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

Background

- Bayesian neural networks have been widely used in many applications.
- Adversarial sensitivity is a common problem of deep neural network models, including Bayesian neural networks.
- Even though Bayesian neural networks have been found more robust to adversarial attacks, their ability to deal with adversarial noises in practice is still limited.

A Theoretical Point of Penetration

- It is proved that a Bayesian neural network model will become more robust if $\mathbb{E}\|W\|_2$ of each layer get restricted.

Theorem

Consider function $f_W(\mathbf{x}) = f(W\mathbf{x} + \mathbf{b})$, *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*

$$\mathbb{E}_W \|f_W(\mathbf{x}+\boldsymbol{\xi}) - f_W(\mathbf{x})\| \leq Lip(f) \cdot \mathbb{E}\|W\|_2 \cdot \|\boldsymbol{\xi}\|, \tag{1}$$

where $\|W\|_2$ *represents the spectral norm of matrix* W, *and it is defined as*

$$\|W\|_2 = \max_{\boldsymbol{\xi}\in\mathbb{R}^n, \boldsymbol{\xi}\neq 0} \frac{\|W\boldsymbol{\xi}\|}{\|\boldsymbol{\xi}\|}. \tag{2}$$

How to restrict $\mathbb{E}\|W\|_2$ in practice? A naive method:

$$\underset{W}{\text{minimize}} \quad \mathcal{L} + \frac{\lambda}{2}\sum_{l=1}^{L}(\mathbb{E}\|W^l\|_2)^2, \tag{3}$$

The expectation is estimated by **Monte Carlo sampling** (K times). The spectral norm is calculated by **Power Iteration** (N iterations) method.

The time complexity is $O(KN)$.

Faster Estimation

A substitution: Estimation of its **upper bound**.

Theorem

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

$$\mathbb{E}\|W\|_2 \leq \|M\|_2 + c\left(\max_i \|A_{i,:}\| + \max_j \|A_{:,j}\| + \mathbb{E}\max_{i,j} |G_{ij}|\right), \tag{4}$$

where c is a constant independent of W.

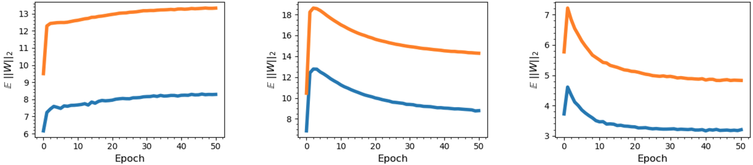
The estimation of the upper bound is faster: $O(K + N)$

Denote $\mathcal{L}_{\mathcal{S}}$ as half of the square of the upper bound of $\mathbb{E}\|W\|_2$ in each layer. Add it into the loss function:

$$\underset{\mathrm{W}}{\text{minimize}}\ \mathcal{L} + \lambda \cdot \mathcal{L}_{\mathcal{S}}. \tag{5}$$

The method is named as Spectral Expectation Bound Regularization (SEBR).

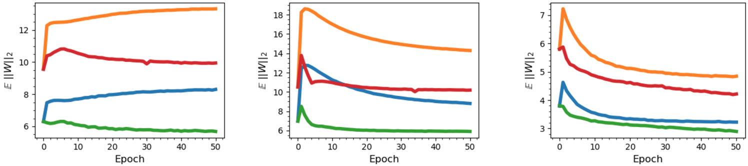
- The upper bounds reflect the variation trends of real values accurately.



(a) Layer 1 (b) Layer 2 (c) Layer 3

MC Estimation Up-Bound Estimation

- The real values get decreased bacause of the usage of SEBR.



(a) Layer 1 (b) Layer 2 (c) Layer 3

MC Estimation MC Estimation (SEBR) Up-Bound Estimation Up-Bound Estimation (SEBR)

Verifications

- The time costs get reduced compared with the naive method.

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

Table 1. Time cost comparison between SEBR and the direct regularization on $\mathbb{E}\|W\|_2$.

Influence on Uncertainty

The epistemic uncertainty of the model output gets reduced by SEBR:

Theorem

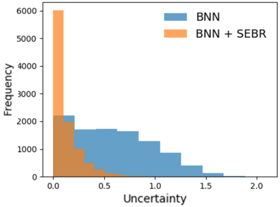
Consider a Bayesian neural network with only a linear layer $f_{\mathrm{W}}(\mathbf{x}) = W\mathbf{x} + \mathbf{b}$, *where* $\mathbf{x} \in \mathbb{R}^n$, $W \in \mathbb{R}^{m \times n}$. *Denote the epistemic uncertainty of the output after one step gradient descent without SEBR as* H_e, *and the epistemic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

$$H_e' \leq H_e. \tag{6}$$

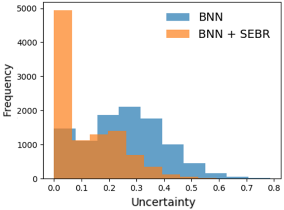
It verifies the robustness of the model from another point of view.

Verification on Uncertainty Decrease

Experiments on the verification of the decrease of the output uncertainty.



(a) Aleatoric Uncertainty



(b) Epistemic Uncertainty

Figure 4. Uncertainties measured by Bayesian neural networks on data with adversarial noises. Models trained with SEBR have lower uncertainty on the predictions. *Best viewed in color.*

Improvement on Adversarial Robustness

Experiments on multiple structures (i.e., MLP and CNN), multiple datasets (MNIST and Fashion-MNIST), and multiple attacks (i.e., FGSM and PGD) verify the efficiency of the proposed method.

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

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

Improvement on Adversarial Robustness

Experiments on more complex structure (i.e., VGG), more complex datasets (CIFAR-10/100) further verify the efficiency of the proposed method.

Dataset	Attack	noise ℓ_∞	w/o. SEBR	w. SEBR
CIFAR10	/	0	91.65	**92.09**
	FGSM	0.005	58.65	**65.74**
		0.01	42.70	**54.78**
		0.02	32.73	**43.76**
	PGD	0.005	46.33	**50.40**
		0.01	9.73	**16.11**
		0.02	2.31	**2.95**
CIFAR100	/	0	**66.94**	66.56
	FGSM	0.002	45.96	**47.67**
		0.01	17.08	**21.18**
		0.02	12.52	**15.97**
	PGD	0.002	44.72	**46.85**
		0.01	2.91	**5.04**
		0.02	0.95	**1.95**

Table S1. Experiments on Bayesian CNN with VGG architecture.

Thanks!