Probabilistic Model-Agnostic Meta-Learning

Chelsea Finn*, Kelvin Xu*, Sergey Levine

*equal contribution
Deep learning requires a lot of data.

Few-shot learning / meta-learning: incorporate prior knowledge from previous tasks for fast learning of new tasks

Given 1 example of 5 classes: few-shot image classification

Classify new examples
Ambiguity in Few-Shot Learning

Can we learn to generate hypotheses about the underlying function?

Important for:
- learning to actively learn
- safety-critical few-shot learning (e.g. medical imaging)
- learning to explore in meta-RL
Background: Model-Agnostic Meta-Learning

Optimize for pretrained parameters

\[
\text{MAML: } \min_{\theta} \sum_{\text{task } i} \mathcal{L}_{\text{test}}^{i}(\theta - \alpha \nabla_{\theta} \mathcal{L}_{\text{train}}^{i}(\theta))
\]

Key idea: Train over many tasks, to learn parameter vector \( \theta \) that transfers

\[
\text{Meta-test time: } \phi_{j} \leftarrow \theta - \alpha \nabla_{\theta} \mathcal{L}_{\text{train}}^{j}(\theta)
\]

How to reason about \textit{ambiguity}?

\[
\text{Finn, Abbeel, Levine ICML '17}
\]
Can we **sample** classifiers?

**Intuition of our approach:** learn a prior where a random kick can put us in different modes

\[
\phi \leftarrow \theta + \epsilon \\
\phi \leftarrow \phi + \alpha \nabla_\phi \mathcal{L}(\phi, D_{\text{train}})
\]
Meta-learning with ambiguity

\[ \theta \sim p(\theta) = \mathcal{N}(\mu_\theta, \Sigma_\theta) \]

\[ \phi_i \sim p(\phi_i | \theta) \]

Goal: sample \( \phi_i \sim p(\phi_i | x_i^{\text{train}}, y_i^{\text{train}}, x_i^{\text{test}}) \)
Sampling parameter vectors

\[ \theta \sim p(\theta) = \mathcal{N}(\mu_\theta, \Sigma_\theta) \quad \log p(y_i^{\text{train}} | x_i^{\text{train}}, \phi_i) \]
\[ \phi_i \sim p(\phi_i | \theta) \quad \log p(y_i^{\text{test}} | x_i^{\text{test}}, \phi_i) \]

Goal: sample \( \phi_i \sim p(\phi_i | x_i^{\text{train}}, y_i^{\text{train}}) \)

\[ p(\phi_i | x_i^{\text{train}}, y_i^{\text{train}}) \propto \int p(\theta) p(\phi_i | \theta) p(y_i^{\text{train}} | x_i^{\text{train}}, \phi_i) d\theta \]

⇒ this is completely intractable!

what if we knew \( p(\phi_i | \theta, x_i^{\text{train}}, y_i^{\text{train}}) \)?

⇒ now sampling is easy! just use ancestral sampling!

**key idea:** \( p(\phi_i | \theta, x_i^{\text{train}}, y_i^{\text{train}}) \approx \delta(\hat{\phi}_i) \)

this is extremely crude

but extremely convenient!

\[ \hat{\phi}_i \approx \theta + \alpha \nabla_\theta \log p(y_i^{\text{train}} | x_i^{\text{train}}, \theta) \]

(Santos '92, Grant et al. ICLR '18)

Training is harder. We use amortized variational inference.
PLATIPUS
Probabilistic LATent model for Incorporating Priors and Uncertainty in few-Shot learning

Ambiguous 5-shot regression:

Ambiguous 1-shot classification:
PLATIPUS
Probabilistic LA汀ent model for Incorporating Priors and Uncertainty in few-Shot learning

<table>
<thead>
<tr>
<th>Ambiguous celebA (5-shot)</th>
<th>Accuracy</th>
<th>Coverage (max=3)</th>
</tr>
</thead>
<tbody>
<tr>
<td>MAML</td>
<td>69.26 ± 2.18%</td>
<td>1.00 ± 0.0</td>
</tr>
<tr>
<td>MAML + noise</td>
<td>54.73 ± 0.8 %</td>
<td>2.60 ± 0.12</td>
</tr>
<tr>
<td>PLATIPUS (ours)</td>
<td>69.97 ± 1.32 %</td>
<td>2.62 ± 0.11</td>
</tr>
</tbody>
</table>
Takeaways

Handling ambiguity is particularly relevant in few-shot learning.

Can use hybrid inference in hierarchical Bayesian model for scalable and uncertainty-aware meta-learning.

Collaborators

Kelvin Xu  Sergey Levine

Questions?
Related concurrent work

Kim et al. “Bayesian Model-Agnostic Meta-Learning”: uses Stein variational gradient descent for sampling parameters

Gordon et al. “Decision-Theoretic Meta-Learning”: unifies a number of meta-learning algorithms under a variational inference framework