gRoMA: a Tool for Measuring the Global Robustness of Deep Neural Networks

Natan Levy[1], Raz Yerushalmi[1,2][0000−0002−0513−3211], and Guy Katz[1]

[1] The Hebrew University of Jerusalem, Jerusalem, Israel
{natan.levy1,gkatz}@mail.huji.ac.il
[2] The Weizmann Institute of Science, Rehovot, Israel
raz.yerushalmi@weizmann.ac.il

Abstract. Deep neural networks (DNNs) are at the forefront of cutting-edge technology, and have been achieving remarkable performance in a variety of complex tasks. Nevertheless, their integration into safety-critical systems, such as in the aerospace or automotive domains, poses a significant challenge due to the threat of *adversarial inputs*: perturbations in inputs that might cause the DNN to make grievous mistakes. Multiple studies have demonstrated that even modern DNNs are susceptible to adversarial inputs, and this risk must thus be measured and mitigated to allow the deployment of DNNs in critical settings. Here, we present gRoMA (global Robustness Measurement and Assessment), an innovative and scalable tool that implements a probabilistic approach to measure the global categorial robustness of a DNN. Specifically, gRoMA measures the probability of encountering adversarial inputs for a specific output category. Our tool operates on pre-trained, black-box classification DNNs, and generates input samples belonging to an output category of interest. It measures the DNN's susceptibility to adversarial inputs around these inputs, and aggregates the results to infer the overall global categorial robustness of the DNN up to some small bounded statistical error. We evaluate our tool on the popular Densenet DNN model over the CIFAR10 dataset. Our results reveal significant gaps in the robustness of the different output categories. This experiment demonstrates the usefulness and scalability of our approach and its potential for allowing DNNs to be deployed within critical systems of interest.

Keywords: Global Robustness · Deep Neural Networks · Adversarial Examples · Categorial Robustness · Regulation · Safety Critical

1 Introduction

Deep neural networks (DNNs) have become fundamental components in many applications that perform classification [2, 25]. Empirically, DNNs often outperform traditional software, and even humans [35, 40]. Nevertheless, DNNs have a significant drawback: they are notoriously susceptible to small input perturbations, called *adversarial inputs* [15], which can cause them to produce erroneous outputs. These adversarial inputs are one of the causes likely to delay the adoption of DNNs in safety-critical domains, such as aerospace [14], autonomous vehicles [26], and medical devices [17].

In the aforementioned critical domains, systems must meet high dependability standards. While strict guidelines exist for certifying that hand-crafted software meets these standards (e.g., the DO-178 standard [13] in the aerospace industry), no such certification guidelines currently exist for systems incorporating DNNs. Several regulatory agencies have recognized the existence of this gap and the importance of addressing it. For example, in its recently published roadmap, the European Union Aviation Safety Administration (EASA) has emphasized the importance of DNN robustness as one of the 7 key requirements for trustworthy artificial intelligence [12]. However, certifying the robustness of DNNs remains an open problem.

The formal methods community has begun addressing this gap by devising methods for rigorously quantifying the *local robustness* of a DNN [21, 37, 42]. Local robustness refers to a DNN's ability to withstand adversarial inputs in the vicinity of a specific point within the input space. Although the rigorous verification approaches proposed to date have had some success in measuring these robustness scores, they typically struggle to scale as network sizes increase [21] — which limits their practical application. To circumvent that limitation, *approximate* methods have been proposed, which can evaluate DNN robustness more efficiently, but often at the cost of reduced precision [4, 11, 19, 31, 36].

Work to date, both on rigorous and on approximate methods, has focused almost exclusively on measuring *local* robustness, which quantifies the DNN's robustness around individual input points within a multi-dimensional, infinite input space. In the context of DNN certification, however, a broader perspective is required — one that measures the *global robustness* of the DNN, over the entire input space, rather than on specific points.

In this paper, we propose a novel approach for approximating the global robustness of a DNN. Our method is computationally efficient, scalable, and can handle various types of adversarial attacks and black-box DNNs. Unlike existing approximate approaches, our approach provides statistical guarantees about the precision of the computed robustness score.

More concretely, our approach (implemented in the gRoMA tool) implements a probabilistic verification approach for performing global robustness measurement and assessment on DNNs. gRoMA achieves this by measuring the *probabilistic global categorial robustness* (*PGCR*) of a given DNN. In this study, we take a conservative approach and consider the DNN as a black-box: gRoMA makes no assumptions, e.g., about the Lipschitz continuity of the DNN, the kinds of activation functions, the hyperparameters it uses, or its internal topology. Instead, gRoMA uses and extends the recently proposed RoMA (*a Method for DNN Robustness Measurement and Assessment*) algorithm [29] for measuring local robustness. gRoMA repeatedly invokes this algorithm on a collection of samples, drawn to represent a specific output category of interest; and then aggregates the results to compute a global robustness score for this category, across the entire input space. As a result, gRoMA is highly scalable, typically taking only a few minutes to run, even for large networks. Further, the tool formally computes an error bound for the estimated PGCR scores using Hoeffding's inequality [18]

to mitigate the drawbacks of using a statistical method. Thus, gRoMA's results can be used in the certification process for components of safety-critical systems, following, e.g., the SAE Aerospace Recommended Practice [27].

For evaluation purposes, we focused on a Densenet DNN [20], trained on the CIFAR10 dataset [25]; and then measured the network's global robustness using one hundred arbitrary images for each CIFAR10 category. gRoMA successfully computed the global robustness scores for these categories, demonstrating, e.g., that the airplane category is significantly more robust than other categories.

To the best of our knowledge, our tool is presently the only scalable solution for accurately measuring the *global categorial robustness* of a DNN, i.e., the aggregated robustness of *all* points within the input space that belong to a category of interest — subject to the availability of a domain expert who can supply representative samples from each category. The availability of such tools could greatly assist regulatory authorities in assessing the suitability of DNNs for integration into safety-critical systems, and in comparing the performance of multiple candidate DNNs.

Outline. We begin with an overview of related work on measuring the local and global adversarial robustness of DNNs, in Section 2. In Section 3, we provide the necessary definitions for understanding our approach. We then introduce the gRoMA tool in Section 4. Next, in Section 5, we evaluate the performance of our tool using a popular dataset and DNN model. Finally, Section 6 concludes our work and discusses future research directions.

2 Related Work

Measuring the local adversarial robustness of DNNs has received significant attention in recent years. Two notable approaches for addressing it are:

- Formal-verification approaches [23, 32, 38], which utilize constraint solving and abstract interpretation techniques to determine a DNN's robustness. These approaches are fairly precise, but generally afford limited scalability, and are applicable only to white-box DNNs.
- Statistical approaches, which evaluate the probability of encountering adversarial inputs. These approaches often need to balance between scalability and accuracy, with prior work [4,11,19,31,36] typically leaning towards scalability.

Recently, the *RoMA* algorithm [29] has been introduced as highly scalability statistical method, but which can also provide rigorous guarantees on accuracy. RoMA is a simple-to-implement algorithm that evaluates local robustness by sampling around an input point of interest; measuring the confidence scores assigned by the DNN to *incorrect* labels on each of the sampled input points; and then using this information to compute the probability of encountering an input on which the confidence score for the incorrect category will be high enough to result in misclassification. In the final step, RoMA assesses robustness using

properties of the normal distribution function [29]. RoMA handles black-box DNNs, without any a priori assumptions; but it can only measure local, as opposed to global robustness.

Due to the limited usefulness of computing local robustness in modern DNNs, initial attempts have been made to compute the *global adversarial robustness* of networks. Prior work formulated and defined the concept of *global adversarial robustness* [21,31]; but in the same breath, noted that global robustness can be hard to check or compute compared to local robustness. More recently, there have been attempts to use formal verification to check global adversarial robustness [24,39,43]; but the reliance on formal verification makes it difficult for these approaches to scale, and requires a white-box DNN with specific activation functions.

Two other recently proposed approaches study an altered version of global robustness. The first work, by Ruan et al. [34], defines global robustness as the expected maximal safe radius around a test data set. It then proposes an approximate method for computing lower and upper bounds on DNN's robustness. The second work, by Zaitang et al. [41], redefines global robustness based on the probability density function, and uses generative models to assess it. These modified definitions of global robustness present an intriguing perspective. However, it is important to note that they differ from common definitions, and whether they will be widely adopted remains to be seen. Another noteworthy recent approach, proposed by Leino et al. [28], advocates for training DNNs that are certifiably robust by construction, assuming that the network is Lipschitz-continuous. This approach can guarantee the global robustness of a DNN without accurate measurements, but it requires the DNN to be white-box, whereas our approach is also compatible with black-box DNNs.

Our work here focuses on measuring and scoring the global robustness of pre-trained black-box DNNs and is the first, to the best of our knowledge, that is scalable and consistent with the commonly accepted definitions.

3 DNNs and Adversarial Robustness

Neural Networks. A DNN $N : \mathbb{R}^n \to \mathbb{R}^m$ is a function that maps input $\boldsymbol{x} \in \mathbb{R}^n$ to output $\boldsymbol{y} \in \mathbb{R}^m$. In classifier DNNs, which are our subject matter here, $\boldsymbol{y}$ is interpreted as a vector of confidence scores, one for each of m possible labels. We say that N classifies $\boldsymbol{x}$ as label l iff $\arg\max(\boldsymbol{y}) = l$, i.e., when y's l'th entry has the highest score. We use L to denote the set of all possible labels, $L = \{1, \ldots, m\}$.

Local Adversarial Robustness. The local adversarial robustness of N around input $\boldsymbol{x}$ is a measure of how sensitive N is to small perturbations around $\boldsymbol{x}$ [5]:

Definition 1. *A DNN N is ϵ-locally-robust at input point $\boldsymbol{x_0}$ iff*

$$\forall \boldsymbol{x}. \quad ||\boldsymbol{x} - \boldsymbol{x_0}||_\infty \leq \epsilon \Rightarrow \arg\max(N(\boldsymbol{x})) = \arg\max(N(\boldsymbol{x_0}))$$

Intuitively, Definition 1 states that the network assigns to $\boldsymbol{x}$ the same label that it assigns to $\boldsymbol{x_0}$, for input $\boldsymbol{x}$ that is within an ϵ-ball around $\boldsymbol{x_0}$. Larger values of ϵ imply a larger ball around $\boldsymbol{x_0}$, and consequently — higher robustness.

The main drawback in Definition 1 is that it considers a single input point in potentially vast input space. Thus, the ϵ-local-robustness of N at $\boldsymbol{x_0}$ does not imply that it is also robust around other points. Moreover, it assumes that DNN robustness is consistent across categories, although it has already been observed that some categories can be more robust than others [29]. To overcome these drawbacks, the notion of *global categorial robustness* has been proposed [22,34]:

Definition 2. *A DNN N is (ϵ, δ)-globally-robust in input region D iff*

$$\forall \boldsymbol{x_1}, \boldsymbol{x_2} \in D.$$

$$||\boldsymbol{x_1} - \boldsymbol{x_2}||_\infty \leq \epsilon \Rightarrow \forall l \in L.\ |N(\boldsymbol{x_1})[l] - N(\boldsymbol{x_2})[l]| < \delta$$

Intuitively, Definition 2 states that for every two inputs $\boldsymbol{x_1}$ and $\boldsymbol{x_2}$ that are at most ϵ apart, there are no spikes greater than δ in the confidence scores that the DNN assigns to each of the labels.

Definitions 1 and 2 are Boolean in nature: given ϵ and δ, the DNN is either robust or not robust. However, in real-world settings, safety-critical systems can still be determined to be sufficiently robust if the *likelihood* of encountering adversarial inputs is sufficiently low [27]. Moreover, it is sometimes more appropriate to measure robustness for specific output categories [29]. To address this, we propose to compute real-valued, *probabilistic global categorial robustness* scores:

Definition 3. *Let N be a DNN, let $l \in L$ be an output label, and let I be a finite set of labeled data representing the input space for N. The (ϵ, δ)-PGCR score for N with respect to l and I, denoted $pgcr_{\delta,\epsilon}(N, l, I)$, is defined as:*

$$pgcr_{\delta,\epsilon}(N, l, I) \triangleq P_{\boldsymbol{x_1} \in I, \boldsymbol{x_2} \in \mathbb{R}^n || \boldsymbol{x_1} - \boldsymbol{x_2} ||_\infty \leq \epsilon} [|N(\boldsymbol{x_1})[l] - N(\boldsymbol{x_2})[l]| < \delta]$$

Intuitively, the definition captures the probability that for an input $\boldsymbol{x_1}$ drawn from I, and for an additional input $\boldsymbol{x_2}$ that it is at most ϵ apart from $\boldsymbol{x_1}$, inputs $\boldsymbol{x_1}$ and $\boldsymbol{x_2}$ will be assigned confidence scores that differ by at most δ for the label l.

4 Introducing the gRoMA Tool

Algorithm. The high-level flow of gRoMA implements Definition 3 in a straightforward and efficient way: it first computes the local robustness for n representative points from each category, and then aggregates the global robustness using Algorithm 1.

The inputs to gRoMA are: (i) a network N; (ii) I, a finite set of labeled data that represents the input space, to draw samples from; (iii) a label l; (iv) n,

the number of representative samples of inputs classified as l to use; and (v) ϵ and δ, which determine the allowed perturbation sizes and differences in confidence scores, as per Definition 3. gRoMA's output consists of the computed $pgcr_{\delta,\epsilon}(N, l, I)$ score and an error term e, both specific to l. We emphasize the reliance of the $pgcr_{\delta,\epsilon}$ score on having representative input samples for each relevant category l. Under that assumption, in which the samples represent the underlying input distribution, our method guarantees that, with some high, predefined probability, the distance of the computed $pgcr_{\delta,\epsilon}$ value from its true value is at most e.

Algorithm 1 gRoMA$(N, I, l, n, \epsilon, \delta)$

```
1: X := drawSamples(I, l, n)
2: for i := 1 to n do
3:    if ( N(X[i]) = l ) then
4:       plr[i] := RoMA(X[i], ε, δ, N)
5:    end if
6: end for
7: pgcr := aggregate(plr)
8: e := computeError(pgcr, plr, X)
9: return (pgcr,e)
```

In line 1, gRoMA begins by creating a vector, $\boldsymbol{X}$, of perturbed inputs — by drawing from I, at random, n samples of inputs labeled as l. Next, for each correctly classified sample (line 3), gRoMA computes the sample's probabilistic local robustness (plr) score using RoMA [29] (line 4). Finally, gRoMA applies statistical aggregation (line 7) to compute the $pgcr$ score and the error bound (line 8); and these two values are then returned on line 9.

gRoMA is modular in the sense that any aggregation method (line 7) and error computation method (line 8) can be used. There are several suitable techniques in the statistics literature for both tasks, a thorough discussion of which is beyond our scope here. We focus here on a few straightforward mechanisms for these tasks, which we describe next.

For score aggregation, we propose to use the numerical average of the local robustness scores computed for the individual input samples. Additional approaches include computing a median score and more complex methods, e.g., methods based on normal distribution properties [9], maximum likelihood methods, Bayesian computations, and others. For computing the PGCR score's probabilistic error bound, we propose to use Hoeffding's Inequality [18], which provides an upper bound on the likelihood that a predicted value will deviate from its expected value by more than some specified amount.

5 Evaluation

Implementation. We implemented gRoMA as a Tensorflow framework [1]. Internally, it uses Google Colab [7,8] tools with 12.7GB system RAM memory, and T4 GPU. It accepts DNNs in Keras H5 format [10], as its input. The gRoMA tool is relatively simple, and can be extended and customized to support, e.g., multiple input distributions of interest, various methods for computing aggregated robustness scores and probabilistic error bounds, and also to accept additional DNN input formats. gRoMA is available online [30].

Setup and Configuration. We conducted an evaluation of gRoMA on a commonly used Densenet model [20] with 797,788 parameters, trained on the CIFAR10 dataset [25]. The model achieved a test accuracy of 93.7% after a standard 200-epoch training period. The code for creating and training the model, as well as the H5 model file, are available online [30]

For gRoMA to operate properly, it is required to obtain a representative sample of the relevant input space I. Creating such a representative sample typically requires some domain-expert knowledge [16,33]. However, random sampling can often serve as an approximation for such sampling [16,33]; a more thorough discussion of that topic goes beyond the scope of this paper. In our experiments here, we used a simple sampling mechanism in order to demonstrate the use of gRoMA. We measured the global categorial robustness of each output category by running the *RoMA* algorithm [29], to calculate the local robustness of one hundred images drawn independently and arbitrarily from the set I, which includes varying angles, lighting conditions, and resolutions. We set ϵ to 0.04 and δ to 0.07, as recommended in that work [29].

Due to our desire to check the approach's applicability to the aerospace industry, we paid special attention to the airplane category. In this category, we focused on Airbus A320-200 commercial airplane images, either airborne or on the ground. This type of airplane exists in the CIFAR10 training set as well, and hence we expected a high level of categorial global robustness for this category. The images, along with our code and dependencies, are available online [30].

Next, for each output category, we used *RoMA* to compute the *probabilistic local robustness (plr)* score for each input sample. We configured gRoMA to use the numerical average as the aggregation method; and for assessing the error of gRoMA, we applied Hoeffding's inequality [18]. Specifically, we aimed for a maximum expected error value of 5%, which is an acceptable error value when calculating a DNN's robustness [19]. We used Hoeffding's inequality to calculate the probability that the actual error is higher than this value. This was achieved by setting the upper and lower bounds of the *plr* values to be plus and minus five standard deviations of the *plr* values, corresponding to a $1 - 1 * 10^{-6}$ accuracy, all in a normal distribution context. These bounds were selected in order to provide a conservative estimate, encompassing a significant portion of the input space. We justify the normal distribution assumption using Anderson-Darling goodness-of-fit test [3] that focuses on the tails of the distribution [6], as detailed in [29].

Results. Running our evaluation took less than 21 minutes for each category, using a Google Colab [7] machine. The various global robustness scores for each category, as well as the calculated probabilistic error, appear in Fig. 1.

In the evaluation, the Airplane category obtained, as expected while focusing on a specific type of airplane, the highest categorial robustness score of 99.91% among all categories; while the Cat and Ship categories obtained the lowest score of 99.52% (the PGCR scores appear in blue in Fig. 1). The statistical error margin (tolerance) was set to 5% for this study. Based on these setting, the Ship category had the highest probability to exceed this bound, at less than 0.16%. On the other hand, the Airplane, Automotive, Bird, Dog, and Frog categories had the lowest error likelihood, all below 0.0005% (the likelihood scores per category scores appear in yellow in Fig. 1).

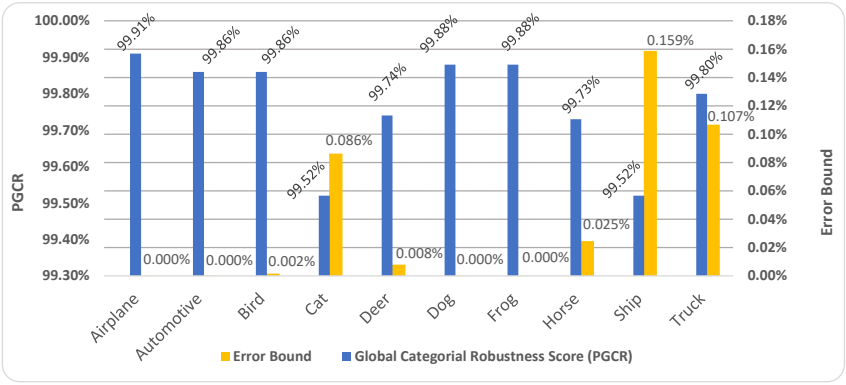

Fig. 1: PGCR scores, per category, for all CIFAR10 categories (blue); and the corresponding statistical errors (yellow).

The PGCR scores calculated are aligned with previous research, that already assessed the local robustness of all the images in the CIFAR10 test set, and which indicate that different categories may obtain different robustness scores [29].

6 Conclusion and Future Work

We introduced here the notion of PGCR and presented the gRoMA tool for probabilistically measuring the global categorial robustness of DNNs, e.g., calculating the $pgcr_{\epsilon,\delta}$ score — which is a step towards formalizing DNN safety and reliability for use in safety-critical applications. Furthermore, we calculate a bound on the statistical error inherent to using a statistical tool. The main contribution of this work is developing a scalable tool for probabilistically measuring categorial global DNN robustness.

Although extensive research has focused on DNN local adversarial robustness, we are not aware of any other scalable tool that can measure the global categorial robustness of a DNN. Therefore, we believe that our tool provides a valuable contribution to the research community.

In future work, we plan to test the accuracy of gRoMA using a range of input distributions and sampling methods, simulating various input spaces used in different applications. Additionally, we intend to extend our tool to other types of DNNs, such as regression networks, to broaden PGCR's applicability. These efforts will enhance our understanding of DNN robustness and facilitate safe and reliable deployment in real-world applications.

6.1 Acknowledgments.

We thank Dr. Or Zuk of the Hebrew University for his valuable contribution and support. This work was partially supported by the Israel Science Foundation (grant number 683/18). The work of Yerushalmi was partially supported by a research grant from the Estate of Harry Levine, the Estate of Avraham Rothstein, Brenda Gruss and Daniel Hirsch, the One8 Foundation, Rina Mayer, Maurice Levy, and the Estate of Bernice Bernath, grant 3698/21 from the ISF-NSFC, and a grant from the Minerva foundation.

References

1. Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G.S., Davis, A., Dean, J., Devin, M., et al.: Tensorflow: Large-Scale Machine Learning on Heterogeneous Distributed Systems. arXiv preprint arXiv:1603.04467 (2016)
2. Al-Saffar, A., Tao, H., Talab, M.: Review of Deep Convolution Neural Network in Image Classification. In: 2017 Int. Conf. on Radar, Antenna, Microwave, Electronics, and Telecommunications (ICRAMET). pp. 26–31 (2017)
3. Anderson, T.: Anderson-Darling Tests of Goodness-of-Fit. Int. Encyclopedia of Statistical Science **1**, 52–54 (2011)
4. Baluta, T., Chua, Z.L., Meel, K.S., Saxena, P.: Scalable Quantitative Verification for Deep Neural Networks. In: 2021 IEEE/ACM 43rd International Conference on Software Engineering (ICSE). pp. 312–323. IEEE (2021)
5. Bastani, O., Ioannou, Y., Lampropoulos, L., Vytiniotis, D., Nori, A., Criminisi, A.: Measuring Neural Net Robustness with Constraints. In: Proc. 30th Conf. on Neural Information Processing Systems (NIPS) (2016)
6. Berlinger, M., Kolling, S., Schneider, J.: A Generalized Anderson-Darling Test for the Goodness-of-Fit Evaluation of the Fracture Strain Distribution of Acrylic Glass. Glass Structures & Engineering **6**(2), 195–208 (2021)
7. Bisong, E.: Building Machine Learning and Deep Learning Models on Google Cloud Platform: A Comprehensive Guide for Beginners. Apress (2019)
8. Bisong, E., Bisong, E.: Google Colaboratory. Building Machine Learning and Deep Learning Models on Google Cloud Platform: a Comprehensive Guide for Beginners pp. 59–64 (2019)
9. Casella, G., Berger, R.: Statistical Inference (2nd Edition). Duxbury (2001)
10. Chollet, F., et al.: Keras (2015), `https://github.com/fchollet/keras`

11. Cohen, J., Rosenfeld, E., Kolter, Z.: Certified Adversarial Robustness via Randomized Smoothing. In: Proc. 36th Int. Conf. on Machine Learning (ICML) (2019)
12. European Union Aviation Safety Agency: EASA Artificial Intelligence Roadmap 2.0 (2023)
13. Federal Aviation Administration: RTCA, Inc., Document RTCA/DO-178B (1993)
14. Gariel, M., Shimanuki, B., Timpe, R., Wilson, E.: Framework for Certification of AI-Based Systems. arXiv preprint arXiv:2302.11049 (2023)
15. Goodfellow, I., Shlens, J., Szegedy, C.: Explaining and Harnessing Adversarial Examples (2014)
16. Grafström, A., Schelin, L.: How to Select Representative Samples. Scandinavian Journal of Statistics **41**(2), 277–290 (2014)
17. Hadar, A., Levy, N., Winokur, M.: Management and Detection System for Medical Surgical Equipment (2022)
18. Hoeffding, W.: Probability Inequalities for Sums of Bounded Random Variables. Journal of the American Statistical Association pp. 13–30 (1963)
19. Huang, C., Hu, Z., Huang, X., Pei, K.: Statistical Certification of Acceptable Robustness for Neural Networks. In: Proc. Int. Conf. on Artificial Neural Networks (ICANN). pp. 79–90 (2021)
20. Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.: Densely Connected Convolutional Networks. In: Proc. 30th IEEE Conf. on Computer Vision and Pattern Recognition (CVPR). pp. 2261–2269 (2017)
21. Katz, G., Barrett, C., Dill, D., Julian, K., Kochenderfer, M.: Reluplex: An Efficient SMT Solver for Verifying Deep Neural Networks. In: Proc. 29th Int. Conf. on Computer Aided Verification (CAV). pp. 97–117 (2017)
22. Katz, G., Barrett, C., Dill, D., Julian, K., Kochenderfer, M.: Towards Proving the Adversarial Robustness of Deep Neural Networks. In: Proc. 1st. Workshop on Formal Verification of Autonomous Vehicles (FVAV). pp. 19–26 (2017)
23. Katz, G., Huang, D., Ibeling, D., Julian, K., Lazarus, C., Lim, R., Shah, P., Thakoor, S., Wu, H., Zeljić, A., Dill, D., Kochenderfer, M., Barrett, C.: The Marabou Framework for Verification and Analysis of Deep Neural Networks. In: Proc. 31st Int. Conf. on Computer Aided Verification (CAV). pp. 443–452 (2019)
24. Khedr, H., Shoukry, Y.: Certifair: A Framework for Certified Global Fairness of Neural Networks. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 37, pp. 8237–8245 (2023)
25. Krizhevsky, A., Hinton, G.: Learning Multiple Layers of Features from Tiny Images (2009)
26. Lan, S., Huang, C., Wang, Z., Liang, H., Su, W., Zhu, Q.: Design Automation for Intelligent Automotive Systems. In: Proc. IEEE Int. Test Conf. (ITC). pp. 1–10 (2018)
27. Landi, A., Nicholson, M.: ARP4754A/ED-79A-Guidelines for Development of Civil Aircraft and Systems-Enhancements, Novelties and Key Topics. SAE International Journal of Aerospace **4**, 871–879 (2011)
28. Leino, K., Wang, Z., Fredrikson, M.: Globally-Robust Neural Networks. In: International Conference on Machine Learning. pp. 6212–6222. PMLR (2021)
29. Levy, N., Katz, G.: RoMA: a Method for Neural Network Robustness Measurement and Assessment. In: Proc. 29th Int. Conf. on Neural Information Processing (ICONIP) (2021)
30. Levy, N., Katz, G.: gRoMA Code (2022), `https://drive.google.com/drive/folders/1cXio-xjcqh1xEyOl5tPM52y4B9wPqAdy?usp=drive_link`

31. Mangal, R., Nori, A., Orso, A.: Robustness of Neural Networks: A Probabilistic and Practical Approach. In: Proc. 41st IEEE/ACM Int. Conf. on Software Engineering: New Ideas and Emerging Results (ICSE-NIER). pp. 93–96 (2019)
32. Muller, M., Makarchuk, G., Singh, G., Puschel, M., Vechev, M.: PRIMA: General and Precise Neural Network Certification via Scalable Convex Hull Approximations. In: Proc. 49th ACM SIGPLAN Symposium on Principles of Programming Languages (POPL) (2022)
33. Omair, A., et al.: Sample Size Estimation and Sampling Techniques for Selecting a Representative Sample. Journal of Health specialties **2**(4), 142 (2014)
34. Ruan, W., Wu, M., Sun, Y., Huang, X., Kroening, D., Kwiatkowska, M.: Global Robustness Evaluation of Deep Neural Networks with Provable Guarantees for the Hamming Distance. In: Proc. 28th Int. Joint Conf. on Artificial Intelligence (IJCAI) (2019)
35. Simard, P., Steinkraus, D., Platt, J.: Best Practices for Convolutional Neural Networks Applied to Visual Document Analysis. In: Proc. 7th Int. Conf. on Document Analysis and Recognition (ICDAR) (2003)
36. Tit, K., Furon, T., Rousset, M.: Efficient Statistical Assessment of Neural Network Corruption Robustness. In: Proc. 35th Conf. on Neural Information Processing Systems (NeurIPS) (2021)
37. Wang, S., Pei, K., Whitehouse, J., Yang, J., Jana, S.: Formal Security Analysis of Neural Networks using Symbolic Intervals. In: Proc. 27th USENIX Security Symposium (USENIX). pp. 1599–1614 (2018)
38. Wang, S., Zhang, H., Xu, K., Lin, X., Jana, S., Hsieh, C.J., Kolter, Z.: Beta-CROWN: Efficient Bound Propagation with Per-Neuron Split Constraints for Complete and Incomplete Neural Network Verification. In: Proc. 35th Conf. on Neural Information Processing Systems (NeurIPS) (2021)
39. Wang, Z., Wang, Y., Fu, F., Jiao, R., Huang, C., Li, W., Zhu, Q.: A Tool for Neural Network Global Robustness Certification and Training (2022)
40. Xu, C., Chai, D., He, J., Zhang, X., Duan, S.: InnoHAR: A Deep Neural Network for Complex Human Activity Recognition. IEEE Access **7**, 9893–9902 (2019)
41. Zaitang, L., Chen, P.Y., Ho, T.Y.: GREAT Score: Global Robustness Evaluation of Adversarial Perturbation using Generative Models. arXiv preprint arXiv:2304.09875 (2023)
42. Zhang, Y., Zhao, Z., Chen, G., Song, F., Chen, T.: BDD4BNN: a BDD-Based Quantitative Analysis Framework for Binarized Neural Networks. In: Computer Aided Verification: 33rd International Conference, CAV 2021, Virtual Event, July 20–23, 2021, Proceedings, Part I 33. pp. 175–200. Springer (2021)
43. Zhang, Y., Wei, Z., Zhang, X., Sun, M.: Using Z3 for Formal Modeling and Verification of FNN Global Robustness. arXiv preprint arXiv:2304.10558 (2023)