarly impacted by the underlying taxonomic structure of the visual concept dataset.

1 Introduction

Neural networks have achieved remarkable success in a variety of complex learning tasks such as speech recognition, object detection, and machine translation (*e.g.,* He et al., 2017; Wu et al., 2016; Xiong et al., 2017). However, these models require a prohibitively large number of labeled examples for good performance. In contrast, humans can learn efficiently from even a small number of examples. To attain human-level intelligence, our artificial agents must also be able to learn from small amounts of data. Recent work in *few-shot* learning has therefore turned to developing methods that work with a small number of examples per target class (*e.g.,* Ravi & Larochelle, 2017; Vinyals et al., 2016; Wang & Hebert, 2016; Finn et al., 2017; Snell et al., 2017).

Yet, even the small data setting as explored in the few-shot classification literature is very different from the real-world settings in which humans learn. People are able to make sharp inferences about conceptual boundaries from only a small number of positive examples of a concept (Feldman, 1997; Lake et al., 2015). Moreover, their inferences benefit from prior knowledge about the relational structure between concepts (Callanan, 1989; Xu & Tenenbaum, 2007). These key features of a naturalistic learning scheme – one-class learning and sensitivity to structured prior knowledge – have not been explored fully in the context of neural network models that learn from small amounts of data. Closing this gap is necessary in order to develop scalable learning systems capable of human-level decision making in the real world.

In this paper, we develop a novel training procedure which enables a standard neural network model to succeed at generalizing concepts in this naturalistic setting, outperforming previous methods for few-shot learning applied to the same task. To the best of our knowledge, no prior deep few-shot method has been reported to handle such a naturalistic concept learning task.

2 Preliminaries

Model-agnostic meta-learning (MAML) (Finn et al., 2017) formulates meta-learning as estimating the parameters $\boldsymbol{\theta}$ of a model so that when one or a few batch gradient descent steps are taken from the initialization at $\boldsymbol{\theta}$ on the training data $\left(\mathbf{X}_{\text{trn}}^{(j)}, \mathbf{y}_{\text{trn}}^{(j)}\right)$, the updated model has good generalization

performance on that task's validation set, $\left(\mathbf{X}_{\text{val}}^{(j)}, \mathbf{y}_{\text{val}}^{(j)}\right)$. In particular, the MAML objective for a model that performs maximum likelihood estimation is

$$\mathcal{L}(\boldsymbol{\theta}, \mathbf{X}, \mathbf{y}) = -\sum_{j} \log p\big(\mathbf{y}_{\text{val}}^{(j)} \mid \mathbf{X}_{\text{val}}^{(j)}, \underbrace{\boldsymbol{\theta} + \alpha \nabla_{\boldsymbol{\theta}} \log p\left(\mathbf{y}_{\text{trn}}^{(j)} \mid \mathbf{X}_{\text{trn}}^{(j)}, \boldsymbol{\theta}\right)}_{\boldsymbol{\phi}^{(j)}}\big) \tag{1}$$

where we use $\boldsymbol{\phi}^{(j)}$ to denote the updated parameters after taking a gradient step with step size α on the likelihood associated with the task $\mathcal{T}^{(j)}$. At meta-test time, a new task from the meta-test set is presented to the model for few-shot adaptation (*i.e.,* batch gradient descent with $\left(\mathbf{X}_{\text{trn}}^{(j)}, \mathbf{y}_{\text{trn}}^{(j)}\right)$) and computation of test-time performance metrics. (*e.g.,* accuracy on $\left(\mathbf{X}_{\text{val}}^{(j)}, \mathbf{y}_{\text{val}}^{(j)}\right)$).

3 CAML: Concept acquisition through meta-learning

MAML is agnostic to the architecture of the model as well as the loss function. We also make a novel observation that MAML is agnostic to the composition of the training samples $\left(\mathbf{X}_{\text{trn}}^{(j)}, \mathbf{y}_{\text{trn}}^{(j)}\right)$ and validation samples $\left(\mathbf{X}_{\text{val}}^{(j)}, \mathbf{y}_{\text{val}}^{(j)}\right)$. That is, unlike standard supervised learning methods, MAML does not assume that the training and validation samples come from the same distribution, though it does assume that the meta-training tasks and meta-test tasks come from the same *task* distribution $p(\mathcal{T})$. We take advantage of this property in order to define a positive-example variant of MAML, which we term *CAML*, to be used for few-shot concept learning from only positive examples. At meta-training time, each of the tasks $\mathcal{T}^{(j)}$ only includes positively labeled examples $\mathbf{X}_{\text{trn}}^{(j)}$, while the validation set $\mathbf{X}_{\text{val}}^{(j)}$ includes both positive and negative examples, with the negatives sampled uniformly at random from other categories.

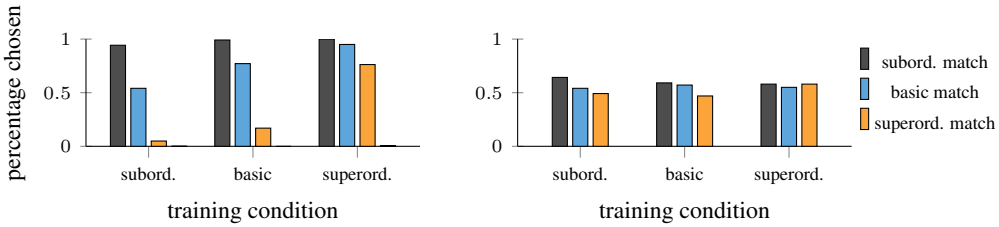
In the meta-level objective in Equation (1), the training examples in the inner gradient computation are strictly positive examples (*i.e.,* $\mathbf{y}_{\text{trn}}^{(j)} = 1$) of a particular concept j, whereas validation examples in the outer gradient computation include both positives and negatives (*i.e.,* $\mathbf{y}_{\text{val}}^{(j)} \in \{0, 1\}$). Meta-training proceeds as follows: A concept index j is sampled from the meta-training set. Then, for K-shot learning, $2K$ positive examples of the concept and K negatives are sampled. $\boldsymbol{\phi}^{(j)}$ is computed using K of the positives, and then meta-gradient is computed using the remaining examples. Finally, the meta-parameters θ are updated by averaging the meta-gradient over a meta-batch of concepts. At meta-test time, the model with trained parameters θ is presented with K positive examples from a new concept in the meta-test set; the model computes $\boldsymbol{\phi}$ and is evaluated on its ability to distinguish new examples of that concept from negatives of held-out concepts.

By training the model to successfully discriminate the members of a category from all other inputs in the validation set, the meta-learning procedure must estimate model parameters that can estimate a decision boundary from only positive examples.

4 Humanlike concept learning

Humans are notable in their ability to identify members of a concept from only a few positive examples, and it is hypothesized that they make use of an inductive bias for taxonomically structured concepts in order to do so (*e.g.,* Tenenbaum, 1999). In this section, we investigate whether CAML is able to learn, via a concept learning meta-learning procedure, a similarly appropriate bias in order to succeed on a concept generalization test that requires sensitivity to an underlying taxonomy of object kinds. To this end, we present the results of a human evaluation that studies behavior on a concept learning task in which we manipulate the taxonomic relationship between the few positive examples of a concept. Following previous work on Bayesian concept learning (Xu & Tenenbaum, 2007; Abbott et al., 2012; Jia et al., 2013), we define a task to investigate how people learn concepts at varying levels of abstraction from only a few positive examples. As stimuli, we constructed 8 image taxonomies, each including a subordinate, a basic, and a superordinate level, following the taxonomies of Rosch et al. (1976). If the exact subordinate node from Rosch et al. (1976) is not found in ImageNet, we find a close semantic match via the WordNet (Fellbaum, 1998) taxonomy.

In our task, a participant observed 5 images of a single concept, sampled from one of the three levels of taxonomic abstraction. For instance: in a subordinate training condition, the examples could be all Dalmatians; in a basic-level training condition, all dogs, or in a superordinate training condition, all animals. To test generalization behavior, participants were then given a test array of 24 images



(a) Behavioral results from our human study.

(b) Our model results on the generalization task.

Figure 1: A comparison of human behavioral data and model behavior on the generalization task of Section 4. The horizontal axis identifies the training condition (*i.e.,* the level of taxonomic abstraction from which the few-shot examples are taken). The vertical axis identifies the proportion of images each of type of match (bar color) chosen from the test array.

and were asked to pick which images also belonged to the learned concept. The test array comprised 2 subordinate matches (*e.g.,* other Dalmatians), 2 basic-level matches (*e.g.,* other breeds of dog), 4 superordinate matches (*e.g.,* other animals), and 16 out-of-domain items (*e.g.,* inanimate objects), following Xu & Tenenbaum (2007). We recruited 900 unique participants from Amazon Mechanical Turk to each complete 8 trials as described above, one randomly sampled for each of the superordinate categories. The test sets was fixed within a superordinate category. Participants were paid $0.40 each.

Figure 1a presents the results of the behavioral experiment for each of the three taxonomic levels. As expected on the basis of previous work, there is an exponentially decreasing generalization gradient as the level of taxonomic abstraction of the test matches (color in the plot) increases. However, this effect diminishes as the intra-class variation of the few-shot examples increases: Moving from the *subordinate* condition to the *basic* condition increases the number of basic-level matches chosen from the test set. The condition in which there is greatest intra-class variation – the superordinate condition – exhibits only a small generalization gradient.

The generalization gradient observed in humans is also exhibited by our model in Figure 1b: When the few-shot examples are taken from a basic-level category (the *basic* condition; *e.g.,* different breeds of dog) as opposed to a subordinate category (the *subord.* condition; *e.g.,* Dalmatians), the model generalizes to more basic-level matches (*e.g.,* different dog breeds) from the test array. In the plot, this can be seen by comparing the ratio of subordinate generalization (**black** column) to basic-level generalization (**blue** column) within each training condition (*i.e.,* the gap between the **black** and **blue** bars is diminished in the *basic* condition *vs.* the *subord.* condition). Furthermore, when the few-shot examples are taken from a superordinate category (*superord.* condition), both the model and humans are equally likely to pick subordinate, basic-level, or superordinate matches from the test array. In Figure 1a, this can be seen as the generalization to all levels of the taxonomy (**black**, **blue**, and **yellow** bars) being close to equal.

Lastly, correlation between human and model judgments shows that a model trained on a conceptual hierarchy provides a better fit (Pearson's r of 0.77) than one trained on a random hierarchy, in which leaf nodes are randomly conjoined and thus there is no underlying hierarchical structure (Pearson's r of 0.68). This is not a problem of underfitting in the model trained on a random hierarchy, as both models achieve similar accuracy on a non-hierarchical (*i.e.,* classic ImageNet classification) test set: 63.8% (hierarchical) *vs.* 62.9% (random).

5 Discussion

We presented a method for few-shot concept learning that combines ideas from human learning and recent advances in deep meta-learning. Central to our approach is the idea that concepts can be learned from only a few positive examples: When humans see an object from a new category, they can quickly infer which other objects belong to that same category, without explicitly being presented with negative examples. We demonstrate that, through a novel training procedure that uses positive training examples and mixed positive and negative validation examples at meta-training time, we can learn from only positive examples at meta-test time using a *discriminative* gradient-based

meta-learning approach. To the best of our knowledge, our method is the first deep few-shot learning method that learns concepts strictly from positive examples.

References

Joshua T Abbott, Joseph L Austerweil, and Thomas L Griffiths. Constructing a hypothesis space from the web for large-scale Bayesian word learning. In *Proceedings of the 34th Annual Meeting of the Cognitive Science Society*, 2012.

Maureen A Callanan. Development of object categories and inclusion relations: Preschoolers' hypotheses about word meanings. *Developmental Psychology*, 25(2):207, 1989.

Jacob Feldman. The structure of perceptual categories. *Journal of Mathematical Psychology*, 41(2): 145 – 170, 1997.

Christiane Fellbaum. *WordNet*. Wiley Online Library, 1998.

Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *Proceedings of the 34th International Conference on Machine Learning*, 2017.

Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask R-CNN. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. IEEE, 2017.

Yangqing Jia, Joshua T Abbott, Joseph L Austerweil, Thomas Griffiths, and Trevor Darrell. Visual concept learning: Combining machine vision and bayesian generalization on concept hierarchies. In *Advances in Neural Information Processing Systems 26*, pp. 1842–1850, 2013.

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

Sachin Ravi and Hugo Larochelle. Optimization as a model for few-shot learning. In *Proceedings of the 5th International Conference on Learning Representations (ICLR)*, 2017.

Eleanor Rosch, Carolyn B Mervis, Wayne D Gray, David M Johnson, and Penny Boyes-Braem. Basic objects in natural categories. *Cognitive Psychology*, 8(3):382–439, 1976.

Jake Snell, Kevin Swersky, and Richard S Zemel. Prototypical networks for few-shot learning. In *Advances in Neural Information Processing Systems 30*, 2017.

Joshua B Tenenbaum. *A Bayesian framework for concept learning*. PhD thesis, Massachusetts Institute of Technology, 1999.

Oriol Vinyals, Charles Blundell, Tim Lillicrap, Daan Wierstra, et al. Matching networks for one shot learning. In *Advances in Neural Information Processing Systems 29*, pp. 3630–3638, 2016.

Yu-Xiong Wang and Martial Hebert. Learning to learn: Model regression networks for easy small sample learning. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 616–634. Springer, 2016.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*, 2016.

Caiming Xiong, Victor Zhong, and Richard Socher. Dynamic coattention networks for question answering. In *Proceedings of the 5th International Conference on Learning Representations (ICLR)*, 2017.

Fei Xu and Joshua B Tenenbaum. Word learning as Bayesian inference. *Psychological Review*, 114 (2):245–272, 2007.