Towards Precise Observations of Neural Model Robustness in Classification

Wenchuan Mu
Singapore University of Technology and Design
Singapore
wenchuan_mu@sutd.edu.sg

Kwan Hui Lim
Singapore University of Technology and Design
Singapore
kwanhui_lim@sutd.edu.sg

ABSTRACT

In deep learning applications, robustness measures the ability of neural models that handle slight changes in input data, which could lead to potential safety hazards, especially in safety-critical applications. Pre-deployment assessment of model robustness is essential, but existing methods often suffer from either high costs or imprecise results. To enhance safety in real-world scenarios, metrics that effectively capture the model's robustness are needed. To address this issue, we compare the rigour and usage conditions of various assessment methods based on different definitions. Then, we propose a straightforward and practical metric utilizing hypothesis testing for probabilistic robustness and have integrated it into the TorchAttacks library. Through a comparative analysis of diverse robustness assessment methods, our approach contributes to a deeper understanding of model robustness in safety-critical applications.

ACM Reference Format:
Wenchuan Mu and Kwan Hui Lim. 2024. Towards Precise Observations of Neural Model Robustness in Classification . In *2024 IEEE/ACM 46th International Conference on Software Engineering: Companion Proceedings (ICSE-Companion '24), April 14–20, 2024, Lisbon, Portugal.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3639478.3643519

1 INTRODUCTION

Deep learning has attained significant accomplishments across a broad range of applications, including in systems critical to security such as self-driving cars, medical diagnosis, and face-recognition-based authentication systems. The reliability and robustness of deep neural networks (DNNs) are important in security-critical systems and for ensuring fair outcomes [2]. In such scenarios, even slight changes in input data can lead to catastrophic consequences, necessitating the pre-deployment assessment of model robustness.

The evaluation of model robustness is a well-established concept, but it comes with significant challenges. Existing robustness evaluation methods like adversarial testing and verification have their limitations. Adversarial testing may not accurately represent real-world scenarios, while verification often faces the issue of incomplete problem formulation [4]. This means that verification methods might not fully capture the diversity of perturbations present in real-world scenarios. Furthermore, these methods may also encounter the problem of high cost, making them impractical for large-scale and resource-intensive applications [10]. Hence, there is a need for broader, practical evaluation methods for robustness assessment.

In order to address these gaps and bolster the safety of deep learning applications, our research focuses on the probabilistic robustness assessment. While some existing probabilistic robustness evaluations resort to approximated methods, these approximations may lead to the omission of critical adversarial instances, consequently overestimating the true robustness of the model.

In our work, we integrate the exact binomial test into the robustness evaluation of deep neural networks (DNNs), implemented within the TorchAttacks library (available at https://github.com/cestwc/precise-robustness). The exact binomial test is a statistical method that precisely measures how small changes in inputs affect the output of DNNs. This technique provides a clear and accurate way to identify vulnerabilities in neural models. Our method is notable for its efficiency, requiring less computational resources compared to traditional methods. It is versatile and can be applied to various DNN architectures, making it a practical solution for assessing robustness in safety-critical applications.

2 PROBABILISTIC ROBUSTNESS FROM BINOMIAL TESTING

There exist multiple interpretations of classifier robustness and we opt for the definition that emphasises the probabilistic nature of adversarial examples. Formally, $P_{\mathbf{x}}(P(h(\boldsymbol{x'}) \neq h(\boldsymbol{x}) \mid \mathbf{x} = \boldsymbol{x}, d(\boldsymbol{x}, \boldsymbol{x'}) \leq \epsilon) \leq \kappa)$, where $\mathbf{x}$ is the random variable input in the distribution, $\boldsymbol{x}, \boldsymbol{x'}$ are specific inputs, d denotes distance, ϵ denotes an imperceptible perturbation. To calculate the probability of any sampled input has less than κ (*e.g.*, 1%) adversarial examples in its neighbourhood, we first formulate this event as a Bernoulli trial $\mathbf{z}$, where the true probability is $P_{\mathbf{x}}(\mathbf{z} = 1 \mid h)$.

Binomial Test With Exact Solution. To get $P_{\mathbf{x}}(\mathbf{z} = 1 \mid h)$, we may first address $P(\mathbf{z} = 1 \mid h, \mathbf{x} = \boldsymbol{x})$ at specific $\boldsymbol{x}$. Given $\boldsymbol{x}$, we want to determine if the probability that h makes an incorrect prediction around $\boldsymbol{x}$ is greater than or equal to κ. This forms the null hypothesis in an exact binomial test. In a right-tail exact binomial test,

$$P(\mathbf{w} = 0 \mid h, \mathbf{x} = \boldsymbol{x}) = \sum_{i=k}^{n} \binom{n}{i} p_{\boldsymbol{x}}^{i} (1 - p_{\boldsymbol{x}})^{n-i} \tag{1}$$

where n denotes sample size, k denotes the number of successes, $p_{\boldsymbol{x}}$ is the true probability of success, and $\mathbf{w}$ is the observed event that the total number of successes is less than k. For the given $\boldsymbol{x}$, we increase the sample size until it rejects either side of the tail, *i.e.*, we would have high confidence $(1 - \alpha)$ to know that

$$P(\mathbf{w} = 1 \mid p_{\boldsymbol{x}} > \kappa) < \alpha \tag{2}$$

ICSE-Companion '24, April 14–20, 2024, Lisbon, Portugal

ACM ISBN 979-8-4007-0502-1/24/04.
https://doi.org/10.1145/3639478.3643519

Table 1: Classification results on CIFAR-10. Our observations of robustness and popular attack failure rates are listed side by side. Our observation gives the minimum probability that the adversarial examples of an arbitrary input account for less than 1 in 10,000.

Training	Accuracy	Attack Failure Rate	Our Observation
ERM [7]	**94.38**	1.25	84.20
ERM+DA[6]	94.21	1.08	84.15
FGSM [1]	84.96	43.50	83.50
PGD [3]	84.38	47.07	82.90
TRADES [9]	80.42	**48.54**	79.12
MART [8]	81.54	48.90	80.21
PRL [5]	93.82	0.71	**90.63**

and similarly on the other tail we have

$$P(\mathbf{w} = 0 \mid p_{\boldsymbol{x}} < \kappa) < \alpha. \tag{3}$$

True Probability Rather Than Observed Events. Existing works also stop at rejecting the null hypothesis and calculate the frequency of right rejection. However, we claim that the frequency of right rejection, *i.e.*, the probability of observation of event $\mathbf{w}$, or $\mathbf{w} = 1$, is not the true probability we are looking for.

Instead, we shall always compute the probability that event $\mathbf{z}$ is true. To achieve that, we apply the law of total probability. Then, the probability of event $\mathbf{w}$ can be expressed as

$$P(\mathbf{w}) = P(\mathbf{w} \mid \mathbf{z})P(\mathbf{z}) + P(\mathbf{w} \mid \neg\mathbf{z})P(\neg\mathbf{z}) \tag{4}$$

If we further write $P(\neg\mathbf{z}) = 1 - P(\mathbf{z})$, $P(\mathbf{w} \mid \neg\mathbf{z}) = 1 - P(\neg\mathbf{w} \mid \neg\mathbf{z})$, we eventually get

$$P(\mathbf{z}) = \frac{P(\mathbf{w}) - P(\mathbf{w} \mid \neg\mathbf{z})}{1 - P(\neg\mathbf{w} \mid \mathbf{z}) + P(\mathbf{w} \mid \neg\mathbf{z})} \tag{5}$$

Now that we know that $0 < P(\mathbf{w} \mid \neg\mathbf{z}), P(\neg\mathbf{w} \mid \mathbf{z}) < \alpha$, we can find the lower and upper limit of $P(\mathbf{z})$ as

$$(P(\mathbf{w}) - \alpha)/(1 + \alpha) < P(\mathbf{z}) < P(\mathbf{w})/(1 - \alpha) \tag{6}$$

which makes sense because $P(\mathbf{z})$ is still predominantly positively related to $P(\mathbf{w})$, while the smaller false positive rate (α) we have the closer $P(\mathbf{z})$ will be to $P(\mathbf{w})$.

In this way, we have made our observation targeted on the true probability, instead of the samples. Conservatively, a simple way is to get the $P(\mathbf{w})$ first, subtract the false positive rate from it, and divide by (1 + false positive rate).

To complete the process, we still need to determine $P(\mathbf{w})$. If in n' times we observed $\mathbf{w}$ k' times and not $\mathbf{w}$ $n' - k'$ times, then we can calculate the probability of $\mathbf{w}$ given these observations using the likelihood $P(\mathbf{w}) = k'/n'$. In summary, we get

$$\frac{k'/n' - \alpha}{1 + \alpha} < P_{\mathbf{x}}(\mathbf{z} = 1 \mid h) < \frac{k'}{n'(1 - \alpha)} \tag{7}$$

3 EXPERIMENTS

We conduct experiments on the CIFAR-10 dataset. We estimate 6 popular robustness improvement models from ERM [7]: ERM+DA [6], FGSM [1], PGDT [3], TRADES [9], MART [8], and PRL [5]. We also compare our robustness estimation (lower bound) with vanilla accuracy and attack-failure rate using projected gradient descent [3].

We use our method to evaluate existing adversarial mitigation methods on the CIFAR-10 dataset, with the result presented in Table 1. ERM leads in accuracy with 94.38%, while MART, known for its state-of-the-art adversarial training, records the highest Attack-Failure Rate at 48.59%. In contrast, the PRL method excels in robustness estimation, achieving a significant score of 90.63%. This performance underscores PRL's capability to improve probabilistic robustness (a critical attribute for models in safety-critical applications). It is important to note the distinct focus of each model: ERM prioritizes accuracy without significant emphasis on robustness, MART leverages adversarial attacks for robustness training, and PRL employs probabilistic methods for robustness training. The respective best performances in their focused areas validate the strengths of our approach, particularly highlighting the balance between robustness and accuracy estimation achieved by our method, which is vital in contexts where neither high accuracy nor attack resistance alone suffices.

Conclusion. This study introduces a new method to improve the assessment of probabilistic robustness in neural networks against adversarial examples, comprising three main elements: an exact binomial test for accurate binomial distribution calculations, a technique to reduce degrees of freedom based on the law of total probability, and standardized failure rate thresholds. Our exact solution addresses potential certification errors caused by approximations. The approach aligns better with the concept of probabilistic robustness by reducing unnecessary false positive rates, using IEC 61508 for certification thresholds to match safety integrity levels.

Acknowledgments. This research is supported by the Ministry of Education, Singapore, under its Academic Research Fund Tier 2 (Award No. MOE-T2EP20123-0015). Any opinions, findings and conclusions, or recommendations expressed in this material are those of the authors and do not reflect the views of the Ministry of Education, Singapore.

REFERENCES

[1] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *ICLR*, 2015.
[2] Sajal Halder, Kwan Hui Lim, Jeffrey Chan, and Xiuzhen Zhang. Capacity-aware fair poi recommendation combining transformer neural networks and resource allocation policy. *Applied Soft Computing*, 147:110720, 2023.
[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*, 2018.
[4] Mark Niklas Mueller, Franziska Eckert, Marc Fischer, and Martin Vechev. Certified training: Small boxes are all you need. In *ICLR*, 2023.
[5] Alexander Robey, Luiz Chamon, George J. Pappas, and Hamed Hassani. Probabilistically robust learning: Balancing average and worst-case performance. In *International Conference on Machine Learning*, pages 18667–18686, 2022.
[6] Connor Shorten and Taghi M. Khoshgoftaar. A survey on image data augmentation for deep learning. *Journal of Big Data*, 6(1):60, Jul 2019.
[7] Vladimir Vapnik. *The nature of statistical learning theory*. Springer, 1999.
[8] Yisen Wang, Difan Zou, Jinfeng Yi, James Bailey, Xingjun Ma, and Quanquan Gu. Improving adversarial robustness requires revisiting misclassified examples. In *International Conference on Learning Representations*, 2019.
[9] Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric Xing, Laurent El Ghaoui, and Michael Jordan. Theoretically principled trade-off between robustness and accuracy. In *ICLR*, volume 97, pages 7472–7482, 2019.
[10] Tianle Zhang, Wenjie Ruan, and Jonathan E. Fieldsend. Proa: A probabilistic robustness assessment against functional perturbations. In *Machine Learning and Knowledge Discovery in Databases*, pages 154–170, 2023.