One Shot Learning in Humans: A Non-Parametric Bayesian Model

Kumar Kshitij Patel (kishinmh@cse.iitk.ac.in)
150348, Project Report PSY-789
Department of Computer Science and Engineering
IIT Kanpur, Kanpur, India

Abstract

Being capable of robust one shot learning sets the human cognition a class apart. It has been deeply studied in the past few years, especially in the field of Artificial Intelligence and cognitive science. In this paper, I present a brief review of some recent attempts to model it, using modern sophisticated probabilistic tools. I also present my own computational model of one shot learning using non parametric Bayesian modelling. It is followed by a some experiments to prove a part of this computational model.

Keywords: One Shot Learning; Bayesian Learning; Cognitive Science; Probabilistic Machine Learning; Computational Mind Model;

Introduction

Humans can learn about their environment from a very small amount of data. As a small baby, while acquiring language to when they predict a stock's behaviour accurately using few data points are all examples of sparse data learning in humans. It is something everyone does on a daily basis, yet we rarely stop to appreciate this human cognitive gem. Only in the last decade, with the re-emergence of statistical machine learning, have we realized how crucial it is for an intelligent system to do efficient and robust sparse data generalization. With the rise of data hungry deep learning models, becoming the state of the art in every other field, it has become a burning need to create algorithms that do one shot learning and do it well. Only then can the dream of social robots be realized. This problem of one shot learning, has been closely tracked in computer vision and machine learning, and it is only in the recent years that there is a revived interest in the community in trying to understand this in humans. With recent developments in probabilistic machine learning, and us trying to model everything around us using sophisticated probabilistic models there has been some attempt to model the one shot learning process too (see next).

In this paper, I review some of these methods, try combining their inferences, and present my own computational mind model of One Shot Learning. It is followed by brief experimentation to prove some of the claims, while most of them have been deferred to a later point in future. I conclude with some motivating ideas about this line of research and how can we benefit from it.

Background

Against Incremental Learning

While considering the general learning paradigm it is assumed that learning rates don't change abruptly between trials. That is one of the factors which define incremental learning. It proceeds and generally increases in proportion to the data we have. One-shot learning however has a marked increase in the learning rate, which facilitates learning from a novel data point, without having to see it multiple times. Such an abrupt and discriminative learning must have been of evolutionary advantage to humans, who had to continuously learn a lot about their surroundings and stay away from possible dangers. One obvious question is that do the general learning principles of US modulation models, offered for classical conditioning apply to one shot learning. That is, how crucial is predictive error is in one shot learning. (Greve, Cooper, Kaula, Anderson, & Henson,2017) found a strong co-relation between the two, which keeps the theory continuous and consistent for one shot learning. Another important aspect is the neural substrates involved in both these forms of learning. (Lee, ODoherty, & Shimojo,2015) showed how the brain actually "switches" between both these modes while learning. These two results at least enable us to explain one shot learning using models similar to the classic Rescorla-Wagner model (Miller, Barnet, & Grahame,1995).

Learning Categories

We have seen how normal category learning can also be reduced to a problem of support evaluation for different stimuli. In other words, updating the beliefs about different stimuli. Thinking the other way round, if we can explain sparse category learning that naturally helps explain many results in psychology and cognitive science (Ashby & Maddox,2005). In fact learning object categories might extend to concept formation in humans, or learning upper ontology which is closely related to language acquisition. The process of language acquisition has been previously modelled in a Bayesian framework; the very task of learning word meanings or categories is one shot (Xu & Tenenbaum,2007). (Kemp, Perfors, & Tenenbaum,2007) demonstrated how hierarchical Bayesian methods can be used to model the hypothesis and meta-hypothesis formation, which are crucial to category learning. The last result is pivotal to the model I propose in this paper. On a similar line of thought (Canini,2007) modelled the categorization process using hierarchical Dirichlet process.

One Shot Learning

Cognitive Science There hasn't been much work in cognitive science to model one shot learning. (Lee et al.,2015) and (Greve et al.,2017) are the only two major results which directly talk about one shot learning. Both of them implicate how predictive error is as crucial in one shot learning as it is in normal incremental learning. Latter is the first study ever to try and provide a neuro-scientific reason for one shot learning.

It verifies a long held belief that there must be some difference on the systems level between these two different kinds of learning. Besides there has been work in probabilistic modelling sparse data learning (Xu & Tenenbaum,2007), discrimination in rapid learning (Fang & Chiang,2016). Bayesian models have mostly crowded the field of cognitive science in the last few decades. They have also covered some ground in learning (Tenenbaum, Griffiths, & Kemp,2006).

Statistical Machine Learning There has been a lot of working in probabilistic machine learning in trying to model the one shot learning. The field started with the seminal work by (Fei-Fei, Fergus, & Perona,2006), establishing the first model for the same. (Salakhutdinov, Tenenbaum, & Torralba,2012) gave a hierarchical non parametric model for the same. The model proposed in this paper heavily relies on this particular research, while using results about human cognition. (Yip & Sussman,1997) Provide another level of sparsity by further presenting sparse representation models for this characterization. (Wang, Zeng, & Xu,2016) gave a neo-cortex based algorithm, which provides interesting insight into system level modelling of this process. Some other notable works are that of (Sudderth, Torralba, Freeman, & Willsky,2008) and (Maas & Kemp,2009).

Deep Learning The works in deep learning are more non-intuitive to providing any insight into the human process, owing to the specific needs they try to cater plus the lack of substantial theory on deep learning itself. Mostly, the research has been restricted to tech giants in desire of improving their customer services and outsourcing the learning, to less data hungry algorithms. Some notable works are (Vinyals, Blundell, Lillicrap, Wierstra, et al.,2016), (Bertinetto, Henriques, Valmadre, Torr, & Vedaldi,2016), (Fang & Chiang,2016), (Santoro, Bartunov, Botvinick, Wierstra, & Lillicrap,2016), (B. M. Lake, Salakhutdinov, & Tenenbaum,2015) and (B. Lake, Salakhutdinov, Gross, & Tenenbaum,2011).

Theory

Non Parametric Bayesian Learning

Unlike usual models in probabilistic machine learning, which have a fixed set if parameters and the a priori or a posterior likelihood of the model is maximized to obtain the value of these parameters, these models don't have a fixed parameter space and it can row with the time as the model learns. This kind of behaviour can be modelled in many ways but one of the simple ways is to model the data using latent variable classes, and the class membership function changes with time, with a finite possibility of an outcome entering a new class in total every time. This makes it possible for the model, to learn more and more about the data as it grows. One simple example of a non-parametric Bayesian modelling is a discrete time stochastic process called Chinese restaurant process. The problem runs as follows: Suppose there are infinite tables in a Chinese restaurant, as soon as a person enters the restaurant he has the choice of either sitting on a new table or sitting on a preexisting table. the probability that he sits on a table is $\frac{|b|}{n+\gamma-1}$, where $|b|$ is the size of the particular table. This leaves the probability of sitting on a new table is $\frac{\gamma}{n+\gamma-1}$. One crucial thing about the process is that it leaves some finite possibility for a new table when infinite tables are possible. Such a modelling is very appropriate for one shot learning, because basically a learner is developing new categories and placing objects in them all the time.

The Computational Model

Here I provide a simple model for how individuals learn in one shot to categorize novel stimuli into categories and eventually super-categories in one shot. I abstain from linking the process to the task of universal one shot learning of say concepts, because that formalization still requires more work. However, it is somewhat similar to the work done by (Kemp et al.,2007).

Model

This and the following section loosely follows the work by (Salakhutdinov et al.,2012). Let us assume our model has a two level hierarchy. Every feature or object that we see ($\{x^1, x^2, ..., x^N\}$, $x^n \in \mathcal{R}$) can be categorized into a category which further gets binned into a super category. Say there are C level-1 categories ($z^b \in \mathcal{R}^N$) is a vector which defines the category for every object i.e. $z_n^b \in \{1, 2...., C\}$) and K super-categories($z^b \in \mathcal{R}$) is a vector which gives super categories for all categories i.e. $z_c^s \in \{1, 2...., K\}$). Assume that the probability distribution of the input objects over level-1 categories is a Gaussian with a mean (μ^c) and variance($\frac{1}{\tau^c}$) suitably defined for each class. Thus the level-1 category parameters are defined by these means and variances respectively. Just like standard Gaussian we place a normal-gamma prior over these parameters. Because of this definition the super-category parameters would have to a mean and a variance for the Gaussian and a rate parameter for the gamma distribution, for every super-category. This is because the nature of the super-category distribution and the prior of the parameters of level-1 has to be same. Hence the involved equations are:

$$\mathcal{P}(x^n|z_n^b = c, \theta^1) = \mathcal{N}(x^n|\mu^c, \frac{1}{\tau^c}) \tag{1}$$

$$P(\mu^c, \tau^c|\theta^2) = \mathcal{P}(\mu^c|\tau^c, \theta^2)\mathcal{P}(\tau^c|\theta^2) \tag{2}$$

$$\mathcal{P}(\mu^c|\tau^c, \theta^2)\mathcal{P}(\tau^c|\theta^2) = \mathcal{N}(\mu^c||mu^k, \frac{1}{\nu\tau^c})\Gamma(\tau^c|\alpha^k, \frac{alpha^k}{\tau^k}) \tag{3}$$

$$\Gamma(\tau^c|\alpha^k, \frac{alpha^k}{\tau^k}) = \frac{(\frac{alpha^k}{\tau^k})^{\alpha^k}}{\Gamma(\alpha^k)}\tau^{\alpha^k - 1}exp(-\tau\frac{\alpha^k}{\tau^k}) \tag{4}$$

Because of the properties of these probability distributions the expected value of the basic level-1 parameters θ^1 are given by the corresponding level-2 parameters θ^2. The parameter α^k further controls the variability of τ^c around its mean. For

Figure 1: An instance of the image set

level-2 parameters too we need to assume a prior on mean, the rate parameter and the variance parameter(See the original paper for some missing details here). The classes are defined using the Chinese restaurant process probabilities we defined above, thus they can grow in number as more data is seen. The referred paper aptly remarks at this point *"This training regime reflects natural language acquisition, where spontaneous category labeling is frequent, almost all spontaneous labeling is at the basic level (Rosch et al., 1976) yet children's generalizations are sensitive to higher superordinate structure (Mandler, 2004), and where new basic-level categories are typically learned with high accuracy from just one or a few labeled examples."*
Now having defined the model we need to sample at every step which can be done very simply because all our probability distributions are conjugate. Moreover, being simple object inputs their dimensionality has also been restricted. The final equations for one shot learning is:

$$p(c*|x^t) = \frac{p(x^t|z*_c)p(z*_c)}{\sum_z p(x^t|z)p(z)} \tag{5}$$

The prior is given by the Chinese restaurant process while the likelihood takes the form:

$$\log p(x^t|c*) = \frac{1}{2}\log\tau* - \frac{1}{2}\tau*(x^t-\mu*)^2 + C \tag{6}$$

Thus our task is finished.

Experiments

In order to reconfirm the hypothesis on the relevance of predictive error in one shot learning and especially to consider the effects of time spacing which were not very well emphasized in the original paper (Lee et al.,2015). In that scope this section is a bit dejected from the other part of the paper which emphasizes on giving a computational model for one shot learning. It is understood that further evidence needs to be given in order to actually show a strong correlation of the model with human judgment but that is deferred to future for now.

Hypothesis

Since the experiment was close to the one conducted by (Lee et al.,2015), the hypothesis was actually the same. It was basically:

> **Higher uncertainty in the causal relationship between stimulus and outcome => Higher Learning rate => One-shot learning**

In terms of predictive error it would be quoted as:

> **Higher Predictive Error => Higher Learning rate => One-shot learning**

Design

Every experiment was divided into two phases:

1. **Learning Phase:** A sequence of fractal images followed by a positive reinforcement or a negative reinforcement (Chocolates were the reinforcement).

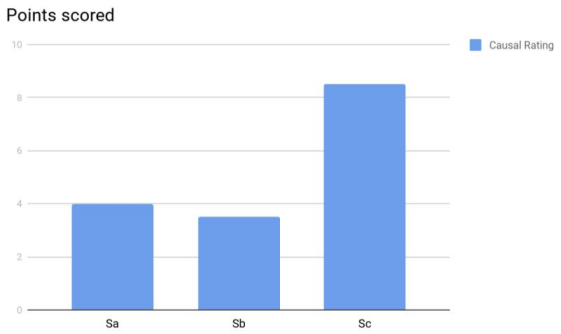
2. **Rating Phase:** Participants had to rate how much they like every image $\in (-5,5)$ and the chances of a particular image being followed by a positive reward $\in (1,10)$.

Now in the learning phase:

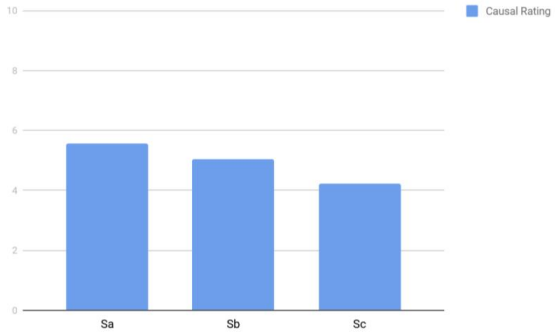
- Every learning phase for an individual had one round.

- Every round had 25 pictures. After every 5 pictures a reinforcement was presented. Every picture was displayed for 1 second. The transition period from one image to another was sampled uniformly from 1-4 seconds.

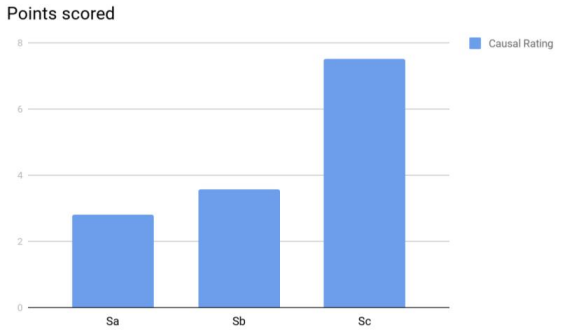
- There were two kinds of pictures non-novel and novel. The non-novel picture could have been displayed 4 teams each in which case it belonged to the set S_a, else it was pictured twice each and it belonged to S_b. Evert novel picture belonging to S_c was pictured exactly once.



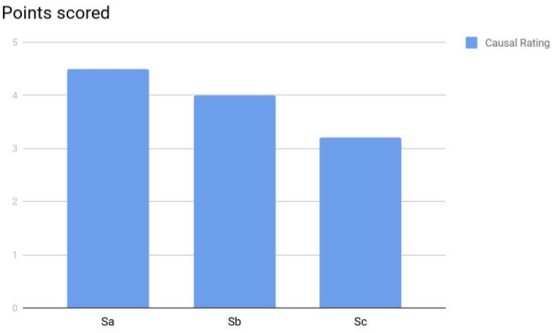
(a) Experiment 1



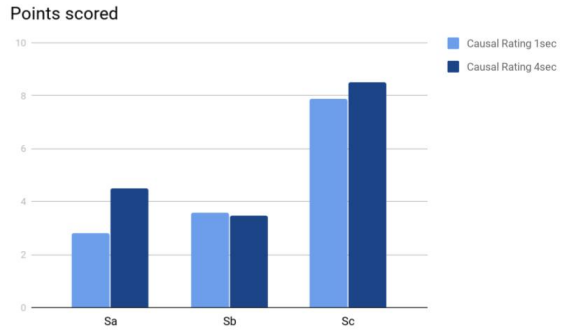
(b) Experiment 2



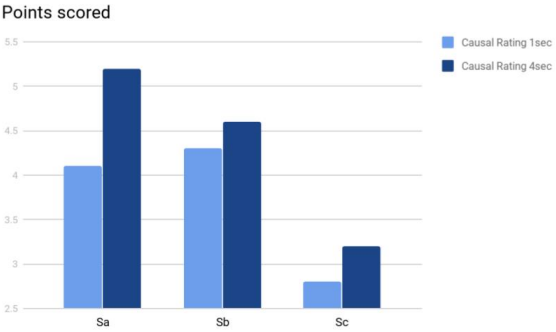
(c) Experiment 3



(d) Experiment 4



(e) Experiment 5-6



(f) Experiment 7-8

Figure 2: Results obtained across different experiments

Table 1: Summary of Experiments

Experiment Number	Type of round	Non-Novel Outcome	Novel Outcome	Time interval between images
1	I	-10	50	$\mathcal{UNF}(1,4)$
2	II	-10	50	$\mathcal{UNF}(1,4)$
3	I	10	-50	$\mathcal{UNF}(1,4)$
4	II	10	-50	$\mathcal{UNF}(1,4)$
5	I	-10	50	1
6	II	-10	50	1
7	I	-10	50	4
8	II	-10	50	4

- The novel picture was necessarily among the last two pictures among-st the 5 pictures batch. The reason for doing this is so that participants build up enough experience of the non-novel images in the earlier trials within a round.
- The reward was +50/+10 and punishments were -10/-50 across different rounds. After every five pictures the result was highlighted using numerals on the screen.
- In every round either the punishment or the reward was given just once to maintain novelty of an outcome against the other.
- There were two types of rounds. In type-1 novel stimuli was followed by a novel outcome, in type-2 non-novel stimuli was followed by a novel outcome.
- Time interval was non-randomized in one setting to test the effects. This was done to see how time given to consolidate an image and/or its result, affects the learning of the individuals.

An instance of the image set has been shown in Figure 1.

Every experiment that is set of two phases was conducted with 5 people each. There were a total of 8 experiments as summarized in Table 1.

Results

The results have been summarized in Fig. 2 and Fig. 3. respectively. As we can see in the first figure all across type II rounds the causal rating by the individuals was almost equal and in fact lower for the novel picture. Such a trend was observed because in absence of one shot learning, the non-novel pictures have a definite advantage because they are visible more often. The reason scores about the liking of images and their ratings were taken together was because I wanted to test if there was any correlation between the liking of the image and the causal rating. However, probably owing to the instructions before the experiment, no string correlation was found (Pearson's $r = 0.123$). Moreover, in all the type I rounds the causal rating is much higher for the novel stimuli group.

As per as significance of the results is considered, as evident in Fig. 3. the average difference between the points scored is much lower than Type I ratings while it is much above the Type II ratings for both experiment 1-2 and 3-4 pairs. p-Values obtained in both the cases were respectively 0.0032 and 0.0063.

Inference

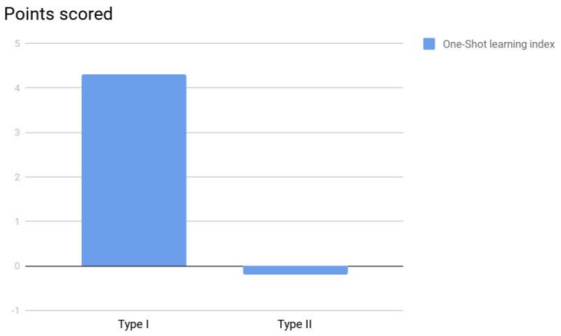
It can be inferred from the results above that there is a significantly higher rating given to the pictures in the Type 1 classes across all experiments while varying the magnitude of total reward, individual reward and the time between the images-images and images-outcomes. This means that people do rate the novel images followed by novel outcomes higher. This is equivalent to having one shot learning, because there learning rate must have increased drastically during the period, they saw this association. Another interesting effect is that change of magnitude and inter stimulus time doesn't bear any consequence to the over all results, which align in a similar direction throughout. Though the experiment our hypothesis that predictive error governs one shot learning is very clearly proved.

Discussion

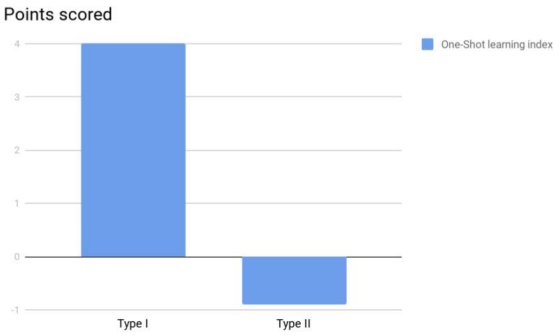
In this paper we discussed a mathematical model for categorization of objects in humans in a one shot manner. It is possible to correlate this process to general one shot learning, but that requires some more characterization and understanding of the process. As far as the experiments are concerned I need to conduct more experiments to prove that humans actually co relate with such a model of one-shot learning. This model has been shown to perform well in the case of machine learning data sets by (Salakhutdinov et al.,2012). Regarding the experiment section it has been proved that predictive error safely governs one shot learning in humans. The results are reassuring and even robust enough to small experimental aberrations. However, one crucial link is to somehow correlate the predictive error with probabilities in our model. As a whole this line of development for one shot learning seems pretty fair and promising.

Future Work

A lot of work needs to be done to establish this model as an independent and self standing model that explains one shot



(a) Experiment 1-2



(b) Experiment 3-4

Figure 3: Statistical Significance across different experiments

learning while also dealing with the surrounding concepts. Some of that is:

1. How does stimulus generalization occur post one-shot learning?

2. How to study the effects of Transfer Learning under the new computation model?

3. How is prediction error associated with confidence in novel learning and how to measure it?

4. How does attention to different dimensions of the stimuli modulate the learning rate? How do we distinguish between the dimensions? i.e. a more CS model based analysis.

5. Completing experiments for the present project, and establishing the relevance of such a model by linking categorization with general one shot learning.

References

Ashby, F. G., & Maddox, W. T. (2005). Human category learning. *Annu. Rev. Psychol.*, *56*, 149–178.

Bertinetto, L., Henriques, J. F., Valmadre, J., Torr, P., & Vedaldi, A. (2016). Learning feed-forward one-shot learners. In *Advances in neural information processing systems* (pp. 523–531).

Canini, K. (2007). *Modeling categorization as a dirichlet process mixture* (Tech. Rep.). Technical Report UCB/EECS-2007-69, Department of Electrical Engineering and Computer Sciences, University of California, Berkeley.

Fang, W.-C., & Chiang, Y.-t. (2016). Cognitive discriminative mappings for rapid learning. *arXiv preprint arXiv:1611.02512*.

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.

Greve, A., Cooper, E., Kaula, A., Anderson, M. C., & Henson, R. (2017). Does prediction error drive one-shot declarative learning? *Journal of Memory and Language*, *94*, 149–165.

Kemp, C., Perfors, A., & Tenenbaum, J. B. (2007). Learning overhypotheses with hierarchical bayesian models. *Developmental science*, *10*(3), 307–321.

Lake, B., Salakhutdinov, R., Gross, J., & Tenenbaum, J. (2011). One shot learning of simple visual concepts. In *Proceedings of the cognitive science society* (Vol. 33).

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

Lee, S. W., ODoherty, J. P., & Shimojo, S. (2015). Neural computations mediating one-shot learning in the human brain. *PLoS biology*, *13*(4), e1002137.

Maas, A., & Kemp, C. (2009). One-shot learning with bayesian networks.

Miller, R. R., Barnet, R. C., & Grahame, N. J. (1995). Assessment of the rescorla-wagner model. *Psychological bulletin*, *117*(3), 363.

Salakhutdinov, R., Tenenbaum, J., & Torralba, A. (2012). One-shot learning with a hierarchical nonparametric bayesian model. In *Proceedings of icml workshop on unsupervised and transfer learning* (pp. 195–206).

Santoro, A., Bartunov, S., Botvinick, M., Wierstra, D., & Lillicrap, T. (2016). One-shot learning with memory-augmented neural networks. *arXiv preprint arXiv:1605.06065*.

Sudderth, E. B., Torralba, A., Freeman, W. T., & Willsky, A. S. (2008). Describing visual scenes using transformed objects and parts. *International Journal of Computer Vision*, *77*(1), 291–330.

Tenenbaum, J. B., Griffiths, T. L., & Kemp, C. (2006). Theory-based bayesian models of inductive learning and reasoning. *Trends in cognitive sciences*, *10*(7), 309–318.

Vinyals, O., Blundell, C., Lillicrap, T., Wierstra, D., et al. (2016). Matching networks for one shot learning. In *Advances in neural information processing systems* (pp.

3630–3638).

Wang, Y., Zeng, Y., & Xu, B. (2016). Shtm: A neocortex-inspired algorithm for one-shot text generation. In *Systems, man, and cybernetics (smc), 2016 ieee international conference on* (pp. 000898–000903).

Xu, F., & Tenenbaum, J. B. (2007). Word learning as bayesian inference. *Psychological review*, *114*(2), 245.

Yip, K., & Sussman, G. J. (1997). Sparse representations for fast, one-shot learning.