Amortized Inference of Bayesian Neural Network parameters for solving Meta-learning problems

Kim Juhyeong○ Heo Jaepil
College of Computing, Sungkyunkwan University

Abstract

In Meta-Learning, we try to extract the knowledge which is compatible and generalizable to multiple tasks, given scarce data for each task. Furthermore, the goal of the Bayesian Meta-learning is to inference posterior conditioned on arbitrary task. In this study, we tackled the problem of estimating task conditional posterior by applying Amortized Variational Inference technique to Bayesian Neural Network posterior. Also, we propose a scheme to reflect the sequential nature of Bayesian Neural Network model to posterior inference. quantitative comparisons with previous baselines in Meta-Learining were performed.

1. Introduction

In Meta-learning problem, we study the methods to extract information which can be utilized among multiple tasks in common. For example, few-shot classification setting requires the abilities to solve newly defined classification problem with extremely limited amount of data for the task.

Bayesian Neural Network(BNN) assumes prior distribution for deterministic Neural Network models and inference posterior distribution of the model given the data. By representing posterior distribution, not point estimates, uncertainty in the model is adequately quantized. However, BNN models has been a minor topic in Machine Learning application due to the over-parameterized and sophisticated nature of modern deep learning architectures.[1] Traditional approximate inference schemes for BNN model have been limited in scalability and accuracy for modern deep models. Thus, the goal of recent literatures are focused on proposing more practical and more exact approximate inference methods.

Historically, baseline approximate inference method for BNN was Markov Chain Monte Carlo variants(MCMC) which demand iterative model sampling. Accuracy was the strength of MCMC method, but time complexity was not. Typical time complexity of the methods largely depends on the number of dimensions of parameters and the number of data.

Another approximate inference method for Bayesian Neural Network model is implemented by using Dropout technique in deep learning. MC-Dropout is performed by applying Dropout at Deterministic Neural Network model both in training time and test time. By doing so, distributions of model which incorporates various architectural consideration is represented. MC-Dropout is concrete baseline among approximate inference method due to its practicality and in-depth theoretical analysis.

However, most active research direction in BNN is done by studying Variational Inference(VI) variants. VI is performed by introducing parameterized distribution which is easy to sample and do inference. In VI, we optimized the parameters of approximate distribution which is called Variational parameters to reduce the approximation error. Mean-field Variational Inference(MFVI) is a standard setting in VI which assumes approximate distribution as factorized Gaussian distribution. However, MFVI limited the practicality of BNN model because of its lack of ability in modeling complex relationship between BNN model parameters. VI has been pointed to be inefficient in computation for BNN models because it also considers the distributions of the redundant parameters in modern over-parameterized deep learning models. However, various schemes for overcoming over-parameterization problem for BNN are intriduced. For example, by finding a subspace of BNN parameter space.[2][3]

Flipout[4] is an method to reduce the variance of the gradient while perform weight perturbation in weight space. It proposed an algorithm to drastically reduce the unstability of training procedure. Generally, number of sampling and variance in estimator have trade-off relationship. Previously, weight space VI required one sampling of parameter set for each individual example. However, while reducing the number of sampling to one time for each minibatch, Flipout enabled moderate level of variance reduction in gradient estimator. It was performed by multiplying random sign to get multiple pseudo-independent sample. Practicality in computation was shown as it requires at most two times of Deterministic Neural Network forward time. This method is a general method which can be applied anytime we want to perform approximate inference in weight space.

Amortized Variational Inference(AVI) is a method which tries to learns a mapping from observed variable to the variational distribution of the latent variable rather than learning variational distribution for each latent variable.[5] To do so, AVI introduces recognition network which stores the information of previous inference results about the latent variables. Since AVI has been a hot keyword in conferences, there exists previous literatures which tries to apply AVI to induce the variational distribution of Bayesian models. To the best of our knowledge, most previous literatures were limited on inferring simple types of Bayesian models. For example, inferencing the posterior distribution of the parameters of the linear Bayesian models or the last layer of deep learning model. However, in this study, we further extend the former works by utilizing AVI to acquire approximate distribution of BNN model.

2. Problem Definition

Assume there exists model f_θ with parameter θ. Then, Meta-Learning problem can be defined as:

$$\min_\theta \mathbb{E}_{T\sim p(T)}\left[\mathbb{E}_{D\sim p_T(D)}[L(D;f_\theta)]\right]$$

In the above formular, $p(T)$ marks task distribution.

For example, in few-shot classification, task distribution can be empirically represented by defining classification problem with subset classes sampled from the entire classes. $p_T(D)$ means data distribution given arbitrary task problem. And L denotes loss function for each task and data given model f_θ. In the similar context such as Multi-task Learning, loss function L may be assumed different. However, it is notable that Few-shot classification setting is a limited setting which assumes the form of task and also its loss function L is identical. In typical Bayesian Meta-Learning, especially in model-based setting, below is the objective we want to obtain.

$$\mathrm{E}_{T\sim p(T)}[p(\theta|T)]$$

In other words, we want to inference posterior distribution of the model parameters conditioned on the task. We applied Amortized Variational Inference to approximate task conditional posterior.

Let z be latent variable. Given the data D, define Variational distribution of latent variable z as $q_\varphi(z|D)$. Then, AVI defines a problem like below:

$$\min_{\varphi} \mathrm{E}_{D\sim p(D)}\left[\mathrm{E}_{z\sim q_\varphi(z|D)}\left[\ln\frac{q_\varphi(z|D)}{p(z,D)}\right]\right]$$

This is a problem which we minimize the approximation error of variational distribution for every data and every tasks with respect to variational parameter. In this study, we tackled the Meta-Learning problem with Amortized Variational Inference. Our final problem formulation is like the below:

$$\min_{\varphi}\mathrm{E}_{T\sim p(T)}\left[\mathrm{E}_{D\sim p_T(D)}\left[KL\left(q_\varphi(\theta|D)||p(\theta|D)\right)\right]\right]$$
$$\equiv \min_{\varphi}\mathrm{E}_{T\sim p(T)}\left[\mathrm{E}_{D\sim p_T(D)}\left[\mathrm{E}_{\theta\sim q_\varphi(\theta|D)}\left[\ln\frac{q_\varphi(\theta|D)}{p(\theta|D)}\right]\right]\right]$$

It is trivial to show this term is a Variational Lower Bound of standard Meta-Learning problem. Detailed description about variational parameter φ and variational distribution $q_\varphi(\theta|T)$ is explained in the methodology part.

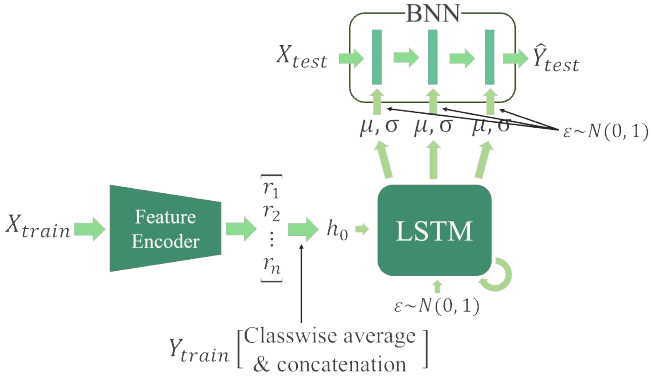

Figure 1. Diagram on our method.

3. Methodology

In this study, we experimented to represent the variational distribution $q_\varphi(\theta|T)$ with Long Short-term Memory(LSTM) model to consider the sequential nature of BNN models. Thus φ indicateds the parameter of LSTM models. We fixed the architecture of BNN model and performed the Amortized Variational Inference to induce the mean and the variance of the each parameter of the BNN model. Each time step in LSTM model considered the variational parameters of each BNN layer. By doing so, we distinct our approximate inference scheme with standard Mean-field Variational Inference for BNN parameter in that our method can inherit the knowledge about the complex relationship among BNN parameters.

The initial state of LSTM model, h_0, was obtained by extracting the permutation-invariant representation from the support set(or task train set). We feed gaussian noise as the input at each LSTM time step. This can be a considered as "Hyperprior" which models the uncertainty in variational parameters. Thus, our method has connection to Hierarchical Bayesian approach and incoporates epistemic uncertainties from multiple sources. This was motivated to induce the robustness, which is a core strength in Hierarchical Bayesian. Generally speaking, BNN models suffer from extreme sensitiveness toward prior distribution selection. To reduce this unstability, we pass the burden of selecting the prior distribution and initialization for BNN models to LSTM models.

For each single optimization step, we Variational parameters for Variational distribution approximating the posterior of BNN are sampled by LSTM and the BNN parameters are sampled from the variational distribution. And then, gradients are estimated from the inference procedure performed by LSTM model to conduct inner gradient step of BNN model. While Inner gradient step is repeated for a pre-defined number of times, gradient information of LSTM models are delivered by backpropagation. Finally, LSTM model are updated utilizing the stacked gradients.

4. Implementation Detail

In our study, we applied weight space inference for BNN approximate inference which requires sampling for entire parameter set for each model sampling. Since this procedure is the bottleneck of computation, we adapt Flipout[4] to maximize the efficiency of sampling.

To extract the information of the task, we utilized another encoder model which calculates permutation-invariant representation from the task data. We first averaged the feature extract from the EfficientNet model grouped by the class. After that, we concatenate those representation per class in randomly permuted order to obtain permutation-invariant representation h_0 for LSTM initial state. This was motivated approach from Neural Process variants to obtain of latent variables.[7]

Our study defines a optimization problem with respect to variational parameter φ and introduces latent variable model to store the knowledge among varying tasks. This type of methods are called Black-box based

Meta-Learning approach. Black-box based Meta-Learning approach performs inner optimization steps to adapt to our task proposal latent variable to new task environment. In this study, original goal was to completely eliminate the inner gradient step(or adaption phase). However, during our study, we observed empirically that inner optimization step is essential step in Black-box based Meta-Learning. For example, without inner optimization step, solving few shot classification problem showed accuracy close to 1/<number of classes> which implies almost random assignment for class prediction. As a result, our algorithm had to include inner optimization step which construct nested loops. But we tried to minimize the numbers of inner optimization update and experimented in settings with 1 and 5 times for update step.

We further searched optimal scale of Kullback-Leiblur divergence between variational distribution with standard normal distribution by multple experiments. It is worth denotable that this scale hyperparmeter has significant effect to the performance of Amortized Variational Inference.

5. Result

This parts records some minor results about the quantitative evaluation result about our method. We compare the benchmark performance in few-shot classification problem in Omniglot and miniImagenet dataset with other strong baseline methods.

Table 1 Comparison on Omniglot dataset

	5-way		20-way	
	1-shot	5-shot	1-shot	5-shot
Matching Network[10]	98.1%	98.9%	93.8%	98.7%
Prototypical Network[11]	98.8%	99.7%	**96.0%**	**98.9%**
MAML[12]	98.7%	**99.9%**	95.8%	**98.9%**
Meta-SGD[13]	**99.5%**	**99.9%**	95.9%	**98.9%**
Our Method (1 Inner Gradient Step)	68.5%	72.6%	29.3%	34.0%
Our Method (5 Inner Gradient Step)	87.4%	86.1%	54.3%	54.7%

Table 2 Comparison on miniImagenet dataset

	5-way	
	1-shot	5-shot
Matching Network[10]	46.6%	60.0%
Prototypical Network[11]	49.4%	**68.2%**
MAML[12]	48.7%	63.1%
Meta-SGD[13]	**50.4%**	64.0%
Our Method (1 Inner Gradient Step)	38.9%	48.9%
Our Method (5 Inner Gradient Step)	35.8%	52.1%

Our experiment result showed lagging performance compared to current State-of-The-Art method. However, result showed that our method extracts meaninful information which can be generalizable to multiple tasks. We expected our method to be more prominent if the BNN model architectures and hyperparameters are tuned to fit the hypothesis of the benchmark dataset. Another intuition we can obtain from the experiment result was that 5-shot performance showed improvement in trend compared to 1-shot performance. Thus our method is utilizing the task data as a useful context for solving few-shot classification problem.

6. Conclusions

Although our study did not showed quantitatively competitive result compared to State-of-the-Art Meta-Learning baselines, we validated that Bayesian Meta-Learning problem can be tackled with Amortized Variational Inference method. And also, we proposed a scheme to model the variational distribution of BNN posterior while considering the sequential structure of Deep Learning model architectures.

Further study can be directed to more sophisticated modification of the BNN architectures which involves latest deep learning modules. Another type of variational distribution rather than LSTM model will be a possible topics to research. Also, testing the Uncertainty Quantification ability for each newly observed task is another direction to validate our work.

7. Reference

[1] Florian Wenzel et al, “How Good is the Bayes Posterior in Deep Neural Networks Really?”, The International Conference on Machine Learning, 2020.

[2] Pavel Izmailov et al, “Subspace Inference for Bayesian Deep Learning”, The Conference on Uncertainty in Artificial Intelligence, 2019.

[3] Michael W. Dusenberry et al, “Efficient and Scalable Bayesian Neural Nets with Rank-1 Factors”, The International Conference on Machine Learning, 2020.

[4] Yeming Wen, Paul Vicol, Jimmy Ba, Dustin Tran and Roger Grosse, “Flipout: Efficient Pseudo-Independent Weight Perturbations on Mini-Batches”, The International Conference on Learning Representations, 2018.

[5] Diederik P Kingma and Max Welling, “Auto-Encoding Variational Bayes”, arXiv:1312.6114v10.

[6] Sachin Ravi and Alex Beatson, “Amortized Bayesian Meta-Learning”, The International Conference on Learning Representations, 2019.

[7] Sebastian W. Ober and Laurence Aitchison, “Global inducing point variational posteriors for Bayesian neural networks and deep Gaussian processes”, arXiv:2005.08140v4.

[8] Cheng Zhang, Judith Butepage, Hedvig Kjellstrom and Stephan Mandt, “Advances in Variational Inference”, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 41, no. 8, 2019.

[9] Garnelo et al, “Neural Processes”, The International Conference on Machine Learning, 2018.

[10] Vinyals et al, “Matching networks for one shot learning”, Advances in Neural Information Processing Systems, 2016.

[11] Jake Snell, Kevin Swersky and Richard S. Zemel, “Prototypical Networks for Few-shot Learning”, Advances in Neural Information Processing Systems, 2017.

[12] Finn et al, “Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks”, The International Conference on Machine Learning, 2017.

[13] Zhenguo Li, Fengwei Zhou, Fei Chen and Hang Li, “Meta-SGD: Learning to Learn Quickly for Few-Shot Learning”, arXiv:1707.09835v2.