Efficient Priors for Scalable Variational Inference in Bayesian Deep Neural Networks

Ranganath Krishnan* ranganath.krishnan@intel.com
Mahesh Subedar* mahesh.subedar@intel.com
Omesh Tickoo omesh.tickoo@intel.com

Intel Labs
Hillsboro, OR (USA)

Abstract

Stochastic variational inference for Bayesian deep neural networks (DNNs) requires specifying priors and approximate posterior distributions for neural network weights. Specifying meaningful weight priors is a challenging problem, particularly for scaling variational inference to deeper architectures involving high dimensional weight space. Based on empirical Bayes approach, we propose Bayesian MOdel Priors Extracted from Deterministic DNN (MOPED) method to choose meaningful prior distributions over weight space using deterministic weights derived from the pretrained DNNs of equivalent architecture. We empirically evaluate the proposed approach on real-world applications including image classification, video activity recognition and audio classification tasks with varying complex neural network architectures. The proposed method enables scalable variational inference with faster training convergence and provides reliable uncertainty quantification.

1. Introduction

Current deep neural networks (DNNs) make overly confident decisions and do not provide uncertainty estimates, which is crucial for informed decision-making. Probabilistic Bayesian models provide principled ways to gain insight about data and capture reliable uncertainty estimates. Bayesian neural networks [20, 6] have allowed bridging deep learning and probabilistic Bayesian theory to leverage the strengths of both methodologies. Variational inference (VI) [12] is an analytical approximation technique to learn the posterior distribution of weights in Bayesian neural networks. VI methods formulate the Bayesian inference problem as an optimization-based approach which lends itself to the stochastic gradient descent (SGD) based optimization used in training DNN models, referred as stochastic variational inference (SVI). Variational inference with generalizable formulations [9, 3, 21] has renewed interest in Bayesian neural networks.

Variational inference for Bayesian neural networks involves choosing prior distributions and approximate posterior distributions for neural network weights. Specifying meaningful weight priors is a challenging problem and is an active area of research [30, 19, 1, 28]. Also, scaling variational inference in deeper model architectures involving high dimensional weight space is an open problem. In [14, 27], hybrid Bayesian DNN architectures are used for complex computer vision tasks to achieve scalable variational inference by balancing complexity of the model, while providing benefits of Bayesian inference.

Our main contribution is to propose MOPED, a simple and yet efficient method for initializing weight priors to enable scalable variational inference in Bayesian DNNs. Inspired by empirical Bayes [22, 4] methods and transfer learning [25] approaches, we propose to specify meaningful priors for Bayesian DNNs based on the deterministic weights of pretrained DNN models obtained from maximum likelihood estimates. The empirical results indicate proposed approach guarantees training convergence for the models along with better uncertainty estimates without sacrificing the accuracies provided by deterministic DNNs.

2. Variational Inference in Bayesian DNNs

Given training dataset $D = \{x, y\}$ with inputs $x = \{x_1, ..., x_N\}$ and their corresponding outputs $y = \{y_1, ..., y_N\}$, in parametric Bayesian setting we would like to infer a distribution over parameters w as a function $y = f_w(x)$ that represents the DNN model. A prior distribution is assigned over the weights $p(w)$ that captures our prior belief as to which parameters would have likely generated the outputs before observing any data. Given the evidence data $p(y|x)$, prior distribution and model likelihood $p(y \mid x, w)$, the goal is to infer the posterior distribution over the weights

*Equal contribution.

$p(w|D)$. Variational inference approximates a complex probability distribution $p(w|D)$ with a simpler distribution $q_\theta(w)$, parameterized by variational parameters θ while minimizing the Kullback-Leibler (KL) divergence [2, 7]. During the training phase in Bayesian DNNs, variational inference optimizes the log evidence lower bound (ELBO) (Equation 1) by performing stochastic sampling of weight parameters from $q_\theta(w)$ and $p(w)$ distributions, while minimizing the KL-divergence between the two distributions.

$$\mathcal{L} := \int q_\theta(w)\, log\, p(y|x,w)\, dw - KL[q_\theta(w)||p(w)] \quad (1)$$

3. MOPED: Efficient priors for Bayesian DNN

The weights derived from pretrained deterministic models through maximum likelihood estimation can provide the best indication of prior knowledge about the probability distributions of weights in Bayesian DNNs before inferring the posteriors of corresponding weights through stochastic variational inference. We propose to initialize parameters of the weight priors in Bayesian DNNs based on the weights obtained from pretrained DNNs of equivalent neural network architectures.

We illustrate our proposed approach on mean-field variational inference. For mean-field variational inference in Bayesian DNNs, each weight is independently sampled from the Gaussian distribution $w = \mathcal{N}(\overline{w}, \sigma)$, where $\overline{w}$ is mean and variance $\sigma = \log(1 + exp(\rho))$. In order to ensure non-negative variance, σ is expressed in terms of softplus function with unconstrained parameter ρ. Here we propose to choose the $\overline{w}$ mean of weight priors from the deterministic weights of pretrained DNN of equivalent architecture.

$$\begin{aligned} \overline{w} &= w_d; \quad \rho \sim \mathcal{N}(\overline{\rho}, \Delta\rho) \\ w &\sim \mathcal{N}(w_d,\, \log(1 + e^{\rho})) \end{aligned} \quad (2)$$

where, w_d represents weights obtained from deterministic DNN model, and ($\overline{\rho}$, $\Delta\rho$) are hyper parameters (mean and variance of Gaussian perturbation for ρ).

For Bayesian DNNs of complex architectures involving very high dimensional weight space, choice of ρ can be sensitive as values of the weights can vary by large margin with each other. So, we propose to initialize the weight priors with w_d and ρ given in Equation 3:

$$\begin{aligned} \overline{w} &= w_d\,; \quad \rho = \log(e^{\delta|w_d|} - 1) \\ w &\sim \mathcal{N}(w_d,\ \delta \mid w_d \mid)) \end{aligned} \quad (3)$$

where, δ is initial perturbation factor for the weight in terms of percentage of the pretrained deterministic weight values.

4. Experiments

We demonstrate the benefits of MOPED method for variational inference with extensive empirical experiments. We showcase the proposed MOPED method helps Bayesian DNN architectures to achieve better model performance, faster training convergence and reliable uncertainty estimates. We evaluate proposed method on real-world applications including image and audio classification, and video activity recognition. We consider multiple architectures with varying complexity to show the scalability of method in training deep Bayesian models. Our experiments include: a) LeNet architecture for MNIST [17] digit classification, b) Simple convolutional neural network (SCNN) consisting of two convolutional layers followed by two dense layers for image classification on Fashion-MNIST (F-MNIST) [31] datasets, b) ResNet-20 and ResNet-56 architectures for the image classification on CIFAR-10 [15] dataset, c) VGGish[11] for audio classification on Urban-Sound8K [24] dataset and d) ResNet-101 C3D [10] for video activity classification on UCF-101[26] dataset.

We implemented above Bayesian DNN models and trained them using Tensorflow and Tensorflow-Probability [5] frameworks. The variational layers are modeled using Flipout [29], an efficient method that decorrelates the gradients within a mini-batch by implicitly sampling pseudo-independent weight perturbations for each input. The model weights obtained from the trained DNN models are used in MOPED method to initialize Gaussian priors over weights (Equation 2 and 3), as described in Section 3.

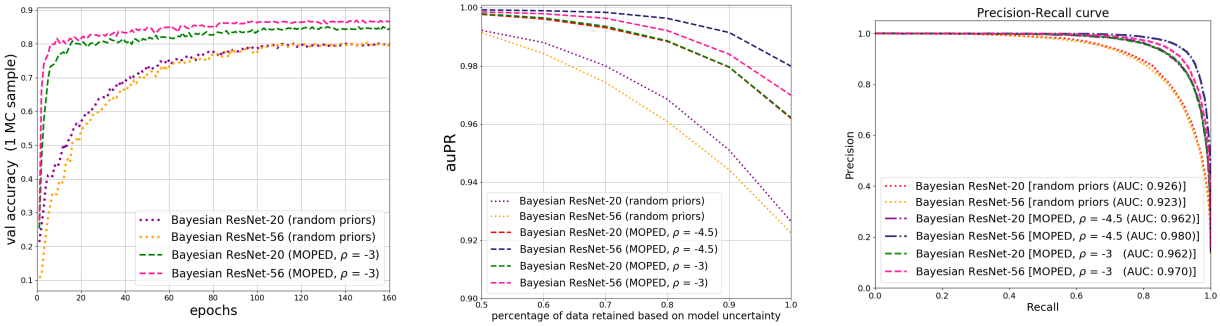
During inference phase, predictive distributions are obtained by performing multiple stochastic forward passes over the network while sampling from posterior distribution of the weights (40 Monte Carlo samples in our experiments). We evaluate the model uncertainty using Bayesian active learning by disagreement (BALD) [13], which quantifies mutual information between parameter posterior distribution and predictive distribution. Following [18], quantitative comparison of uncertainty estimates are made by calculating area under the curve (AUC) of precision-recall (auPR) values by retaining different percentages (0.5 to 1.0) of most certain test samples (i.e. ignoring most uncertain predictions based on model uncertainty estimates).

4.1. Weight priors with MOPED

In Table 1, classification accuracies for architectures with increasing complexity are presented. Bayesian DNNs with priors initialized with MOPED method achieves similar or better predictive accuracies as compared to equivalent DNN models. Bayesian DNNs with random initialization of Gaussian priors has difficulty in converging to optimal solution for the complex architectures (ResNet-101 C3D and VGGish) with hundreds of millions of trainable parameters. It is evident from these results that MOPED method guarantees the training convergence even for the complex models when trained DNN model of equivalent architecture is available.

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

Table 1: Comparison of the accuracies for architectures with different complexities and input modalities. MOPED method obtain reliable uncertainty estimates from Bayesian DNNs while achieving similar or better accuracy as deterministic DNNs.



(a) Training convergence curves (b) AuPR curves (c) Precision-recall

Figure 1: Comparison of MOPED and random initialization of priors for Bayesian ResNet-20 and ResNet-56 architectures. (a) training convergence, (b) AuPR for different percentage of retained data based on model uncertainty and (c) precision-recall plots.

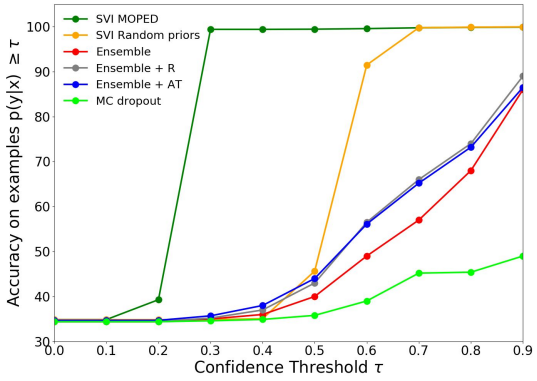
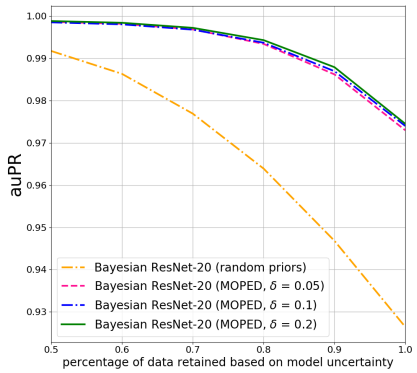


Figure 2: Accuracy vs Confidence curves: Networks trained on MNIST and tested on both MNIST and the NotMNIST (out-of-distribution) test sets.

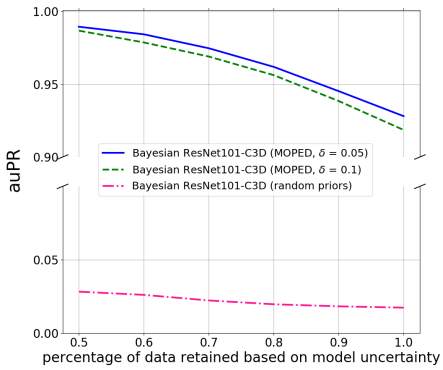
In Figure 1, comparison of MOPED and random initialization of priors is shown for Bayesian ResNet-20 and ResNet-56 architectures trained on CIFAR-10 dataset. The auPR plots [18] capture the precision-recall AUC values for different percentage of most certain predictions based on the model uncertainty estimates. Figure 1 (a) shows the faster convergence of MOPED method, while achieving the better accuracy values. Figure 1 (b) shows that MOPED method provides higher auPR values than the random initialization and also that auPR increases as most uncertain predictions are ignored based on the model uncertainty, indicating the method enables superior performance and reliable uncertainty estimates. In Figure 1 (c), the precision-recall plots show MOPED method provides better precision-recall values compared to random initialization of priors. We show the results for different selection of ρ (details are in Section 3) values. In Figure 3, we show auPR plots with different δ values as mentioned in Equation 3.

4.2. Robustness to out-of-distribution data

In order to evaluate robustness of our method (SVI MOPED) and usefulness of statistical inference for decision making, we compare state-of-the-art probabilistic deep

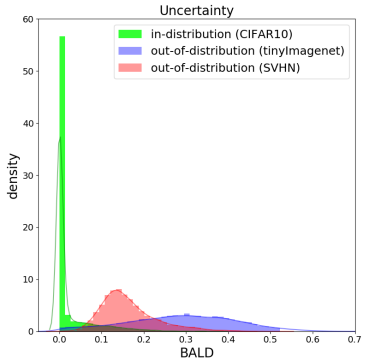


(a) Bayesian ResNet-20 (CIFAR-10)

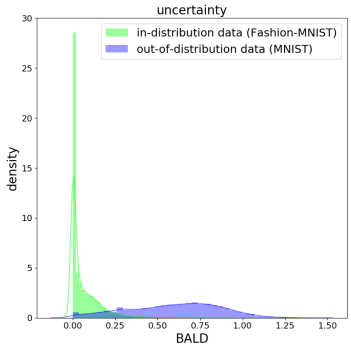


(b) Bayesian ResNet-101 C3D (UCF-101)

Figure 3: Precision-recall AUC (auPR) plots with different δ scale factors for initializing variance values in MOPED method.



(a) Model uncertainty (Bayesian ResNet-56)



(b) Model uncertainty (Bayesian SCNN)

Figure 4: Density histograms obtained from in- and out-of-distribution samples.

learning methods for prediction accuracy as a function of model confidence. Following the experiments in [16], we trained our model on MNIST training set and tested it on a mix of examples from MNIST and NotMNIST[1] (out-of-distribution) test set. The accuracy as a function of confidence plots should increase monotonically, as higher accuracy is expected for more confident results. A robust model should provide low confidence for out-of-distribution samples while providing high confidence for correct prediction from in-distribution samples. The proposed variational inference method with MOPED priors provides more robust results as compared to the MC Dropout [7] and deep model ensembles [16] approaches (shown in Figure 2).

We evaluate the uncertainty estimates obtained from MOPED method to detect out-of-distribution data. Out-of-distribution samples are data points which fall far off from the training data distribution. We evaluate two sets of out-of-distribution detection experiments. In the first set, we use CIFAR-10 as the in-distribution samples trained using ResNet-56 Bayesian DNN model. TinyImageNet [23] and SVHN [8] datasets are used as out-of-distribution samples which were not seen during the training phase. The density histograms (area under the histogram is normalized to one) for uncertainty estimates obtained from the Bayesian DNN models are plotted in Figure 4. The density histograms in Figure 4 (a) & (b) indicate higher uncertainty estimates for the out-of-distribution samples and lower uncertainty values for the in-distribution samples. A similar trend is observed in the second set using Fashion-MNIST([31])) as in-distribution and MNIST as the out-of-distribution data. These results confirm the uncertainty estimates obtained from proposed MOPED method are reliable and can identify out-of-distribution data.

5. Conclusions

We proposed a simple and efficient method to choose weight priors for variational inference in Bayesian DNN. We demonstrated with thorough empirical experiments that MOPED enables scalable variational inference for Bayesian DNNs while achieving faster training convergence, and provides reliable uncertainty quantification without compromising on the accuracy provided by the deterministic DNNs. We showed the uncertainty estimates obtained from the proposed method are reliable to identify out-of-distribution data.

[1] http://yaroslavvb.blogspot.co.uk/2011/09/notmnist-dataset.html

References

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

[2] Christopher M Bishop. Pattern recognition and machine learning (information science and statistics) springer-verlag new york. *Inc. Secaucus, NJ, USA*, 2006.

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

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

[5] Joshua V Dillon, Ian Langmore, Dustin Tran, Eugene Brevdo, Srinivas Vasudevan, Dave Moore, Brian Patton, Alex Alemi, Matt Hoffman, and Rif A Saurous. Tensorflow distributions. *arXiv preprint arXiv:1711.10604*, 2017.

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

[7] 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.

[8] Ian J Goodfellow, Yaroslav Bulatov, Julian Ibarz, Sacha Arnoud, and Vinay Shet. Multi-digit number recognition from street view imagery using deep convolutional neural networks. *arXiv preprint arXiv:1312.6082*, 2013.

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

[10] 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.

[11] 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.

[12] 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.

[13] Neil Houlsby, Ferenc Huszár, Zoubin Ghahramani, and Máté Lengyel. Bayesian active learning for classification and preference learning. *arXiv preprint arXiv:1112.5745*, 2011.

[14] Ranganath Krishnan, Mahesh Subedar, and Omesh Tickoo. Bar: Bayesian activity recognition using variational inference. *arXiv preprint arXiv:1811.03305*, 2018.

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

[16] 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.

[17] 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.

[18] Christian Leibig, Vaneeda Allken, Murat Seçkin Ayhan, Philipp Berens, and Siegfried Wahl. Leveraging uncertainty information from deep neural networks for disease detection. *Scientific reports*, 7(1):17816, 2017.

[19] 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.

[20] Radford M Neal. *Bayesian learning for neural networks*, volume 118. Springer Science & Business Media, 2012.

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

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

[23] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015.

[24] 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.

[25] 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.

[26] 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.

[27] Mahesh Subedar, Ranganath Krishnan, Paulo Lopez Meyer, Omesh Tickoo, and Jonathan Huang. Uncertainty aware audiovisual activity recognition using deep bayesian variational inference. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 103–108, 2019.

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

[29] 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.

[30] 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.

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