Robust Bayesian Neural Networks by Spectral Expectation Bound Regularization

Jiaru Zhang[1] Yang Hua[2] Zhengui Xue[1] Tao Song[1] Chengyu Zheng[1] Ruhui Ma[1] Haibing Guan[1]

[1]Shanghai Jiao Tong University [2]Queen's University Belfast

CVPR VIRTUAL JUNE 19-25

Introduction

Background: Bayesian neural networks have been widely used in many applications because of the distinctive probabilistic representation framework. Even though Bayesian neural networks have been found more robust to adversarial attacks compared with vanilla neural networks, their ability to deal with adversarial noises in practice is still limited.

Goal: Improve the adversarial robustness of Bayesian neural networks.

Key Contributions:

- Apply the Lipschitz constraint in Bayesian neural networks, and propose Spectral Expectation Bound Regularization (SEBR) method to enhance the adversarial robustness.
- Prove that SEBR reduces the uncertainty effectively in theoretical analysis, and provide another explanation of the model robustness.
- Verify the theory and the effectiveness of the proposed method by experiments under multiple situations.

Spectral Expectation Bound Regularization

Theorem 1 presents that the expectation of disturbance of the output in a layer of Bayesian neural network is bounded by the expectation of the spectral norm of parameter matrix $\mathbb{E}\|W\|_2$, the length of the perturbation vector $\|\xi\|$, and the Lipschitz constant of the activation function $Lip(f)$.

Theorem 1 *Consider function $f_{\mathbf{W}}(\mathbf{x}) = f(W\mathbf{x} + \mathbf{b})$, where the activation function $f(\cdot)$ is Lipschitz continuous with Lipschitz constant $Lip(f)$. For any perturbation $\boldsymbol{\xi}$ with norm $\|\boldsymbol{\xi}\|$, we have*

$$\mathbb{E}_{\mathbf{W}} \|f_{\mathbf{W}}(\mathbf{x} + \boldsymbol{\xi}) - f_{\mathbf{W}}(\mathbf{x})\| \le Lip(f) \cdot \mathbb{E}\|W\|_2 \cdot \|\boldsymbol{\xi}\|,$$

Where $\|W\|_2$ represents the spectral norm of matrix W.

To accelerate the training process, we propose a method to fast estimate the upper bound of $\mathbb{E}\|W\|_2$ analytically.

Theorem 2 *Consider a Gaussian random matrix $W \in \mathbb{R}^{m\times n}$, where $W_{ij} \sim N(M_{ij}, A_{ij}^2)$ with $M, A \in \mathbb{R}^{m\times n}$. Suppose $G \in \mathbb{R}^{m\times n}$ is a zero-mean Gaussian random matrix with the same variance, i.e., $G_{ij} \sim N(0, A_{ij}^2)$. We have*

$$\mathbb{E}\|W\|_2 \le \|M\|_2 + c\left(\max_i \|A_{i,:}\| + \max_j \|A_{:,j}\| + \mathbb{E}\max_{i,j}|G_{ij}|\right),$$

where c is a constant independent of W.

Adding the upper bound of $\mathbb{E}\|W\|_2$ in each layer as a regularisation term into the loss function, we propose our Spectral Expectation Bound Regularization (SEBR) method.

Influence on Uncertainties

Theoretical Analysis:

Our SEBR method can reduce the epistemic uncertainty on the output of a Bayesian neural network model.

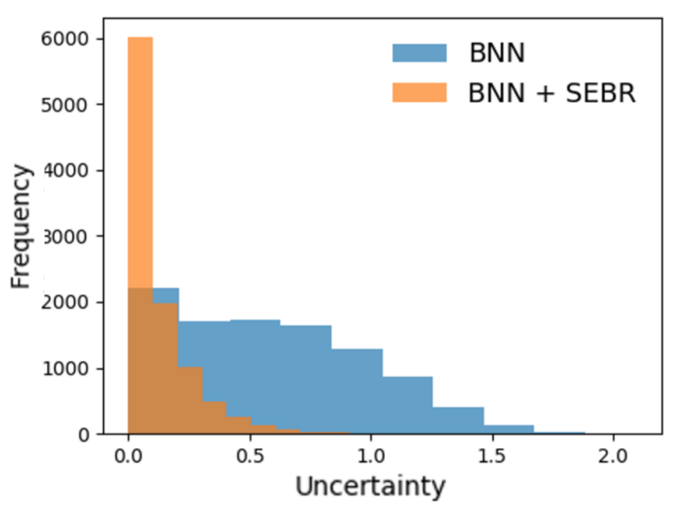
Theorem 3 *Consider a Bayesian neural network with only a linear layer $f_{\mathbf{W}}(\mathbf{x}) = W\mathbf{x} + \mathbf{b}$, where $\mathbf{x} \in \mathbb{R}^n$, $W \in \mathbb{R}^{m\times n}$. Denote the epistemic uncertainty (following the definition in Equation of the output after one step gradient descent without SEBR as H_e, and the epistemic uncertainty after one step gradient descent with SEBR as H_e'. With sufficient sampling times, we have*

$$H_e' \le H_e. \quad (1)$$

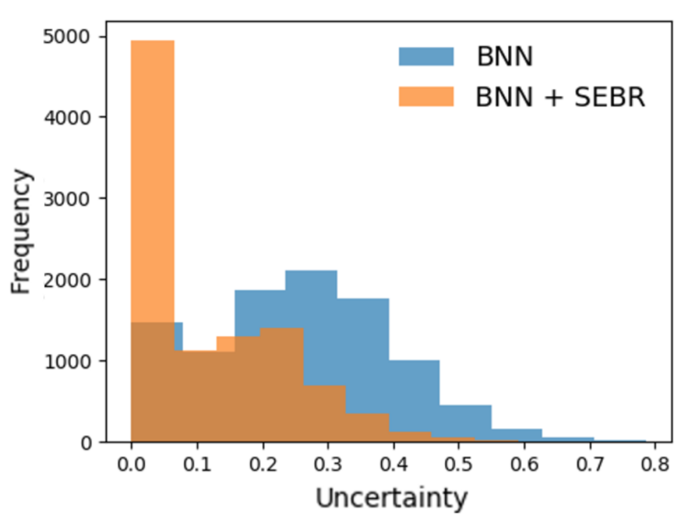
Additionally, the aleatoric uncertainty is also reduced because of the optimization on the spectral norm of the mean matrix $\|M\|_2$.

Experimental Verifications:

Models trained with SEBR have both lower aleatoric uncertainties and lower epistemic uncertainties.



(a) Aleatoric Uncertainty



(b) Epistemic Uncertainty

Experiments & Results

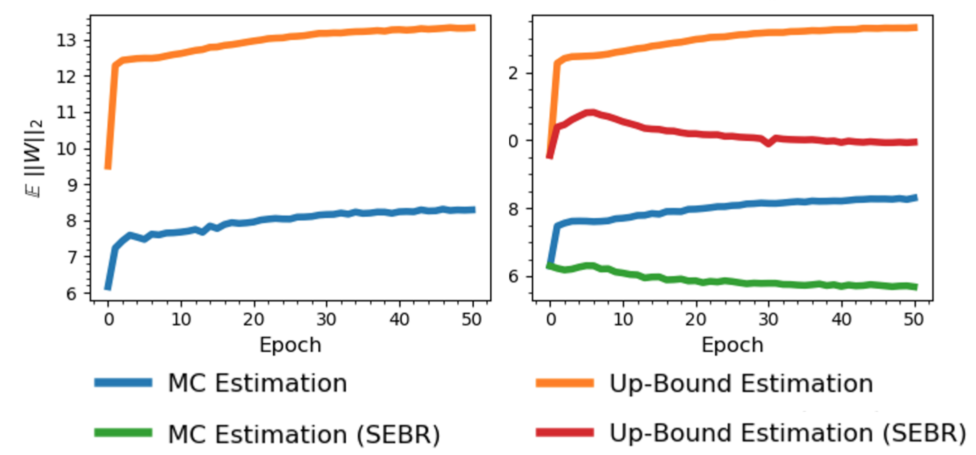
Verification

- The difference between the upper-bound and the un-biased estimated value keeps stable and their variation trends are synchronous.
- SEBR not only reduces the upper bound itself but also reduces the un-biased evaluated value.
- The training with SEBR significantly reduces the amount of time cost for training compared with the direct optimization method.

Time cost comparison

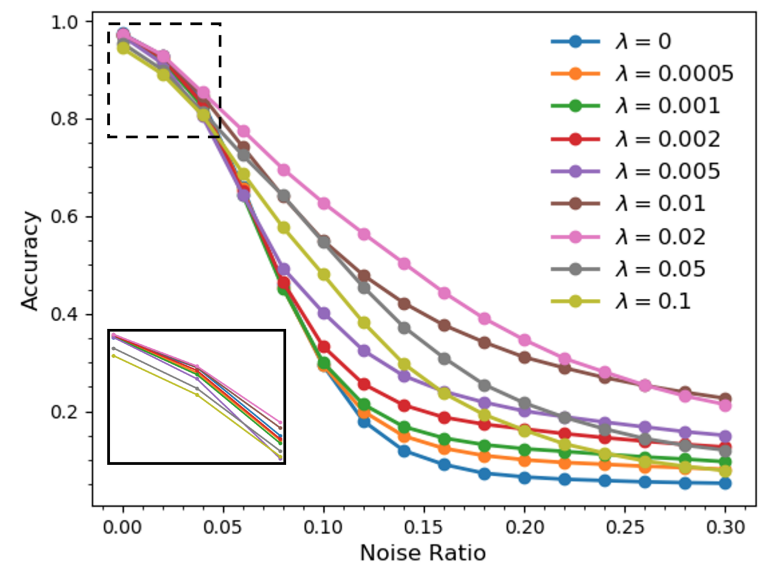
Method	Avg. time per epoch
Reg. on $\mathbb{E}\|W\|_2$	1654.8 (s)
SEBR	410.5 (s)

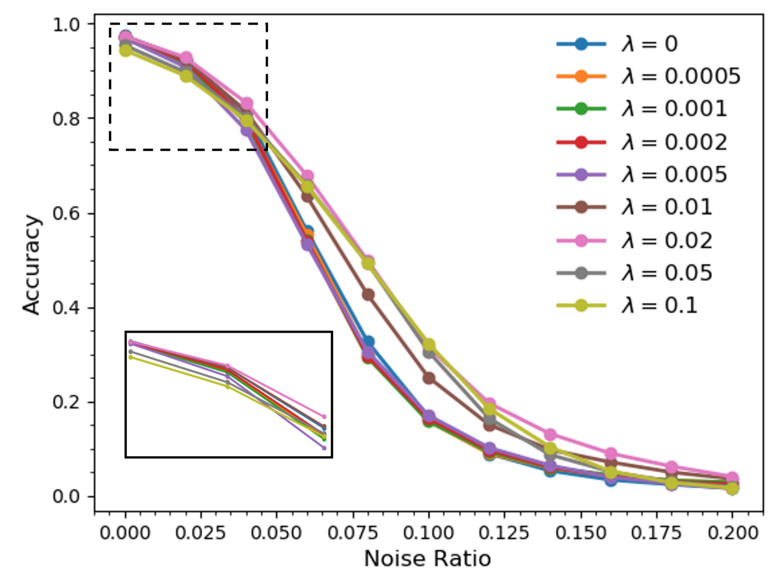
Variation trends in training



Influence of the parameter λ

Using a suitable λ is important.





Improvements on Adversarial Robustness

Model	Dataset	Attack	Norm	Acc. w/o. SEBR	Acc. w. SEBR	Δ
Bayesian MLP	MNIST	/	0	97.05 ± 0.38	96.83 ± 0.48	−0.22
		FGSM	0.04	83.83 ± 0.51	85.74 ± 0.64	**+ 1.91**
			0.16	8.97 ± 0.28	43.69 ± 5.92	**+ 34.72**
			0.3	5.06 ± 0.21	24.54 ± 8.65	**+ 19.48**
		PGD	0.04	81.99 ± 1.05	83.67 ± 0.67	**+ 1.68**
			0.16	4.20 ± 0.84	9.54 ± 2.82	**+ 5.34**
			0.22	1.55 ± 0.35	3.18 ± 1.52	**+ 1.63**
Bayesian CNN	MNIST	/	0	98.88 ± 0.27	98.70 ± 0.04	−0.18
		FGSM	0.04	85.64 ± 2.52	86.14 ± 2.76	**+ 0.50**
			0.08	55.98 ± 4.40	60.27 ± 8.65	**+ 4.29**
			0.14	18.16 ± 0.57	22.55 ± 11.23	**+ 4.39**
		PGD	0.04	82.91 ± 2.63	85.10 ± 2.96	**+ 2.19**
			0.08	36.53 ± 5.85	49.20 ± 10.75	**+ 12.67**
			0.14	9.88 ± 2.02	12.33 ± 5.31	**+ 2.45**
Bayesian MLP	Fashion MNIST	/	0	84.38 ± 0.37	78.75 ± 0.83	−5.63
		FGSM	0.04	60.96 ± 0.24	62.06 ± 1.15	**+ 1.10**
			0.1	24.29 ± 1.16	31.65 ± 1.25	**+ 7.36**
			0.2	1.99 ± 0.57	4.59 ± 0.75	**+ 2.60**
		PGD	0.04	59.86 ± 0.34	61.80 ± 1.13	**+ 1.94**
			0.1	19.18 ± 1.01	29.67 ± 1.22	**+ 10.49**
			0.2	0.44 ± 0.14	2.71 ± 0.60	**+ 2.27**
Bayesian CNN	Fashion MNIST	/	0	87.45 ± 0.57	84.83 ± 0.33	−2.62
		FGSM	0.04	40.82 ± 1.86	46.03 ± 4.22	**+ 5.21**
			0.08	15.89 ± 0.97	18.96 ± 5.00	**+ 3.07**
			0.1	10.24 ± 0.31	11.97 ± 3.95	**+ 1.73**
		PGD	0.04	32.81 ± 1.70	39.92 ± 3.25	**+ 7.11**
			0.06	15.03 ± 2.03	20.87 ± 4.00	**+ 5.84**
			0.08	5.62 ± 0.73	9.27 ± 1.62	**+ 3.65**

The models trained with SEBR are more robust on defending all kinds of noises.

Model	Dataset	Attack	Norm	Acc. w/o. SEBR	Acc. w. SEBR	Δ
Bayesian MLP + Adv. Training	MNIST	/	0	97.22 ± 0.27	96.94 ± 0.39	−0.28
		FGSM	0.04	92.87 ± 0.27	92.08 ± 0.12	−0.79
			0.16	54.56 ± 1.71	57.63 ± 1.08	**+ 3.07**
			0.3	9.94 ± 0.13	33.09 ± 8.23	**+ 23.15**
		PGD	0.04	92.57 ± 0.40	91.87 ± 0.26	−0.70
			0.16	40.05 ± 5.32	40.66 ± 4.18	**+ 0.61**
			0.22	11.15 ± 5.70	16.47 ± 3.57	**+ 5.32**
Bayesian CNN + Adv.Training	MNIST	/	0	98.89 ± 0.19	98.77 ± 0.08	−0.12
		FGSM	0.04	96.23 ± 0.40	95.96 ± 0.23	−0.27
			0.2	62.34 ± 4.70	63.20 ± 4.10	**+ 0.86**
			0.44	11.36 ± 2.17	14.18 ± 0.82	**+ 2.82**
		PGD	0.04	95.98 ± 0.40	95.79 ± 0.24	−0.19
			0.2	26.17 ± 4.39	30.06 ± 3.92	**+ 3.89**
			0.44	6.85 ± 1.67	8.78 ± 1.07	**+ 1.93**

GitHub Repository: AISIGSJTU/SEBR