Review: Loss-Calibrated Approximate Inference in Bayesian Neural Networks

Carlos Cuevas Villarmín, Javier Alejadro Lopetegui González, José Felipe Espinosa Orjuela

Université Paris-Saclay
M2-MVA, Course: Bayesian Machine Learning
carlos.cuevas-villarmin@universite-paris-saclay.fr,
javier-alejandro.lopetegui-gonzalez@universite-paris-saclay.fr,
jose.espinosa-orjuela@universite-paris-saclay.fr

Abstract

The well-known approach of minimise Kullback-Leibler divergence in approximate inference for Bayesian Neural Networks (BNNs) does not use the final purpose of the model. Therefore, the authors of (1) include a novel loss-calibrated evidence lower bound for Bayesian Neural Networks for supervised learning models. Their contribution is translated into higher utility for a given task than traditional approaches. We will summarize the novelty of the paper highlighting the strong and possible weak points. Finally, we will illustrate its potential through real data of interest.

1 Introduction

BNNs are a powerful tool which is able to capture uncertinty. However, there are particular applications that have asymmetric utility functions. Bayesian decision theory combines uncertainty with task-specific utility functions to make rational predictions. In order to deal with asymmetric penalization of errors in BNNs, introducing an additional utility-dependent penalty term will guide predictions in the presence of uncertainty.

It is important to notice the difference between loss and utility function. In this work, the loss function corresponds to a log likelihood that describes the noise model. On the other hand, the utility function will determine the consequences of making an incorrect prediction.

Our report is organised as follow: in Section 2 we summary the theoretical framework that justifies the introduction of the utility-dependent lower bound. Section 3 we explain briefly the results reported in the paper and debate about the strong and weak points of the paper. All the reported figures have been done by ourselves in an attempt to replicate the original results except for the ones of the autonomous driving experiment due to lack of computing resources. Then, in Section 4 we provide an additional experiment with the purpose of justifying the use of approach when using simple models. Finally, in 5 we present the conclusions.

2 Theoretical Framework

In this section we assume that the reader is acquainted with Bayesian Neural Networks and Bayesian decision theory (2). Otherwise, we recommend first to read the provided articles where the authors combine both fields to introduce the proposed loss-calibrated Bayesian neural network. Our objective in this section is to summarize the most important part of the contribution as well as highlight the advantages and disadvantages or limitations that this approach may present.

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

The prediction of which label is assigned to a given input x^* depends on the specific task and uncertainty. Then, it is introduced a utility function $u(\mathbf{h} = \mathbf{c}, \mathbf{y}^* = \mathbf{c'})$ where $\mathbf{y}^* = \mathbf{c'}$ is the target label and u defines the gain from predicting different labels **h**.

The conditional-gain in assigning a label **h** conditioned on a test input $\mathbf{x}^*$ is defined as the average of the utility over the predictions $\mathbf{y}^*$, i.e.,

$$\mathcal{G}(\mathbf{h} = \mathbf{c}|\mathbf{x}^*) = \int_{\mathbf{y}^*} u(\mathbf{h} = \mathbf{c}, \mathbf{y}^* = \mathbf{c'}) p(\mathbf{y}^* = \mathbf{c'}|\mathbf{x}^*, \mathbf{X}, \mathbf{Y}) d\mathbf{c'}, \tag{1}$$

and it could be rewritten in terms of an integretion with respect to w as

$$\mathcal{G}(\mathbf{h} = \mathbf{c}|\mathbf{x}^*) = \int_w \mathcal{G}(\mathbf{h} = \mathbf{c}|\mathbf{x}^*, w) p(w|\mathbf{X}, \mathbf{Y}) dw \tag{2}$$

The label **h** that maximises the conditional-gain is the optimal prediction $\mathbf{h}^*$ for the given input $\mathbf{x}^*$ conditioned on the dataset $\{\mathbf{X}, \mathbf{Y}\}$.

$$\mathbf{h}^*(\mathbf{x}^*) = \text{argmax}_{c \in \mathcal{C}} \log(\mathcal{G}(\mathbf{h} = \mathbf{c}|\mathbf{x}^*)). \tag{3}$$

The marginal conditional-gain is

$$\mathcal{G}(\mathbf{H}|\mathbf{X}) := \int_w \mathcal{G}(\mathbf{H}|\mathbf{X}, w) p(w|\mathbf{X}, \mathbf{Y}) dw \tag{4}$$

by assuming conditional independence over the inputs and that given the model parameters, the optimal prediction depends only on the input x_j.

The value of this conditional-gain provides good information about the quality of the model. If the value is large, it has been assigned high predictive probability to class labels that give a high task-specific utility across our data. On the other hand, low values mean that the choice of **H** leads to unpleasant low task-specific utility over the data. Consequently, as the purpose is to assign labels in a way that the utility is maximized it can be done by maximizing the conditional-gain. To do that, they define a lower bound to the log conditional-gain which later will be maximize:

$$\begin{aligned} &\log(\mathcal{G}(\mathbf{H}|\mathbf{X})) = \log\left(\int_w q(w) \frac{p(w|\mathbf{X}, \mathbf{Y}) \mathcal{G}(\mathbf{H}|\mathbf{X}, w)}{q(w)} dw\right) \geq \\ &\geq \int_w q(w) \log\left(\frac{p(w|\mathbf{X}, \mathbf{Y}) \mathcal{G}(\mathbf{H}|\mathbf{X}, w)}{q(w)} dw\right) := \mathcal{L}(q(w), \mathbf{H}), \end{aligned} \tag{5}$$

where $q(w)$ is the approximate posterior distribution.

Based on the equivalence proved between the maximization of the lower bound and the minimization of the KL divergence $KL(q||\tilde{p_h}) = \log(\mathcal{G}(\mathbf{H}|\mathbf{X}) - \mathcal{L}(q, \mathbf{H})$ where the probability distribution $\tilde{p_h}$ is the true posterior scaled by the conditional-gain the authors calibrate the approximate posterior to take into account the utility.

Then, the loss-calibrated ELBO fot the BNN is defined by expanding the lower bound

$$\mathcal{L}(q(w), \mathbf{H}) = \int_w q(w) \log p(\mathbf{Y}|\mathbf{X}, w) dw + KL(q(w)||p(w)) + \underbrace{\int_w q(w) \log \mathcal{G}(\mathbf{H}|\mathbf{X}, w) dw}_{\text{New term, requieres optimal prediction } \mathbf{H}} + \text{const.} \tag{6}$$

Finally, using MC integration and $q(w)$ it can be implemented as the standard objective loss of a dropout NN with the **new penalty term**:

$$\underbrace{-\sum_i \log p(\mathbf{y}_i|\mathbf{x}_i, \hat{w}_i) + ||w||^2}_{\text{Equivalent to standard dropout loss}} \underbrace{-\sum_i \left(\log \sum_c u(\mathbf{h}_i, \mathbf{c}) p(\mathbf{y}_i = \mathbf{c'}|\mathbf{x}_i, \hat{w}_i) \right)}_{\text{Additional utility-dependent penalty term}}, \tag{7}$$

where $\hat{w}_i \sim q_\Theta(w)$. This utility-dependent penalty term allows the model to learn in a way that maximizes the utility of the model for a given task. Therefore, this approach will be more useful than others as it uses the prior knowledge of the application.

3 Reported results

During this section we will summarize the conclusions of the three experiments done in the paper[1]: Illustrative example: Diabetes diagnosis (classification), Robustness to label corruption and network capacity and Scalability to real world applications: Semantic segmentation in Autonomous Driving.

We will present the strong points of each of them and enhance its importance to conclude the relevance of the novel approach.

3.1 Illustrative example: Diabetes diagnosis classification

The goal is to classify the diagnosis of a patient into three possible classes: severe diabetes, moderate diabetes and healty. The purpose of the experiment is to compare the novel approach with standard BNN and weighted cross entropy in terms of how each of the networks misclassify and how they enforce that behaviour. Figure 1 shows the confusion matrices for the three approaches. Based on how is defined the utility function all the models try to not diagnose unhealthy patients as healthy.

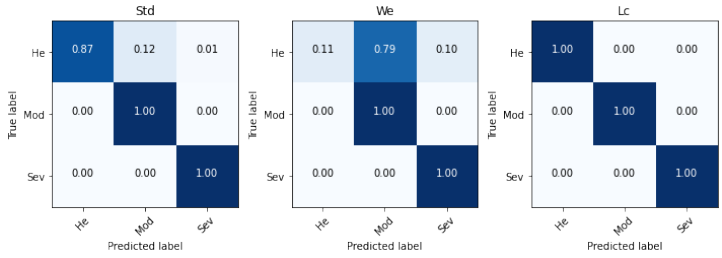

Figure 1: Each confusion matrix shows the diagnosis when averaging the utility function with respect to the dropout samples of each network.

Additionally, it is seen that the weighted cross-entropy case tends to diagnoses a patient as moderate diabetes when it is healthy. This enhance that the model predicts well the healthy class in extreme cases due to the high weight in moderate class. It can be conclude that the new approach outperforms the others and captures better the utility function information.

3.2 Robustness to Label Corruption and Network Capacity

In this experiment the authors consider the MNIST dataset and change arbitrarily 50% of labels in order to have a corrupted dataset. The utility function is define to focus on maximising the utility for digits 3 and 8 (two classes that usually have ambiguity) by giving the utility of 1 for true positives, 0.3 for false positives and 0 for false positives for all other digits. Figure 2 shows that the utility-dependent lower bound does not have a bad effect on the performance in the ideal scenario with no label noise. On the other hand, Figure 3 enhance the robustness of the novel approach in mislabelled training data scenarios. Therefore, it can be concluded that the utility function forces the network to prioritise certain classes to maximize the utility when it is impossible to obtain good performance across all classes due to the limitation of the capacity of the chosen model, in this example they consider a small neural network with one hidden layer.

[1] the original repository containing the paper's experiments is available at GitHub repository

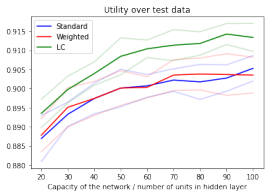

Figure 2: No label noise

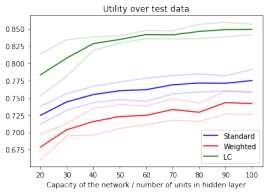

Figure 3: 50% uniformly corrupted label noise

3.3 Scalability to real world applications: Semantic segmentation in Autonomous Driving

In this task the utility function is defined to have preferences for capturing pedestrians, cyclists, and cars over other classes. Figure 4 shows the importance of relying on the framework of Bayesian decision theory (Optimal prediction case) and also the benefits in including the utility into the lower bound that can be translated into a higher utility than models without using prior knowledge of the final task.

Models	Standard Pred.		Optimal Pred.	
	Acc.	Exp. Util.	Acc.	Exp. Util.
Standard	81.1	0.619	78.1	0.676
Weighted	82.1	0.633	79.6	0.682
LC BNN	82.5	**0.652**	81.8	**0.685**

Figure 4: Results over test data for segmentation problem. Standard pred. before integrating over the utility and Optimal pred. are the results after the integration. The utility on test data increases when labels are assigned cording to optimal prediction.

Additionally, it is shown in the paper that in terms of IOU the loss-calibrated approach achieves similar metric values for the classes with lower utility and increases the performance in the priority classes.

3.4 Strong and weak points

Having said that, the contribution has been proved to outperform traditional approaches, to be more robust in complex scenarios with label corruption and limited models. However, it could have been interesting to report initially the performance of a more complex model (e.g. CNNs) capable of solving the required task accurately without Bayesian approach. Then, by comparing this results with the performance of the selected simple architecture, also without the Bayesian approach, it could be proved that there is actually a limitation in the expected performance of the model due to its capacity. This will justify the necessity of applying the prior knowledge about the task in the small model to achieve closer performance to the one reported by the complex model. Finally, with the last experiment it has been seen that the new approach is applicable to real data.

4 Medical Image Classification

In this section we will provide our own results using the paper approach in a real dataset in the health care field. Concretely, we will focus on Chest X-ray images classification. The goal is to classify the images into one of the following classes: Healthy patient/Normal, bacterial, viral or COVID disease. We are going to test the necessity of applying this approach in such a complex task when there is a limitation in the capacity of the model due to its capacity. The data has been extracted from (3) and (4). Moreover, this dataset has been used in other works such as (5).

We have considered two different models. The first one is a Neural Network with three convolutional layers and then a smaller one, similar to the one used in the paper for MNIST dataset experiments.

Firstly, we trained both models without using a prior knowledge of the given task, i.e., we do not use the Bayesian approach. With this experiment we aim to justify the necessity of consider prior

knowledge when using an easier model. The confusion matrices of both classifiers can be seen in 5 and 6. We can see the gap between the performance of the models and then the necessity of applying alternatives to improve the performance of the non-convolutional model.

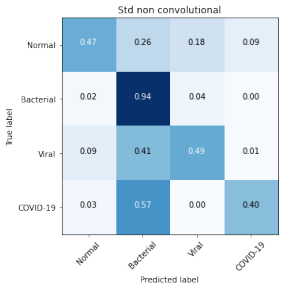

Figure 5: Non-convolutional model

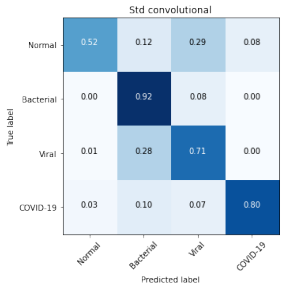

Figure 6: Convolutional model

Then, we trained the simple model defining a utility function that captures the prior knowledge of the task. The utility function can be seen as a matrix seen in Table 1.

Utility function	Normal	Bacterial	Viral	COVID
Normal	1	0	0	0
Bacterial	0	1	0.1	0
Viral	0.1	0.3	1	0
COVID	0.1	0.4	0.7	1.3

Table 1: Utility function for Medical Image Classification

The confusion matrices of the classification, maximizing the conditional gain considering this specific utility function, can be seen in 7. The results enhance the importance of implementing prior knowledge if a simple model is considered. If we compare the results using the Loss-calibration approach with those obtained with the convolutional architecture (6) we can see that the performances are similar.

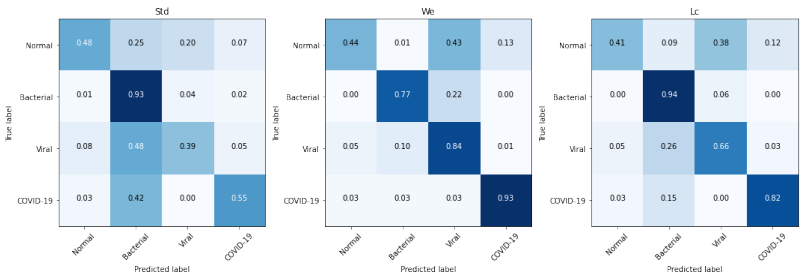

Figure 7: Results for Chest X-rays dataset.

5 Conclusion

In this work we have done a review of the principal aspects of the paper Loss-Calibrated Approximate Inference in Bayesian Neural Networks (1). We explained the contribution of the paper and the results obtained in the experiments they presented proving the point that using a utility-dependent lower bound for training BNNs allows the models to attain better performance for applications that have asymmetric utility function. Furthermore, we applied the paper approach in a real dataset consisting of chest X-rays images and we proved the strength of the proposed methods to overcome limitations due to model capacity for a given task. The implementation for our experiments is available in the GitHub repository BayesianML-project.

References

[1] Cobb, A. D., Roberts, S., & Gal, Y. (2018). Loss-Calibrated Approximate Inference in Bayesian Neural Networks. arXiv (Cornell University). https://arxiv.org/pdf/1805.03901.pdf

[2] Arbel, J., Pitas, K., Vladimirova, M., & Fortuin, V. (2023). A Primer on Bayesian Neural Networks: Review and Debates. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2309.16314

[3] Cohen, J., Morrison, P., & Dao, L. (2020). COVID-19 Image data collection. arXiv (Cornell University). https://github.com/ieee8023/covid-chestxray-dataset

[4] Chest X-Ray images (Pneumonia). (2018, 24 marzo). Kaggle. https://www.kaggle.com/datasets/paultimothymooney/chest-xray-pneumonia

[5] Ghoshal, B., & Tucker, A. (2022). On Calibrated Model Uncertainty in Deep Learning. arXiv. https://arxiv.org/abs/2206.07795