Supplementary Material for the Paper: Amortized Bayesian Prototype Meta-learning: A New Probabilistic Meta-learning Approach to Few-shot Image Classification

1 Overview

In this document, we present details of experimental settings, including hyper-parameters (batch size, learning rate, etc.). We also provide pseudo-code for meta-validation/meta-testing and detailed statistics in plots and figures. All experiments are implemented with PyTorch.

2 Pseudo-code for Meta-validation/Meta-testing

Algorithm 2 Meta-validation/Meta-testing of the proposed method

Require: Input Meta-trained model $\hat{\mathcal{M}}$. Set $\tilde{\mathcal{D}} = \mathcal{D}_{val}$ or $\mathcal{D}_{te}$.

- **for** i from 1 to E **do**:
  - Generate a task $\mathcal{T}_i = \mathcal{S}_i \cup \mathcal{Q}_i$ from $\tilde{\mathcal{D}}$.
  - Initialize $\phi_i \leftarrow \theta$.
  - **for** d from 1 to D **do**:
    - Compute $q_{\phi_i}(z|\mathcal{S}_i)$.
    - Approximate KL.
    - Update variational parameters $\phi_i \leftarrow \phi_i - \alpha \nabla_{\phi_i} \{\mathcal{L}_{PR}(\mathcal{S}_i|z) + KL[q_{\phi_i}(z|\mathcal{S}_i) \mid\mid p(z|\theta)]\}$.
  - Predict for an image x: $\hat{y} = \arg\max_c \Pr(\mu(x)|q_{\phi_i}(z_c|\mathcal{S}_{i,c})),\ c \in [C]$.
  - Compute prediction accuracy a_i for $\mathcal{Q}_i$.
- Output mean accuracy $\frac{1}{E}\sum_{i=1}^{E} a_i$ as $\hat{\mathcal{M}}$'s performance.

3 Proofs

3.1 Proof for Eq.2

In this section, we provide a detailed derivation of the evidence lower bound of $\log p_\theta(\mathcal{S})$.

$$\log p_\theta(\mathcal{S}) \geq \mathbb{E}_{z\sim q_\phi(z)}[\log p_\theta(\mathcal{S}|z)] - \mathbb{KL}[q_\phi(z) \mid\mid p_\theta(z)]$$

Proof:

$$\begin{aligned}
\log p_\theta(\mathcal{S}) &= \log \int p_\theta(\mathcal{S}, z) dz \\
&= \log \int p_\theta(\mathcal{S}, z) \frac{q_\phi(z)}{q_\phi(z)} dz \\
&= \log \mathbb{E}_q \left[\frac{p_\theta(\mathcal{S}, z)}{q_\phi(z)} \right] \\
&\geq \mathbb{E}_q \left[\log p_\theta(\mathcal{S}|z) + \log \frac{p_\theta(z)}{q_\phi(z)} \right] \quad \text{, by Jensen's inequality} \\
&= \mathbb{E}_q \left[\log p_\theta(\mathcal{S}|z) \right] - \mathbb{KL}\left[q_\phi(z) || p_\theta(z) \right]
\end{aligned}$$

3.2 Unbiased Estimator Scaled by A Constant

Although we replace the $\mathbb{KL}$ term in the evidence lower bound of $\log p_\theta(\mathcal{S})$ with Eq.7 as our proposed prior distribution of z is now dependent on the support set $\mathcal{S}$, our estimator to Eq.4 is still an unbiased estimator to evidence lower bound of $\log p_\theta(\mathcal{S})$ in Eq.2 (scaled by a constant). Therefore the proposed method still learns to learn the approximate posteriors of latent z conditional on $\mathcal{S}$ properly. To appreciate this, note that during the inference stage τ is the support set $\mathcal{S}$ (and we have $|\tau| = CK$). Then, after putting the unbiased estimator of Eq.7 and Eq.9 into Eq.4, we can rewrite the loss in Eq.4 as

$$\mathcal{L}(\mathcal{S}) = \frac{1}{CK}\sum_{c=1}^{C}\sum_{i=1}^{K}\Big(-\log\Big(\frac{\Pr\left[\mu(x_i^{(\mathcal{S}_c)})|z_c\right]}{\sum_{k=1}^{C}\Pr\left[\mu(x_i^{(\mathcal{S}_c)})|z_k\right]}\Big) + \mathbb{KL}[q_\phi(z_c|\mathcal{S}_c)||p_\theta(z_c;\mu(x_i^{(\mathcal{S}_c)}),\Sigma(x_i^{(\mathcal{S}_c)}))]\Big)$$

, where $\mathcal{S}_c$ is the subset of $\mathcal{S}$ and only contains all support images from the class $c \in \{1,\dots,C\}$, and $(x_i^{(\mathcal{S}_c)}, y_i^{(\mathcal{S}_c)} = c)$ is the i^{th} image in $\mathcal{S}_c$. This immediately tells that $-\mathcal{L}(\mathcal{S}) = \frac{1}{CK}\sum_{c=1}^{C}\sum_{i=1}^{K}(\log p_\theta(y_i^{(\mathcal{S}_c)}|x_i^{(\mathcal{S}_c)}, z_c) - \mathbb{KL}[q_\phi(z_c|\mathcal{S}_c)||p_\theta(z_c;\mu(x_i^{(\mathcal{S}_c)}),\Sigma(x_i^{(\mathcal{S}_c)}))])$, where the terms inside the double summation is an unbiased estimator of the evidence lower bound of $\log p_\theta(y_i^{(\mathcal{S}_c)}|x_i^{(\mathcal{S}_c)})$. Since $\frac{1}{CK}\sum_{c,i}\log p_\theta(y_i^{(\mathcal{S}_c)}|x_i^{(\mathcal{S}_c)}) = \frac{1}{CK}\log p_\theta(\mathcal{S})$, it tells that $-\mathcal{L}(\mathcal{S})$ is an unbiased estimator of the evidence lower bound of $\log p_\theta(\mathcal{S})$ scaled by a factor of $1/CK$.

4 Experimental Details

At the meta-training stage, except that the maximum training epoch is 12000 for 1-shot classification on *mini-ImageNet*, the maximum training epoch is set to be 3500 epochs for all the other experiments. We use a mini-batch of tasks consisting T tasks to update the shared θ during meta-training.

We select the optimal meta-training epoch on the meta-validation set according to classification accuracy. At the meta-testing stage, we randomly sample 600 novel tasks from the meta-testing set, and report the mean accuracy with its 95% confidence interval, i.e., mean acc. $\pm\, 1.96\frac{\text{std}}{\sqrt{600}}$. For C-way K-shot, a task is constructed by sampling C classes and then subsequently sampling $K+M$ instances for each class, with K being the number of support images in each class. In our experiments,

- *Omniglot*: $M = 15$ for meta-training/meta-validation/meta-testing;
- *mini-ImageNet*: $M = 16$ for meta-training and meta-validation, $M = 15$ for meta-testing;
- *CUB-200-2011*: $M = 16$ for meta-training and meta-validation, $M = 15$ for meta-testing;
- *Stanford-dogs*: $M = 16$ for meta-training and meta-validation, $M = 15$ for meta-testing.

The values of T, D, α and β in **Alg. 1** and **Alg. 2** are set to be

- *Omniglot*: $T = 32$, $D = 1$, $\alpha = 0.1$, $\beta = 0.001$;
- *mini-ImageNet*: $T = 4$, $D = 5$, $\alpha = 0.01$, $\beta = 0.001$;
- *CUB-200-2011*: $T = 4$, $D = 5$, $\alpha = 0.01$, $\beta = 0.001$;
- *Stanford-dogs*: $T = 4$, $D = 5$, $\alpha = 0.01$, $\beta = 0.001$.

In addition, we use standard stochastic gradient descent to generate variational parameters ϕ_i, during meta-training/meta-validation/meta-testing, for a task $\mathcal{T}_i$ and for all i. We use the *Adam* optimizer to update the shared parameter θ at meta-training stage.

5 Details of Figures

In this section, we present detailed statistics in **Fig. 2**.

Ablation study in **Fig.2-a**.

	Meta-training conditions	
C-way at meta-testing	5-way 5-shot (%)	10-way 5-shot (%)
$C = 5$	99.45 ± 0.09	99.44 ± 0.08
$C = 10$	98.97 ± 0.08	99.14 ± 0.08
$C = 15$	98.45 ± 0.09	98.80 ± 0.09
$C = 20$	98.14 ± 0.09	98.52 ± 0.08
$C = 25$	97.85 ± 0.09	98.20 ± 0.08
$C = 30$	97.44 ± 0.09	97.87 ± 0.08
$C = 35$	97.17 ± 0.09	97.63 ± 0.08
$C = 40$	96.84 ± 0.08	97.34 ± 0.08
$C = 45$	96.57 ± 0.08	97.12 ± 0.08
$C = 50$	96.30 ± 0.08	96.85 ± 0.08

Ablation study in **Fig.2-b**.

	Meta-training conditions	
K-shot at meta-testing	5-way 5-shot (%)	10-way 5-shot (%)
$K = 2$	98.65 ± 0.27	98.38 ± 0.15
$K = 4$	99.47 ± 0.11	99.00 ± 0.11
$K = 5$	99.60 ± 0.10	99.17 ± 0.09
$K = 6$	99.53 ± 0.11	99.19 ± 0.10
$K = 8$	99.59 ± 0.10	99.06 ± 0.13
$K = 10$	99.61 ± 0.09	99.32 ± 0.10
$K = 12$	99.60 ± 0.09	99.34 ± 0.09

- *Omniglot*: *Dropout* with a keep probability of 0.9.
- *mini-ImageNet*: *Dropout* with a keep probability of 0.5.

Ablation study in **Fig.2-c**.

KL	*Dropout*	*Omniglot* (%)	*mini-ImageNet* (%)
-	-	96.16 ± 0.28	43.08 ± 0.62
✓	-	99.54 ± 0.08	70.44 ± 0.72
✓	✓	99.50 ± 0.08	69.92 ± 0.67

6 Comparisons of Convolution Networks

Here, we present details of shallow convolution networks used in the probabilistic meta-learning methods listed in **Table 1**. CONV-X means a convolution network with X convolution blocks.

Convolution networks of methods in **Table 1**.

	Omniglot	*mini-ImageNet*
BMAML	CONV-5	CONV-5
PLATIPUS	CONV-4	CONV-4
VAMPIRE	CONV-4	CONV-4
ABML	CONV-4	CONV-4
Amortized VI	CONV-4	CONV-5
VERSA	CONV-4	CONV-5
Meta-Mixture	CONV-4	CONV-4
DKT	CONV-4	CONV-4
Ours	CONV-4	CONV-4

7 Effect of D

We also take the effect of D into account. Recall that D is the number of updates of the inner loop for the approximate inference. We consider the cases when $D = 1$, $D = 3$ and $D = 5$. Performance for each choice of D is measured on the meta-testing set.

Effect of D.

mini-ImageNet	$D = 1$(%)	$D = 3$(%)	$D = 5$(%)
5-way 1-shot	52.79 ± 0.94	53.29 ± 0.89	53.28 ± 0.91
5-way 5-shot	69.63 ± 0.70	70.56 ± 0.70	70.44 ± 0.72