PRBench:
A Primer on the Probabilistic Robustness Benchmark

Kenneth

Abstract

PRBench is the first comprehensive benchmark dedicated to *probabilistic robustness* (PR) of deep learning models. While adversarial robustness (AR) examines worst-case perturbations, PRBench evaluates the *likelihood* that a model withstands random perturbations within a specified budget. This tutorial-style article introduces the core concepts, metrics, and design of PRBench in an accessible yet academically rigorous manner.

1 Background: From Adversarial to Probabilistic Robustness

Modern classifiers achieve near-human performance on many tasks but can be fooled by imperceptible, worst-case perturbations called *adversarial examples* [**?**, **?**]. Adversarial robustness (AR) focuses on the *maximum* loss under any perturbation $\|\delta\|_p \leq \epsilon$:

$$\min_\theta \mathbb{E}_{(x,y)\sim\mathcal{D}}\Big[\max_{\|\delta\|_p\leq\epsilon} \ell\big(f_\theta(x+\delta), y\big)\Big].$$

In contrast, *probabilistic robustness* measures the *probability* that a random perturbation causes a misclassification:

$$\mathrm{PR}_{p,\epsilon}(f, x) \;=\; \Pr_{\delta\sim\mathcal{P}_p(\epsilon)}\big[f(x+\delta) = y\big],$$

where $\mathcal{P}_p(\epsilon)$ is a distribution (e.g. uniform or Gaussian) over the ℓ_p-ball of radius ϵ.

2 Why PRBench?

- **Practicality.** Real-world noise often follows stochastic patterns rather than worst-case. PR captures this.
- **Complementarity.** PR complements AR by quantifying *average-case* rather than *worst-case* behavior.
- **Method Development.** Few methods are tailored to PR; PRBench drives progress by providing a standardized evaluation.

3 Key Metrics in PRBench

PRBench evaluates each model under multiple metrics, aggregated over datasets and architectures:

Clean Accuracy (Acc.) Percentage of correctly classified clean inputs.

Probabilistic Robustness PR(γ) For perturbation radius γ,

$$\mathrm{PR}(\gamma) = \Pr_{\|\delta\|\leq\gamma}\big[f(x+\delta) = y\big] \times 100\%.$$

ProbAcc(ρ, γ) Probability that confidence remains above threshold ρ under perturbation γ.

Generalisation Error (GE_PR(γ)**)** Difference between training- and test-time PR at radius γ.

4 Benchmark Design

4.1 Datasets and Architectures

PRBench covers common vision datasets (e.g. CIFAR-10, CIFAR-100, ImageNet subsets) and diverse network families (e.g. ResNet, DenseNet).

4.2 Perturbation Distributions

Three families of random perturbations are evaluated:

- Uniform noise on the ℓ_∞-ball.
- Gaussian noise truncated to ℓ_2-ball.
- Laplace noise within ℓ_1-ball.

4.3 Methods Compared

PRBench compares:

- *Standard training* (ERM).
- *Adversarial training* (PGD-based AT).
- *PR-targeted training* (e.g. corruption training with uniform, Gaussian, Laplace noise).
- *Hybrid methods* combining AR and PR objectives.

5 Using PRBench

1. **Select dataset & architecture.**
2. **Choose perturbation radius γ.**
3. **Compute metrics.** DataTables and plots summarize Acc., PR(γ), ProbAcc, GE_PR(γ).
4. **Analyse trade-offs.** Compare how methods balance clean accuracy, robustness, and generalisation.

6 Interpreting Results

- *High PR but low Acc.* Method may overfit to random noise.
- *Low GE_PR.* Indicates stable generalisation of PR across train/test split.
- *Method ranking.* A unified score aggregates multiple metrics to propose an overall ranking.

7 Conclusion

PRBench fills a critical gap by standardizing evaluation of probabilistic robustness. It encourages development of methods that not only guard against worst-case but also maintain high reliability under realistic, stochastic perturbations.

References

- Szegedy, C. et al. (2014). Intriguing properties of neural networks. *ICLR*.
- Madry, A. et al. (2018). Towards deep learning models resistant to adversarial attacks. *ICLR*.
- Cohen, J. et al. (2019). Certified robustness to adversarial examples via randomized smoothing. *ICML*.