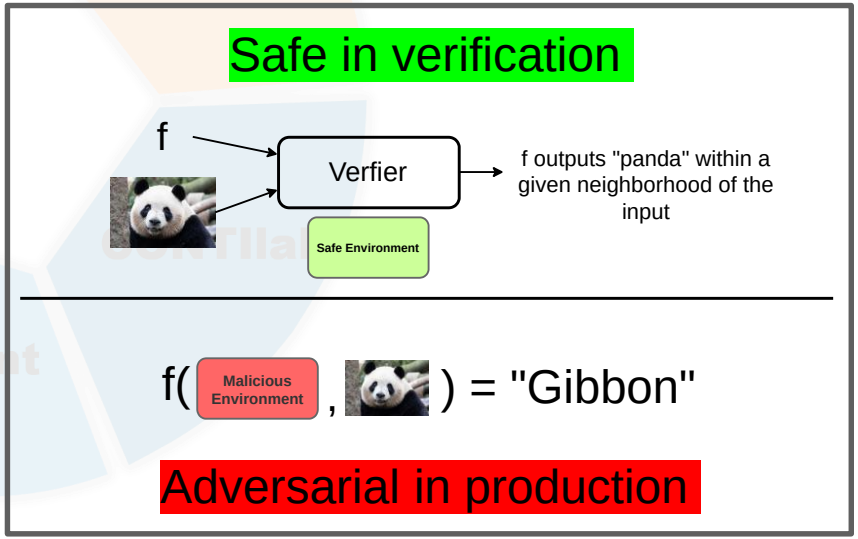
Name: Attila Szász, PhD student

Project type: laboratory project

Topic: On the Challenges of Sound Verification of Neural Networks

Supervisors: Dr. Balázs Bánhelyi

- Quantizing models to lower precision yields significant performance gains; however, many studies have shown that quantized models often suffer from considerable **accuracy degradation**.
- The main cause of this phenomenon is the effect of numerical representation on the accumulation of numerical errors, which highlights the importance of **clearly defining different environments** (e.g., training, evaluation, etc.).
- In this work, we highlighted the importance of precisely defining environments in the context of **verification**.
- Main idea: **The verifier operates under the assumption of an environment that differs from the one used in production.**
- This is a serious problem that can be exploited by an attacker.

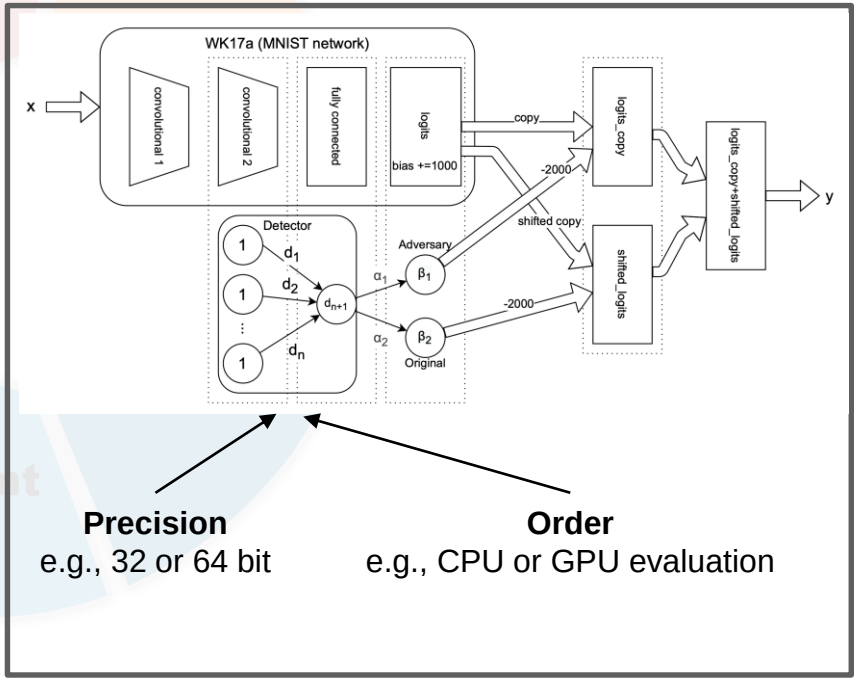


Backdoors

- We defined backdoors that detect the underlying implementation (e.g., numerical representation or the order of summation) and **trigger malicious behavior only under specific environments**. (Typically **differs from the one used during verification**.)
- Main idea: Bound propagation algorithms used in verification are sensitive to numerical precision and operation order.

Expression tree	$(2^{53}+1)+1$	$(1+2^{53})+1$	$(1+1)+2^{53}$
FP64 result	2^{53}	2^{53}	$2^{53}+2$
Exact value	$2^{53}+2$	$2^{53}+2$	$2^{53}+2$
Interval bounds	$[2^{53}, 2^{53}+4]$	$[2^{53}, 2^{53}+4]$	$[\mathbf{2^{53}+2}, \mathbf{2^{53}+2}]$

- Precision detector: $2^{24}+1-2^{24}$
- Order detector: $\underbrace{1+\cdots+1}_{h\times}+\omega-\omega$



Results – ICML 2025 Spotlight - Vancouver

Verifier	Ver. Env.	Bounding	Precision	Order1	Order2	Order3
MIPVerify (Tjeng et al., 2017)	64-bit, CPU	IBP	**unsound**	sound	**unsound**	**unsound**
MN-BAB (Ferrari et al., 2022)	64-bit, GPU	Polyhedra	**unsound**	sound	**unsound**	**unsound**
β-CROWN BaB (Wang et al., 2021)	32-bit, CPU	Polyhedra	**unsound**	sound	**unsound**	**unsound**
β-CROWN BaB (Wang et al., 2021)	64-bit, CPU	Polyhedra	**unsound**	sound	**unsound**	**unsound**
β-CROWN BaB (Wang et al., 2021)	64-bit, GPU	Polyhedra	**unsound**	sound	**unsound**	**unsound**
GCP-CROWN (Zhang et al., 2022)	64-bit, CPU	Polyhedra	**unsound**	sound	**unsound**	**unsound**
DeepPoly (Singh et al., 2019a)	64-bit, CPU	Polyhedra	**unsound**	sound	sound	**unsound**
RefinePoly (Singh et al., 2019a)	64-bit, CPU	Polyhedra	**unsound**	sound	sound	**unsound**
DeepZono (Singh et al., 2018)	64-bit, CPU	Zonotope	**unsound**	sound	sound	**unsound**
RefineZono (Singh et al., 2019b)	64-bit, CPU	Zonotope	**unsound**	sound	sound	**unsound**

**No Soundness in the Real World:
On the Challenges of the Verification of Deployed Neural Networks**

Attila Szász [1] Balázs Bánhelyi [1] Márk Jelasity [1 2]

Acknowledgements

This work was supported by the European Union project RRF-2.3.1-21-2022-00004 within the framework of the Artificial Intelligence National Laboratory and project TKP2021-NVA-09, implemented with the support provided by the Ministry of Culture and Innovation of Hungary from the National Research, Development and Innovation Fund, financed under the TKP2021-NVA funding scheme. We thank the support by the PIA Project, a collaboration between the University of Szeged and Continental Autonomous Mobility Hungary Ltd. with the goal of supporting students' research in the field of deep learning and autonomous driving. We thank Andras Balogh for his initial feedback and our anonymous reviewers for the relevant

Continental

PIA PROJECT PROFESSIONAL INTELLIGENCE FOR AUTOMOTIVE

UNIVERSITY OF SZEGED INSTITUTE OF INFORMATICS

Upcoming work and publication

Verifier	Ver. Env.	Bounding	Pr.	O1	O2	O3	Zombori et al. (2021)
MIPVerify (Tjeng et al. (2017))	64-bit, CPU	IBP	U	S	U	U	U
MN-BAB (Ferrari et al. (2022))	64-bit, GPU	Symbolic	U	S	U	U	U
β-CROWN BaB (Wang et al. (2021))	32-bit, CPU	Symbolic	U	S	U	U	[no 32-bit model]
β-CROWN BaB (Wang et al. (2021))	64-bit, CPU	Symbolic	U	S	U	U	S
β-CROWN BaB (Wang et al. (2021))	64-bit, GPU	Symbolic	U	S	U	U	S
GCP-CROWN (Zhang et al. (2022))	64-bit, CPU	Symbolic	U	S	U	U	S
DeepPoly (Singh et al. (2019a))	64-bit, CPU	Symbolic	U	S	S	U	S
RefinePoly (Singh et al. (2019a))	64-bit, CPU	Symbolic	U	S	S	U	U
DeepZ (Singh et al. (2018))	64-bit, CPU	Symbolic	U	S	S	U	S
RefineZono (Singh et al. (2019b))	64-bit, CPU	Symbolic	U	S	S	U	[Gurobi error]
FPSoundIBP (ours)	32-bit, CPU	IBP	S	S	S	S	S
FPSoundSymbolic (ours)	32-bit, CPU	Symbolic	S	S	S	S	S

- We developed an algorithm that is capable of handling the non-associative nature of floating-point operations.
- We implemented a neural network verifier incorporating this algorithm (FPSoundIBP and FPSoundSymbolic) and demonstrated that our approach is uniquely capable of bounding the aforementioned numerical error-based backdoors.
- This work is currently being prepared for publication.