The Omniglot Challenge: A 3-Year Progress Report

Brenden M. Lake¹, Ruslan Salakhutdinov², and Joshua B. Tenenbaum³
¹Department of Psychology and Center for Data Science, New York University
²Machine Learning Department, Carnegie Mellon University
³Department of Brain and Cognitive Sciences and Center for Brains Minds and Machines, MIT

Abstract

Three years ago, we released the Omniglot dataset for developing more human-like learning algorithms. Omniglot is a one-shot learning challenge, inspired by how people can learn a new concept from just one or a few examples. Along with the dataset, we proposed a suite of five challenge tasks and a computational model based on probabilistic program induction that addresses them. The computational model, although powerful, was not meant to be the final word on Omniglot; we hoped that the machine learning community would both build on our work and develop novel approaches to tackling the challenge. In the time since, we have been pleased to see the wide adoption of Omniglot and notable technical progress. There has been genuine progress on one-shot classification, but it has been difficult to measure since researchers have adopted different splits and training procedures that make the task easier. The other four tasks, while essential components of human conceptual understanding, have received considerably less attention. We review the progress so far and conclude that neural networks are still far from human-like concept learning on Omniglot, a challenge that requires performing all of the tasks with a single model. We also discuss new tasks to stimulate further progress.

Introduction

Three years ago, we released the Omniglot dataset of handwritten characters from 50 different alphabets [Lake et al., 2015]. The dataset was developed to study how humans and machines perform one-shot learning – the ability to learn a new concept from just a single example. The domain of handwritten characters provides a large set of novel concepts that people learn and use in the real world. Compared to other common concepts, handwritten characters are simple and tractable enough to hope that machines, in the near future, will see most of the structure in the images that people do. For these reasons, Omniglot is an ideal testbed for developing more human-like learning algorithms, and it was released as a challenge to the machine learning, artificial intelligence (AI), and cognitive science communities.

In this paper, we review the progress made since Omniglot’s release. Our review is organized through the lens of the dataset itself, since datasets have been instrumental in driving progress in AI. New larger datasets contributed to the resurgence of interest in neural networks, such as the ImageNet dataset for objection recognition that provides 1,000 classes with about 1,200 examples each [Deng et al., 2009] [Krizhevsky et al., 2012] and the Atari benchmark that typically provides 900 hours of experience playing each game [Bellemare et al., 2013] [Mnih et al., 2015]. These datasets opened important new lines of work, but they offer far more experience than human learners require. People can learn a new concept from just one or a handful of examples, and then use this concept for a range of tasks beyond recognition (Fig. 1). Similarly, people can learn a new Atari game in minutes rather than hundreds of hours, and then generalize to game variants beyond those that were trained [Lake et al., 2017]. Given the wide gap between human and machine learning and the trend toward unrealistically large datasets, a new benchmark was needed to challenge machines to learn concepts more like people do.

The Omniglot challenge is to build a single model that can perform five concept learning tasks at a human level (Fig. 1). In the same paper, we introduced a framework called Bayesian Program Learning (BPL) that represents concepts as probabilistic programs and utilizes three key ingredients – compositionality, causality, and learning to learn – to learn programs from just one or a few examples [Lake et al., 2015]. Programs allow concepts to be built “compositionally” from simpler primitives, while capturing real “causal” structure about how the data was formed. The model “learns to learn” by using experience with related concepts to accelerate the learning of new concepts, through the formation of priors over programs and by re-using sub-programs to build new concepts. Finally, probabilistic modeling handles noise and facilitates creative generalizations. BPL produces human-like behavior on all five tasks, and lesion analyses confirm that each of the three ingredients contribute to the model’s success. But we did not see our work as the final word on Omniglot. We hoped that the machine learning, AI, and cognitive science communities would build on our work to develop more neurally-grounded learning models that address the Omniglot challenge. In fact, we anticipated that new models could meet the challenge by incorporating compositionality, causality, and learning to learn.

We have been pleased to see that the Omniglot dataset has been widely adopted and that the challenge has been
well-received by the community. There has been genuine progress on one-shot classification, but it has been difficult to gauge since researchers have adopted different splits and training procedures that make the task easier. The other four tasks have received less attention, and critically, no new algorithm has attempted to perform all of the tasks together. Human-level understanding requires developing a single model that can do all of these tasks, acquiring conceptual representations that support fast and flexible, task-general learning. We conjectured that compositionality and causality are essential to this capability (Lake et al., 2017), yet most new approaches aim to “learn from scratch,” utilizing learning to learn in ingenious new ways while incorporating compositionality and causality only to the extent that they can be learned from images. People never learn anything from scratch in this way, and we believe that incorporating all three principles is the most promising route to achieving human-level concept learning. We are re-releasing the Omniglot dataset with the drawing data in a new format to facilitate research in this direction, and we encourage future work on algorithms that can tackle the full suite of tasks together.

One-shot classification

One-shot classification was evaluated in Lake et al. (2015) through a series of 20-way within-alphabet classification problems. Two classification trials are illustrated in Fig. 1A. A single image of a new character is presented, and the goal is to select another example of that character from a set of 20 images produced by a typical drawer of that alphabet. Human participants are skilled one-shot
networks specifically for one-shot classification (Vinyals et al., 2016). Meta-learning algorithms can train discriminative neural networks to learn by training directly on many randomly generated one-shot classification problems (known as episodes) from the background set. They do not incorporate compositional or causal structure of how characters are formed, beyond what is learned implicitly through tens of thousands of episodes of character discrimination. Unfortunately it has been difficult to compare performance with the original results, since most meta-learning algorithms were evaluated on alternative variants of the classification challenge. Vinyals et al. (2016) introduced a one-shot classification task that requires discriminating characters from different Omniglot alphabets (between-alphabet classification), rather than the more challenging task of discriminating characters from within the same alphabet (within-alphabet classification; Fig. 1A). This setup also used a different split with more background characters and applied class augmentation to further increase the number of background characters four-fold through rotations. On this augmented between-alphabets problem with effectively 4,800 background characters (Table 1, column 4), meta-learning approaches have performed well (Table 1), achieving 6.2% using matching networks (Vinyals et al., 2016), 4.0% using prototypical networks (Snell et al., 2017), and 4.2% using model-agnostic meta-learning (MAML; Finn et al., 2017).

To compare our results with these recent methods, we retrained and evaluated a top-performing method, prototypical networks, on the original one-shot classification task released with Omniglot. Note that for one-shot classification, matching networks and prototypical networks are equivalent up to the choice of distance metric, and we modified the implementation from Snell et al. (2017) for within-alphabet classification. The neural network performs with an error rate of 13.7% (Table 1, column 1), which is substantially worse than the 4.0% error for the between-alphabet problem. Using class augmentation to expand the effective number of characters four-fold within each alphabet, the network achieves 6.0% error. Even with considerable augmentation, this is still substantially higher than BPL, which like children can learn to learn from quite limited amounts of background experience (Smith et al., 2002), perhaps familiarity with only one or a few alphabets along with related drawing experience. In fact, Omniglot was released with two more challenging “minimal” splits containing only five

| # alphabets | # classes | Original | Augmented |
|-------------|----------|----------|-----------|
|             |          | Within | Within | Between |
|             |          | alphabet (minimal) | alphabet | alphabet |
| Humans      |          | 30     | 5      | 30      |
| BPL         |          | 146    | 3,856  | 4,800   |
| Simple ConvNet | 13.5% | 23.2% |
| Siamese Net |          | 8.0%*  |        |         |
| Prototypical Net | 13.7% | 30.1% |
| Matching Net |          | 6.0%*  | 4.0%    |
| MAML        |          | 4.2%   |        |         |
| Graph Net   |          | 2.6%   |        |         |
| ARC         |          | 1.5%*  | 2.5%*  |
| RCN         |          | 7.3%   |        |         |
| VHE         |          | 18.7%  | 4.8%   |

Table 1: One-shot classification error rate for both within-alphabet classification (Lake et al., 2015) and between-alphabet classification (Vinyals et al., 2016), either with the “Original” background set or with an “Augmented” set that uses more alphabets (# alphabets) and character classes for learning to learn (# classes). The best results for each problem formulation are bolded, and the results for “minimal” setting are the average of two different splits. * results used additional data augmentation beyond class expansion.

classifiers, achieving an error rate of 4.5%, although this is an upper bound since they responded quickly and were not incentivized for performance. The goal for computational models is to perform similarly or better.

Models learn to learn from a set of 964 background characters spanning 30 alphabets, including both images and drawing demonstrations for learning general domain knowledge. These characters and alphabets are not used during the subsequent evaluation problems, which provide only images. BPL performs comparably to people, achieving an error rate of 3.3% on one-shot classification (Table 1, column 1). Lake et al. (2015) also trained a simple convolutional neural network (ConvNet) to perform the same task, achieving a one-shot error rate of 13.5% by using the features learned on the background set through 964-way image classification. The most successful neural network at the time was a deep Siamese ConvNet that achieves an 8.0% error (Koch et al., 2015), still about twice as high as people and BPL and requiring substantial data augmentation. As both ConvNets are discriminative models, they were not applicable to the other tasks beyond classification, an ability critical to how the Omniglot challenge was formulated.

In the time since Omniglot was released, the machine learning community has embraced the one-shot classification challenge. Table 1 shows a summary of notable results. Among the most successful new approaches, meta-learning algorithms can train discriminative neural networks specifically for one-shot classification (Vinyals et al., 2016; Snell et al., 2017; Finn et al., 2017). Rather than training on a single auxiliary problem (e.g., 964-way classification), meta-learning networks utilize learning to learn by training directly on many randomly generated one-shot classification problems (known as episodes) from the background set. They do not incorporate compositional or causal structure of how characters are formed, beyond what is learned implicitly through tens of thousands of episodes of character discrimination. Unfortunately it has been difficult to compare performance with the original results, since most meta-learning algorithms were evaluated on alternative variants of the classification challenge. Vinyals et al. (2016) introduced a one-shot classification task that requires discriminating characters from different Omniglot alphabets (between-alphabet classification), rather than the more challenging task of discriminating characters from within the same alphabet (within-alphabet classification; Fig. 1A). This setup also used a different split with more background characters and applied class augmentation to further increase the number of background characters four-fold through rotations. On this augmented between-alphabets problem with effectively 4,800 background characters (Table 1, column 4), meta-learning approaches have performed well (Table 1), achieving 6.2% using matching networks (Vinyals et al., 2016), 4.0% using prototypical networks (Snell et al., 2017), and 4.2% using model-agnostic meta-learning (MAML; Finn et al., 2017).

To compare our results with these recent methods, we retrained and evaluated a top-performing method, prototypical networks, on the original one-shot classification task released with Omniglot. Note that for one-shot classification, matching networks and prototypical networks are equivalent up to the choice of distance metric, and we modified the implementation from Snell et al. (2017) for within-alphabet classification. The neural network performs with an error rate of 13.7% (Table 1, column 1), which is substantially worse than the 4.0% error for the between-alphabet problem. Using class augmentation to expand the effective number of characters four-fold within each alphabet, the network achieves 6.0% error. Even with considerable augmentation, this is still substantially higher than BPL, which like children can learn to learn from quite limited amounts of background experience (Smith et al., 2002), perhaps familiarity with only one or a few alphabets along with related drawing experience. In fact, Omniglot was released with two more challenging “minimal” splits containing only five

---

1The code’s default parameters use 60-way classification for training and 20-way classification for evaluation. The default was used for the augmented within-alphabet task, but 20-way training was used for the original task since there are not enough characters within alphabets. Background alphabets with less than the required n-way classes were excluded during training. The number of training epochs was determined by the code’s default early stopping train/validation procedure, except for the five alphabet case where it was trained for 200 fixed epochs.
background alphabets (Table 1 column 2), and BPL still performed well in this setting (4.3% and 4.0% errors). In contrast, the metalearner shows substantial degradation with minimal background training (30.8% and 29.3% errors). Meta-learning with neural networks remains a promising approach to one-shot classification, but none of the current methods solve the challenge.

There are two other noteworthy recent architectures for constructing context-sensitive image representations. Meta-learning can be combined with graph neural networks to learn embeddings that are sensitive to the other items in the episode, achieving 2.6% error on the betweenalphabets classification task with four-fold class augmentation as before (Graph Net; Garcia and Bruna 2018). Attentive recurrent comparators (ARCs) use a learned attention mechanism to make repeated targeted comparisons between two images (Shyam et al., 2017), achieving strong results (2.5% error between-alphabets and 1.5% error within-alphabets) but with four-fold class augmentation plus several additional types of data augmentation through random scaling, rotations, shearing, translations, etc. These more complex methods are likely to be especially prone to overfitting without substantial class and data augmentation, as has been noted for the ARC network (Shyam, 2018). As researchers interested in human-level learning in AI systems, we want to develop machine learning algorithms that require even less data, not more, than 30 alphabets and 964 characters in the lake et al., 2015 background set. For the goal of reaching humanlevel performance with minimal training, given a rough estimate of what minimal means for people, there is a need to explore settings with fewer training examples per class and fewer background classes for learning to learn.

Another serious limitation of discriminative methods is that they only perform the task they were trained for. Human conceptual representations are far more flexible and task-general, and thus discriminative learning is not a plausible account, at least not on its own. Generative models capture more causal structure about images and can perform multiple tasks, and deep generative models have recently been applied to one-shot classification, including the neural statistician (12% error betweenalphabets; Edwards and Storkey, 2016) and recursive cortical networks (RCNs), a more explicitly compositional architecture (7.3% error within-alphabets; George et al., 2017). The variational homoencoder (VHE; Hewitt et al., 2018) performs well on the Vinyals et al., 2016 betweenalphabet classification task with 4.8% error, but performs much worse on the original within-alphabet classification problem (18.7% error), which is a harder task with less background training available. Deep generative models have made important progress but they have not solved the one-shot classification problem either; they have only the barest form of causality and do not understand how real-world characters are generated, a point we discuss further in the next section.

Generating new exemplars

The Omniglot challenge is about more than classification; when a human learner acquires a new concept, the representation endows a realm of capabilities beyond mere recognition (Murphy, 2002). Lake et al. (2015) studied one-shot exemplar generation – how people and models generate new examples given just a single example of a new concept. Human participants and computational models were compared through visual Turing tests, in which human judges attempt to determine which drawings were produced by humans and which by machines (Fig. 1). Models were evaluated using the identification (ID) level of the judges, where ideal model performance was an ID level of 50%. BPL can generate new examples that can pass for human, achieving an average 52% ID level where only 3 of 48 judges were reliably above chance.

There has been substantial interest in developing generative models on the Omniglot dataset, including new neural network approaches that build on variational autoencoders, adversarial training, and reinforcement learning. Some models can generate high-quality unconditional samples from Omniglot, but it is unclear how these approaches would produce new examples of a particular concept (Gregor et al., 2015; Esami et al., 2016; Gregor et al., 2016; Ganin et al., 2018). Other generative models have been applied to one-shot (or few-shot) learning problems, examining only generative tasks (Rezende et al., 2016) or both classification and exemplar generation (Edwards and Storkey, 2016; George et al., 2017; Hewitt et al., 2018). These approaches generate compelling new examples of a character in some cases, while in other cases they produce examples that are not especially human-like. So far, deep generative models tend to produce unarticulated strokes (Fig. 2A and B), samples with too much variation (Fig. 2A), and samples with too little variation (Fig. 2B). These machine-generated examples have not been quantitatively compared to human generated examples, but we are doubtful they would pass a visual Turing test.

Deep generative architectures could perform in more human-like ways by incorporating stronger forms of compositionality and causality. Current neural network models use only image data for background training, unlike BPL which learns to learn from images and drawing demonstrations. As a consequence, the networks learn to generate images in ways unrelated to how the data was actually produced, although some notable neural network approaches have taken more causal approaches in the past (Hinton and Nair, 2006). In contrast, people have rich causal and compositional knowledge of this and many other domains in which they can rapidly learn and use new concepts (Lake et al., 2012). BPL has rich domain knowledge too and does not try to learn everything from scratch: some of these causal and compositional components are built into the architecture, while other com-
components are learned by training on drawing demonstrations. Several recent deep generative models applied to Omniglot have taken initial steps toward incorporating causal knowledge, including using a pen or pen-like attentional window for generating characters (Gregor et al., 2015; Ganin et al., 2018). Stronger forms of compositionality and causality could be incorporated by training on the Omniglot drawing demonstrations rather than just the images. To encourage further explorations in this direction, we are re-releasing the Omniglot drawing demonstrations (trajectory data) in a more accessible format.

The drawing demonstrations can be used in other predictive tasks, such as predicting people’s motor programs for producing novel letters. BPL draws in realistic enough ways to confuse most judges in a visual Turing test of this task (Fig. 1B), although there is room for improvement since the average ID level was 59%. We believe that building generative models with genuine causal and compositional components, whether learned or built in, is key to solving the five Omniglot tasks.

Generating new concepts

In addition to generating new examples, the Omniglot challenge includes generating whole new concepts (Fig. 1D and E). To examine this productive capability, human participants were shown a few characters from a novel foreign alphabet, and they were asked to quickly generate new characters that could plausibly belong to that alphabet (Fig. 1D). BPL performs this task by placing a non-parametric prior on its programs, and judges in a visual Turing test had only a 49% ID level in discriminating human versus machine produced letters (Lake et al., 2015). This ability has been explored in several deep generative architectures but with limited success, often producing blurry and unarticulated novel characters (Rezende et al., 2016; Hewitt et al., 2018). This task remains wide open challenge for deep neural networks.

The final task examines generating new concepts without constraints (Fig. 1E). This task has received more attention and can be performed through unconditional sampling from a generative model trained on Omniglot. Many new approaches produce high-quality unconditional samples (Gregor et al., 2015; Eslami et al., 2016; Gregor et al., 2016; Ganin et al., 2018), although they have not been evaluated for their generative creativity, as opposed to merely copying characters in the training set. Nonetheless we believe this task is within reach of current neural network approaches, and the greater challenge is developing new architectures than can perform all of the tasks together.

---

*https://github.com/brendenlake/omniglot
Discussion

There are many promising new models that have advanced the state-of-the-art in one-shot learning, yet they are still far from solving the Omniglot challenge. There has been evident progress on neural architectures for one-shot classification and one-shot exemplar generation, but these algorithms do not yet solve the most difficult versions of these problems. BPL, which incorporates more compositional and causal structure than subsequent approaches, achieves a one-shot classification error rate of 4.5% on the original task, while the best neurally-grounded architecture achieves 7.3% (Table 1). On the same task with minimal background training, BPL achieves 4.2% while the best neural network results are 23.2% errors. The more creative tasks can be evaluated with visual Turing tests (Lake et al., 2015), where ideal model performance is a 50% ID level based on human judges. BPL achieves an ID level of 52% on one-shot exemplar generation (and 55% with minimal background training), 59% on parsing new examples, 49% on generating new concepts from a type, and 51% on generating new concepts without constraints. The Omniglot challenge is to achieve similar success with a single model across all of these tasks jointly.

Some of the most exciting advances in the last three years have come from using learning to learn in innovative ways. A similar sustained and creative focus on compositionality and causality will lead to substantial further advances. We have yet to see deep learning approaches that can achieve human-level of performance without explicitly making use of this structure, and we hope researchers will take up the challenge of incorporating compositionality and causality into more neurally-grounded architectures. This is a promising avenue for addressing the Omniglot challenge and for building more domain-general and more powerful human-like learning algorithms (Lake et al., 2017).

New algorithms in AI and machine learning usually aim for domain breadth at the expense of domain depth, targeting a narrow task and measuring performance across many datasets. As a representative case, matching networks were applied to three datasets (Omniglot, miniImageNet, and language modeling) but only the one task of one-shot classification (Vinyals et al., 2016). Human learners are certainly remarkable for their domain breadth, but they are equally remarkable for their domain depth – individual concepts are rich, flexible, and deep enough that people can use them for action, imagination, and explanation. Even simple concepts such as handwritten characters have this richness; indeed Holsinger (1985) famously argued that learning to recognize the characters in all the ways that people do contains most of the fundamental challenges of AI.

Omniglot is a dataset for evaluating algorithms on domain depth. Far more than a benchmark for one-shot classification, Omniglot challenges researchers to build algorithms for learning genuine concepts for handwritten characters. Further progress is needed on each task individually and especially on task-general representations that can be applied to all of them. Importantly, we see these goals as closely linked, such that progress can be made on individual tasks by learning deeper and more flexible conceptual representations.

A model that achieves human-level understanding would be capable of performing the five tasks discussed here and many more. These five representative tasks are surely an important start, yet more tasks and benchmarks would further accelerate this scientific and engineering challenge of developing more human-like learning algorithms. While we especially encourage tasks that go beyond classification, several novel and interesting classification tasks have already been contributed: Santoro et al. (2016) and Rae et al. (2016) studied sequential one-shot classification with Omniglot where the stimuli arrive sequentially and agents must remember and generalize the category labels, and Woodward and Finn (2016) studied an active learning version of the same task. Other more challenging versions of within-alphabet one-shot classification should be studied too. Additional Omniglot tasks could include filling in occluded images, understanding CAPTCHAs constructed with novel characters, classifying new characters by alphabet, or constructing new characters from verbal descriptions. Each new task offers an additional bridge between machine and human learning, with potential for human behavior and cognitive principles to inform the development of new algorithms. We are excited to see what additional progress the next three years will bring.

Acknowledgements

We are grateful to Jason Gross for his essential contributions to Omniglot, and we thank the Omniglot.com encyclopedia of writing systems for helping to make this dataset possible. We thank Kelsey Allen, Reuben Feinman, and Tammy Kwan for valuable feedback on earlier drafts, and Pranav Shyam for his helpful correspondence regarding the ARC model.

References

Bellemare, M. G., Naddaf, Y., Veness, J., and Bowling, M. (2013). The arcade learning environment: An evaluation platform for general agents. Journal of Artificial Intelligence Research, 47:253–279.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). ImageNet: A large-scale hierarchical image database. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

Edwards, H. and Storkey, A. (2016). Towards a Neural Statistician. Advances in Neural Information Processing Systems (NIPS).
