Critically Assessing the State of the Art in CPU-based Local Robustness Verification

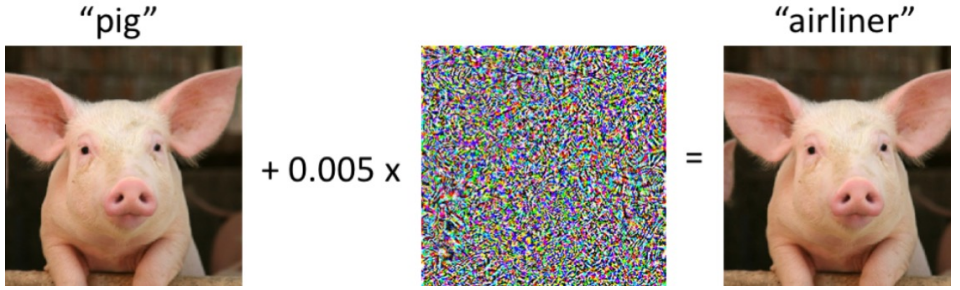
Matthias König[1], Annelot W. Bosman[1], Holger H. Hoos[1,2,3], Jan N. van Rijn[1] | ADA Research Group

Washington, D.C., 14.2.23

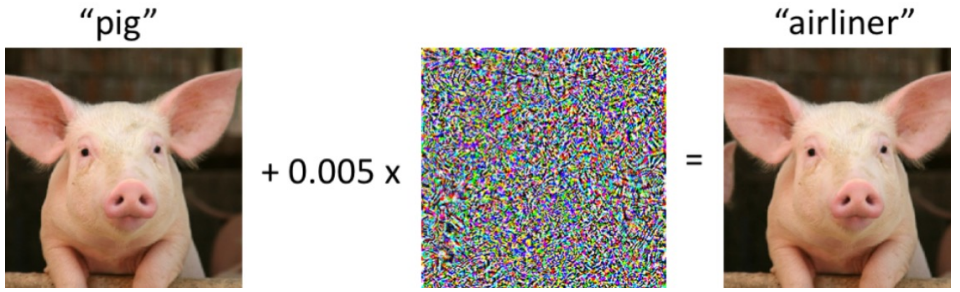
Universiteit Leiden
The Netherlands

[1] LIACS, Leiden University, The Netherlands
[2] Chair for AI Methodology, RWTH Aachen University, Germany
[3] CS Department, University of British Columbia, Canada

Local Robustness Verification



Local Robustness Verification



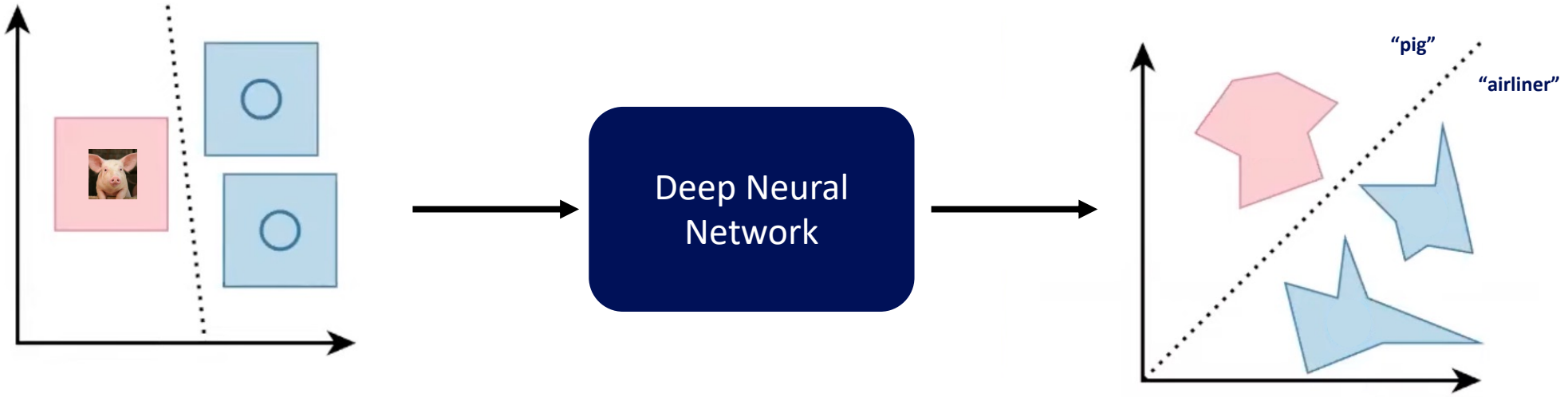
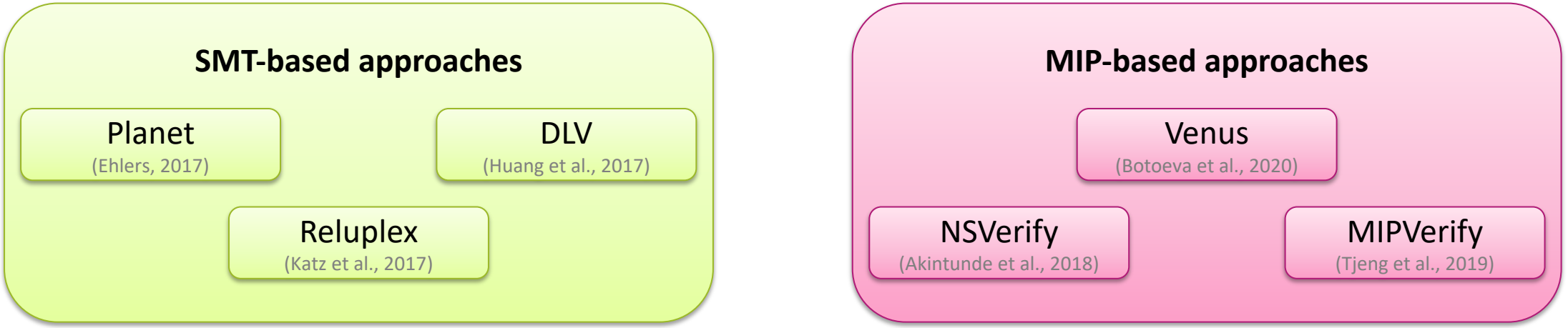


Figure inspired by Stanford AI Safety Seminar

Verification methods are highly diverse



Network architectures are highly diverse

Can differ in terms of…

- activation function (ReLU, Tanh,…)
- layer operations (pooling, convolutions,…)
- input specifications (number of inputs, dimensionality,…)

Which verification method is most suitable to verify a given network?

No clear answer can be derived from literature

Neural network verification methods are typically evaluated…

- on small number of benchmarks
- against ambiguous baselines

No clear answer can be derived from literature

Neural network verification methods are typically evaluated…

- on small number of benchmarks
- against ambiguous baselines

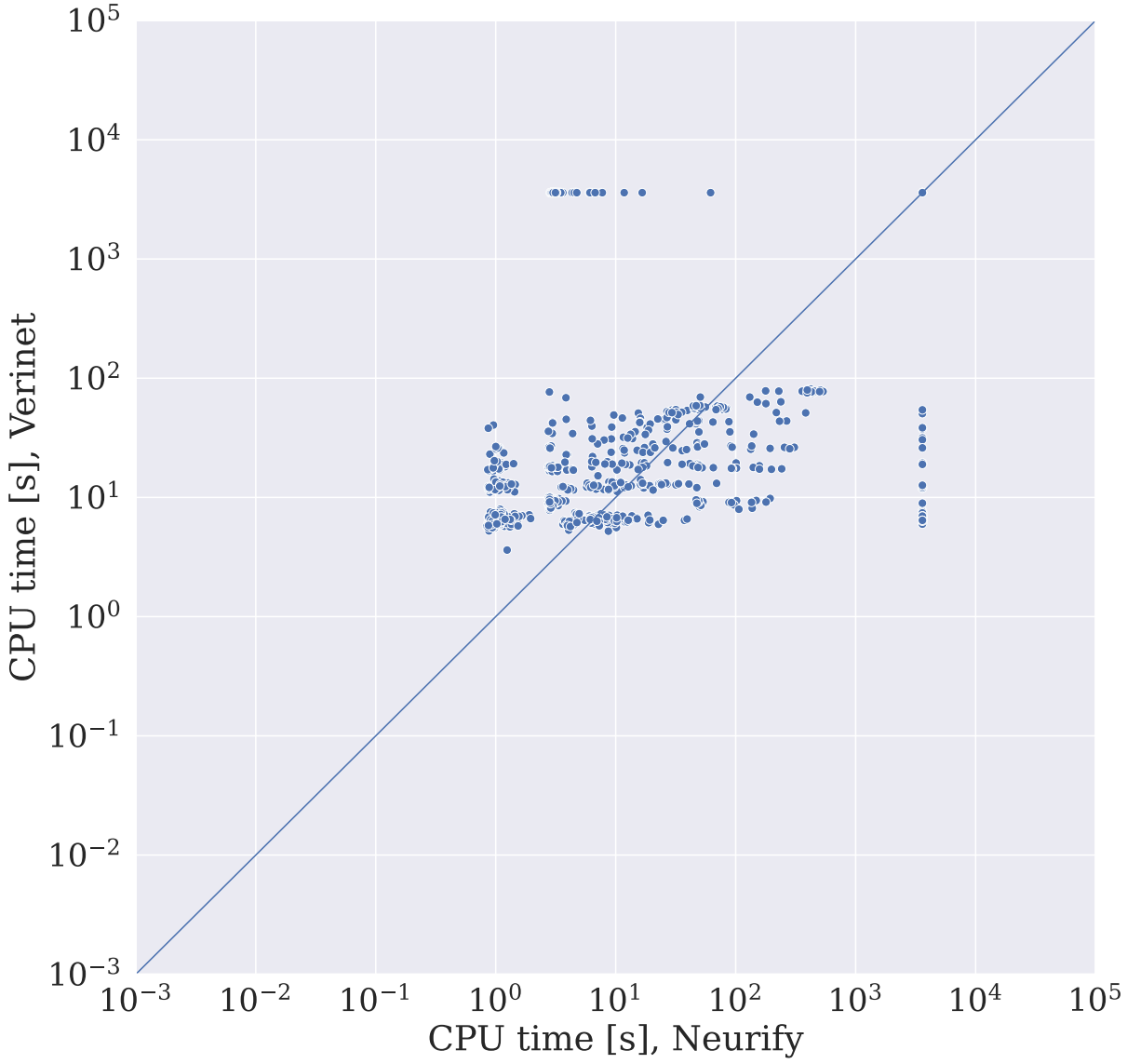
VNN Competition…

- seeks to determine “a winner” based on performance ranking

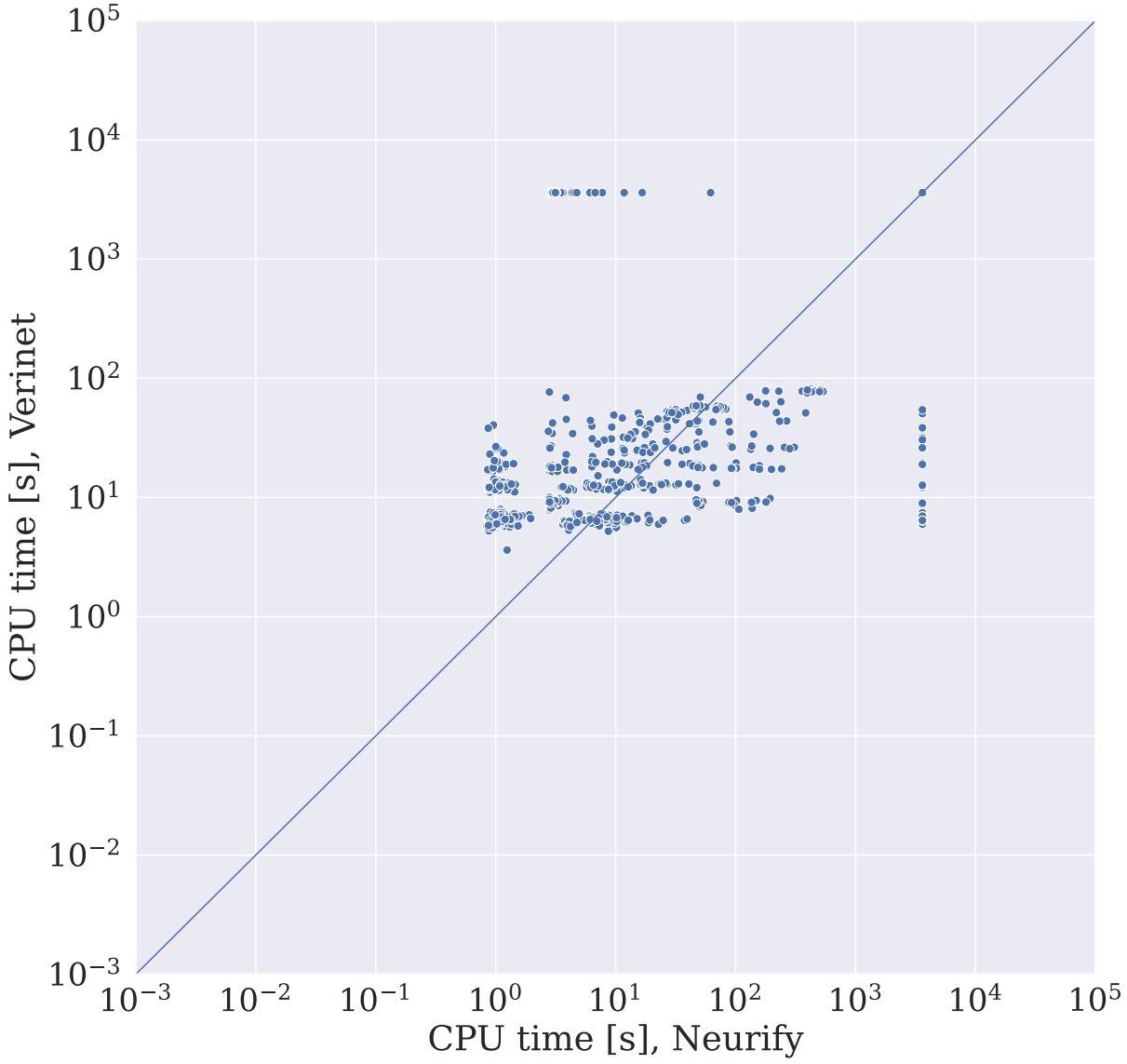
Verifier	#Solved (*n*=1 500)
BaBSB (Bunel et al., 2018)	307
Marabou (Katz et al., 2018)	400
Neurify (Wang et al., 2018)	915
nnenum (Bak et al., 2020)	76
Verinet (Henriksen & Lomuscio, 2020)	841

Verifier	#Solved (*n*=1 500)
BaBSB (Bunel et al., 2018)	307
Marabou (Katz et al., 2018)	400
Neurify (Wang et al., 2018) 	915
nnenum (Bak et al., 2020)	76
Verinet (Henriksen & Lomuscio, 2020) 	841

Verifier	#Solved
BaBSB (Bunel et al., 2018)	307
Marabou (Katz et al., 2018)	400
Neurify (Wang et al., 2018)	915
nnenum (Bak et al., 2020)	76
Verinet (Henriksen & Lomuscio, 2020)	841



Verifier	#Solved
BaBSB (Bunel et al., 2018)	307
Marabou (Katz et al., 2018)	400
Neurify (Wang et al., 2018)	915
nnenum (Bak et al., 2020)	76
Verinet (Henriksen & Lomuscio, 2020)	841



Our proposed benchmarking approach

How does each algorithm contribute to SOTA?

Measured by means of…

- marginal contribution

How does each algorithm contribute to SOTA?

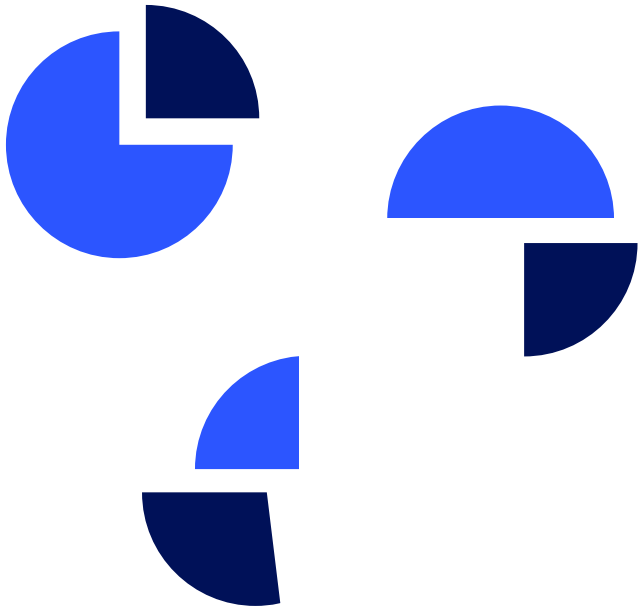
Measured by means of…

- marginal contribution

How does each algorithm contribute to SOTA?

Measured by means of…

- marginal contribution
- Shapley value (Fréchette et al., 2016)



How does each algorithm contribute to SOTA?

Measured by means of…

- marginal contribution
- Shapley value (Fréchette et al., 2016)

→ Measure performance of a given algorithm in relation to others

Conclusions

- Strong complementarity between state-of-the-art verification algorithms

[König, Bosman, Hoos, van Rijn. Critically Assessing the State of the Art in CPU-based Local Robustness Verification. *Workshop on Artificial Intelligence Safety @AAAI*. 2023]

Conclusions

- Strong complementarity between state-of-the-art verification algorithms

Verifier	#Solved (n=1 500)	Shapley
BaBSB (Bunel et al., 2018)	307	86
Marabou (Katz et al., 2018)	400	117
Neurify (Wang et al., 2018)	915	411
nnenum (Bak et al., 2020)	76	28
Verinet (Henriksen & Lomuscio, 2020)	841	330

[König, Bosman, Hoos, van Rijn. Critically Assessing the State of the Art in CPU-based Local Robustness Verification. *Workshop on Artificial Intelligence Safety @AAAI*. 2023]

Conclusions

- Strong complementarity between state-of-the-art verification algorithms
- Large performance differences between image datasets

[König, Bosman, Hoos, van Rijn. Critically Assessing the State of the Art in CPU-based Local Robustness Verification. *Workshop on Artificial Intelligence Safety @AAAI*. 2023]

Conclusions

- Strong complementarity between state-of-the-art verification algorithms
- Large performance differences between image datasets

Future work involves…

- including GPU-based verifiers
- analysing broader set of perturbation radii
- studying failure modes of verification systems

[König, Bosman, Hoos, van Rijn. Critically Assessing the State of the Art in CPU-based Local Robustness Verification. *Workshop on Artificial Intelligence Safety @AAAI*. 2023]

Conclusions

- Strong complementarity between state-of-the-art verification algorithms
- Large performance differences between image datasets

→How should we evaluate verification algorithms?

[König, Bosman, Hoos, van Rijn. Critically Assessing the State of the Art in CPU-based Local Robustness Verification. *Workshop on Artificial Intelligence Safety @AAAI*. 2023]