Probabilistic Model-Agnostic Meta Learning

Chelsea Finn*, Kelvin Xu*, Sergey Levine

BERKELEY ARTIFICIAL INTELLIGENCE RESEARCH

Motivation

1. Learning a prior is one avenue to enable **few-shot learning**, which is a key aspect of human intelligence.
2. For some tasks however, **task ambiguity** in the few shot regime is unavoidable. For active learning, reinforcement learning or safety critical tasks (e.g., in medical domains), **modeling ambiguity** is key.

Key takeaways of this work:

1) Extends model-agnostic meta learning (MAML) [1] by modeling uncertainty in a principled manner using amortized variational inference in a graphical model formulation of meta-learning (also see [2]).

2) Proposes a scalable probabilistic meta-learning algorithm that can be used to improve performance in *regression*, *active learning*, and *ambiguous classification settings*.

Background: the Meta-Learning Problem

- Assume access to a set of tasks drawn from the task distribution $\mathcal{T}_i \sim p(\mathcal{T})$ which can be split into a training set $\mathcal{D}^{\text{tr}}_{\mathcal{T}_i} := \{(\mathbf{x}^{tr}_{i,j}, \mathbf{y}^{tr}_{i,j}) \mid \forall j\}$ and test set $\mathcal{D}^{\text{test}}_{\mathcal{T}_i} := \{(\mathbf{x}^{tr}_{i,j}, \mathbf{y}^{tr}_{i,j}) \mid \forall j\}$

Model-Agnostic Meta-Learning (MAML) [1]

$$\min_\theta \sum_{\mathcal{T}\sim p(\mathcal{T})} \mathcal{L}(\theta - \alpha\nabla_\theta \mathcal{L}(\theta, \mathcal{D}^{\text{tr}}_{\mathcal{T}_i}), \mathcal{D}^{\text{test}}_{\mathcal{T}_i}) = \min_\theta \sum_{\mathcal{T}_i\sim p(\mathcal{T})} \mathcal{L}(\phi_i, \mathcal{D}^{\text{test}}_{\mathcal{T}_i})$$

Probabilistic Gradient Based Meta-Learning

- We start with a probabilistic deep network and the conditional likelihoods of both the meta-train and meta-test set

$$\theta \sim p(\theta) = \mathcal{N}(\mu_\theta, \Sigma_\theta) \qquad \log p(y^{\text{tr}}_i \mid x^{\text{tr}}_i, \phi_i)$$
$$\phi_i \sim p(\phi_i \mid \theta) \qquad \log p(y^{\text{test}}_i \mid x^{\text{test}}_i, \phi_i)$$

- Goal: sample $\phi_i \sim p(\phi_i \mid x^{tr}_i, y^{tr}_i, x^{test}_i)$

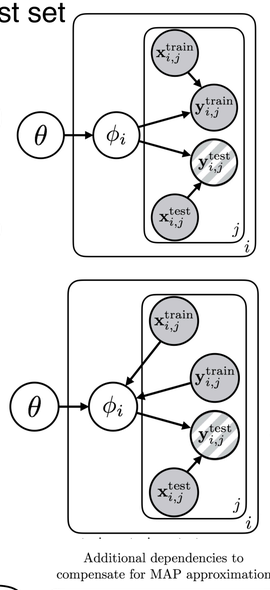
$$p(\phi_i \mid x^{tr}_i, y^{tr}_i) \propto \int p(\theta)p(\phi_i|\theta)p(y^{tr}_i|x^{tr}_i, \phi_i)d\theta$$

Completely intractable!

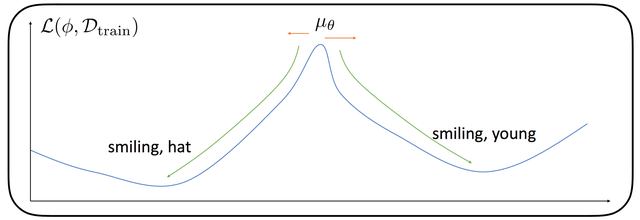
- If we knew $p(\phi_i \mid x^{tr}_i, y^{tr}_i)$, we could use ancestral sampling

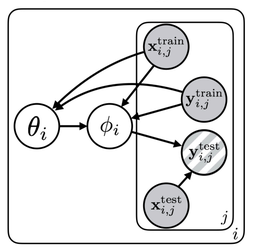
Key idea: use the MAP approximation which is crude but extremely convenient [2]

$$p(\phi_i \mid \theta, x^{tr}_i, y^{tr}_i) \approx \delta(\hat{\phi}_i)$$
$$\hat{\phi}_i \approx \theta + \alpha\nabla_\theta \log p(y^{tr}_i \mid x^{tr}_i, \theta)$$



Additional dependencies to compensate for MAP approximation





Learning with amortized inference

- To train our model parameters θ, we can derive a variational bound

$$E_{\theta\sim q}\left[\log p(y^{\text{test}}_i \mid x^{\text{test}}_i, \hat{\phi}_i(\theta))\right] - D_{\text{KL}}(q(\theta|x^{\text{test}}_i, y^{\text{test}}_i)) \mid\mid p(\theta))$$

- Training a network to output parameters not scalable, instead parameterize the q distribution by embedding gradient descent:

$$q(\theta \mid x^{\text{test}}_i, y^{\text{test}}_i) = \mathcal{N}(\mu_\theta + \alpha\nabla_{\mu_\theta} \log p(y^{\text{test}}_i \mid x^{\text{test}}_i, \mu_\theta), \Sigma_q)$$

Intuition: *cheat* at meta-training time using the test set, but minimize a divergence with a prior which is used for sampling at meta-test time

Additional dependencies

- In the 2nd graphical model, $x^{\text{tr}}_i, y^{\text{tr}}_i, \theta$ are conditionally independent

- Since we used a crude MAP approximation, these independences may not hold. Allow the model to compensate by learning a task specific prior $p(\theta \mid x^{\text{tr}}_i, y^{\text{tr}}_i) = \mathcal{N}(\mu_\theta + \alpha\nabla_{\mu_\theta} \log p(y^{\text{tr}}_i \mid x^{\text{tr}}_i, \mu_\theta), \Sigma_p)$

Meta-training

(1) $\theta \sim q(\theta) = \mathcal{N}(\mu_\theta + \alpha\nabla_{\mu_\theta} \log p(y^{\text{test}}_i \mid x^{\text{test}}_i, \mu_\theta), \Sigma_q)$

(2) $\phi_i \sim p(\phi_i \mid x^{\text{tr}}_i, y^{\text{tr}}_i) \approx \hat{\phi}_i = \theta + \alpha\nabla_\theta \log p(y^{\text{tr}}_i \mid x^{\text{tr}}_i, \theta)$

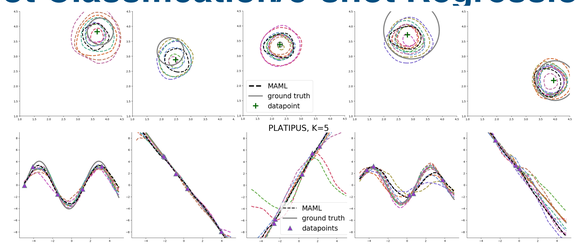
(3) take gradient step with respect to variational bound

Meta-testing

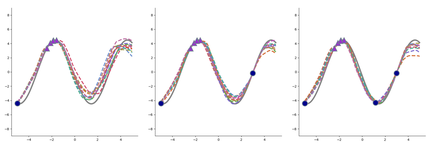
(1) $\theta \sim p(\theta) = \mathcal{N}(\mu_\theta + \alpha\nabla_{\mu_\theta} \log p(y^{\text{tr}}_i \mid x^{\text{tr}}_i, \mu_\theta), \Sigma_p)$

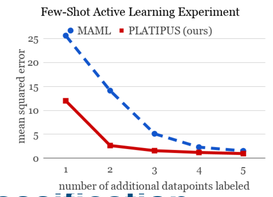
(2) $\phi_i \sim p(\phi_i \mid x^{\text{tr}}_i, y^{\text{tr}}_i) \approx \hat{\phi}_i = \theta + \alpha\nabla_\theta \log p(y^{\text{tr}}_i \mid x^{\text{tr}}_i, \theta)$

1-Shot Classification/5-shot Regression



Active Learning





Ambiguous Attributes Classification

Ambiguous celebA (5-shot)			
	Accuracy	Coverage (max=3)	Average NLL
MAML	**89.00 ± 1.78%**	1.00 ± 0.0	0.73 ± 0.06
MAML + noise	84.3 ± 1.60 %	1.89 ± 0.04	0.68 ± 0.05
PLATIPUS (ours) (KL weight = 0.05)	**88.34 ± 1.06 %**	1.59 ± 0.03	0.67± 0.05
PLATIPUS (ours) (KL weight = 0.15)	**87.8 ± 1.03 %**	**1.94 ± 0.04**	**0.56 ± 0.04**

Mini-Imagenet

MiniImagenet	5-way, 1-shot Accuracy
MAML [8]	48.70 ± 1.84%
LLAMA [14]	49.40 ± 1.83%
Reptile [28]	**49.97 ± 0.32%**
PLATIPUS (ours)	**50.13 ± 1.86%**
Meta-SGD [24]	**50.71 ± 1.87%**

- Slight boost over MAML, while comparable to other approaches, (*similar architecture)

Future Directions

- Studying how ambiguity and uncertainty can guide data-acquisition
- Data-dependent posterior variances for tasks with differing levels of uncertainty

[1] Finn, Abbeel, Levine. *Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks*. ICML '17
[2] Grant, Finn, Levine, Darrell, Griffiths. *Recasting gradient-based meta-learning as hierarchical Bayes*. ICLR '18