Understanding RBench: A Benchmark for Robustness Evaluation

Kenneth*

August 8, 2025

1 Introduction

In the rapidly evolving field of machine learning, *robustness* has become a central theme. Models deployed in the real world often encounter inputs that differ from their training distribution due to noise, adversarial attacks, or environmental changes. Measuring and comparing robustness across methods, datasets, and architectures requires well-designed benchmarks.

RBench is a conceptual benchmark framework designed to evaluate a model's robustness in a systematic and reproducible way. Although not yet a standardized public benchmark in the academic literature, the RBench idea illustrates how a single platform can unify diverse robustness metrics into a consistent evaluation protocol.

2 Motivation

Traditional benchmarks like ImageNet [1] and CIFAR [2] focus on clean accuracy under standard conditions. However, robust machine learning research demands more:

- **Consistency:** Use a fixed evaluation protocol across different robustness methods.
- **Comparability:** Ensure that results for different models and training regimes can be fairly compared.
- **Comprehensiveness:** Capture multiple aspects of robustness, not just one metric.

RBench aims to address these needs by providing a standardized robustness evaluation suite.

3 Core Components of RBench

An RBench-style framework typically includes:

*This article is intended as an educational blog post. If RBench is an in-development or internal project, please treat all descriptions as illustrative until official documentation is released.

3.1 Multiple Threat Models

Robustness is context-dependent. RBench can incorporate:

1. **Adversarial Robustness (AR):** Performance under worst-case perturbations generated by attacks such as PGD [3], CW [4], or AutoAttack [5].
2. **Probabilistic Robustness (PR):** Accuracy under stochastic perturbations (e.g., Gaussian noise, blur, or compression artifacts) [6].
3. **Corruption Robustness:** Performance on datasets with synthetic corruptions, such as CIFAR-C or ImageNet-C [7].

3.2 Comprehensive Metrics

RBench may track:

- Clean accuracy (**Acc.**)
- Adversarial accuracy under various ℓ_p norms
- PR metrics like $PR(\gamma)$ and $\text{ProbAcc}(\rho, \gamma)$
- Generalization error in robustness (GEAR, GEPR)
- Training cost metrics (e.g., seconds per epoch)

3.3 Unified Reporting

Results are aggregated into a standardized table format. Composite scores may be computed by weighting metrics according to their importance for a given application.

4 Why RBench Matters

The benefits of a benchmark like RBench include:

1. **Fair Comparison:** All models face the same evaluation conditions.
2. **Transparency:** Publicly released evaluation code and datasets promote reproducibility.
3. **Guidance for Practitioners:** Helps select models not only for accuracy but also for robustness in real-world settings.

5 Example Workflow

A typical RBench evaluation might follow:

1. Train a model using a chosen method (ERM, adversarial training, PR-targeted method).
2. Submit the trained model to RBench's evaluation server or script.
3. Receive a detailed report including all robustness metrics and a composite score.

6 Limitations and Considerations

- **Metric Selection Bias:** Choice of metrics can influence rankings.
- **Attack Strength:** If adversarial attacks are too weak, robustness scores may be misleading.
- **Relevance to Deployment:** Benchmarks should simulate perturbations relevant to the intended real-world application.

7 Conclusion

While RBench as described here is a conceptual benchmark, its principles align with the growing need for standardized, multi-metric robustness evaluation in machine learning. Whether implemented as an internal tool or released as a public resource, a well-designed benchmark can greatly advance our understanding of robust ML.

Disclaimer

If RBench is your internal or in-progress project, the descriptions here are illustrative. Any claims about performance, novelty, or "first-of-its-kind" status should be backed by a public release or clearly labeled as preliminary.

References

[1] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255. IEEE, 2009.

[2] Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. *Technical Report*, 2009.

[3] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations (ICLR)*, 2018.

[4] Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *IEEE Symposium on Security and Privacy (SP)*, 2017.

[5] Francesco Croce and Matthias Hein. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In *International Conference on Machine Learning (ICML)*, 2020.

[6] Jeremy Cohen, Elan Rosenfeld, and J. Zico Kolter. Certified adversarial robustness via randomized smoothing. In *International Conference on Machine Learning (ICML)*, 2019.

[7] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *International Conference on Learning Representations (ICLR)*, 2019.