Specifying Weight Priors in Bayesian Deep Neural Networks with Empirical Bayes

Ranganath Krishnan[*]
Intel Labs
ranganath.krishnan@intel.com

Mahesh Subedar[*]
Intel Labs
mahesh.subedar@intel.com

Omesh Tickoo
Intel Labs
omesh.tickoo@intel.com

Abstract

Stochastic variational inference for Bayesian deep neural network (DNN) requires specifying priors and approximate posterior distributions over neural network weights. Specifying meaningful weight priors is a challenging problem, particularly for scaling variational inference to deeper architectures involving high dimensional weight space. We propose MOdel Priors with Empirical Bayes using DNN (MOPED) method to choose informed weight priors in Bayesian neural networks. We formulate a two-stage hierarchical modeling, 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 empirically evaluate the proposed approach on real-world tasks including image classification, video activity recognition and audio classification with varying complex neural network architectures. We also evaluate our proposed approach on diabetic retinopathy diagnosis task and benchmark with the state-of-the-art Bayesian deep learning techniques. We demonstrate MOPED method enables scalable variational inference and provides reliable uncertainty quantification.

1 Introduction

Uncertainty estimation in deep neural network (DNN) predictions is essential for designing reliable and robust AI systems. Bayesian deep neural networks (Neal 1995; Gal 2016) has allowed bridging deep learning and probabilistic Bayesian theory to quantify uncertainty by borrowing the strengths of both methodologies. Variational inference (VI) (Blei, Kucukelbir, and McAuliffe 2017) is an analytical approximation technique to infer the posterior distribution of model parameters. VI methods formulate the Bayesian inference problem as an optimization-based approach which lends itself to the stochastic gradient descent based optimization used in training DNN models. VI with generalized formulations (Graves 2011; Blundell et al. 2015) has renewed interest in Bayesian neural networks.

The recent research in Bayesian Deep Learning (BDL) is focused on scaling the VI to more complex models. The

---

[*]Equal Contribution

scalability of VI in Bayesian DNNs to practical applications involving deep models and large-scale datasets is an open problem. Hybrid Bayesian DNN architectures (Subedar et al. 2019; Krishnan, Subedar, and Tickoo 2018) are used for complex computer vision tasks to balance complexity of the model while providing benefits of Bayesian inference. DNNs are shown to have structural benefits (Bengio, Courville, and Vincent 2013) which helps them in learning complex models on larger datasets. The convergence speed and performance (Goodfellow, Bengio, and Courville 2016) of DNN models heavily depend on the initialization of model weights and other hyper parameters. The transfer learning approaches (Shin et al. 2016) demonstrate the benefit of fine tuning the pretrained DNN models from adjacent domains in order to achieve faster convergence and better accuracies.

Variational inference for Bayesian DNN involves choosing prior distributions and approximate posterior distributions over neural network weights. In a pure Bayesian approach, prior distribution is specified before any data is observed. Specifying meaningful priors in large Bayesian DNN models with high dimensional weight space is an active area of research (Wu et al. 2019; Nalisnick, Hernández-Lobato, and Smyth 2019; Sun et al. 2019; Atanov et al. 2019), as it is practically difficult to have prior belief on millions of parameters. Empirical Bayes (Robbins 1956; Casella 1992) methods estimates prior distribution from the data. Based on Empirical Bayes and transfer learning approaches, we propose MOdel Priors with Empirical Bayes using DNN (MOPED) method to initialize the weight priors in Bayesian DNNs, which in our experiments have shown to achieve better training convergence for larger models.

Our main contributions include:

- We propose MOPED method to specify informed weight priors in Bayesian neural networks using Empirical Bayes framework. MOPED advances the current state-of-the-art by enabling scalable variational inference for large models applied to real-world tasks.

- We demonstrate with thorough empirical experiments on multiple real-world tasks that the MOPED method helps training convergence and provides better model performance, along with reliable uncertainty estimates. We also evaluate MOPED on diabetic retinopathy diagnosis task

using BDL benchmarking framework (Filos et al. 2019) and demonstrate it outperforms state-of-the-art Bayesian deep learning methods.

The rest of the document is organized as below. We provide background material in Section 2. The details of proposed method for initializing the weight priors in Bayesian DNN models is presented in Section 3, and related work in Section 4. Followed by empirical experiments and results supporting the claims of proposed method in Section 5.

2 Background

2.1 Bayesian neural networks

Bayesian neural networks provide a probabilistic interpretation of deep learning models by placing distributions over the neural network weights (Neal 1995). 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 weights w as a function $y = f_w(x)$ that represents the neural network 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 $p(w|D)$:

$$p(w|D) = \frac{p(y \mid x, w) \, p(w)}{\int p(y \mid x, w) \, p(w) \, dw} \quad (1)$$

Computing the posterior distribution $p(w|D)$ is often intractable, some of the previously proposed techniques to achieve an analytically tractable inference include Markov Chain Monte Carlo (MCMC) sampling based probabilistic inference (Neal 2012; Welling and Teh 2011), variational inference (Graves 2011; Ranganath, Gerrish, and Blei 2013; Blundell et al. 2015), expectation propagation (Minka 2001) and Monte Carlo dropout approximate inference (Gal and Ghahramani 2016) .

Predictive distribution is obtained through multiple stochastic forward passes on the network while sampling from the weight posteriors using Monte Carlo estimators. Equation 2 shows the predictive distribution of the output y^* given new input x^*:

$$p(y^*|x^*, D) = \int p(y^*|x^*, w) \, p(w \mid D) dw$$
$$p(y^*|x^*, D) \approx \frac{1}{T} \sum_{i=1}^{T} p(y^*|x^*, w_i) \,, \quad w_i \sim p(w \mid D) \quad (2)$$

where, T is number of Monte Carlo samples.

2.2 Variational inference

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. Minimizing the KL divergence is equivalent to maximizing the log evidence lower bound (ELBO) (Bishop 2006), as shown in Equation 3.

$$\mathcal{L} := \int q_\theta(w) \log p(y|x, w) \, dw - KL[q_\theta(w)||p(w)] \quad (3)$$

In mean field variation inference, weights are modeled with fully factorized Gaussian distribution parameterized by variational parameters μ and σ.

$$q_\theta(w) := \mathcal{N}(w \mid \mu, \sigma) \quad (4)$$

The variational distribution $q_\theta(w)$ and its parameters μ and σ are learnt while optimizing the cost function ELBO with the stochastic gradient steps.

(Graves 2011) proposed fully factorized Gaussian posteriors and a differentiable loss function. (Blundell et al. 2015) proposed a Bayes by Backprop method which learns probability distribution on the weights of the neural network by minimizing loss function. (Wen et al. 2018) proposed a Flipout method to apply pseudo-independent weight perturbations to decorrelate the gradients within mini-batches.

2.3 Empirical Bayes

Empirical Bayes (EB) (Casella 1992) methods lie in between frequentist and Bayesian statistical approaches as it attempts to leverage strengths from both methodologies. EB methods are considered as approximation to a fully Bayesian treatment of a hierarchical Bayes model. EB methods estimates prior distribution from the data, which is in contrast to typical Bayesian approach. The idea of Empirical Bayes is not new and the original formulation of Empirical Bayes dates back to 1950s (Robbins 1956), which is non-parametric EB. Since then, many parametric formulations has been proposed and used in wide variety of applications.We use parametric Empirical Bayes approach in our proposed method for mean field variational inference in Bayesian deep neural network, where weights are modeled with fully factorized Gaussian distribution.

Parametric EB specifies a family of prior distributions $p(w|\lambda)$ where λ is a hyper-parameter. Analogous to Equation 1, posterior distribution can be obtained with EB as given by Equation 5.

$$p(w|D, \lambda) = \frac{p(y \mid x, w) \, p(w \mid \lambda)}{\int p(y \mid x, w) \, p(w \mid \lambda) \, dw} \quad (5)$$

2.4 Uncertainty Quantification

Uncertainty estimation is essential to build reliable and robust AI systems, which is pivotal to understand system's confidence in predictions and decision-making. Bayesian DNNs enable to capture different types of uncertainties: "Aleatoric" and "Epistemic" (Gal 2016). Aleatoric uncertainty captures noise inherent with observation. Epistemic uncertainty, also known as model uncertainty captures lack of knowledge in representing model parameters, specifically in the scenario of limited data.

We evaluate the model uncertainty using Bayesian active learning by disagreement (BALD) (Houlsby et al. 2011; Gal 2016), which quantifies mutual information between parameter posterior distribution and predictive distribution.

$$BALD := H(y^*|x^*, D) - \mathbb{E}_{p(w|D)}[H(y^*|x^*, w)] \quad (6)$$

Dataset	Modality	Architecture	Bayesian DNN Complexity (# parameters)	DNN	Validation Accuracy Bayesian 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**
Diabetic Retinopathy	Images	VGG	21,242,689	0.842	0.843	**0.857**
CIFAR-10	Images	Resnet-56	1,714,250	0.926	0.896	**0.927**
		Resnet-20	546,314	0.911	0.878	**0.916**
MNIST	Images	LeNeT	1,090,856	0.994	0.993	**0.995**
Fashion-MNIST	Images	SCNN	442,218	0.921	0.906	**0.923**

Table 1: Accuracies for architectures with different complexities and input modalities. Mean field variational inference with MOPED initialization (MOPED_MFVI) obtains reliable uncertainty estimates from Bayesian DNNs while achieving similar or better accuracy as the deterministic DNNs. Mean field variational inference with random priors (MFVI) has convergence issues (shown in italics) for complex architectures, while the proposed method achieves model convergence. DNN and MFVI accuracy numbers for diabetic retinopathy dataset are obtained from BDL-benchmarks.

where, $H(y^*|x^*, D)$ is the predictive entropy as shown in Equation 7. Predictive entropy captures a combination of input uncertainty and model uncertainty.

$$H(y^*|x^*, D) := -\sum_{i=0}^{K-1} p_{i\mu} * log\, p_{i\mu} \qquad (7)$$

and $p_{i\mu}$ is predictive mean probability of ith class from T Monte Carlo samples, and K is total number of output classes.

3 MOPED: informed weight priors

MOPED advances the current state-of-the-art in variational inference for Bayesian DNNs by providing a way for specifying meaningful prior and approximate posterior distributions over weights using Empirical Bayes framework. Empirical Bayes framework borrows strengths from both classical (frequentist) and Bayesian statistical methodologies.

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

We illustrate our proposed approach on mean-field variational inference (MFVI). For MFVI in Bayesian DNNs, weights are modeled with fully factorized Gaussian distributions parameterized by variational parameters, i.e. 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. We propose to set the weight priors in Bayesian neural networks based on the MLE obtained from standard DNN of equivalent architecture. We set the prior with mean equals w_{MLE} and unit variance respectively, and initialize the variational parameters in approximate posteriors as given in Equation 8.

$$\overline{w} := w_{MLE}; \quad \rho \sim \mathcal{N}(\overline{\rho}, \Delta\rho)$$
$$w \sim \mathcal{N}(w_{MLE}, \log(1 + e^\rho)) \qquad (8)$$

where, w_{MLE} represents maximum likelihood estimates of 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 (hundreds of millions of parameters), 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 variational parameters in approximate posteriors as given in Equation 9.

$$\overline{w} := w_{MLE}; \quad \rho := \log(e^{\delta|w_{MLE}|} - 1)$$
$$w \sim \mathcal{N}(w_{MLE}, \delta \mid w_{MLE} \mid)) \qquad (9)$$

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

In the next section, 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 along with reliable uncertainty estimates.

4 Related Work

Deterministic pretraining (Molchanov, Ashukha, and Vetrov 2017; Sønderby et al. 2016) has been used to improve model training for variational probabilistic models. Molchanov, Ashukha, and Vetrov use a pretrained deterministic network for Sparse Variational Dropout method. Sønderby et al. use a warm-up method for variational-auto encoder by rescaling the KL-divergence term with a scalar term β, which is increased linearly from 0 to 1 during the first N epochs of training. Whereas, in our method we use the point-estimates from pretrained standard DNN of the same architecture to set the informed priors and the model is optimized with MFVI using full-scale KL-divergence term in ELBO.

Choosing weight priors in Bayesian neural networks is an active area of research. Atanov et al. propose implicit priors for variational inference in convolutional neural networks that exploit generative models. Nguyen et al. 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

Dataset	Bayesian DNN Archiectures	AUPR		AUROC	
		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 2: Comparison of AUC of precision-recall (AUPR) and ROC (auROC) for models with varying complexities. MOPED method outperforms training with random initialization of weight priors.

(a) Training convergence curves   (b) AUPR curves   (c) Precision-recall
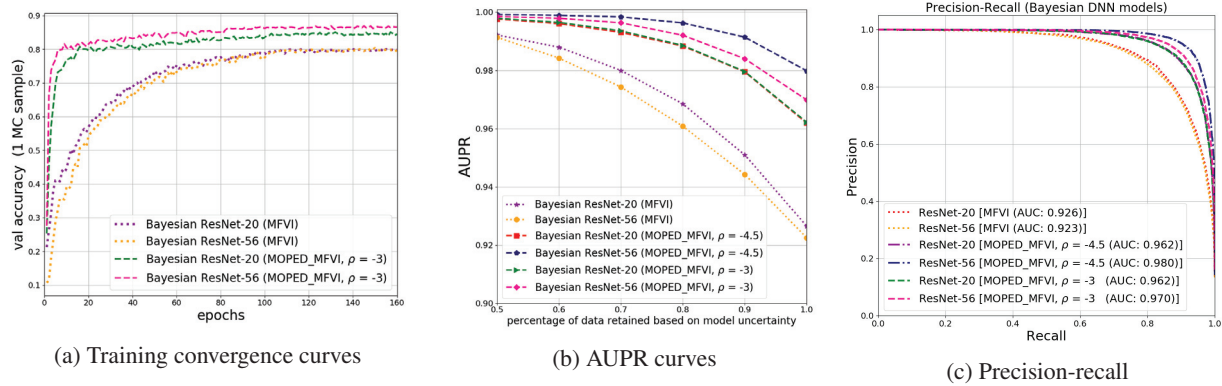
Figure 1: Comparison of MOPED and MOPED_MFVI for Bayesian ResNet-20 and ResNet-56 architectures. (a) training convergence, (b) AUPR as a function of retained data based on model uncertainty and (c) precision-recall plots.

for small-scale MNIST experiments. Wu et al. modify ELBO with a deterministic approximation of reconstruction term and use Empirical Bayes procedure for selecting variance of prior in KL term (with zero prior mean). The authors also caution about inherent scaling of their method could potentially limit its practical use for networks with large hidden size. All of these works have been demonstrated only on small-scale models and simple datasets like MNIST. In our method, we retain the stochastic property of the expected log-likelihood term in ELBO, and specify both mean and variance of weight priors based on pretrained DNN with Empirical Bayes. Further, we demonstrate our method on large-scale Bayesian DNN models with complex datasets on real-world tasks.

5 Experiments

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: (i) ResNet-101 C3D (Hara, Kataoka, and Satoh 2018) for video activity classification on UCF-101(Soomro, Zamir, and Shah 2012) dataset, (ii) VGGish(Hershey et al. 2017) for audio classification on UrbanSound8K (Salamon, Jacoby, and Bello 2014) dataset, (iii) Modified version of VGG(Filos et al. 2019) for diabetic retinopathy detection (Kaggle 2015), (iv) ResNet-20

and ResNet-56 (He et al. 2016) for CIFAR-10 (Krizhevsky and Hinton 2009), (v) LeNet architecture for MNIST (LeCun et al. 1998) digit classification, and (vi) Simple convolutional neural network (SCNN) consisting of two convolutional layers followed by two dense layers for image classification on Fashion-MNIST (Xiao, Rasul, and Vollgraf 2017) datasets.

We implemented above Bayesian DNN models and trained them using Tensorflow and Tensorflow-Probability (Dillon et al. 2017) frameworks. The variational layers are modeled using Flipout (Wen et al. 2018), an efficient method that decorrelates the gradients within a mini-batch by implicitly sampling pseudo-independent weight perturbations for each input. The MLE weights obtained from the pretrained DNN models are used in MOPED method to set the priors and initialize the variational parameters in approximate posteriors (Equation 8 and 9), as described in Section3.

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 and predictive uncertainty using Bayesian active learning by disagreement (BALD) (Equation 6) and predictive entropy (Equation 7), respectively. Quantitative comparison of uncertainty estimates are made by calculating area under the curve 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
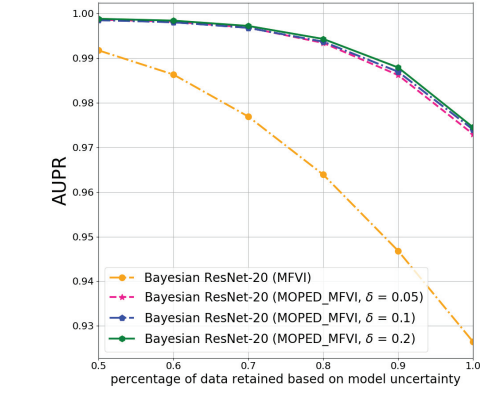
(a) Bayesian ResNet-20 (CIFAR-10)
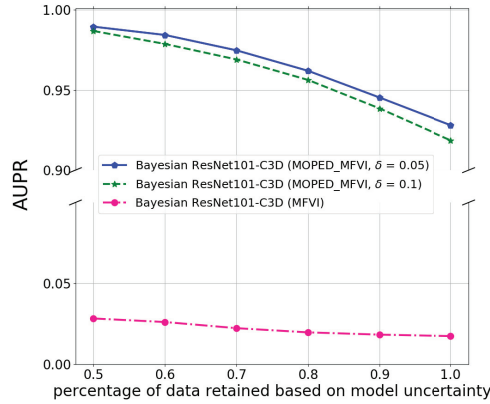
(b) Bayesian ResNet-101 C3D (UCF-101)

Figure 2: Precision-recall AUC (AUPR) plots with different δ scale factors for initializing variance values in MOPED method.

based on uncertainty estimates).

In Table 1, classification accuracies for architectures with various model 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 larger models (ResNet-101 C3D and VGGish). It is evident from these results that MOPED method guarantees the training convergence even for the complex models.

In Figure 1, comparison of mean field variational inference with MOPED method (MOPED_MFVI) and mean field variational inference with random initialization of priors (MFVI) is shown for Bayesian ResNet-20 and ResNet-56 architectures trained on CIFAR-10 dataset. The AUPR plots capture the precision-recall AUC values as a function of retained data based on the model uncertainty estimates. Figure 1 (b) & (c) show that MOPED_MFVI provides better performance than MFVI. AUPR increases as most uncertain predictions are ignored based on the model uncertainty, indicating reliable uncertainty estimates. We show the results for different selec-

tion of ρ values (as shown in Equation 8).

In Figure 2, we show AUPR plots for CIFAR-10 and UCF-101 with different δ values as mentioned in Equation 9.

5.1 Benchmarking uncertainty estimates

Bayesian Deep Learning (BDL) benchmarks (Filos et al. 2019) 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. The proposed MOPED_MFVI method is compared with state-of-the-art baseline methods available in BDL-benchmarks suite on diabetic retinopathy detection task (Kaggle 2015). The evaluation methodology assesses the techniques by their diagnostic accuracy and area under receiver-operating-characteristic (AUC-ROC) curve, as a function of percentage of retained data based on predictive uncertainty estimates. It is expected that the models with well-calibrated uncertainty improve their performance (detection accuracy and AUC-ROC) as most certain data is retrained.

We have followed the evaluation methodology presented in the BDL-benchmarks to compare the accuracy and uncertainty estimates obtained from our method. We used the same model architecture (VGG) and hyper-parameters as used by other baselines for evaluating MOPED_MFVI. The results for the BDL baseline methods are obtained from (Filos et al. 2019). In Table 3 and Figure 3, quantitative evaluation of AUC and accuracy values for BDL baseline methods and MOPED_MFVI are presented. The proposed MOPED_MFVI method outperforms other state-of-the-art BDL techniques.

5.2 Robustness to out-of-distribution data

We evaluate the uncertainty estimates obtained from MOPED_MFVI 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 (Russakovsky et al. 2015) and SVHN (Goodfellow et al. 2013) 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 5. The density histograms in Figure 5 (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 UCF-101 (Soomro, Zamir, and Shah 2012) and Moments-in-Time (MiT) (Monfort et al. 2018) video activity recognition datasets as the in- and out-of-distribution data, respectively. These results confirm the uncertainty estimates obtained from proposed method are reliable and can identify out-of-distribution data.

In order to evaluate robustness of our method (MOPED_MFVI), we compare state-of-the-art probabilistic deep learning methods for prediction accuracy as a function of model confidence. Following the experiments in (Lakshminarayanan, Pritzel, and Blundell 2017), we trained our

Method	50% data retrained		75% data retrained		100% data retrained	
	AUC	Accuracy	AUC	Accuracy	AUC	Accuracy
MC Dropout	0.878	0.913	0.852	0.871	0.821	0.845
Mean-field VI	0.866	0.881	0.84	0.850	0.821	0.843
Deep Ensembles	0.872	0.899	0.849	0.861	0.818	0.846
Deterministic	0.849	0.861	0.823	0.849	0.82	0.842
Ensemble MC Dropout	0.881	0.924	0.854	0.881	0.825	0.853
MOPED Mean-field VI	**0.912**	**0.937**	**0.885**	**0.914**	**0.883**	**0.857**
Random Referral	0. 818	0.848	0.820	0.843	0.820	0.842

Table 3: Comparison of Area under the receiver-operating characteristic curve (AUC) and classification accuracy as a function of retained data from the BDL benchmark suite. The proposed method demonstrates an improvement of superior performance compared to all the baseline models.

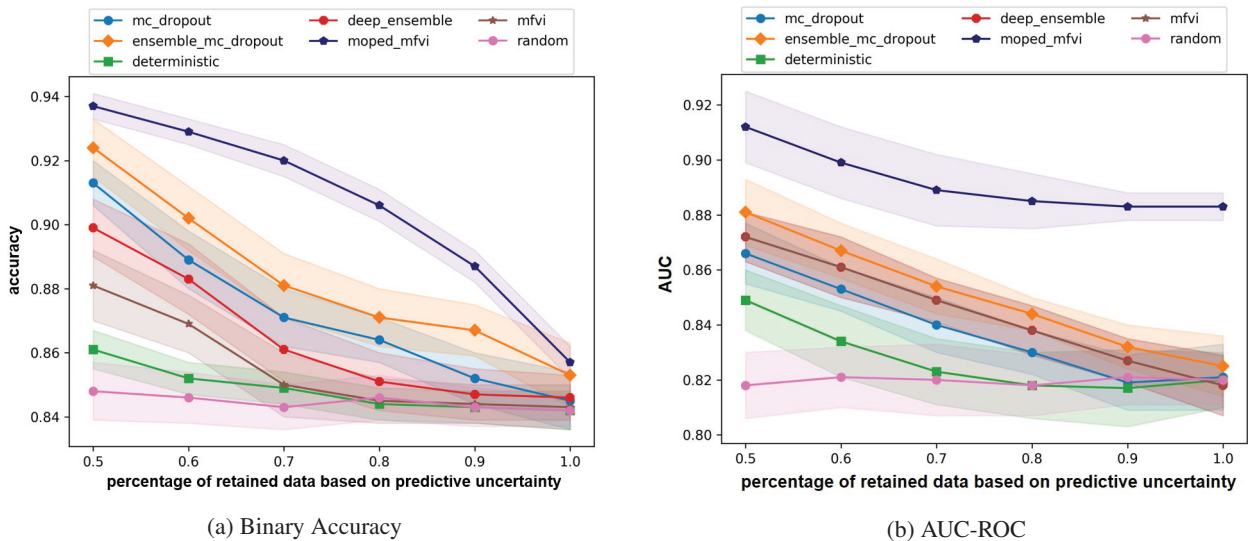

(a) Binary Accuracy

(b) AUC-ROC

Figure 3: Benchmarking MOPED_MFVI with state-of-art Bayesian deep learning techniques on diabetic retinopathy diagnosis task using BDL-benchmarks. Accuracy and area under the receiver-operating characteristic curve (AUC-ROC) plots for varied percentage of retained data based on predictive uncertainty. MOPED_MFVI performs better than the other baselines from BDL-benchmarks suite. Shading shows the standard error.
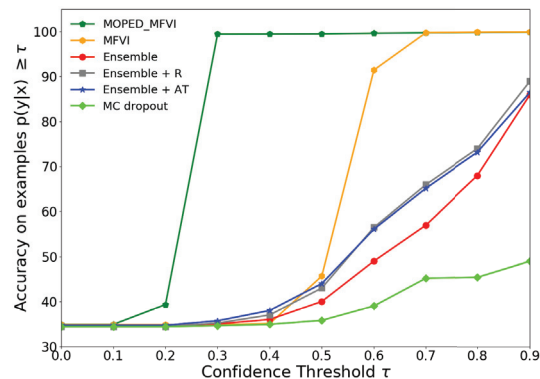

Figure 4: Accuracy vs Confidence curves: Networks trained on MNIST and tested on both MNIST and the NotMNIST (out-of-distribution) test sets.

model on MNIST training set and tested it on a mix of examples from MNIST and NotMNIST (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 (Gal and Ghahramani 2016) and deep model ensembles (Lakshminarayanan, Pritzel, and Blundell 2017) approaches (shown in Figure 4).

6 Conclusions

We proposed MOPED method that specifies informed weight priors in Bayesian deep neural networks with Empirical Bayes approach. We demonstrated with thorough empirical experiments that MOPED enables scalable variational

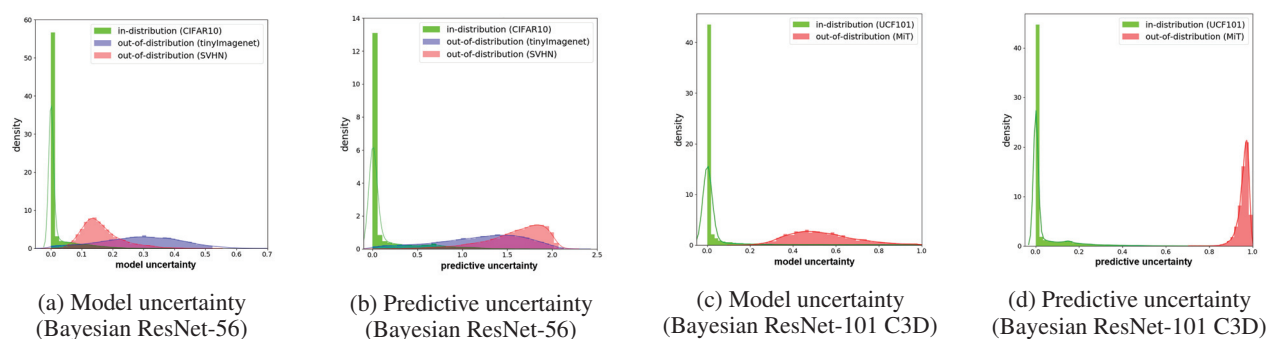| (a) Model uncertainty (Bayesian ResNet-56) | (b) Predictive uncertainty (Bayesian ResNet-56) | (c) Model uncertainty (Bayesian ResNet-101 C3D) | (d) Predictive uncertainty (Bayesian ResNet-101 C3D) |

Figure 5: Density histograms obtained from in- and out-of-distribution samples. Bayesian DNN model uncertainty estimates indicate higher uncertainty for out-of-distribution samples as compared to the in-distribution samples.

inference for Bayesian DNNs. We demonstrated the proposed method outperforms state-of-the-art Bayesian deep learning techniques using BDL-benchmarks framework. We also showed the uncertainty estimates obtained from the proposed method are reliable to identify out-of-distribution data. The results support proposed approach provides better model performance and reliable uncertainty estimates on real-world tasks with large scale complex models.

References

Atanov, A.; Ashukha, A.; Struminsky, K.; Vetrov, D.; and Welling, M. 2019. The deep weight prior. *In International Conference on Learning Representations*.

Bengio, Y.; Courville, A.; and Vincent, P. 2013. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence* 35(8):1798–1828.

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

Blei, D. M.; Kucukelbir, A.; and McAuliffe, J. D. 2017. Variational inference: A review for statisticians. *Journal of the American Statistical Association* 112(518):859–877.

Blundell, C.; Cornebise, J.; Kavukcuoglu, K.; and Wierstra, D. 2015. Weight uncertainty in neural networks. *arXiv preprint arXiv:1505.05424*.

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

Dillon, J. V.; Langmore, I.; Tran, D.; Brevdo, E.; Vasudevan, S.; Moore, D.; Patton, B.; Alemi, A.; Hoffman, M.; and Saurous, R. A. 2017. Tensorflow distributions. *arXiv preprint arXiv:1711.10604*.

Filos, A.; Farquhar, S.; Gomez, A. N.; Rudner, T. G. J.; Kenton, Z.; Smith, L.; Alizadeh, M.; de Kroon, A.; and Gal, Y. 2019. Benchmarking bayesian deep learning with diabetic retinopathy diagnosis.

Gal, Y., and Ghahramani, Z. 2016. Dropout as a bayesian approximation: Representing model uncertainty in deep learn-

ing. In *international conference on machine learning*, 1050–1059.

Gal, Y. 2016. Uncertainty in deep learning. *University of Cambridge*.

Goodfellow, I.; Bengio, Y.; and Courville, A. 2016. *Deep learning*. MIT press.

Goodfellow, I. J.; Bulatov, Y.; Ibarz, J.; Arnoud, S.; and Shet, V. 2013. Multi-digit number recognition from street view imagery using deep convolutional neural networks. *arXiv preprint arXiv:1312.6082*.

Graves, A. 2011. Practical variational inference for neural networks. In *Advances in neural information processing systems*, 2348–2356.

Hara, K.; Kataoka, H.; and Satoh, Y. 2018. 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)*, 6546–6555.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Hershey, S.; Chaudhuri, S.; Ellis, D. P.; Gemmeke, J. F.; Jansen, A.; Moore, R. C.; Plakal, M.; Platt, D.; Saurous, R. A.; Seybold, B.; et al. 2017. Cnn architectures for large-scale audio classification. In *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*, 131–135. IEEE.

Houlsby, N.; Huszár, F.; Ghahramani, Z.; and Lengyel, M. 2011. Bayesian active learning for classification and preference learning. *arXiv preprint arXiv:1112.5745*.

Kaggle. 2015. Diabetic retinopathy detection challenge. https://www.kaggle.com/c/diabetic-retinopathy-detection/overview/description.

Krishnan, R.; Subedar, M.; and Tickoo, O. 2018. Bar: Bayesian activity recognition using variational inference. *arXiv preprint arXiv:1811.03305*.

Krizhevsky, A., and Hinton, G. 2009. Learning multiple layers of features from tiny images. Technical report, Citeseer.

Lakshminarayanan, B.; Pritzel, A.; and Blundell, C. 2017. Simple and scalable predictive uncertainty estimation using

deep ensembles. In *Advances in Neural Information Processing Systems*, 6402–6413.

LeCun, Y.; Bottou, L.; Bengio, Y.; Haffner, P.; et al. 1998. Gradient-based learning applied to document recognition. *Proceedings of the IEEE* 86(11):2278–2324.

Minka, T. P. 2001. Expectation propagation for approximate bayesian inference. In *Proceedings of the Seventeenth conference on Uncertainty in artificial intelligence*, 362–369. Morgan Kaufmann Publishers Inc.

Molchanov, D.; Ashukha, A.; and Vetrov, D. 2017. Variational dropout sparsifies deep neural networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, 2498–2507. JMLR. org.

Monfort, M.; Zhou, B.; Bargal, S. A.; Andonian, A.; Yan, T.; Ramakrishnan, K.; Brown, L.; Fan, Q.; Gutfruend, D.; Vondrick, C.; et al. 2018. Moments in time dataset: one million videos for event understanding. *arXiv preprint arXiv:1801.03150*.

Nalisnick, E.; Hernández-Lobato, J. M.; and Smyth, P. 2019. Dropout as a structured shrinkage prior. In *International Conference on Machine Learning*, 4712–4722.

Neal, R. M. 1995. *BAYESIAN LEARNING FOR NEURAL NETWORKS*. Ph.D. Dissertation, Citeseer.

Neal, R. M. 2012. *Bayesian learning for neural networks*, volume 118. Springer Science & Business Media.

Nguyen, C. V.; Li, Y.; Bui, T. D.; and Turner, R. E. 2017. Variational continual learning. *arXiv preprint arXiv:1710.10628*.

Ranganath, R.; Gerrish, S.; and Blei, D. M. 2013. Black box variational inference. *arXiv preprint arXiv:1401.0118*.

Robbins, H. 1956. An empirical bayes approach to statistics. *Herbert Robbins Selected Papers* 41–47.

Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; et al. 2015. Imagenet large scale visual recognition challenge. *International journal of computer vision* 115(3):211–252.

Salamon, J.; Jacoby, C.; and Bello, J. P. 2014. A dataset and taxonomy for urban sound research. In *Proceedings of the 22nd ACM international conference on Multimedia*, 1041–1044. ACM.

Shin, H.-C.; Roth, H. R.; Gao, M.; Lu, L.; Xu, Z.; Nogues, I.; Yao, J.; Mollura, D.; and Summers, R. M. 2016. Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning. *IEEE transactions on medical imaging* 35(5):1285–1298.

Sønderby, C. K.; Raiko, T.; Maaløe, L.; Sønderby, S. K.; and Winther, O. 2016. Ladder variational autoencoders. In *Advances in neural information processing systems*, 3738–3746.

Soomro, K.; Zamir, A. R.; and Shah, M. 2012. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*.

Subedar, M.; Krishnan, R.; Meyer, P. L.; Tickoo, O.; and Huang, J. 2019. Uncertainty-aware audiovisual activity recognition using deep bayesian variational inference. In *The IEEE International Conference on Computer Vision (ICCV)*.

Sun, S.; Zhang, G.; Shi, J.; and Grosse, R. 2019. Functional variational bayesian neural networks. *arXiv preprint arXiv:1903.05779*.

Welling, M., and Teh, Y. W. 2011. Bayesian learning via stochastic gradient langevin dynamics. In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, 681–688.

Wen, Y.; Vicol, P.; Ba, J.; Tran, D.; and Grosse, R. 2018. Flipout: Efficient pseudo-independent weight perturbations on mini-batches. *arXiv preprint arXiv:1803.04386*.

Wu, A.; Nowozin, S.; Meeds, E.; Turner, R. E.; Hernández-Lobato, J. M.; and Gaunt, A. L. 2019. Deterministic variational inference for robust bayesian neural networks. *International Conference on Learning Representations (ICLR)*.

Xiao, H.; Rasul, K.; and Vollgraf, R. 2017. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*.