1 Introduction

1. If we want system that exhibits the generality of human intelligence, they must not require millions of datapoints for each and every new task, concept or environment.
2. Current system is fed data in a way that is very different from how humans are fed data. A MNIST model sees 6k instances of 10 different objects. While human probably sees the transpose: 10 instances of 6k different objects.
3. Training a model on diverse data is not enough to ensure fast adaptability.
4. The performance of pre-training techniques is limited when only fine-tuning with a small number of examples.
5. Her thesis invents methods that explicitly trains model for fast adaptability, referred to as training the model for the ability to quickly learn new concepts (learn how to learn).
6. Her approach can be be seen as learning the right prior from existing data that when combined with a small amount of data in a new task leads to fast adaptability. Interesting connection to hierarchical Bayesian modeling, which is useful later to reason about uncertainty in learning.
7. There are two main approaches to meta-learning: using a black-box NN that learn from data that is passed in, and approach that incorporates structure of known optimization procedure into the learner.
8. In the first, a NN takes as input the dataset and a new unlabelled point and either predicts its label or the weights of another NN model that can solve the task.
9. However, this approach can be inefficient because they do not incorporate any structure.
10. In the second approach, prior works that seek to incorporate structure into the meta-learning process includes: comparing examples in a learned metric space (but difficult to extend to RL).

2 Problem Statement

2.1 Problem and terminology

1. She would like to apply her framework to a variety of learning problems, so she introduces a generic notion of a learning task.
2. Each task

$$\mathcal{T} = \{\mathcal{L}(\theta, \mathcal{D}), \rho\left(\mathbf{x}_1\right), \rho\left(\mathbf{x}_{\mathrm{t}+1}|\mathbf{x}_{\mathrm{t}}, \mathbf{y}_{\mathrm{t}}\right), \mathrm{H}\}$$

consists of a loss function $\mathcal{L}$ that takes as input the model's parameters θ and dataset $\mathcal{D}$, a distribution over initial observations $\rho\left(\mathbf{x}_1\right)$, a transition distribution $\rho\left(\mathbf{x}_{\mathrm{t}+1}|\mathbf{x}_{\mathrm{t}}, \mathbf{y}_{\mathrm{t}}\right)$ and an episode length H.

3. In supervised learning problem, $\mathrm{H} = 1$ and the dataset consists of labelled input-output pairs $\mathcal{D} = \left\{(\mathbf{x}_1, \mathbf{y}_1)^{(k)}\right\}$.
4. In RL setting, $\mathrm{H} > 1$ and $\mathcal{D} = \left\{(\mathbf{x}_1, \hat{\mathbf{y}}_1, \ldots, \mathbf{x}_{\mathrm{H}}, \hat{\mathbf{y}}_{\mathrm{H}})^{(\mathrm{k})}\right\}$ where $\hat{\mathbf{y}}$ represents the action chosen by the policy at each timestep.
5. The loss $\mathcal{L}(\theta, \mathcal{D})$ provides task-specific feedback for the model f_θ, which can be classification loss in supervised or cost function in RL setting.
6. In her meta-learning scenario, she considers a distribution of task $\mathrm{p}(\mathcal{T})$ that the model should be able to quickly adapt to.
7. In the K-shot learning setting, during meta-training, the task $\mathcal{T}_i$ is sampled from $\mathrm{p}(\mathcal{T})$, K samples are sampled for this task $\mathcal{D}_{\mathrm{i}}^{\mathrm{tr}}$, the model is trained with the corresponding loss $\mathcal{L}_{\mathcal{T}_{\mathrm{i}}}$, and then tested on new samples from the same task $\mathcal{T}_i$, denoted as $\mathcal{D}_{\mathrm{i}}^{\mathrm{test}}$. The model f is then improved by considering how the test error on new data $\mathcal{D}_{\mathrm{i}}^{\mathrm{test}}$ changes wrt to the parameters.

8. At the end of meta-training, new tasks are sampled from $p(\mathcal{T})$ and meta-performance is measured by the model's performance after learning from K samples.

3 Desirable properties of meta-learning algorithms

3.1 Expressive power of the learning algorithms

1. One intuitive way to define learning procedure would be something that takes in a dataset and output a vector of parameters that are used to make predictions about new data-points.

2. However, this definition is bad because: 1. It can only represents learning procedure that takes a parametric approach, 2. It is overcomplete, i.e. there is often more than one parameter vectors that can lead to the same underlying function.

3. She thus defines a learning algorithm as something that takes as input both a dataset $\mathcal{D}$ and a test observation $\mathbf{x}^{\star}$ and outputs a prediction , i.e. $\hat{\mathbf{y}}^{\star} = \mathrm{f}\left(\mathcal{D}, \mathbf{x}^{\star}\right)$.

4. Now that she has defined a learning procedure, she wants to measure the set of learning procedures that a particular meta-learning algorithms can represent.

5. She defines a universal learning procedure approximator as a universal function approximator for the function mapping from $\mathcal{D}$ and $\mathbf{x}^{\star}$ to $\hat{\mathbf{y}}^{\star}$.

6. This is a binary measure, a learning algorithm either has maximal expressive power or not.

3.2 Consistent learning algorithms

1. A learning algorithm is consistent if it finds the true function when provided infinite data.

$$\lim_{|\mathcal{D}| \to \infty} f\left(\mathcal{D}, \mathbf{x}_{\mathrm{i}}^{\star}\right) \to \mathbf{y}_{\mathrm{i}}^{\star} \forall \left(\mathbf{x}_{\mathrm{i}}^{\star}, \mathbf{y}_{\mathrm{i}}^{\star}\right)$$