ECAI 2023
K. Gal et al. (Eds.)

doi:10.3233/FAIA230617

Information Bound and Its Applications in Bayesian Neural Networks

Jiaru Zhang[a], Yang Hua[b], Tao Song[a], Hao Wang[c], Zhengui Xue[a], Ruhui Ma[a;*] and Haibing Guan[a]

[a]Shanghai Jiao Tong University
[b]Queen's University Belfast
[c]Louisiana State University

Abstract. Bayesian neural networks have drawn extensive interest because of their distinctive probabilistic representation framework. However, despite its recent success, little work focuses on the information-theoretic understanding of Bayesian neural networks. In this paper, we propose Information Bound as a metric of the amount of information in Bayesian neural networks. Different from mutual information on deterministic neural networks where modification of network structure or specific input data is usually necessary, Information Bound can be easily estimated on current Bayesian neural networks without any modification of network structures or training processes. By observing the trend of Information Bound during training, we demonstrate the existence of the "critical period" in Bayesian neural networks. Besides, we show that the Information Bound can be used to judge the confidence of the model prediction and to detect out-of-distribution datasets. Based on these observations of model interpretation, we propose Information Bound regularization and Information Bound variance regularization methods. The Information Bound regularization encourages models to learn the minimum necessary information and improves the model generality and robustness. The Information Bound variance regularization encourages models to learn more about complex samples with low Information Bound. Extensive experiments on KMNIST, Fashion-MNIST, CIFAR-10, and CIFAR-100 verify the effectiveness of the proposed regularization methods.

1 Introduction

Bayesian neural networks are a unique class of neural networks that typically refer to stochastic artificial neural networks obtained through training with Bayesian inference methods [15, 29]. BNNs can be effectively trained using Variational Inference [5, 6], which has proven particularly useful in large-scale practical applications. One of the primary advantages of Bayesian neural networks is their probabilistic representation of network parameters, hidden representations, and outputs. This characteristic endows Bayesian neural networks with enhanced interpretability for both model weights and predictions. Consequently, Bayesian neural networks have demonstrated considerable potential and have been extensively employed across various tasks, including computer vision [19, 22, 32], natural language processing [39], active learning [20], semi-supervised learning [3], continual learning [25], and reinforcement learning [11].

Although Bayesian neural networks offer a natural probabilistic representation of network parameters and model predictions, there has been limited research focusing on information theory and its applications within the context of Bayesian neural networks. Jae Oh Woo [37] proposed an analytical solution for calculating the mutual information between model parameters and predictive output in Bayesian neural networks. This approach has been employed in active learning to select the most informative data points from datasets. However, to the best of our knowledge, no existing work has addressed the calculation of mutual information between input and intermediate features in Bayesian neural networks. Estimating such mutual information in deterministic neural networks typically necessitates specific input data [33] or modifications to the network structure [2]. This gap in the literature highlights the need for further exploration of information theory in the context of Bayesian neural networks, which could potentially lead to novel insights and applications in various domains.

In this paper, we introduce Information Bound, a novel approach to estimate the mutual information between input and hidden representations in Bayesian Neural Networks. Information Bound serves as an upper bound for the mutual information between the network input and the hidden representation. Within the context of Bayesian neural networks, Information Bound can be viewed as a metric quantifying the amount of information about a specific input. By judiciously selecting the prior distribution $q(\mathbf{z})$ as a Unit Gaussian distribution, $\mathcal{N}(0, \mathbf{I})$, Information Bound can be easily estimated in Bayesian neural networks without necessitating any modifications to the network structure or training process. This innovative approach paves the way for a deeper understanding of the relationship between input and hidden representations in Bayesian neural networks, potentially leading to improved model performance and interpretability.

Information Bound plays a crucial role in interpreting Bayesian neural network models. By monitoring the variation trends of Information Bound during model training, we can confirm the existence of the two periods, "learning" and "forgetting", in the training process of Bayesian neural networks. Furthermore, both theoretical analysis and experimental verification indicate that Bayesian neural network models tend to make more accurate predictions on samples with higher Information Bounds. As a result, Information Bound can be employed to evaluate the confidence of model predictions. Since Information Bound represents the knowledge of a Bayesian neural network on inputs, it can also be utilized to detect out-of-distribution datasets. This capability further emphasizes the importance of In-

[*] Corresponding Author. Email: ruhuima@sjtu.edu.cn

formation Bound in understanding and interpreting Bayesian neural network models.

Based on the model interpretation with the critical period, we propose Information Bound variance regularization to encourage the network to learn the necessary minimum information during training. This technique aims to encourage the network to focus on learning the essential minimum information during the training process. By doing so, it effectively reduces the amount of redundant information captured by the model, leading to improved robustness and generality. Moreover, based on the analysis that models tend to make incorrect predictions on samples with low Information Bound, we introduce Information Bound variance regularization. It is specifically designed to enforce the network to learn more information on complex samples, thereby enhancing the model's performance on challenging instances. To validate the effectiveness of these two regularization methods, we conducted extensive experiments comparing models trained with and without the proposed techniques. The results demonstrate that models incorporating Information Bound variance regularization outperform their original counterparts, highlighting the potential benefits of incorporating these methods into the training process.

In summary, the main contributions of this paper are listed as follows:

- We introduce Information Bound as a metric to measure the quantity of information in Bayesian neural networks. This metric can be easily estimated on Bayesian neural networks without requiring any modifications to the training process or network structure.
- We demonstrate that Information Bound can be used to interpret Bayesian neural networks. By tracking the trend of Information Bound during training, we provide evidence for the existence of a "critical period". Furthermore, we show that Information Bound can be employed to estimate the confidence of model predictions and detect out-of-distribution datasets.
- Building on our model interpretation with Information Bound, we propose two regularization methods: Information Bound regularization and Information Bound variance regularization. Our experiments confirm the effectiveness of these methods in enhancing model performance. The codes are available at `https://github.com/AISIGSJTU/IBBNN`.

2 Related Work

2.1 Bayesian Neural Networks

Unlike deterministic deep neural networks, which obtain a point estimate of model parameters by optimizing a specific objective function, Bayesian neural networks [6, 7, 29] aim to find the posterior distribution of the parameters instead of a point estimate. Variational Inference [5, 6, 13, 18] and Markov Chain Monte Carlo [4, 8] are two mainstream methods for training Bayesian neural networks. In practice, Variational Inference scales better than the Markov Chain Monte Carlo approach and is gaining popularity [21].

Bayesian neural networks have been applied in various domains, such as computer vision and natural language processing. In computer vision, they are commonly used to model the uncertainties of predictions. Gustafsson, Fredrik K et al. applied Bayesian neural networks to enhance model robustness in computer vision [22, 32]. In natural language processing, Xiao et al. [39] studied the benefits of characterizing model and data uncertainties using Bayesian neural networks. Bae et al. proposed detecting and avoiding out-of-distribution data in semi-supervised learning [21]. Ebrahimi et al. introduced a continual learning framework with Bayesian neural networks by retaining the most influential parameters and re-initializing the rest [12]. Similarly, Li et al. employed Bayesian neural networks to enable learning from new tasks without forgetting previously learned tasks [25].

Despite these applications, there is limited work on the information-theoretic exploration of Bayesian neural networks. Mutual information has been shown to quantify epistemic uncertainty in Bayesian neural networks [28], and Jae Oh Woo presented an analytical calculation method for the mutual information between model parameters and the output [37].

2.2 Information Bottleneck Theory

The Information Bottleneck theory, first proposed by Tishby et al. in 2000, aims to identify the minimum necessary information for a given task [35]. This theory has been widely applied to analyze and explain deep neural networks, providing insights into their inner workings and performance [33, 36]. To incorporate the Information Bottleneck model into a neural network, Alemi et al. introduced a variational approximation [2]. This approximation allows the Information Bottleneck to be used as a regularization term, leading to improved generalization performance and robustness in the trained models.

In recent years, researchers have continued to explore the applications of the Information Bottleneck theory in deep learning [31]. For example, Peng et al. developed an efficient data selection method based on the Information Bottleneck to accelerate the training process of deep networks. Similarly, Xu et al. proposed selecting hard examples with high mutual information of the input to enhance Adversarial Training, improving the model's ability to withstand adversarial attacks [40]. Zhai et al. presented an adversarial Information Bottleneck (AIB) method by introducing an adversarial regularization term to estimate the information [43]. This approach combines the Information Bottleneck with adversarial training, further expanding its applicability in deep learning.

To some extent, our proposed Information Bound can be seen as a natural extension of the Information Bottleneck methods applied to Bayesian neural networks. Unlike other methods, which typically require specially designed inputs or modifications to the network structure, our approach does not necessitate any additional changes, making it more practical and versatile for a wide range of applications.

2.3 Critical Periods in Neural Networks

In biology, critical periods refer to specific timeframes in the early stages of postnatal development when sensory deficiencies might lead to long-term skill impairment. Achille et al. demonstrated that critical periods also exist in deep neural networks by using the Fisher Information of the weight matrices to measure the effective connectivity between layers [1]. They found that the Fisher information increases in the initial period of training and then decreases for the remainder of the training.

Furthermore, Golatkar et al. showed that weight decay and data augmentation significantly affect the critical period, and there is also a "critical period" for regularization [16]. On the other hand, Frankle et al. highlighted the importance of the early phase in neural network training [14]. Maennel et al. explained the critical period phenomenon by showing that inactive ReLU units at later layers demonstrate how early specialization occurs in the early layers during training [27].

Yan et al. revealed that the final accuracy of Federated Learning is affected by the early phase of training, which verifies the existence of a critical period in Federated Learning [41]. For applications, You et al. discovered that the critical sub-network can be identified at a very early critical stage [42]. De et al. introduced a new learning target for curricularized learning in the early critical period in reinforcement learning [10].

However, all mentioned above focus on the critical period in deterministic neural networks, and it remains unknown whether the critical period exists in Bayesian neural network training. To the best of our knowledge, our work is the first to demonstrate the critical periods in Bayesian neural networks.

3 Information Bound in Bayesian Neural Networks

3.1 Bayesian Neural Networks with Variational Inference

Suppose we have observations $\mathcal{D} = \{(\mathbf{x}_1, \mathbf{y}_1), (\mathbf{x}_2, \mathbf{y}_2), \dots\}$, Bayesian neural networks parameterized by $\mathbf{W}$ aim to model the real posterior distribution $P(\mathbf{W} \mid \mathcal{D})$ by Bayesian theorem:

$$P(\mathbf{W} \mid \mathcal{D}) = \frac{P(\mathcal{D} \mid \mathbf{W}) P(\mathbf{W})}{P(\mathcal{D})}. \tag{1}$$

Since the formula above is intractable in practice, Bayesian neural networks [6] trained with Variational Inference use a variational distribution $Q_\theta(\mathbf{W})$ to approximate the real posterior probability $P(\mathbf{W}|\mathcal{D})$. The training process involves minimizing the Kullback–Leibler (KL) divergence between the variational distribution and the true posterior distribution:

$$\begin{aligned} KL(Q_\theta(\mathbf{W})||P(\mathbf{W}|\mathcal{D}))) &= -\int Q_\theta(\mathbf{W}) \log \frac{P(\mathbf{W}|\mathcal{D})}{Q_\theta(\mathbf{W})} d\mathbf{W} \\ &= \log P(\mathcal{D}) - \int Q_\theta(\mathbf{W}) \log \frac{P(\mathbf{W}, \mathcal{D})}{Q_\theta(\mathbf{W})} d\mathbf{W}. \end{aligned} \tag{2}$$

Since $\log P(\mathcal{D})$ is a constant for given observations $\mathcal{D}$, minimizing the KL divergence is equivalent to minimizing the following objective function:

$$\begin{aligned} \mathcal{L} &= -\int Q_\theta(\mathbf{W}) \log \frac{P(\mathbf{W}, \mathcal{D})}{Q_\theta(\mathbf{W})} d\mathbf{W} \\ &= \underbrace{-\mathbb{E}_{\mathbf{W} \sim Q_\theta(\mathbf{W})} \log P(\mathcal{D} \mid \mathbf{W})}_{\mathcal{L}_p} + \underbrace{KL(P(\mathbf{W})||Q_\theta(\mathbf{W}))}_{\mathcal{L}_r}. \end{aligned} \tag{3}$$

Note that $-\mathcal{L}$ is a lower bound of $logP(\mathcal{D})$, thus $\mathcal{L}$ is usually called the Evidence Lower Bound (ELBO) loss [5]. It can be divided into two terms. Following previous work [44], the first term is directly related to the predictions, and it is named the prediction loss $\mathcal{L}_p$. The second term can be seen as a regularization of the model parameters, and it is named the regularization loss $\mathcal{L}_r$.

The target of the training process is to find the parameters θ of the variational distribution $Q_\theta(\mathbf{W})$ to minimize $\mathcal{L}$:

$$\theta = \underset{\theta}{\operatorname{argmin}} \mathcal{L}. \tag{4}$$

The prediction $\mathbf{y}$ of a given input $\mathbf{x}$ is obtained by multiple stochastic forward passes through sampling K times from the probability distribution:

$$p(\mathbf{y}|\mathbf{x}, \mathcal{D}) \approx \frac{1}{K} \sum_{k=1}^{K} p(\mathbf{y}|\mathbf{x}, \mathbf{W}_k), \mathbf{W}_k \sim Q_\theta(\mathbf{W}). \tag{5}$$

3.2 Definition of Information Bound

A Bayesian neural network with L layers can be represented in the form of multiple layers as below:

$$\mathbf{z}_0 = \mathbf{x}, \tag{6}$$

$$\mathbf{z}_l \sim P_{\theta_l, \mathbf{z}_{l-1}}(\mathbf{z}_i), \forall l \in 1, \dots L, \tag{7}$$

$$\mathbf{y} = \mathbf{z}_L, \tag{8}$$

where $\mathbf{x}$ and $\mathbf{y}$ represent the input and the output of the Bayesian neural network separately. P_{θ_l} corresponds to a random function of the i-th layer in the Bayesian neural network, where θ_i represents the trainable parameters. This means that for a fixed middle variable $\mathbf{z}_{i-1}$ and parameter θ_l, the output $\mathbf{z}_l$ is a random variable whose distribution is determined by $\mathbf{z}_{l-1}$ and θ_l. Therefore, each run is equivalent to sampling from this distribution.

For a given Bayesian neural network with parameter θ, the hidden variable $\mathbf{z}_l$ in layer l only depends on the output of the previous layer $\mathbf{z}_{l-1}$. Therefore, all hidden representations in the Bayesian neural network generate a Markov chain. From the perspective of information theory, we have

$$I(\mathbf{x}, \mathbf{z}_1) > I(\mathbf{x}, \mathbf{z}_2) > \cdots > I(\mathbf{x}, \mathbf{z}_L), \tag{9}$$

where $I(\cdot, \cdot)$ represents the mutual information between two random variables. Therefore, the mutual information $I(\mathbf{x}, \mathbf{z}_l)$ between input $\mathbf{x}$ and the hidden representation $\mathbf{z}_l$ represents the amount of information in the output of the l-th layer.

In practice, calculating $I(\mathbf{x}, \mathbf{z}_l)$ analytically is infeasible because the distribution $P(\mathbf{x})$ is unknown. To address this issue, we propose estimating an upper bound for $I(\mathbf{x}, \mathbf{z}_l)$ as a substitution. From the definition of mutual information, we have

$$I(\mathbf{x}, \mathbf{z}) = \iint p(\mathbf{z} \mid \mathbf{x}) p(\mathbf{x}) \log \frac{p(\mathbf{z} \mid \mathbf{x})}{p(\mathbf{z})} d\mathbf{x} d\mathbf{z}. \tag{10}$$

As $p(\mathbf{z})$ is intractable, it is impossible to calculate the integral. As a substitution, we suppose $q(\mathbf{z})$ is a knowable distribution, hence

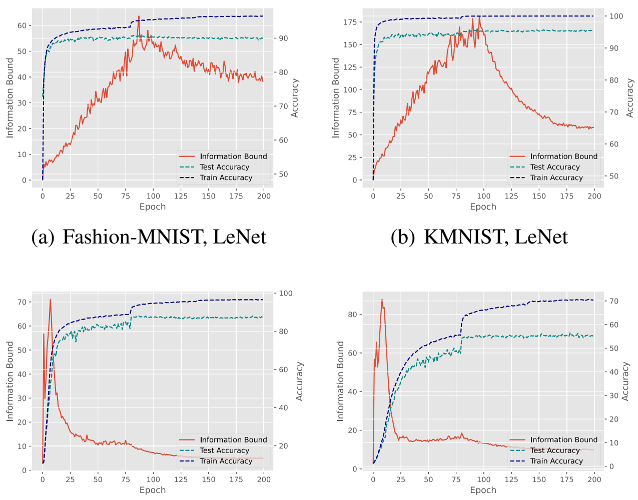
$$\begin{aligned} I(\mathbf{x}, \mathbf{z}) &= \iint p(\mathbf{z} \mid \mathbf{x}) p(\mathbf{x}) \log \frac{p(\mathbf{z} \mid \mathbf{x})}{q(\mathbf{z})} dxdz \\ &\quad + \iint p(\mathbf{z} \mid \mathbf{x}) p(\mathbf{x}) \log \frac{q(\mathbf{z})}{p(\mathbf{z})} dxdz \\ &= \int p(\mathbf{x}) KL(p(\mathbf{z} \mid \mathbf{x}) \| q(\mathbf{z})) d\mathbf{x} - KL(p(\mathbf{z}) \| q(\mathbf{z})) \\ &< \int p(\mathbf{x}) KL(p(\mathbf{z} \mid \mathbf{x}) \| q(\mathbf{z})) \\ &= \mathbb{E}_{\mathbf{x}} KL(p(\mathbf{z} \mid \mathbf{x}) \| q(\mathbf{z})). \end{aligned} \tag{11}$$

Therefore, $\mathbb{E}_{\mathbf{x}} KL(p(\mathbf{z} \mid \mathbf{x}) \| q(\mathbf{z}))$ is an upper bound of $I(\mathbf{x}, \mathbf{z})$. We name $KL(p(\mathbf{z} \mid \mathbf{x} \| q(\mathbf{z})))$ as **Information Bound** in Bayesian neural networks. It depends on the input $\mathbf{x}$ and the hidden representation variable distribution $\mathbf{z}$, and can be used to estimate the amount of information in Bayesian neural networks.

4 Model Interpretation with Information Bound

4.1 Information Bound Calculation

Bayesian neural networks model the probabilistic distribution of $p(\mathbf{z}_l|\mathbf{x})$ explicitly as the output of layer l. Consider a Bayesian linear layer with input $\mathbf{x} \in \mathbb{R}^n$ and output $\mathbf{z} \in \mathbb{R}^m$. Suppose its weight



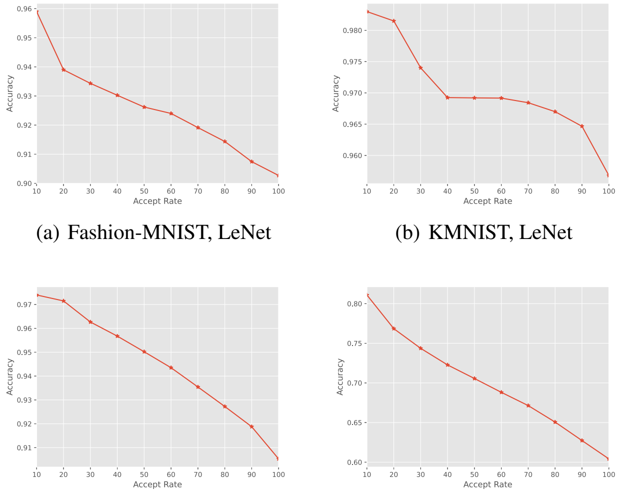
(a) Fashion-MNIST, LeNet

(b) KMNIST, LeNet

(c) CIFAR-10, VGG

(d) CIFAR-100, VGG

Figure 1. Trends of Information Bound and Accuracy during Bayesian neural networks training. *Best viewed in color.*



(a) Fashion-MNIST, LeNet

(b) KMNIST, LeNet

(c) CIFAR-10, VGG

(d) CIFAR-100, VGG

Figure 2. Trends of accuracy with varying accept rates according to Information Bounds. It verifies the effectiveness of Information Bound as a metric of model confidence. *Best viewed in color.*

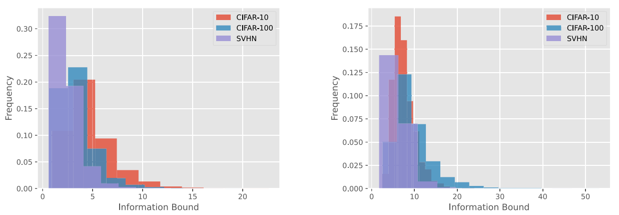
matrix W is a random matrix with $W_{ij} \sim N(M_{ij}, A_{ij}^2)$, where $M, A \in \mathbb{R}^{m \times n}$ are matrices representing the expectation and standard error, respectively. Ignoring the bias term for the convenience of our derivation, we have

$$\mathbf{z}_i = \sum_j W_{ij}\mathbf{x}_j. \tag{12}$$

According to the adding rule of Gaussian variables, we have

$$p(\mathbf{z}_i|\mathbf{x}) \sim \mathcal{N}(\sum_j M_{ij}\mathbf{x}_j, \sum_j A_{ij}^2\mathbf{x}_j^2). \tag{13}$$

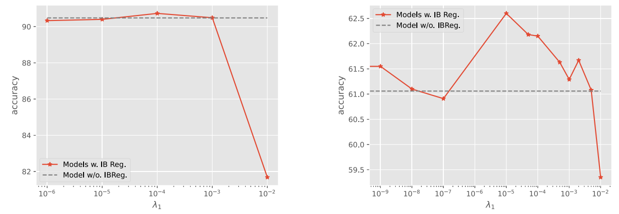
In practice, we set the knowable distribution $q(\mathbf{z})$ as a unit Gaus-



(a) VGG trained on CIFAR-10

(b) VGG trained on CIFAR-100

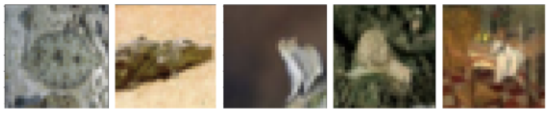
Figure 3. Information Bounds of VGG models trained on CIFAR-10 and CIFAR-100 with in-distribution data and out-of-distribution dataset. *Best viewed in color.*



(a) Fashion-MNIST, LeNet

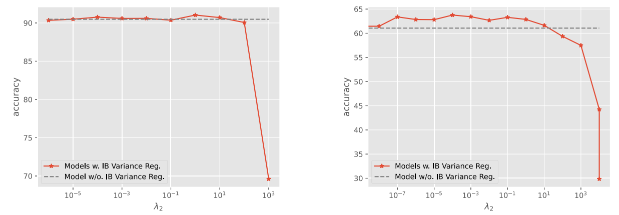
(b) CIFAR-100, VGG

Figure 4. Trends of accuracy of models trained with Information Bound regularization with different hyperparameter λ_1 in Equation (16. *Best viewed in color.*



Label	flatfish	crocodile	squirrel	snail	table
Old pred, IB	ray, 2.52	lizard, 2.77	possum, 2.72	snail, 2.84	table, 2.86
New pred, IB	flatfish, 7.32	crocodile, 8.12	squirrel, 5.30	porcupine, 8.12	table, 7.03

Figure 5. The five images with the lowest Information Bounds in CIFAR-100, and their labels, predictions, and Information Bounds. The model trained with Information Bound variance regularization keeps more Information and predicts more accurately. *Best viewed in color.*



(a) Fashion-MNIST, LeNet

(b) CIFAR-100, VGG

Figure 6. Trends of the accuracy of models trained with Information Bound variance regularization with different hyperparameter λ_2 in Equation (17). *Best viewed in color.*

Algorithm 1: Information Bound Calculation on a Bayesian Layer

```
Input: Input $\mathbf{x}$, Layer Weight $W$, Layer Bias $\mathbf{b}$, Layer
       operator $f$
Output: Information Bound $IB$ for the Input $\mathbf{x}$
Denote $M$, $A$ as the expectation and standard deviation
  matrix of $W$.
Denote $\mu_{\mathbf{b}}$, $\sigma_{\mathbf{b}}$ as the expectation and standard deviation of $\mathbf{b}$.
$\mu_{\mathbf{z}} = f(\mathbf{x}, M, \mu_{\mathbf{b}})$
$\sigma_{\mathbf{z}} = \sqrt{f(\mathbf{x}^2, A^2, (\sigma_{\mathbf{b}})^2)}$
$IB = -\log(\sigma_{\mathbf{z}}) + 0.5(\sigma_{\mathbf{z}}.pow(2) + \mu_{\mathbf{z}}.pow(2)) - 0.5$
```

sian distribution $\mathcal{N}(\mathbf{0}, \mathbf{I})$. Therefore, the Information Bound can be obtained by analytically calculating the KL divergence of two Gaussian Distributions:

$$\begin{aligned} IB(\mathbf{x}, \mathbf{z}) &= KL(p(\mathbf{z}|\mathbf{x}) \| q(\mathbf{z})) \\ &= -\frac{1}{2}\sum_{i=1}^{m}\Bigg(1 + \log \sum_j A_{ij}^2 \mathbf{x}_j^2 - \\ &\qquad \sum_j A_{ij}^2 \mathbf{x}_j^2 - \sum_j M_{ij}\mathbf{x}_j\Bigg). \end{aligned} \tag{14}$$

Similarly, for convolutional layers with a kernel matrix W which is a random matrix with $W_{ij} \sim \mathcal{N}(M_{ij}, A_{ij}^2)$, we have

$$p(\mathbf{z}|\mathbf{x}) \sim \mathcal{N}(conv(\mathbf{x}, M), conv(\mathbf{x}^2, A^2)). \tag{15}$$

Therefore, the Information Bound with the output of convolutional layers can be calculated accordingly. The procedure to calculate the Informaton Bound in Bayesian neural networks is provided in Algorithm 1.

4.2 Critical Periods in BNNs

In Section 2.3, we demonstrate that the critical period is a fundamental concept in traditional neural networks. The Information Bound serves as a metric for estimating the amount of information contained within Bayesian neural networks model. By monitoring the Information Bound during training, we can assess the trend of information in the model. Consequently, we can validate the presence of the Critical Period in Bayesian Neural Networks training by incorporating the Information Bound.

We examine the trends of Information Bound and accuracy during the training of Bayesian LeNet [24] and Bayesian VGG [34] on four widely-used datasets, namely Fashion-MNIST [38], KMNIST [9], CIFAR-10, and CIFAR-100 [23]. As depicted in Fig. 1, the Information Bound experiences a sharp increase during the initial phase of training. Subsequently, the Information Bound consistently decreases throughout the remainder of the training process. Interestingly, both training accuracy and test accuracy continue to rise during the entire training period, despite the decline in Information Bound. This observation suggests that the network initially acquires a substantial amount of information before gradually forgetting most of it, ultimately retaining only the essential information relevant to the task at hand.

4.3 Model Confidence Evaluation

In Bayesian neural networks, the Information Bound values for different inputs can vary significantly. This variation implies that a model possesses more knowledge about an input image when the corresponding Information Bound is large. Consequently, predictions with higher Information Bounds are likely to be more accurate, while those with lower Information Bounds tend to be incorrect. Our experimental results, as illustrated in Fig. 2 using the CIFAR-10 and CIFAR-100 datasets, support this analysis.

To further demonstrate the relationship between Information Bound and prediction accuracy, we present an experiment where only a subset of predictions is accepted based on their corresponding Information Bound values, while the remaining predictions are rejected. As more predictions are retained, the overall accuracy decreases, confirming that predictions with lower Information Bound values are more prone to being incorrect. This observation underscores the value of Information Bound as a reliable and meaningful metric for estimating the confidence level of model predictions. By leveraging the Information Bound metric, practitioners can gain insights into the reliability of the predictions from Bayesian neural networks and make more informed decisions when using these models in real-world applications.

4.4 Out-of-Distribution Dataset Detection

A low Information Bound indicates that the Bayesian neural network model has limited understanding of the input data, suggesting that the model may not be capturing the underlying patterns and relationships within the data effectively. This property of Information Bound can be leveraged to detect out-of-distribution datasets. For example, if a dataset contains a large number of samples with Information Bounds substantially lower than those observed in in-distribution samples, it is highly probable that the dataset is out-of-distribution. This implies that the model may not generalize well to this new dataset, as it is likely to be significantly different from the data the model was trained on.

The experimental results presented in Fig. 3 support our analysis. We train two VGG models on the CIFAR-10 and CIFAR-100 datasets. For each model, we display the Information Bound of the in-distribution datasets (CIFAR-10 and CIFAR-100) and the out-of-distribution datasets (CIFAR-10, CIFAR-100, and SVHN [30]). The results reveal that the distribution of Information Bound for in-distribution datasets is higher than that of out-of-distribution datasets. Furthermore, the Information Bound of SVHN is smaller than that of the CIFAR datasets, indicating a more significant difference between the SVHN dataset and the CIFAR datasets than the difference between the CIFAR datasets themselves. All experimental results validate that Information Bound can be effectively utilized for out-of-distribution dataset detection.

Table 1. Comparison of models trained **with Information Bound regularization** and **without Information Bound regularization**. The mean value and maximum deviation of three runs are reported.

Model	Dataset	Acc. w/o. IB Reg.	Acc. w. IB Reg.
LeNet	KMNIST	95.49 ± 0.26	$\mathbf{95.73 \pm 0.39}$
LeNet	Fashion-MNIST	90.48 ± 0.37	$\mathbf{90.87 \pm 0.14}$
VGG	CIFAR-10	91.03 ± 0.12	$\mathbf{91.46 \pm 0.20}$
VGG	CIFAR-100	61.06 ± 0.86	$\mathbf{62.13 \pm 0.50}$

Table 2. Comparison on the Robustness of Models without Information Bound Regularization and with Information Bound Regularization. The mean value and maximum deviation of three runs are reported.

Model	Dataset	Attack	ℓ_∞ norm	Acc. w/o. IB Reg. (%)	Acc. w. IB Reg. (%)	Δ (%)
LeNet	KMNIST	/	0	95.49 ± 0.26	95.73 ± 0.39	**+ 0.24**
		FGSM	$1/255$	94.61 ± 0.38	94.89 ± 0.20	**+ 0.27**
			$2/255$	93.46 ± 0.51	93.89 ± 0.03	**+ 0.43**
			$4/255$	90.49 ± 0.92	91.06 ± 0.61	**+ 0.57**
		PGD	$1/255$	94.62 ± 0.36	94.92 ± 0.12	**+ 0.30**
			$2/255$	93.37 ± 0.50	93.83 ± 0.12	**+ 0.46**
			$4/255$	89.88 ± 0.98	90.32 ± 0.60	**+ 0.44**
LeNet	Fashion-MNIST	/	0	90.48 ± 0.37	90.87 ± 0.14	**+ 0.39**
		FGSM	$1/255$	82.49 ± 0.14	83.53 ± 0.19	**+ 1.04**
			$2/255$	73.06 ± 0.63	74.48 ± 0.88	**+ 1.42**
			$4/255$	57.67 ± 2.39	57.95 ± 2.25	**+ 0.28**
		PGD	$1/255$	82.04 ± 0.21	83.30 ± 0.23	**+ 1.26**
			$2/255$	69.39 ± 0.92	71.09 ± 1.10	**+ 1.70**
			$4/255$	41.92 ± 1.47	40.58 ± 1.76	-1.34
VGG	CIFAR-10	/	0	91.03 ± 0.12	91.46 ± 0.20	**+ 1.43**
		FGSM	$1/255$	64.17 ± 0.31	64.50 ± 0.99	**+ 0.33**
			$2/255$	39.53 ± 0.62	40.41 ± 0.44	**+ 0.88**
			$4/255$	20.00 ± 0.93	19.86 ± 0.28	-0.14
		PGD	$1/255$	63.05 ± 0.51	63.36 ± 0.95	**+ 0.31**
			$2/255$	15.54 ± 0.28	16.08 ± 0.49	**+ 0.54**
			$4/255$	0.16 ± 0.06	0.40 ± 0.63	**+ 0.24**
VGG	CIFAR-100	/	0	61.06 ± 0.86	62.13 ± 0.50	$+$**1.07**
		FGSM	$1/255$	26.68 ± 1.83	28.29 ± 0.78	**+ 1.61**
			$2/255$	13.90 ± 1.08	14.71 ± 0.66	**+ 0.81**
			$4/255$	7.34 ± 0.63	7.40 ± 0.46	**+ 0.06**
		PGD	$1/255$	25.13 ± 1.68	26.48 ± 0.40	**+ 1.35**
			$2/255$	2.93 ± 0.45	3.14 ± 0.42	**+ 0.21**
			$4/255$	0.04 ± 0.01	0.05 ± 0.02	**+ 0.01**

5 Regularization Methods based on Information Bound

5.1 Information Bound Regularization

Definition. As shown in Sec. 4.2, the Bayesian neural networks models initially acquire a substantial amount of information but subsequently forgets most of it during the whole learning process, ultimately retaining only the least necessary information. Inspired by this observation and the information constraint method presented in 1, we propose to incorporate Information Bound as a regularization term. This leads to the formulation of the following objective function, which we aim to minimize during the training of Bayesian Neural Networks:

$$\mathcal{L} = \mathcal{L}_p + \mathcal{L}_r + \lambda_1 \cdot \frac{1}{n} \sum_{i=0}^{n} IB(X_i, Z_i), \tag{16}$$

where $\mathcal{L}_p$ and $\mathcal{L}_r$ are defined in Equation (3), and λ_1 is a hyperparameter to control the ratio of Information Bound regularization. A larger λ_1 encourages the model to discard more useless information while increasing the risk of discarding necessary information.

Ablation Study. In order to investigate the impact of different parameter settings for λ_1, we conduct an analysis of the performance of Bayesian LeNet and Bayesian VGG models trained using Information Bound regularization. Our experiments are carried out on two widely-used benchmark datasets, Fashion-MNIST and CIFAR-100. The experimental results are depicted in Fig. 4, where we systematically vary the value of λ_1 and observe its influence on the accuracy of the trained models. The outcomes demonstrate that the proposed Information Bound regularization maintains its effectiveness across a broad spectrum of λ_1 values. The insensitivity of our method to the specific value of λ_1 implies that it can be easily adopted in various practical scenarios without the need for extensive hyperparameter tuning, making it a valuable addition to the arsenal of tools available for training Bayesian neural networks models.

Performance Improvements. We present the classification accuracy in Table 1 to substantiate the claim that the Information Bound regularization method significantly improves the performance of the models. We employ Bayesian neural networks with LeNet and VGG architectures, trained on a diverse set of datasets, including KMNIST, Fashion-MNIST, CIFAR-10, and CIFAR-100. This selection of datasets ensures a comprehensive evaluation of the proposed approach across various domains and challenges. The comparative analysis of the models trained with Information Bound regularization and their original counterparts reveals that our method consistently outperforms the latter in all scenarios. This observation confirms the effectiveness of the proposed approach and highlights its generality.

Adversarial Robustness. According to the observations and analysis in previous work [5], the Variational Information Bottleneck method makes models more robust to adversarial samples. Similarly, Information Bound regularization improves the model's adversarial robustness because of the reduced redundant information. The Fast Gradient Sign Method (FGSM) [17] is a related simple attack method, and the Projected Gradient Descent method (PGD) [26] is a more sophisticated and powerful adversarial attack method. We test the adversarial robustness of original models and models trained with Information Bound regularization on defending FGSM and PGD attacks. To display the adversarial robustness more comprehensively, we present the accuracy of models in defending against adversarial attacks with ℓ_∞ norms of $1/255$, $2/255$, and $4/255$. The experimental results are presented in Table 2. Models trained with Information Bound regularization are more robust in defending all the noises, further verifying the generality of the proposed method.

Table 3. Comparison of models trained **with Information Bound variance regularization** and **without Information Bound variance regularization**. The mean value and maximum deviation of three runs are reported.

Model	Dataset	Acc. w/o. IB Var. Reg.	Acc. w. IB Var. Reg.
LeNet	KMNIST	95.49 ± 0.26	$\mathbf{95.67 \pm 0.34}$
LeNet	Fashion-MNIST	$\mathbf{90.48 \pm 0.37}$	90.43 ± 0.39
VGG	CIFAR-10	91.03 ± 0.12	$\mathbf{91.05 \pm 0.10}$
VGG	CIFAR-100	61.06 ± 0.86	$\mathbf{63.13 \pm 0.63}$

5.2 *Information Bound Variance Regularization*

Definition. As discussed in Sec. 4.3, Bayesian neural networks tend to make incorrect predictions on examples characterized by lower Information Bound values, primarily due to the limited information retained by the models for such instances. Drawing inspiration from these observations, we introduce Information Bound variance regularization, a novel approach designed to encourage the model to acquire more knowledge about examples with low information content. To implement the Information Bound variance regularization, we formulate the following objective function:

$$\mathcal{L} = \mathcal{L}_p + \mathcal{L}_r + \lambda_2 \cdot Var_i(IB(X_i, Z_i)), \tag{17}$$

where $\mathcal{L}_p$ and $\mathcal{L}_r$ are defined in Equation (3), $Var_i(\cdot)$ means the variance across a batch, and λ_2 is a hyperparameter to control the ratio of Information Bound variance regularization. The Information Bound variance regularization method aims to promote a more balanced learning process across all samples. By doing so, it ensures that the model pays greater attention to the hard samples, which might have been overlooked by the original model. As a result, the overall performance of the model is enhanced, leading to a more robust and accurate representation of the underlying data distribution.

Ablation Study. In this section, we aim to investigate the influence of varying the parameter λ_2 on the performance of Bayesian LeNet and Bayesian VGG models trained using the Information Bound variance regularization technique. To this end, we conduct experiments on two benchmark datasets, Fashion-MNIST and CIFAR-100, and present the accuracy results for different λ_2 settings in Fig. 4. Our findings reveal that the proposed Information Bound variance regularization method consistently enhances the model performance across a wide range of λ_2 values, spanning from 10^{-6} to 10^{1}. This observation suggests that our method exhibits robustness and is not highly sensitive to the choice of the hyperparameter λ_2. Among the tested values, the model with $\lambda_2 = 10^{-4}$ demonstrates a marginally superior performance compared to its counterparts. Consequently, we adopt $\lambda_2 = 10^{-4}$ as the optimal hyperparameter setting for the subsequent experiments in this paper.

Performance Improvements. We showcase the impact of Information Bound variance regularization by presenting the five images with the smallest amounts of Information Bound in the CIFAR-100 dataset on the Bayesian VGG model. The model trained without Information Bound variance regularization misclassifies three of these images. In contrast, the model trained with Information Bound variance regularization maintains a higher Information Bound on these images and successfully classifies four of them. This result validates our motivation and demonstrates that improving Information Bound on complex examples enhances model performance.

To further verify the effectiveness of the Information Bound variance regularization method, we present the classification accuracy of Bayesian neural networks with LeNet and VGG structures trained on KMNIST, Fashion-MNIST, CIFAR-10, and CIFAR-100 datasets. The models trained with Information Bound variance regularization outperform the original models, confirming the effectiveness and generality of our proposed approach. This evidence highlights the potential of Information Bound variance regularization as a valuable technique for improving the performance of Bayesian neural networks across various tasks and datasets.

6 Conclusion

In this paper, we introduce a novel metric called Information Bound, which aims to quantify the amount of information present in Bayesian neural networks. This metric is applicable to both linear and convolutional layers within these networks, and it can be calculated without the need for modifying the network structures or altering the training processes. The introduction of Information Bound allows us to make several key observations and discoveries. Firstly, we observe and prove that the critical period phenomenon, which has been previously identified in other learning systems, also exists during the training of Bayesian neural networks. This finding has significant implications for understanding the learning dynamics of these networks. Secondly, we find that models tend to make incorrect predictions on examples with lower Information Bound values. This suggests that the Information Bound metric can also serve as an indicator of the confidence level of model predictions, providing valuable insights into the reliability of the model's outputs. Building on these observations, we propose two regularization methods that leverage the Information Bound metric: Information Bound regularization and Information Bound variance regularization. The Information Bound regularization method enforces models to focus on learning the most essential information, thereby improving their overall performance. On the other hand, the Information Bound variance regularization method encourages models to acquire more information on challenging examples with low Information Bound values, which can lead to better generalization on difficult cases. Extensive experiments validate our proposal and show that models trained with the two regularization methods outperform the original models. The Information Bound regularization method also improves the adversarial robustness of models, which further expands their application scenarios.

References

[1] Alessandro Achille, Matteo Rovere, and Stefano Soatto, 'Critical Learning Periods in Deep Networks', in *ICLR*, (2018).

[2] Alexander A. Alemi, Ian Fischer, Joshua V. Dillon, and Kevin Murphy, 'Deep Variational Information Bottleneck', in *ICLR*, (2017).

[3] Jinsoo Bae, Minjung Lee, and Seoung Bum Kim, 'Safe Semi-Supervised Learning Using a Bayesian Neural Network', *Information Sciences*, **612**, 453–464, (2022).

[4] Rémi Bardenet, Arnaud Doucet, and Chris Holmes, 'On Markov Chain Monte Carlo Methods for Tall Data', *The Journal of Machine Learning Research*, **18**(1), 1515–1557, (2017).

[5] 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).

[6] Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra, 'Weight Uncertainty in Neural Network', in *ICML*, (2015).

[7] Wray L Buntine, 'Bayesian Backpropagation', *Complex Systems*, **5**, 603–643, (1991).

[8] Tianqi Chen, Emily Fox, and Carlos Guestrin, 'Stochastic Gradient Hamiltonian Monte Carlo', in *ICML*, (2014).

[9] Tarin Clanuwat, Mikel Bober-Irizar, Asanobu Kitamoto, Alex Lamb, Kazuaki Yamamoto, and David Ha, 'Deep Learning for Classical Japanese Literature', *arXiv preprint arXiv:1812.01718*, (2018).

[10] Roy De Kleijn, Deniz Sen, and George Kachergis, 'A Critical Period for Robust Curriculum-Based Deep Reinforcement Learning of Sequential Action in a Robot Arm', *Topics in Cognitive Science*, **14**(2), 311–326, (2022).

[11] Stefan Depeweg, José Miguel Hernández-Lobato, Finale Doshi-Velez, and Steffen Udluft, 'Learning and Policy Search in Stochastic Dynamical Systems with Bayesian Neural Networks', in *ICLR*, (2017).

[12] Sayna Ebrahimi, Mohamed Elhoseiny, Trevor Darrell, and Marcus Rohrbach, 'Uncertainty-guided Continual Learning with Bayesian Neural Networks', in *ICLR*, (2020).

[13] Nick Foti, Jason Xu, Dillon Laird, and Emily Fox, 'Stochastic Variational Inference for Hidden Markov Models', in *NeurIPS*, (2014).

[14] Jonathan Frankle, David J. Schwab, and Ari S. Morcos, 'The Early Phase of Neural Network Training', in *ICLR*, (2020).

[15] Yarin Gal, 'Uncertainty in Deep Learning', *Ph.D.'s Thesis, University of Cambridge*, (2016).

[16] Aditya Sharad Golatkar, Alessandro Achille, and Stefano Soatto, 'Time Matters in Regularizing Deep Networks: Weight Decay and Data Augmentation Affect Early Learning Dynamics, Matter Little Near Convergence', in *NeurIPS*, (2019).

[17] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy, 'Explaining and Harnessing Adversarial Examples', in *ICLR*, (2015).

[18] Alex Graves, 'Practical Variational Inference for Neural Networks', in *NeurIPS*, (2011).

[19] Fredrik K Gustafsson, Martin Danelljan, and Thomas B Schön, 'Evaluating Scalable Bayesian Deep Learning Methods for Robust Computer Vision', in *CVPR Workshops*, (2020).

[20] José Miguel Hernández-Lobato and Ryan Adams, 'Probabilistic Backpropagation for Scalable Learning of Bayesian Neural Networks', in *ICML*, (2015).

[21] Laurent Valentin Jospin, Hamid Laga, Farid Boussaid, Wray Buntine, and Mohammed Bennamoun, 'Hands-On Bayesian Neural Networks—A Tutorial for Deep Learning Users', *IEEE Computational Intelligence Magazine*, **17**(2), 29–48, (2022).

[22] Alex Kendall and Yarin Gal, 'What Uncertainties Do We Need in Bayesian Deep Learning for Computer Vision?', in *NeurIPS*, (2017).

[23] A Krizhevsky, 'Learning Multiple Layers of Features from Tiny Images', *Master's Thesis, University of Toronto*, (2009).

[24] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner, 'Gradient-Based Learning Applied to Document Recognition', *Proceedings of the IEEE*, **86**(11), 2278–2324, (1998).

[25] Honglin Li, Payam Barnaghi, Shirin Enshaeifar, and Frieder Ganz, 'Continual Learning using Bayesian Neural Networks', *IEEE Transactions on Neural Networks and Learning Systems*, **32**(9), 4243–4252, (2020).

[26] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu, 'Towards Deep Learning Models Resistant to Adversarial Attacks', in *ICLR*, (2018).

[27] Hartmut Maennel, Ibrahim M Alabdulmohsin, Ilya O Tolstikhin, Robert Baldock, Olivier Bousquet, Sylvain Gelly, and Daniel Keysers, 'What Do Neural Networks Learn When Trained with Random Labels?', in *NeurIPS*, (2020).

[28] Hermann G Matthies, 'Quantifying Uncertainty: Modern Computational Representation of Probability and Applications', in *Extreme Man-Made and Natural Hazards in Dynamics of Structures*, 105–135, (2007).

[29] Radford M Neal, *Bayesian Learning for Neural Networks*, volume 118, Springer Science & Business Media, 2012.

[30] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y. Ng, 'Reading Digits in Natural Images with Unsupervised Feature Learning', in *NeurIPS Workshop*, (2011).

[31] Xinyu Peng, Jiawei Zhang, Fei-Yue Wang, and Li Li, 'Drill the Cork of Information Bottleneck by Inputting the Most Important Data', *IEEE Transactions on Neural Networks and Learning Systems*, **33**(11), 6360–6372, (2021).

[32] Buu Truong Phan, *Bayesian Deep Learning and Uncertainty in Computer Vision*, Master's thesis, University of Waterloo, 2019.

[33] Ravid Shwartz-Ziv and Naftali Tishby, 'Opening the Black Box of Deep Neural Networks via Information', *arXiv preprint arXiv:1703.00810*, (2017).

[34] Karen Simonyan and Andrew Zisserman, 'Very Deep Convolutional Networks for Large-Scale Image Recognition', in *ICLR*, (2015).

[35] Naftali Tishby, Fernando C Pereira, and William Bialek, 'The Information Bottleneck Method', *arXiv preprint physics/0004057*, (2000).

[36] Naftali Tishby and Noga Zaslavsky, 'Deep Learning and the Information Bottleneck Principle', in *ITW*, (2015).

[37] Jae Oh Woo, 'Analytic Mutual Information in Bayesian Neural Networks', *arXiv preprint arXiv:2201.09815*, (2022).

[38] Han Xiao, Kashif Rasul, and Roland Vollgraf, 'Fashion-MNIST: A Novel Image Dataset for Benchmarking Machine Learning Algorithms', *arXiv preprint arXiv:1708.07747*, (2017).

[39] Yijun Xiao and William Yang Wang, 'Quantifying Uncertainties in Natural Language Processing Tasks', in *AAAI*, (2019).

[40] Mengting Xu, Tao Zhang, Zhongnian Li, and Daoqiang Zhang, 'InfoAT: Improving Adversarial Training Using the Information Bottleneck Principle', *IEEE Transactions on Neural Networks and Learning Systems*, (2022). Early Access.

[41] Gang Yan, Hao Wang, and Jian Li, 'Seizing Critical Learning Periods in Federated Learning', in *AAAI*, (2022).

[42] Haoran You, Chaojian Li, Pengfei Xu, Yonggan Fu, Yue Wang, Xiaohan Chen, Richard G Baraniuk, Zhangyang Wang, and Yingyan Lin, 'Drawing Early-Bird Tickets: Toward More Efficient Training of Deep Networks', in *ICLR*, (2019).

[43] Penglong Zhai and Shihua Zhang, 'Adversarial Information Bottleneck', *IEEE Transactions on Neural Networks and Learning Systems*, (2022). Early Access.

[44] Jiaru Zhang, Yang Hua, Tao Song, Hao Wang, Zhengui Xue, Ruhui Ma, and Haibing Guan, 'Improving Bayesian Neural Networks by Adversarial Sampling', in *AAAI*, (2022).