Attacking Bayes: Are Bayesian Neural Networks Inherently Robust?

Yunzhen Feng * Tim G. J. Rudner * Nikolaos Tsilivis Julia Kempe

New York University * Corresponding authors: yf2231@nyu.edu, tim.rudner@nyu.edu

Summary

- We investigate whether BNNs are robust to adversarial attacks and able to detect adversarial examples.
- We identify various conceptual and experimental errors in previous works that claim inherent adversarial robustness.
- We show that unsophisticated attacks like PGD-variants could break BNNs in three tasks: (1) label prediction under the posterior predictive mean, (2) adversarial example detection with Bayesian predictive uncertainty, and (3) semantic shift detection.

Background

Bayesian Inference Methods

We evaluate four Bayesian inference methods:

- Hamiltonian Monte Carlo (HMC, the gold standard)
- Monte Carlo Dropout (MCD)
- Parameter-Space Variational Inference (mean-field VI)
- Function-Space Variational Inference (FSVI)

Adversarial Robustness

- Adversarial Objective for generating adversarial examples,

$$\boldsymbol{\eta} = \arg\max_{\|\boldsymbol{\eta}\|_\infty \leq \epsilon} \mathcal{L}(f(\mathbf{x}+\boldsymbol{\eta}), y), \tag{1}$$

- Attacking stochastic neural network with expected gradient ascent, FGSM with one step and PGD with multiple steps,

$$\tilde{\mathbf{x}} = \mathbf{x} + \epsilon \cdot \operatorname{sign}\left(\nabla_{\mathbf{x}} \mathcal{L}(\mathbb{E} f(\mathbf{x}), y)\right). \tag{2}$$

  In the standard FGSM and PGD attack, $\mathcal{L}$ is the standard cross-entropy loss. We also attack both uncertainty and cross-entropy with PGD+ attack.
- We follow standard adversarial perturbation ϵ used in the robustness community. MNIST(0.3), FashionMNIST(0.1), and CIFAR-10(8/255).

Evaluating Previous Claims

[1, 2, 3, 4] present empirical evidence supporting the superior robustness of BNNs in prediction and AE detection. However, upon careful inspection, we found various implementation errors, such as double-softmax application (oversmoothing) and vanishing gradients caused by numerical instabilities. After correcting these issues, the previously claimed advantages in BNN robustness disappear.

Empirical Evaluation

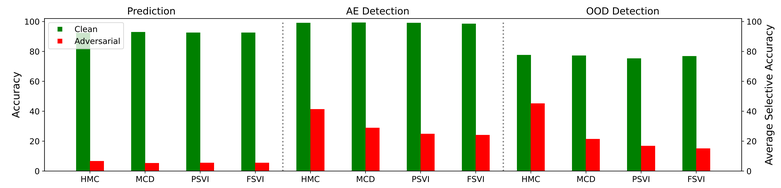

Fig 1. Summary statistics for Fashion-MNIST with a four-layer CNN. For prediction, the y-axis is the accuracy and for AE detection and OOD detection, the y-axis is the averaged selective accuracy. We could successfully attack all three tasks with adversarial predictive accuracy close to zero.

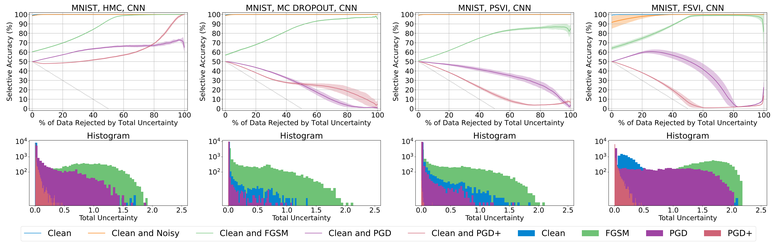

Fig 2. AE detection statistics for all four methods on MNIST with a four-layer CNN. We mix clean samples and adversarial samples together and report the selective accuracy. A model that randomly rejects samples would yield a flat line around 50%. Interestingly, most of the PGD curves lie below this 50% line, suggesting that the adversarial examples deceive the model to reject clean samples, without directly targeting the model's uncertainty. Furthermore, the PGD+ method could push the curves closer to the lower bound (the dashed line) when attacking the uncertainty estimation.

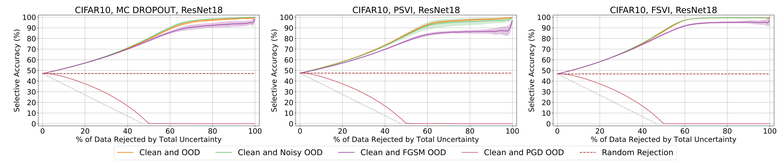

Fig 1. Out-of-distribution detection for CIFAR-10 (Fashion-MNIST) with a ResNet-18. We perform PGD attack on the total uncertainty only for the out-of-distribution (OOD) samples. We are the first to show that OOD detection with BNNs is equally vulnerable: minor perturbations applied to OOD samples cause the model to primarily classify in-distribution (ID) samples as out-of-distribution.

Discussion

- Even BNNs trained with HMC, the gold standard for Bayesian inference in stochastic neural networks, do not withstand adversarial attacks and exhibit a significant deterioration in robust accuracy, average selective accuracy, and semantic shift detection.
- Our investigation of HMC was limited to small CNNs. Investigating the robustness of larger BNNs trained with HMC remains an open question.
- We hope to draw the attention of the Bayesian learning community toward devising Bayesian defenses against adversarial attacks.

References

[1] Luca Bortolussi, Ginevra Carbone, Luca Laurenti, Andrea Patane, Guido Sanguinetti, and Matthew Wicker. On the robustness of Bayesian neural networks to adversarial attacks. *arXiv preprint arXiv:2207.06154*, 2022.

[2] Ginevra Carbone, Matthew Wicker, Luca Laurenti, Andrea Patane, Luca Bortolussi, and Guido Sanguinetti. Robustness of Bayesian neural networks to gradient-based attacks. In *Advances in Neural Information Processing Systems*, volume 33, pages 15602–15613, 2020.

[3] Lewis Smith and Yarin Gal. Understanding measures of uncertainty for adversarial example detection. In Amir Globerson and Ricardo Silva, editors, *Proceedings of the Thirty-Fourth Conference on Uncertainty in Artificial Intelligence, UAI 2018, Monterey, California, USA, August 6-10, 2018*, pages 560–569. AUAI Press, 2018.

[4] Jiaru Zhang, Yang Hua, Zhengui Xue, Tao Song, Chengyu Zheng, Ruhui Ma, and Haibing Guan. Robust Bayesian neural networks by spectral expectation bound regularization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3815–3824, June 2021.