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

Ranganath Krishnan[1] **Mahesh Subedar**[1] **Omesh Tickoo**[1] **Angelos Filos**[2] **Yarin Gal**[2]

[1]Anticipatory Computing Lab, Intel Labs
[2]Oxford 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, MOdel 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

Fourth workshop on Bayesian Deep Learning (NeurIPS 2019), Vancouver, Canada.

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 μ and σ 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. $w_{_{MLE}}$ represents weights obtained through maximum likelihood estimation from DNN model of equivalent architecture. The prior mean and variance is set at $w_{_{MLE}}$ and unit variance respectively. The variational parameters μ and σ in q(w) is initialized using $w_{_{MLE}}$ and δ as mentioned below.

$$\begin{aligned} p(w) &= \mathcal{N}(w_{_{MLE}},\, I) \\ q(w) &= \mathcal{N}(\mu,\, \sigma) \\ \mu := w_{_{MLE}}\,; \quad \sigma &:= \delta\, |w_{_{MLE}}| \end{aligned} \tag{1}$$

where, δ 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 δ=0.3). In order to obtain statistically significant results, we trained five independent models with MOPED-MFVI method.

Method	50% data retained		70% data retained		100% data retained	
	AUC	**Accuracy**	**AUC**	**Accuracy**	**AUC**	**Accuracy**
MC Dropout	87.8±1.1	91.3±0.7	85.2±0.9	87.1±0.9	82.1±0.9	84.5±0.9
Mean-field VI	86.6±1.1	88.1±1.1	84.0±1.0	85.0±1.0	82.1±1.2	84.3±0.7
Deep Ensembles	87.2±0.9	89.9±0.9	84.9±0.8	86.1±1.0	81.8±1.1	84.6±0.7
Deterministic	84.9±1.1	86.1±0.6	82.3±1.2	84.9±0.5	82.0±1.0	84.2±0.6
Ensemble MC Dropout	**88.1±1.2**	92.4±0.9	**85.4±1.0**	88.1±1.0	**82.5±1.1**	85.3±1.0
MOPED Mean-field VI	87.3±0.8	**93.4±0.4**	84.4±0.6	**91.8±0.5**	82.1±0.2	**85.5±0.7**
Random referral	81.8±1.2	84.8±0.9	82.0±1.3	84.3±0.7	82.0±0.9	84.2±0.5

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.

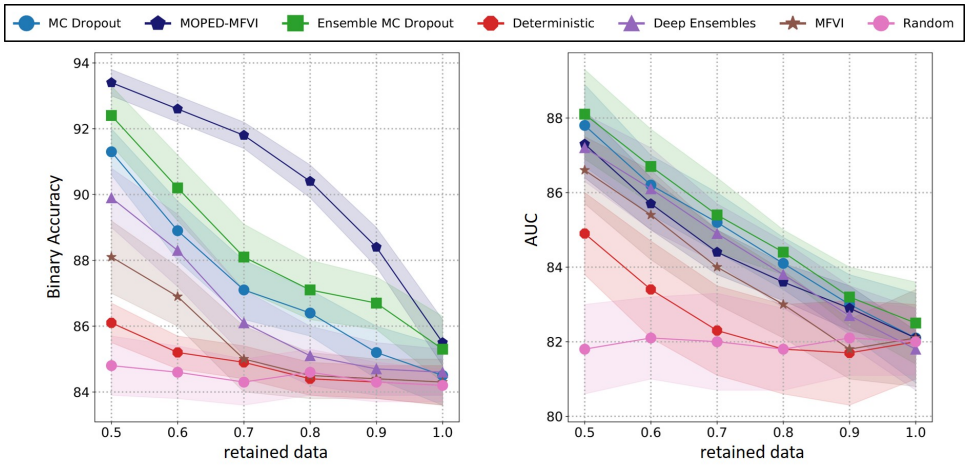

Figure 1: Benchmarking MOPED-MFVI with state-of-the-art BDL techniques on diabetic retinopathy diagnosis task in BDL-benchmarks [18]. Accuracy and area under the receiver-operating characteristic curve (AUC) plots for varied percentage of retained data based on predictive uncertainty. Shading shows the standard error.

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

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.

			Bayesian DNN Complexity	Validation Accuracy		
					Bayesian DNN	
Dataset	**Modality**	**Architecture**	**(# parameters)**	**DNN**	**MFVI**	**MOPED-MFVI**
UCF-101	Video	ResNet-101 C3D	170,838,181	0.851	**0.029**	**0.867**
UrbanSound8K	Audio	VGGish	144,274,890	0.817	**0.143**	**0.819**
CIFAR-10	Image	ResNet-56	1,714,250	0.926	0.896	**0.927**
		ResNet-20	546,314	0.911	0.878	**0.916**
MNIST	Image	LeNeT	1,090,856	0.994	0.993	**0.995**
Fashion-MNIST	Image	SCNN	442,218	0.921	0.906	**0.923**

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.

	Bayesian DNN	**AUPR**		**AUROC**	
Dataset	**Archiecture**	MFVI	MOPED-MFVI	MFVI	MOPED-MFVI
UCF-101	ResNet-101 C3D	**0.0174**	**0.9186**	**0.6217**	**0.9967**
Urban Sound 8K	VGGish	**0.1166**	**0.8972**	**0.551**	**0.9811**
CIFAR-10	ResNet-20	0.9265	**0.9622**	0.9877	**0.9941**
	ResNet-56	0.9225	**0.9799**	0.987	**0.9970**
MNIST	LeNet	**0.9996**	**0.9997**	**0.9999**	**0.9999**
Fashion-MNIST	SCNN	0.9722	**0.9784**	0.9962	**0.9969**

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 [28] for video activity classification on UCF-101[29] dataset, 2) VGGish[30] for audio classification on UrbanSound8K [31] dataset, 3) ResNet-20 and ResNet-56 [32] architectures for the image classification on CIFAR-10 [33] dataset, 4) LeNet for MNIST [34] 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 [35] 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.

	50% data retained		70% data retained		100% data retained	
Method	**AUC**	**Accuracy**	**AUC**	**Accuracy**	**AUC**	**Accuracy**
MOPED Mean-field VI	91.2±1.3	96.4±0.2	88.9±1.2	94.9±0.3	88.3±0.5	88.9±0.5

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].

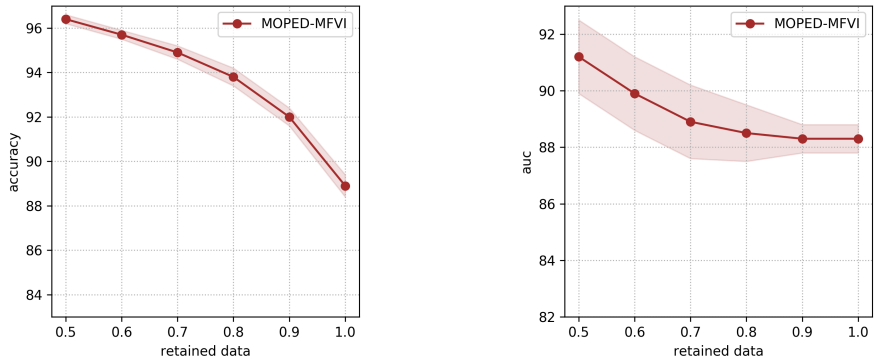

Figure 2: MOPED-MFVI on diabetic retinopathy diagnosis task in real-world experiment setting. Accuracy and AUC plots for varied percentage of retained data based on predictive uncertainty. Shading shows the standard error from multiple MOPED-MFVI models that was trained independently.

References

[1] Michael I Jordan, Zoubin Ghahramani, Tommi S Jaakkola, and Lawrence K Saul. An introduction to variational methods for graphical models. *Machine learning*, 37(2):183–233, 1999.

[2] Matthew D Hoffman, David M Blei, Chong Wang, and John Paisley. Stochastic variational inference. *The Journal of Machine Learning Research*, 14(1):1303–1347, 2013.

[3] David M Blei, Alp Kucukelbir, and Jon D McAuliffe. Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112(518):859–877, 2017.

[4] Alex Graves. Practical variational inference for neural networks. In *Advances in neural information processing systems*, pages 2348–2356, 2011.

[5] Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. Weight uncertainty in neural networks. *arXiv preprint arXiv:1505.05424*, 2015.

[6] Rajesh Ranganath, Sean Gerrish, and David M Blei. Black box variational inference. *arXiv preprint arXiv:1401.0118*, 2013.

[7] John Paisley, David Blei, and Michael Jordan. Variational bayesian inference with stochastic search. *arXiv preprint arXiv:1206.6430*, 2012.

[8] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pages 1050–1059, 2016.

[9] Yarin Gal. Uncertainty in deep learning. *University of Cambridge*, 2016.

[10] João Ferdinando Gomes de Freitas. *Bayesian methods for neural networks*. PhD thesis, University of Cambridge, 2003.

[11] Herbert Robbins. An empirical bayes approach to statistics. *Herbert Robbins Selected Papers*, pages 41–47, 1956.

[12] George Casella. Illustrating empirical bayes methods. *Chemometrics and intelligent laboratory systems*, 16(2):107–125, 1992.

[13] Bradley P Carlin and Thomas A Louis. *Bayes and empirical Bayes methods for data analysis*. Chapman and Hall/CRC, 2010.

[14] Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828, 2013.

[15] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT press, 2016.

[16] Hoo-Chang Shin, Holger R Roth, Mingchen Gao, Le Lu, Ziyue Xu, Isabella Nogues, Jianhua Yao, Daniel Mollura, and Ronald M Summers. Deep convolutional neural networks for computer-aided detection: Cnn architectures, dataset characteristics and transfer learning. *IEEE transactions on medical imaging*, 35(5): 1285–1298, 2016.

[17] Ranganath Krishnan, Mahesh Subedar, and Omesh Tickoo. Moped: Efficient priors for scalable variational inference in bayesian deep neural networks. *arXiv preprint arXiv:1906.05323*, 2019.

[18] Angelos Filos, Sebastian Farquhar, Aidan N. Gomez, Tim G. J. Rudner, Zachary Kenton, Lewis Smith, Milad Alizadeh, Arnoud de Kroon, and Yarin Gal. Benchmarking bayesian deep learning with diabetic retinopathy diagnosis, 2019.

[19] Anqi Wu, Sebastian Nowozin, Edward Meeds, Richard E Turner, José Miguel Hernández-Lobato, and Alexander L Gaunt. Deterministic variational inference for robust bayesian neural networks. 2018.

[20] Eric Nalisnick, José Miguel Hernández-Lobato, and Padhraic Smyth. Dropout as a structured shrinkage prior. In *International Conference on Machine Learning*, pages 4712–4722, 2019.

[21] Andrei Atanov, Arsenii Ashukha, Kirill Struminsky, Dmitriy Vetrov, and Max Welling. The deep weight prior. 2018.

[22] Shengyang Sun, Guodong Zhang, Jiaxin Shi, and Roger Grosse. Functional variational bayesian neural networks. *arXiv preprint arXiv:1903.05779*, 2019.

[23] Gintare Karolina Dziugaite and Daniel M Roy. Computing nonvacuous generalization bounds for deep (stochastic) neural networks with many more parameters than training data. *arXiv preprint arXiv:1703.11008*, 2017.

[24] Cuong V Nguyen, Yingzhen Li, Thang D Bui, and Richard E Turner. Variational continual learning. *arXiv preprint arXiv:1710.10628*, 2017.

[25] Yeming Wen, Paul Vicol, Jimmy Ba, Dustin Tran, and Roger Grosse. Flipout: Efficient pseudo-independent weight perturbations on mini-batches. *arXiv preprint arXiv:1803.04386*, 2018.

[26] Kaggle. Diabetic retinopathy detection challenge. `https://www.kaggle.com/c/diabetic-retinopathy-detection/overview/description`.

[27] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems*, pages 6402–6413, 2017.

[28] Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh. Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet? In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6546–6555, 2018.

[29] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012.

[30] Shawn Hershey, Sourish Chaudhuri, Daniel PW Ellis, Jort F Gemmeke, Aren Jansen, R Channing Moore, Manoj Plakal, Devin Platt, Rif A Saurous, Bryan Seybold, et al. Cnn architectures for large-scale audio classification. In *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*, pages 131–135. IEEE, 2017.

[31] Justin Salamon, Christopher Jacoby, and Juan Pablo Bello. A dataset and taxonomy for urban sound research. In *Proceedings of the 22nd ACM international conference on Multimedia*, pages 1041–1044. ACM, 2014.

[32] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[33] Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. Technical report, Citeseer, 2009.

[34] Yann LeCun, Léon Bottou, Yoshua Bengio, Patrick Haffner, et al. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[35] Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.