Hypernetwork Approach to Bayesian MAML (Student Abstract)

Piotr Borycki Piotr Kubacki Marcin Przewięźlikowski
Tomasz Kuśmierczyk Jacek Tabor Przemysław Spurek

Jagiellonian University in Kraków
piotr.borycki@student.uj.edu.pl

Abstract

The main goal of Few-Shot learning algorithms is to enable learning from small amounts of data. One of the most popular and elegant Few-Shot learning approaches is Model-Agnostic Meta-Learning (MAML). In this paper, we propose a novel framework for Bayesian MAML called BH-MAML, which employs Hypernetworks for weight updates. It learns the universal weights point-wise, but a probabilistic structure is added when adapted for specific tasks. In such a framework, we can use simple Gaussian distributions or more complicated posteriors induced by Continuous Normalizing Flows.

Introduction

Few-Shot learning models easily adapt to previously unseen tasks based on a few labeled samples. One of the most popular and elegant among them is Model-Agnostic Meta-Learning (MAML) (Finn, Abbeel, and Levine 2017). The main idea behind this method is to produce universal weights θ, which can be rapidly updated to θ' dedicated to solving new small tasks. However, limited data sets lead to two main problems. First, the method tends to overfit to training data, preventing us from using deep architectures with large numbers of weights. Second, it lacks good quantification of uncertainty, e.g., the model does not know how reliable its predictions are. Both problems can be addressed by employing Bayesian Neural Networks (BNNs), which learn distributions in place of point-wise estimates.

There exist a few Bayesian modifications of the classical MAML algorithm. Bayesian MAML (Yoon et al. 2018), Amortized bayesian meta-learning (Ravi and Beatson 2018) learn distributions for the common universal weights, which are then updated to per-task local weights distributions. The above modifications of MAML, similar to the original MAML, rely on gradient-based updates. Weights specialized for small tasks are obtained by taking a fixed number of gradient steps from the standard universal weights. Such a procedure requires two levels of Bayesian regularization, and the universal distribution is usually employed as a priori for the per-task specializations. However, the hierarchical structure complicates the optimization procedure and limits updates in the MAML procedure.

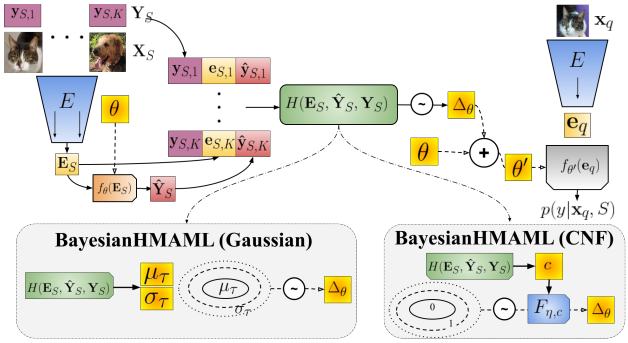

Figure 1: BH-MAML: First, the support set is transformed by an encoder into $\mathbf{E}_{\mathcal{S}}$. It is then concatenated with the original labels $\mathbf{Y}$ and predictions obtained using universal weights $\hat{\mathbf{Y}}$. Next, the hypernetwork produces posterior distributions $H(\mathbf{E}_{\mathcal{S}}, \hat{\mathbf{Y}}_{\mathcal{S}}, \mathbf{Y}_{\mathcal{S}})$. We consider two posterior variants: Gaussian and CNF-based. Using the hypernetwork for weight updates θ allows for larger and smarter adaptations of the posterior parameters. In the end, we sample weight updates $\Delta\theta$ to obtain weights $\theta' = \theta + \Delta\theta$ dedicated to a specific task.

Method

The paper presents BH-MAML – a new framework for Bayesian Few-Shot learning. It simplifies the explained above weight-adapting procedure and thanks to the use of hypernetworks, enables learning more complicated posterior updates. Similar to the previous approaches, the final weight posteriors are obtained by updating from the universal weights. However, we avoid learning the aforementioned hierarchical structure by point-wise modeling of the universal weights. The probabilistic structure is added only later when specializing the model for a specific task. In BH-MAML updates from the universal weights to the per-task specialized ones are generated by hypernetworks instead of the previously used gradient-based optimization. Because hypernetworks can easily model more complex structures, they allow for better adaptations. In particular, we tested the standard Gaussian posteriors against more general posteriors induced by Continuous Normalizing Flows (CNF).

BH-MAML is a Bayesian extension of the classical MAML using Hypernets. Bayesian treatment for such a model is to pose priors for the model parameters and learn

their posteriors. In particular, for MAML one needs to learn posterior distributions for both universal θ and specialized θ' weights. This naturally hints at a hierarchical Bayesian model: $\theta \rightarrow \theta'_i \rightarrow \mathcal{T}_i$ (where $\mathcal{T}_i$ denotes i-th task's data), which was previously presented by Chen and Chen (2022).

We propose an alternative approach that alleviates the problems of previous attempts at Bayesian MAML. Contrary to them, we do not learn distributions for universal parameters θ, but instead learn them in a pointwise manner, and distributional posteriors we learn only for task-specialized θ'_i. Additionally, we remove the coupling prior between θ and θ'_i. Finally, we propose a basic non-hierarchical prior $p(\theta'_i)$, where we used the standard normal priors for the weights of the neural network, i.e., $p(\theta'_i) = \mathcal{N}(\theta'_i|0, \mathbb{I})$. The learning objective takes the following form:

$$\mathcal{L}^{our}(\{\mathcal{T}_i\}) = \sum_i^N \mathbb{E}_{q(\theta'_i|\lambda_i(\theta,\mathcal{S}_i))}[\log p(\mathcal{T}_i|\theta'_i) - \gamma \cdot KL\left(q(\theta'_i|\theta, \lambda_i(\theta, \mathcal{S}_i))\|p(\theta'_i)\right)],$$

The hyperparameter γ allows controlling the impact of the priors and compensating for model misspecification. Overall, the proposed modifications enable better optima for the objective and simplify the optimization landscape helping convergence.

When adapting a function f_θ with parameters θ to a task $\mathcal{T}_i$, the updated model's parameters are $\theta'_i \sim q(\theta|\lambda_i(\theta, S_i))$, where the posterior parameters λ_i are modeled by a hypernetwork as $\lambda_i(\theta, S_i) := H_\phi(\theta, S_i)$. BH-MAML takes the support set S_i and universal weights θ, and combines these weights θ with $\Delta\theta'_i$ to produce samples from the posterior distribution (see Fig. 1). Thanks to the hypernetwork, we obtain unconstrained updates and can model arbitrary posterior distributions, potentially improving upon all previous non-hypernetwork models. Furthermore, the hypernetwork H has a fixed number of parameters, regardless of the number of tasks $\mathcal{T}_i$ used for training. The amortized learning scheme provides two additional benefits: (1) faster training; (2) regularization of learned parameters through a shared architecture and common weights ϕ.

We implemented two variants of BH-MAML:

Gaussian version. BH-MAML-G is a simple realization of BH-MAML. Here, the hypernetwork H_ϕ returns the mean update and covariance matrix of a Gaussian posterior: $(\mu_\theta(\mathcal{S}_i), \sigma_\theta(\mathcal{S}_i)) := H_\phi(\mathcal{S}_i, \theta)$. Weights are then sampled from the induced posterior: $\theta'_i \sim \mathcal{N}(\theta + \mu_\theta(\mathcal{S}_i), \sigma_\theta(\mathcal{S}_i))$. We apply here the mean-field assumption, but the standard deviations σ are not entirely independent. Due to the used amortization scheme, they are tied together and to the means μ by the shared weights ϕ of the hypernetwork. Posterior means however are additionally explicitly dependent by the universal weights θ. Similar as for the classic MAML, any change of θ affects all the values of θ'_i.

CNF version. BH-MAML-CNF is a generalization of BH-MAML-G, where we use conditional flows to produce weight posteriors for the specialized tasks. Similar to BH-MAML-G, we employ a hypernetwork to amortize updates of the target model parameters. However, in the above model the hypernetwork outputs parameters of a Gaussian distribution, whereas in BH-MAML-CNF it is responsible for conditioning a flow $F_{\eta,C(\theta,\mathcal{S}_i)}(\cdot)$ (see Fig. 1): $C(\theta, \mathcal{S}_i) := H_\phi(\mathcal{S}_i, \theta)$.

The conditioning vector C is added to each layer of the flow to parameterize function $g_{\theta,C}$, so in the end, the flow $F_{\eta,C}$ depends on trainable parameters η and conditioning parameters C. Then, the posterior for a task $\mathcal{T}_i$ is obtained by a two-stage process: $\Delta\theta'_i \sim F_{\eta,C(\theta,\mathcal{S}_i)}$ and $\theta'_i = \theta + \Delta\theta'_i$, where the shape of the posterior distribution is determined by the flow F, but its position, similarly to BH-MAML-G, mainly by the universal weights. From the implementation point of view, sampling from the conditioned flow also happens in two stages. First, we sample some z from a flow prior and then, push this z through a chain of deterministic transformations to obtain the final sample. Formally, $\Delta\theta'_i := F_{\eta,C(\theta,\mathcal{S}_i)}(z)$, where $z \sim \mathrm{N}(0, t \cdot \mathbb{I})$, where t is a hyperparameter, we used $t = 0.1$.

Method	Omniglot→EMNIST 1-shot	Omniglot→EMNIST 5-shot	mini-ImageNet→CUB 1-shot	mini-ImageNet→CUB 5-shot
Feature Transfer	64.22 / 1.24	86.10 / 0.84	32.77 / 0.35	50.34 / 0.27
ProtoNet	72.04 / 0.82	87.22 / 1.01	33.27 / 1.09	*52.16 / 0.17*
MAML	74.81 / 0.25	83.54 / 1.79	34.01 / 1.25	48.83 / 0.62
DKT	75.40 / 1.10	**90.30/0.49**	**40.14/0.18**	**56.40/1.34**
Bayesian MAML	63.94/0.47	65.26/0.30	33.52/0.36	51.35/0.16
HyperMAML	79.07/1.09	89.22/0.78	36.32/0.61	49.43/0.14
BH-MAML (G)	*80.95/0.46*	89.21/0.27	36.90/0.34	49.24/0.38
BH-MAML (G)+ada.	**81.05/0.47**	*89.76/0.26*	*37.23/0.44*	50.79/0.59
BH-MAML(CNF)	72.02/0.56	82.36/0.12	33.77/0.30	44.09/0.32
BH-MAML(CNF)+ada.	72.54/0.36	82.63/0.37	34.67/0.35	45.14/0.27

Table 1: Classification accuracy for inference on cross-domain data sets.

Experiments

In our experiments, we evaluate the model in a cross-domain adaptation setting, and the model is evaluated on tasks from a different distribution than the one on which it had been trained. We report the results in Table 1. In the task of 1-shot Omniglot→EMNIST classification, BH-MAML-G achieves the best result. The 5-shot Omniglot→EMNIST classification task BH-MAML-G yields comparable results to baseline methods. In the mini-ImageNet→CUB classification, our method performs comparably to baseline methods such as MAML and ProtoNet.

Conclusions

In this work, we introduced BH-MAML – a novel Bayesian Meta-Learning algorithm strongly motivated by MAML. In BH-MAML, we have universal weights trained in a pointwise manner, similar to MAML, and Bayesian updates modeled with hypernetworks. Such an approach allows for significantly larger updates in the adaptation phase and better uncertainty quantification.

Acknowledgements

The work of P. Borycki was supported by the National Centre of Science (Poland) Grant No. 2021/41/B/ST6/01370. The work of P. Spurek was supported by the National Centre of Science (Poland) Grant No. 2023/50/E/ST6/00068. The research of Marcin Przewięźlikowski was supported by the National Science Centre (Poland), grant no. 2023/49/N/ST6/03268.

This research has been supported by the flagship project entitled "Artificial Intelligence Computing Center Core Facility" from the Priority Research Area DigiWorld under the Strategic Programme Excellence Initiative at Jagiellonian University.

This research is part of the project No. 2022/45/P/ST6/02969 co-funded by the National Science Centre and the European Union Framework Programme for Research and Innovation Horizon 2020 under the Marie Sklodowska-Curie grant agreement No. 945339. For the purpose of Open Access, the author has applied a CC-BY public copyright licence to any Author Accepted Manuscript (AAM) version arising from this submission.

References

Chen, L.; and Chen, T. 2022. Is Bayesian Model-Agnostic Meta Learning Better than Model-Agnostic Meta Learning, Provably? In *International Conference on Artificial Intelligence and Statistics*, 1733–1774. PMLR.

Finn, C.; Abbeel, P.; and Levine, S. 2017. Model-agnostic meta-learning for fast adaptation of deep networks. In *International Conference on Machine Learning*, 1126–1135. PMLR.

Ravi, S.; and Beatson, A. 2018. Amortized bayesian meta-learning. In *International Conference on Learning Representations*.

Yoon, J.; Kim, T.; Dia, O.; Kim, S.; Bengio, Y.; and Ahn, S. 2018. Bayesian model-agnostic meta-learning. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, 7343–7353.