Benchmarking Local Robustness of High-Accuracy Binary Neural Networks for Enhanced Traffic Sign Recognition

Andreea Postovan, Mădălina Erașcu

FROM 2023

Friday 22nd September, 2023

This work was supported by a grant of the Romanian National Authority for Scientific Research and Innovation, CNCS/CCCDI - UEFISCDI, project number PN-III-P1-1.1-TE-2021-0676, within PNCDI III.

Overview

Contents

Motivation

Traffic sign classification is an integral part of any vision system for autonomous driving.

Motivation

Traffic sign classification is an integral part of any vision system for autonomous driving.

Steps for traffic sign classification:

Motivation

Traffic sign classification is an integral part of any vision system for autonomous driving.
Steps for traffic sign classification:

- isolating the traffic sign in a bounding box

Motivation

Traffic sign classification is an integral part of any vision system for autonomous driving.
Steps for traffic sign classification:

- isolating the traffic sign in a bounding box
- classifying the sign into a specific traffic class.

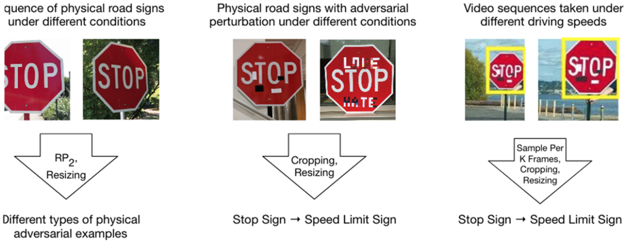
Motivation (cont'd)

Well-know problem of the classifiers: the lack of robustness[1] [2].

[1]Szegedy, Christian, et al. "Intriguing properties of neural networks." arXiv preprint arXiv:1312.6199 (2013).
[2]Guo, Xingwu, et al. "OccRob: Efficient SMT-Based Occlusion Robustness Verification of Deep Neural Networks." TACAS 2023.

Motivation (cont'd)

Well-know problem of the classifiers: the lack of robustness[1] [2].



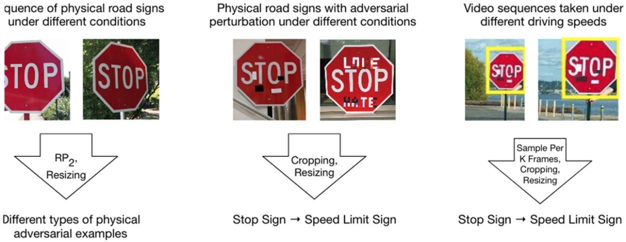
Modified from https://deepdrive.berkeley.edu

[1] Szegedy, Christian, et al. "Intriguing properties of neural networks." arXiv preprint arXiv:1312.6199 (2013).

[2] Guo, Xingwu, et al. "OccRob: Efficient SMT-Based Occlusion Robustness Verification of Deep Neural Networks." TACAS 2023.

Motivation (cont'd)

Well-know problem of the classifiers: the lack of robustness[1] [2].



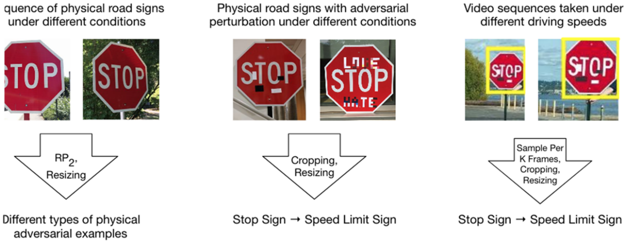
Modified from https://deepdrive.berkeley.edu

Solution:

[1] Szegedy, Christian, et al. "Intriguing properties of neural networks." arXiv preprint arXiv:1312.6199 (2013).
[2] Guo, Xingwu, et al. "OccRob: Efficient SMT-Based Occlusion Robustness Verification of Deep Neural Networks." TACAS 2023.

Motivation (cont'd)

Well-know problem of the classifiers: the lack of robustness[1] [2].



Modified from https://deepdrive.berkeley.edu

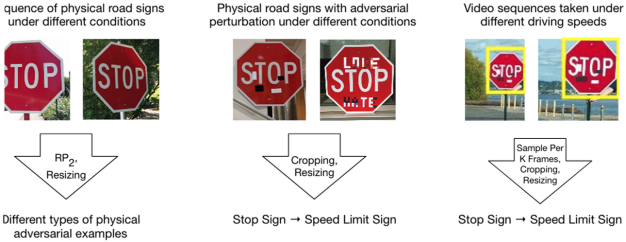
Solution:

- probabilistic methods: traditionally used, have proven limitations

[1] Szegedy, Christian, et al. "Intriguing properties of neural networks." arXiv preprint arXiv:1312.6199 (2013).
[2] Guo, Xingwu, et al. "OccRob: Efficient SMT-Based Occlusion Robustness Verification of Deep Neural Networks." TACAS 2023.

Motivation (cont'd)

Well-know problem of the classifiers: the lack of robustness[1] [2].



Modified from https://deepdrive.berkeley.edu

Solution:

- probabilistic methods: traditionally used, have proven limitations
- logical methods: recently explored, scalability issues ⇝ this presentation, our long time goal

[1] Szegedy, Christian, et al. "Intriguing properties of neural networks." arXiv preprint arXiv:1312.6199 (2013).
[2] Guo, Xingwu, et al. "OccRob: Efficient SMT-Based Occlusion Robustness Verification of Deep Neural Networks." TACAS 2023.

Motivation (cont'd)

Well-know limitation in autonomous driving: computationally limited and energy-constrained devices.

[3] Hubara, Itay, et al. "Binarized neural networks." Advances in neural information processing systems 29 (2016).

Motivation (cont'd)

Well-know limitation in autonomous driving: computationally limited and energy-constrained devices.

Solution: Binary neural network (BNN)[3] - a feedforward network where weights and activations are mainly binary.

[3] Hubara, Itay, et al. "Binarized neural networks." Advances in neural information processing systems 29 (2016).

Motivation (cont'd)

Well-know limitation in autonomous driving: computationally limited and energy-constrained devices.

Solution: Binary neural network (BNN)[3] - a feedforward network where weights and activations are mainly binary.

The absence of BNN models specifically tailored for traffic sign recognition poses a significant gap and a unusual situation, knowing the benefits of BNNs $\rightsquigarrow$ we constructed BNN models with high accuracy.

[3] Hubara, Itay, et al. "Binarized neural networks." Advances in neural information processing systems 29 (2016).

Motivation (cont'd)

Well-know limitation in autonomous driving: computationally limited and energy-constrained devices.

Solution: Binary neural network (BNN)[3] - a feedforward network where weights and activations are mainly binary.

The absence of BNN models specifically tailored for traffic sign recognition poses a significant gap and a unusual situation, knowing the benefits of BNNs $\rightsquigarrow$ we constructed BNN models with high accuracy.

These models should have high accuracy while amenable for formal verification.

[3] Hubara, Itay, et al. "Binarized neural networks." Advances in neural information processing systems 29 (2016).

Motivation (cont'd)

Characteristics that count in machine learning and formal verification:

Motivation (cont'd)

Characteristics that count in machine learning and formal verification:

- Layers' type: convolution (Conv), sign (Sgn), max pooling (MP), batch normalization (BN), fully connected (FC)

Motivation (cont'd)

Characteristics that count in machine learning and formal verification:

- Layers' type: convolution (Conv), sign (Sgn), max pooling (MP), batch normalization (BN), fully connected (FC)
- Number of parameters

Motivation (cont'd)

Characteristics that count in machine learning and formal verification:

- Layers' type: convolution (Conv), sign (Sgn), max pooling (MP), batch normalization (BN), fully connected (FC)
- Number of parameters
- Sparsity

Motivation (cont'd)

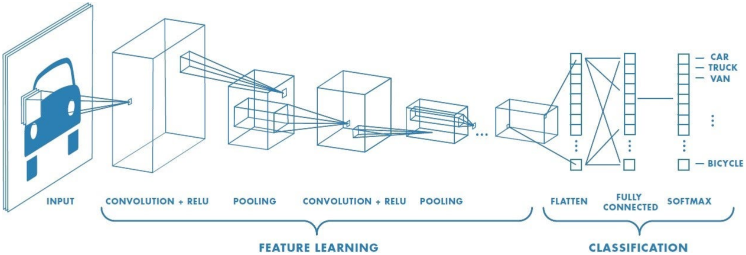
Characteristics that count in machine learning and formal verification:

- Layers' type: convolution (Conv), sign (Sgn), max pooling (MP), batch normalization (BN), fully connected (FC)
- Number of parameters
- Sparsity
- Number of classes

Motivation (cont'd)

Characteristics that count in machine learning and formal verification:

- Layers' type: convolution (Conv), sign (Sgn), max pooling (MP), batch normalization (BN), fully connected (FC)
- Number of parameters
- Sparsity
- Number of classes



From https://saturncloud.io/blog/a-comprehensive-guide-to-convolutional-neural-networks-the-eli5-way/

Contents

Problem Specification: Verification of BNNs.

Given a trained model and a property to be verified, does the model satisfy that property?

[4] Katz, G., Barrett, C., Dill, D., Julian, K., Kochenderfer, M.: Reluplex: An efficient SMT solver for verifying deep neural networks. Supplementary Material (2017). https://arxiv.org/abs/1702.01135

Problem Specification: Verification of BNNs.

Given a trained model and a property to be verified, does the model satisfy that property?

Approach:

[4] Katz, G., Barrett, C., Dill, D., Julian, K., Kochenderfer, M.: Reluplex: An efficient SMT solver for verifying deep neural networks. Supplementary Material (2017). https://arxiv.org/abs/1702.01135

Problem Specification: Verification of BNNs.

Given a trained model and a property to be verified, does the model satisfy that property?

Approach:

- The verification problem is translated into a constrained satisfaction problem.

---

[4] Katz, G., Barrett, C., Dill, D., Julian, K., Kochenderfer, M.: Reluplex: An efficient SMT solver for verifying deep neural networks. Supplementary Material (2017). https://arxiv.org/abs/1702.01135

Problem Specification: Verification of BNNs.

Given a trained model and a property to be verified, does the model satisfy that property?

Approach:

- The verification problem is translated into a constrained satisfaction problem.
- Existing verification tools can be used to solve it.

[4] Katz, G., Barrett, C., Dill, D., Julian, K., Kochenderfer, M.: Reluplex: An efficient SMT solver for verifying deep neural networks. Supplementary Material (2017). https://arxiv.org/abs/1702.01135

Problem Specification: Verification of BNNs.

Given a trained model and a property to be verified, does the model satisfy that property?

Approach:

- The verification problem is translated into a constrained satisfaction problem.
- Existing verification tools can be used to solve it.

Challenges:

[4] Katz, G., Barrett, C., Dill, D., Julian, K., Kochenderfer, M.: Reluplex: An efficient SMT solver for verifying deep neural networks. Supplementary Material (2017). https://arxiv.org/abs/1702.01135

Problem Specification: Verification of BNNs.

Given a trained model and a property to be verified, does the model satisfy that property?

Approach:

- The verification problem is translated into a constrained satisfaction problem.
- Existing verification tools can be used to solve it.

Challenges:

- NP-complete problem[4]

[4] Katz, G., Barrett, C., Dill, D., Julian, K., Kochenderfer, M.: Reluplex: An efficient SMT solver for verifying deep neural networks. Supplementary Material (2017). https://arxiv.org/abs/1702.01135

Problem Specification: Verification of BNNs.

Given a trained model and a property to be verified, does the model satisfy that property?

Approach:

- The verification problem is translated into a constrained satisfaction problem.
- Existing verification tools can be used to solve it.

Challenges:

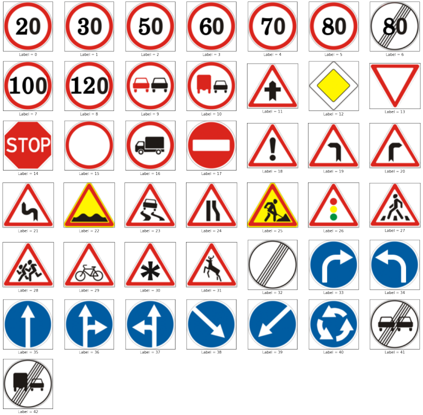
- NP-complete problem[4]
- How to formalize the property to be verified

[4] Katz, G., Barrett, C., Dill, D., Julian, K., Kochenderfer, M.: Reluplex: An efficient SMT solver for verifying deep neural networks. Supplementary Material (2017). https://arxiv.org/abs/1702.01135

Contents

Contents

Data collection



Training:

- **GTSRB (German) traffic sign dataset.**
  - Classes: 43,
  - Size: from 25 $\times$ 25 to 243 $\times$ 225, and not all of them are square.
  - Each class: 210 - 2250 images
  - 39209 images used for training and validation with ratio 80:20

Testing:

- **GTSRB (German) traffic sign dataset.**
  - 12630 images used for testing
- **Belgium traffic sign dataset.**
  - Number of images = 4533.
  - Only 23 classes match the one from GTSRB.
- **Chinese traffic sign dataset.**
  - Number of images = 1818.
  - Only 15 classes match the one from GTSRB.

Contents

Data analysis

Difference between Belgium (left) and GTSRB (right) dataset

Difference between Chinese (left) and GTSRB (right) dataset

Contents

BNNs Architectures with Best Accuracy[5]

The architectures below were obtained by a bottom-up approach, starting with simple layers (fully connected) and stacking new more complicated ones for higher accuracy.

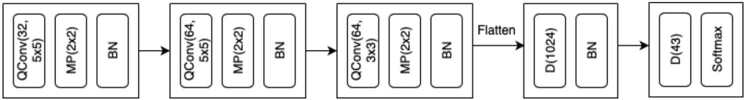


Figure: Architecture with Best Accuracy for GTSRB (96.45%) and Belgium (88.17%) dataset. Input: 64 px x 64 px

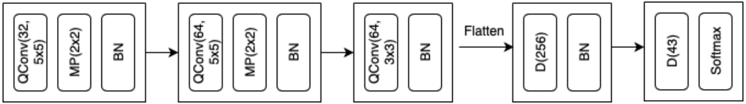


Figure: Architecture with Best Accuracy (83.9%) for Chinese dataset. Input: 48 px x 48 px

[5] More details in: A. Postovan, M, Erașcu. Architecturing binarized neural networks for traffic sign recognition. to appear in ICANN 2023

XNOR Architecture

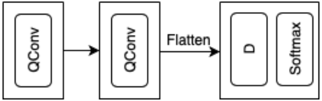


Figure: XNOR(QConv) architecture

Table: XNOR(QConv) architecture. Image size: 30px $\times$ 30px. Dataset for train and test: GTSRB.

Model description	Acc	#Binary Params	Model Size (in KiB)	
			Binary	Float-32
QConv(16, 3×3), QConv(32, 2×2), D(43)	81.54	1005584	122.75	3932.16

Contents

Contents

Definition of the Property to be Verified

Property to be verified: robustness – refers to their ability to maintain stable and accurate outputs in the presence of perturbations or adversarial inputs. Adversarial inputs are intentionally crafted inputs designed to deceive or mislead the network's predictions.

Definition of the Property to be Verified

Property to be verified: robustness – refers to their ability to maintain stable and accurate outputs in the presence of perturbations or adversarial inputs. Adversarial inputs are intentionally crafted inputs designed to deceive or mislead the network's predictions.

- Local robustness ensures that for a given input x from a set χ, the neural network F remains unchanged within a specified perturbation radius ϵ, implying that small variations in the input space do not result in different outputs. The output for the input x is represented by its label l_x. We consider L_∞ norm defined as $||x||_\infty = \sup_n |x_n|$.

Definition of the Property to be Verified

Property to be verified: robustness – refers to their ability to maintain stable and accurate outputs in the presence of perturbations or adversarial inputs. Adversarial inputs are intentionally crafted inputs designed to deceive or mislead the network's predictions.

- Local robustness ensures that for a given input x from a set χ, the neural network F remains unchanged within a specified perturbation radius ϵ, implying that small variations in the input space do not result in different outputs. The output for the input x is represented by its label l_x. We consider L_∞ norm defined as $||x||_\infty = \sup_n |x_n|$.
- Global robustness is extension of the local robustness and it is defined as the expected maximum safe radius over a given test dataset, representing a collection of inputs.

Definition of the Property to be Verified

Property to be verified: robustness – refers to their ability to maintain stable and accurate outputs in the presence of perturbations or adversarial inputs. Adversarial inputs are intentionally crafted inputs designed to deceive or mislead the network's predictions.

- Local robustness ensures that for a given input x from a set χ, the neural network F remains unchanged within a specified perturbation radius ϵ, implying that small variations in the input space do not result in different outputs. The output for the input x is represented by its label l_x. We consider L_∞ norm defined as $||x||_\infty = \sup_n |x_n|$.
- Global robustness is extension of the local robustness and it is defined as the expected maximum safe radius over a given test dataset, representing a collection of inputs.

Definition of local robustness useful in a computational setting. A network is ϵ-locally robust in the input x if for every x', such that $||x - x'||_\infty \leq \epsilon$, the network assigns the same label to x and x'.

Contents

Property Specification

In VNN-LIB standard which uses the SMT-LIB format.

Property Specification

In VNN-LIB standard which uses the SMT-LIB format.

A VNN-LIB file is structured as follows:

Property Specification

In VNN-LIB standard which uses the SMT-LIB format.

A VNN-LIB file is structured as follows:

1. definition of input variables representing the values of the pixels X_i ($i = \overline{1, P}$, where P is the dimension of the input image: $N \times M \times 3$ pixels).

Property Specification

In VNN-LIB standard which uses the SMT-LIB format.

A VNN-LIB file is structured as follows:

1. definition of input variables representing the values of the pixels X_i ($i = \overline{1, P}$, where P is the dimension of the input image: $N \times M \times 3$ pixels).
2. definition of the output variables representing the values Y_j ($j = \overline{1, L}$, where L is the number of classes of the images in the dataset).

Property Specification

In VNN-LIB standard which uses the SMT-LIB format.

A VNN-LIB file is structured as follows:

1. definition of input variables representing the values of the pixels X_i ($i = \overline{1, P}$, where P is the dimension of the input image: $N \times M \times 3$ pixels).
2. definition of the output variables representing the values Y_j ($j = \overline{1, L}$, where L is the number of classes of the images in the dataset).
3. bounding constraints for the input variables: local robustness definition is used for generating the property taking into account that vector x (its elements are the values of the pixels of the image) and ε (perturbation) are known.

```
(assert (<= X_2699 34.00000000))
(assert (>= X_2699 14.00000000))
```

Property Specification

In VNN-LIB standard which uses the SMT-LIB format.

A VNN-LIB file is structured as follows:

1. definition of input variables representing the values of the pixels X_i ($i = \overline{1, P}$, where P is the dimension of the input image: $N \times M \times 3$ pixels).
2. definition of the output variables representing the values Y_j ($j = \overline{1, L}$, where L is the number of classes of the images in the dataset).
3. bounding constraints for the input variables: local robustness definition is used for generating the property taking into account that vector x (its elements are the values of the pixels of the image) and ε (perturbation) are known.
   ```
   (assert (<= X_2699 34.00000000))
   (assert (>= X_2699 14.00000000))
   ```
4. constraints involving the output variables assessing the value of the output label.
   ```
   (assert (or (>= Y_0 Y_38)
               ...
               (>= Y_37 Y_38)
               (>= Y_39 Y_38)
               ...
               (>= Y_42 Y_38)))
   ```

Model Representation: Open Neural Network Exchange (ONNX)

- storage and organization of large amounts of data, including the parameters and architecture of machine learning models
- vendor-neutral

Model Representation: Open Neural Network Exchange (ONNX)

- storage and organization of large amounts of data, including the parameters and architecture of machine learning models
- vendor-neutral
- ONNX representation of the neural network is transformed into a constraint satisfaction problem in the VNN-LIB format

Contents

Benchmarks Proposal

VNN-COMP 2023:

- neural network models in ONNX format
- property specification in VNN-LIB format

Benchmarks Proposal

VNN-COMP 2023:

- neural network models in ONNX format
- property specification in VNN-LIB format

Characteristics of the previous models to be verified

# of Params	Input Dimension	Sparsity	# of Regions
905k-1.7M	2.7k-12k	0%	43 or 38

Benchmarks Proposal

VNN-COMP 2023:

- neural network models in ONNX format
- property specification in VNN-LIB format

Characteristics of the previous models to be verified

# of Params	Input Dimension	Sparsity	# of Regions
905k-1.7M	2.7k-12k	0%	43 or 38

Adversarial robustness property: property specifications encompass perturbations within the infinity norm around zero, with radius denoted as $\epsilon = \{1, 3, 5, 10, 15\}$.

Benchmarks Proposal

VNN-COMP 2023:

- neural network models in ONNX format
- property specification in VNN-LIB format

Characteristics of the previous models to be verified

# of Params	Input Dimension	Sparsity	# of Regions
905k-1.7M	2.7k-12k	0%	43 or 38

Adversarial robustness property: property specifications encompass perturbations within the infinity norm around zero, with radius denoted as $\epsilon = \{1, 3, 5, 10, 15\}$.

Randomly selected 3 distinct images from the test set of the GTSRB dataset for each model and have generated the VNN-LIB files for each epsilon in the set, in the way we ended up having 45 VNN-LIB files in total.

Benchmarks Proposal

VNN-COMP 2023:

- neural network models in ONNX format
- property specification in VNN-LIB format

Characteristics of the previous models to be verified

# of Params	Input Dimension	Sparsity	# of Regions
905k-1.7M	2.7k-12k	0%	43 or 38

Adversarial robustness property: property specifications encompass perturbations within the infinity norm around zero, with radius denoted as $\epsilon = \{1, 3, 5, 10, 15\}$.

Randomly selected 3 distinct images from the test set of the GTSRB dataset for each model and have generated the VNN-LIB files for each epsilon in the set, in the way we ended up having 45 VNN-LIB files in total.

Timeout of 480 seconds was allocated for each instance, in total 6 hours for all instances.

Benchmarks Proposal

VNN-COMP 2023:

- neural network models in ONNX format
- property specification in VNN-LIB format

Characteristics of the previous models to be verified

# of Params	Input Dimension	Sparsity	# of Regions
905k-1.7M	2.7k-12k	0%	43 or 38

Adversarial robustness property: property specifications encompass perturbations within the infinity norm around zero, with radius denoted as $\epsilon = \{1, 3, 5, 10, 15\}$.

Randomly selected 3 distinct images from the test set of the GTSRB dataset for each model and have generated the VNN-LIB files for each epsilon in the set, in the way we ended up having 45 VNN-LIB files in total.

Timeout of 480 seconds was allocated for each instance, in total 6 hours for all instances.

Our benchmark was used for scoring the competing tools but different images were chosen in order to avoid tuning of the solvers for precise instances.

Experimental Results of the VNN-COMP 2023

Table: VNN-COMP 2023 Results for Traffic Signs Recognition Benchmark

#	Tool	Verified	Falsified	Fastest	Penalty	Score	Percent
1	Marabou	0	18	0	1	30	100%
2	PyRAT	0	7	0	1	-80	0%
3	NeuralSAT	0	31	0	4	-290	0%
4	alpha-beta-CROWN	0	39	0	3	-60	0%

- **Verified** is number of instances that were UNSAT (no counterexample) and proven by the tool.
- **Falsified** is number that were SAT (counterexample was found) and reported by the tool.
- **Fastest** is the number where the tool was fastest (this did not impact the scoring in this year competition). Penalty is the number where the tool gave the incorrect result or did not produce a valid counterexample.
- **Score** is the sum of scores (10 points for each correct answer and -150 for incorrect ones).
- **Percent** is the score of the tool divided by the best score for the benchmark (so the tool with the highest score for each benchmark gets 100) and was used to determine final scores across all benchmarks.

Contents

Conclusion and Future Work

Conclusion

- Proposal of BNNs benchmarks for local robustness verification.

Conclusion and Future Work

Conclusion

- Proposal of BNNs benchmarks for local robustness verification.
- VNN-COMP 2023 evaluation: 4 out of 7 competing tools produced results.

Conclusion and Future Work

Conclusion

- Proposal of BNNs benchmarks for local robustness verification.
- VNN-COMP 2023 evaluation: 4 out of 7 competing tools produced results.

Future Work

Conclusion and Future Work

Conclusion

- Proposal of BNNs benchmarks for local robustness verification.
- VNN-COMP 2023 evaluation: 4 out of 7 competing tools produced results.

Future Work

- Investigate for which architectures the previous results were obtained.

Conclusion and Future Work

Conclusion

- Proposal of BNNs benchmarks for local robustness verification.
- VNN-COMP 2023 evaluation: 4 out of 7 competing tools produced results.

Future Work

- Investigate for which architectures the previous results were obtained.
- Investigate the potential for solving more instances by extending the time limit (currently set at 8 minutes).

Conclusion and Future Work

Conclusion

- Proposal of BNNs benchmarks for local robustness verification.
- VNN-COMP 2023 evaluation: 4 out of 7 competing tools produced results.

Future Work

- Investigate for which architectures the previous results were obtained.
- Investigate the potential for solving more instances by extending the time limit (currently set at 8 minutes).
- Understand the factors contributing to incorrect outputs from the tools on specific benchmark tasks.