Improving MFVI in Bayesian Neural Networks with Empirical Bayes: a Study with Diabetic Retinopathy Diagnosis

Ranganath Krishnan¹ Mahesh Subedar¹ Omesh Tickoo¹ Angelos Filos² Yarin Gal²

¹Anticipatory Computing Lab, Intel Labs
²Oxford Applied and Theoretical Machine Learning, University of Oxford

Abstract

Specifying meaningful weight priors for variational inference in Bayesian deep neural network (DNN) is a challenging problem, particularly for scaling to larger models involving high dimensional weight space. We evaluate the recently proposed, MOPed Priors with Empirical Bayes using DNN (MOPED) method for Bayesian DNNs within the Bayesian Deep Learning (BDL) benchmarking framework. MOPED enables scalable VI in large models by providing a way to choose informed prior and approximate posterior distributions for Bayesian neural network weights using Empirical Bayes framework. We benchmark MOPED with mean field variational inference on a real-world diabetic retinopathy diagnosis task and compare with state-of-the-art BDL techniques. We demonstrate MOPED method provides reliable uncertainty estimates while outperforming state-of-the-art methods, offering a new strong baseline for the BDL community to compare on complex real-world tasks involving larger models.

1. Introduction

Variational inference (VI) [¹⁻³] is an approximation technique to learn the posterior distribution. VI formulates the Bayesian inference problem as an optimization-based approach which lends itself to the stochastic gradient descent based optimization used in training DNN models. The generalized formulations of VI [⁴⁻⁹] has renewed interest in Bayesian neural networks.

Variational inference for Bayesian DNN involves specifying prior distributions and approximate posterior distributions for weights. In a pure Bayesian approach, a prior distribution is specified before any data is observed. But specifying meaningful priors in large Bayesian DNN models is a challenging problem [¹⁰], as it is practically difficult to have prior belief on millions of parameters that we intend to estimate through Bayesian inference. Empirical Bayes [¹¹⁻¹³] methods estimate prior distribution from the data, which is in contrast to typical Bayesian approach. Also, scaling variational inference in Bayesian DNNs to practical applications involving large-scale datasets and deeper models in high dimensional weight space is an ongoing research problem. On the contrary, DNNs are shown to have structural benefits [¹⁴] which helps them in learning complex models on larger datasets. The convergence speed and performance [¹⁵] of DNN models heavily depend on the initialization of model weights and other hyperparameters. The transfer learning approaches [¹⁶] demonstrate the benefit of fine tuning the pretrained DNN models from adjacent domains in order to achieve faster convergence and better accuracies.

Based on Empirical Bayes (EB) and transfer learning approaches, we have proposed MOPED [¹⁷] to scale VI to large Bayesian DNN models. MOPED is a simple and yet efficient method to specify informed priors and approximate posteriors, which in our experiments with complex real-world tasks has shown to provide good initialization for weights. The original formulation of Empirical Bayes dates back to 1950s [¹¹], since then many parametric formulations have been proposed and used in wide variety of applications. We use a parametric EB approach in our method for
mean field variational inference (MFVI) in Bayesian DNN, where weights are modeled with fully factorized Gaussian distribution. We evaluate MOPED-MFVI within Bayesian Deep Learning Benchmarks (BDL-benchmarks) \[18\] framework on a real-world diabetic retinopathy diagnosis task. BDL-benchmarks is an open-source framework for evaluating deep probabilistic machine learning models and their application to real-world problems. BDL-benchmarks assess both the scalability and effectiveness of different techniques for uncertainty estimation.

**Contribution:** Our main contribution in this paper is to propose a new strong baseline (MOPED-MFVI) for Bayesian DNNs evaluated within the BDL-benchmarks framework \[18\] on Diabetic Retinopathy diagnosis. Our empirical results indicate our method provides reliable uncertainty estimates and achieves better model performance than state-of-the-art baselines in BDL-benchmarks, offering a strong baseline for the BDL community to compare on real-world tasks.

2. MOPED: specifying weights in Bayesian DNN using Empirical Bayes framework

MOPED (MOdel Priors from Empirical Bayes using DNN) provides a way for specifying meaningful prior distributions and approximate posterior distributions over weights in Bayesian DNNs using Empirical Bayes (EB) framework. EB methods intend to combine the strengths of frequentist and Bayesian statistical approaches, and considered as an approximation to a fully Bayesian treatment of a hierarchical Bayes model.

We formulate a two-stage hierarchical modeling approach, first find the maximum likelihood estimates of weights with DNN, and then set the weight priors using empirical Bayes approach to infer the posterior with variational inference.

We illustrate our approach on mean-field variational inference (MFVI). For MFVI in Bayesian DNNs, weights are modeled with fully factorized Gaussian distribution i.e. each weight is independently sampled from the Gaussian distribution \( w \sim \mathcal{N}(\mu, \sigma) \). In Bayesian DNNs of complex architectures involving very high dimensional weight space, initial choice of \( \mu \) and \( \sigma \) can be very sensitive for variational inference. MOPED method proposes to specify prior \( p(w) \) and approximate posterior \( q(w) \) for each weight as mentioned in Equations 1.

\[
p(w) = \mathcal{N}(w_{\text{MLE}}, I) \\
q(w) = \mathcal{N}(\mu, \sigma) \\
\mu := w_{\text{MLE}}; \quad \sigma := \delta |w_{\text{MLE}}|
\]

where, \( \delta \) is initial perturbation factor (hyperparameter) in terms of decimal percentage of the mean.

**Related work:** Choosing weight priors in Bayesian DNN is an active area of research \[19\]-\[22\]. Dziugaite et al. \[23\] use a method of setting prior mean as maximum likelihood estimate, which gives a tight generalization bound. Nguyen et al. \[24\] use prior with zero mean and unit variance, and initialize the optimizer at the mean of the MLE model and a very small initial variance for small-scale MNIST experiments. Wu et al. \[19\] introduced hierarchical prior for parameters and a novel EB procedure for automatically selecting prior variances.

3. Experiments and Results

We evaluate our method on diabetic retinopathy diagnosis task in BDL-benchmarks \[18\]. We use the same model architecture and experiment setup described in BDL-benchmarks framework. The Bayesian DNN model used in experiments is a variant of VGG architecture (same as MFVI baseline in BDL-benchmarks), where the weights in variational layers are modeled with mean-field Gaussian distribution using Flipout \[25\]. The number of trainable parameters is twice as compared to a deterministic DNN. The model is trained using Adam adaptive optimizer with an initial learning rate of \( 4e^{-4} \) and batch size of 64. We trained for 60 epochs (40 epochs to obtain the maximum likelihood estimates for weights using DNN, and 20 epochs with MFVI after specifying approximate posteriori and priori using MOPED method as given by Equations 1, with \( \delta=0.3 \)). In order to obtain statistically significant results, we trained five independent models with MOPED-MFVI method.
Table 1: Comparison of Area under the receiver-operating characteristic curve (AUC) and classification accuracy as a function of retained data (based on predictive uncertainty). The results for the BDL baseline methods other than MOPED-MFVI are presented from [18]. MOPED Mean-field VI outperforms other baselines in terms of accuracy vs retained data.

<table>
<thead>
<tr>
<th>Method</th>
<th>50% data retained</th>
<th>70% data retained</th>
<th>100% data retained</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>AUC</td>
<td>Accuracy</td>
<td>AUC</td>
</tr>
<tr>
<td>MC Dropout</td>
<td>87.8±1.1</td>
<td>91.3±0.7</td>
<td>85.2±0.9</td>
</tr>
<tr>
<td>Mean-field VI</td>
<td>86.6±1.1</td>
<td>88.1±1.1</td>
<td>84.0±1.0</td>
</tr>
<tr>
<td>Deep Ensembles</td>
<td>87.2±0.9</td>
<td>89.9±0.9</td>
<td>84.9±0.8</td>
</tr>
<tr>
<td>Deterministic</td>
<td>84.9±1.1</td>
<td>86.1±0.6</td>
<td>82.3±1.2</td>
</tr>
<tr>
<td>Ensemble MC Dropout</td>
<td><strong>88.1±1.2</strong></td>
<td>92.4±0.9</td>
<td><strong>85.4±1.0</strong></td>
</tr>
<tr>
<td>MOPED Mean-field VI</td>
<td>87.3±0.8</td>
<td><strong>93.4±0.4</strong></td>
<td>84.4±0.6</td>
</tr>
<tr>
<td>Random referral</td>
<td>81.8±1.2</td>
<td>84.8±0.9</td>
<td>82.0±1.3</td>
</tr>
</tbody>
</table>

MOPED-MFVI is compared with state-of-the-art baseline methods in BDL-benchmarks suite on diabetic retinopathy detection task [26]. The baselines include Monte Carlo (MC) dropout [8], MFVI with randomly initialized weight parameters [5, 4], deep ensembles [27], ensemble of MC dropout and deterministic DNN models. The results for baselines from BDL-benchmarks [18] is currently provided for medium scale experiment setup, so we have compared MOPED-MFVI in the same setting for fair comparison. In appendix section A.2, we provide the results for MOPED-MFVI in real-world experiment setting along with other real-world tasks in Section A.1. The evaluation methodology expects that the models with well-calibrated uncertainty improve their performance (accuracy and AUC) as most certain data is retained, while referring the uncertain predictions to expert doctors. Table 1 and Figure 1 provides quantitative evaluation of AUC and accuracy values for MOPED-MFVI and other BDL baseline methods. MOPED-MFVI outperforms other state-of-the-art BDL techniques in terms of accuracy with respect to retained data based on predictive uncertainty.

**Conclusion:** We offer a new strong baseline for the Bayesian Deep Learning community to compare on large scale real-world tasks. We demonstrated MOPED-MFVI improves MFVI in Bayesian neural networks, and outperforms state-of-the-art Bayesian deep learning techniques in BDL-benchmarks. The results support that new baseline provides better model performance and reliable uncertainty estimates on a real-world diabetic retinopathy diagnosis task. We will integrate the code and setup for MOPED-MFVI baseline into BDL-benchmarks framework.

†[https://github.com/OATML/bdl-benchmarks](https://github.com/OATML/bdl-benchmarks)
Acknowledgment: We like to thank Sebastian Farquhar (University of Oxford) for the valuable discussions and comments.

A Appendix

A.1 Additional experiments with various datasets and large-scale models

We evaluated MOPED-MFVI on additional real-world applications including video activity recognition, audio and image classification. We consider multiple model architectures with varying complexity to show the scalability of method in training deep Bayesian models.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Modality</th>
<th>Architecture</th>
<th>Complexity (# parameters)</th>
<th>Validation Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>UCF-101 Video</td>
<td>ResNet-101 C3D</td>
<td>170,838,181</td>
<td>0.851</td>
<td><strong>0.029</strong></td>
</tr>
<tr>
<td>UrbanSound8K Audio</td>
<td>VGGish</td>
<td>144,274,890</td>
<td>0.817</td>
<td><strong>0.143</strong></td>
</tr>
<tr>
<td>CIFAR-10 Image</td>
<td>ResNet-56</td>
<td>1,714,250</td>
<td>0.926</td>
<td>0.896</td>
</tr>
<tr>
<td>MNIST Image</td>
<td>LeNet</td>
<td>1,090,856</td>
<td>0.994</td>
<td>0.993</td>
</tr>
<tr>
<td>Fashion-MNIST Image</td>
<td>SCNN</td>
<td>442,218</td>
<td>0.921</td>
<td>0.906</td>
</tr>
</tbody>
</table>

Table 2: Accuracies for model architectures with different complexities and input modalities. MOPED-MFVI achieves similar or better accuracy as the deterministic DNNs while providing reliable uncertainty estimates. MFVI with random initialization has difficulty in converging to optimal solution (shown in red) for larger models, while MOPED-MFVI enables scalable VI in larger models.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Architecture</th>
<th>AUPR</th>
<th>AUROC</th>
</tr>
</thead>
<tbody>
<tr>
<td>UCF-101</td>
<td>ResNet-101 C3D</td>
<td><strong>0.0174</strong></td>
<td><strong>0.9967</strong></td>
</tr>
<tr>
<td>Urban Sound 8K</td>
<td>VGGish</td>
<td><strong>0.1166</strong></td>
<td><strong>0.9811</strong></td>
</tr>
<tr>
<td>CIFAR-10</td>
<td>ResNet-20</td>
<td><strong>0.9265</strong></td>
<td><strong>0.9941</strong></td>
</tr>
<tr>
<td>MNIST</td>
<td>LeNet</td>
<td><strong>0.9996</strong></td>
<td><strong>0.9999</strong></td>
</tr>
<tr>
<td>Fashion-MNIST</td>
<td>SCNN</td>
<td>0.9722</td>
<td><strong>0.9969</strong></td>
</tr>
</tbody>
</table>

Table 3: Comparison of area under precision-recall curve (AUPR) and receiver operating characteristic curve (AUROC) for models with varying complexities. MOPED-MFVI outperforms MFVI that was trained with random initialization.

Our experiments include: 1) ResNet-101 C3D for video activity classification on UCF-101 dataset, 2) VGGish for audio classification on UrbanSound8K dataset, 3) ResNet-20 and ResNet-56 architectures for the image classification on CIFAR-10 dataset, 4) LeNet for MNIST digit classification, and 5) Simple convolutional neural network (SCNN) consisting of two convolutional layers followed by two dense layers for image classification on Fashion-MNIST datasets. We implemented all of these Bayesian DNN models and trained them with MFVI and MOPED-MFVI.

In Table 2 classification accuracies for model architectures with various complexity are presented. Table 3 compares AUPR and AUROC for MFVI and MOPED-MFVI on various model architectures and datasets. Bayesian DNNs trained with MOPED-MFVI achieves similar or better accuracies as...
compared to equivalent DNN models. MFVI with random initialization has difficulty in converging to optimal solution for larger models (ResNet-101 C3D and VGGish) with hundreds of millions of trainable parameters. These results show that MOPED enables scalable VI and guarantees the training convergence even for the larger models.

A.2 Real-world experiment setting on Diabetic Retinopathy Diagnosis

In these experiments, visual inputs consisting RGB images of retinas with resolution 512x512 pixels is considered. Where as in medium scale experiment setting, RGB images of retinas with resolution 256x256 pixels are used. The AUC and accuracy evaluation for MOPED-MFVI at different percentage of retained data based on predictive uncertainty is shown in Table 4 and Figure 2.

<table>
<thead>
<tr>
<th>Method</th>
<th>50% data retained</th>
<th>70% data retained</th>
<th>100% data retained</th>
</tr>
</thead>
<tbody>
<tr>
<td>MOPED Mean-field VI</td>
<td>91.2±1.3</td>
<td>96.4±0.2</td>
<td>88.9±1.2</td>
</tr>
<tr>
<td></td>
<td>88.9±0.3</td>
<td>94.9±0.3</td>
<td>88.3±0.5</td>
</tr>
<tr>
<td></td>
<td>88.9±0.5</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 4: AUC and classification accuracy as a function of retained data (based on predictive uncertainty) for MOPED-MFVI in real-world experiment setting as described in [18].

Figure 2: MOPED-MFVI on diabetic retinopathy diagnosis task in real-world experiment setting. Accuracy and AUC plots for varied percentage of retained data based on predictive uncertainty. Shading shows the standard error from multiple MOPED-MFVI models that was trained independently.

References


