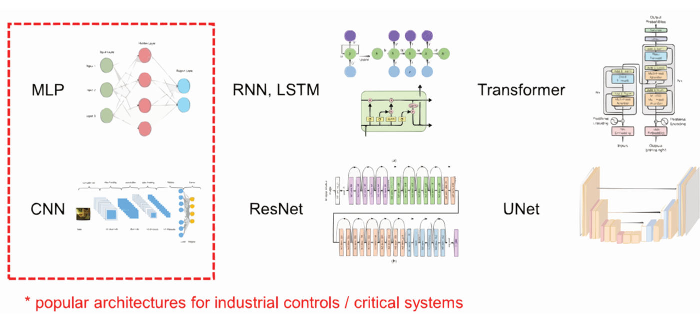
Formal Verification of Neural Networks

Thanapong Sommart[1], Borja Fernández Adiego[2]

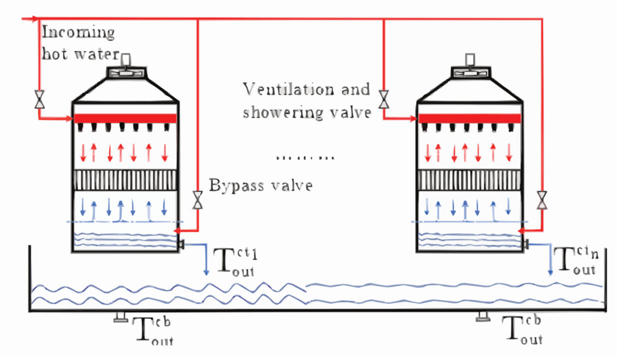
[1] Sirindhorn International Institute of Technology, Thammasat University, Pathum Thani, Thailand
[2] CERN, Geneva, Switzerland
thanapong.net@gmail.com, borja.fernandez.adiego@cern.ch

Introduction



Neural networks (NNs) with various architectures and sizes are becoming ubiquitous in many fields and applications.

At CERN, several NNs are being developed specifically for control systems for LHC, such as cooling tower control systems and BLM sensor instance segmentation.



However, reliability and safety of NNs are heavily concerned due to the "black box" nature of their behavior. What if they give unexpected outputs? Can it be guaranteed or verified that a certain scenario will never happen?

Verification is essential for NNs in critical systems.

Background and Tools

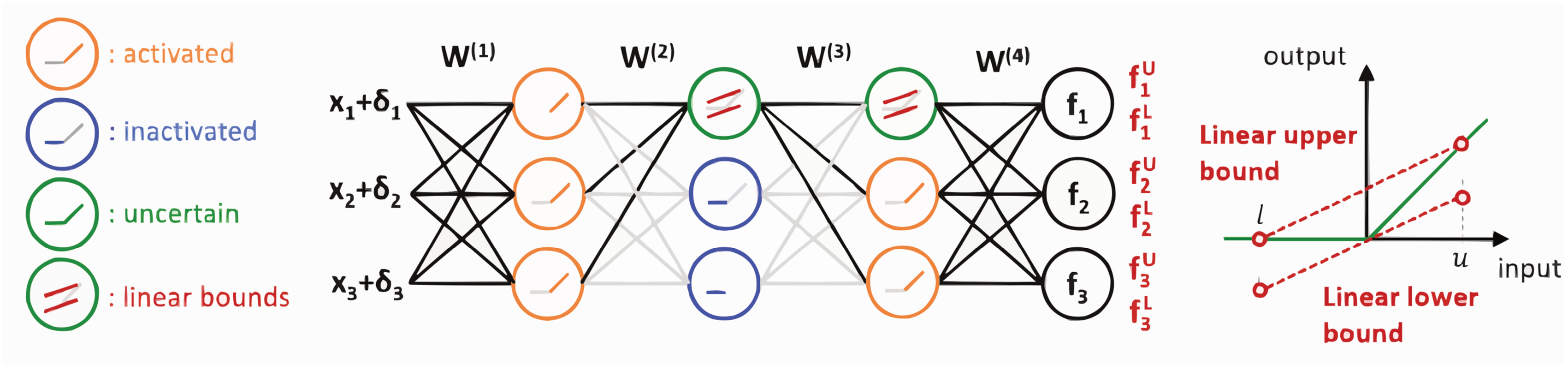


Figure from [1] illustrating one of the concepts for deriving convex approximation of ReLU activation function using linear lower and upper bounds.

A common approach is to efficiently compute the output bounds of neural network outputs by relaxing activation functions and non-linear operations.

Here, top-performing tools in the *VNN-COMP 2022* [2] are investigated.

alpha-beta CROWN
Linear Bound Propagation
\+ Branch and Bound
Pros:
- Has the lowest runtime
- Works with custom built networks

Cons:
- Needs complicated configurations
- Provides some unknown results

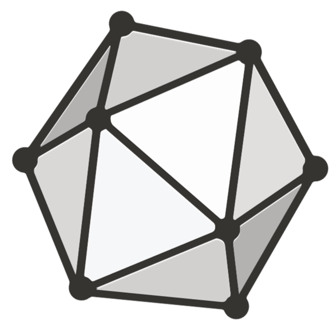
nnenum
Zonotope Over-Approximation
\+ Geometric Path Enumeration
Pros:
- Is fast and easy to use
- Barely provides unknown results

Cons:
- Only works with ReLU activation

VeriNet
Symbolic Interval Propagation
\+ Branch and Bound
Pros:
- Works with most activations
- Can utilize multiprocessing

Cons:
- Does not support certain operations
- Has higher runtime

Pipeline

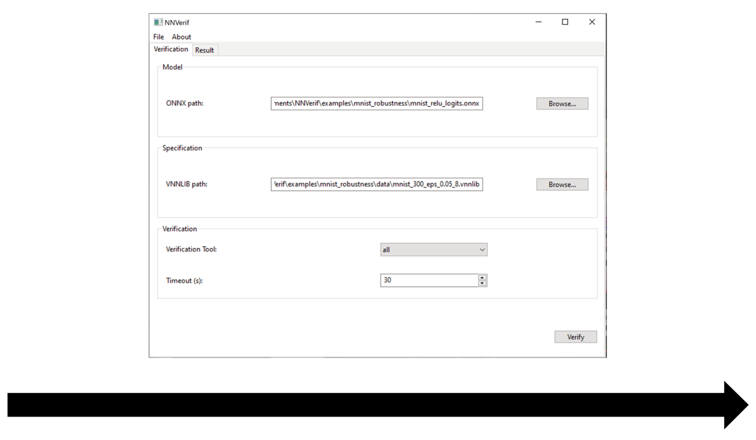


ONNX Model

\+

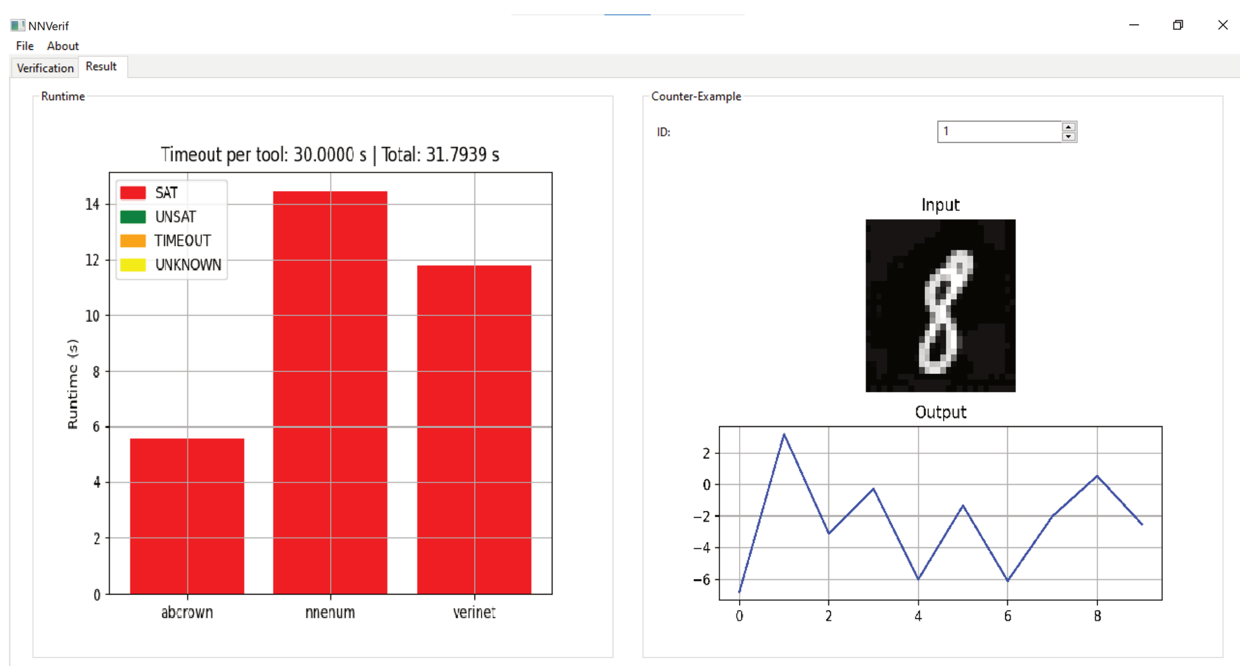
```
(declare-const X_0 Real)
(declare-const X_1 Real)
(declare-const X_2 Real)
(declare-const Y_0 Real)
(declare-const Y_1 Real)
(declare-const Y_2 Real)

(assert (>= X_0 23.6))
(assert (<= X_0 25.0))
(assert (>= X_1 23))
(assert (<= X_1 27))
(assert (>= X_2 14.1))
(assert (<= X_2 21.0))

(assert (or
    (and (>= Y_0 Y_2))
    (and (>= Y_1 Y_2)(>= Y_1 Y_0))
))
```

VNNLIB Specification



NN Verification Tools

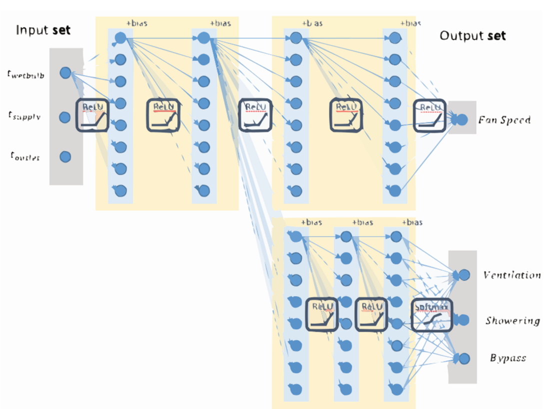
Each tool works by trying to find a single counter-example which satisfies all conditions given in the specification.

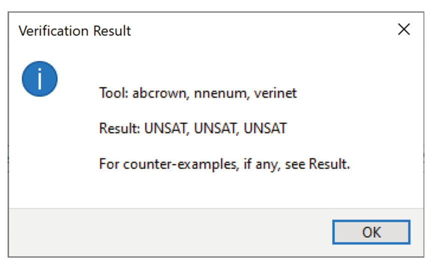


Verification Result

Example Use Cases

LHC Cooling Tower System

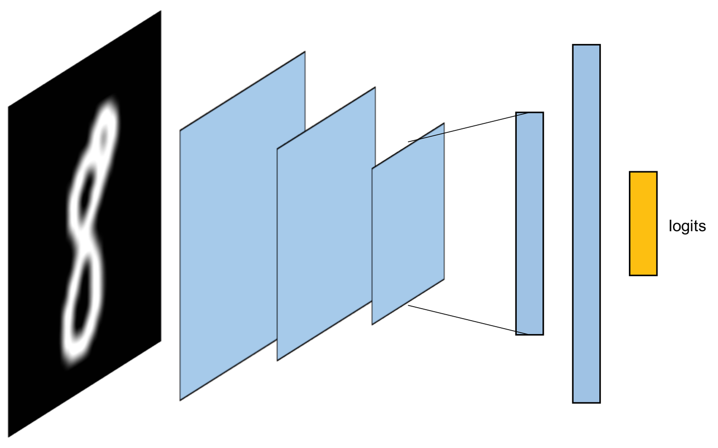


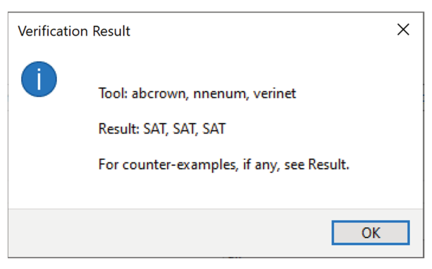


Reachability

$f: R^3 \to R$; $y = f(x)$
$v_{pre} \equiv \forall i \in \{0,1,2\}: x_{li} \le x_i \le x_{ui}$
$v_{post} \equiv 0.0 \le y \le 0.2$
$H \equiv \exists x: v_{pre} \to v_{post}$

MNIST Classification





L_∞-Norm Robustness

$f: [0,1]^{784} \to R^{10}$; $y = f(x)$; $\text{argmax}_i\,(f(x')_i) = 8$
$v_{pre} \equiv x' - \epsilon \le x \le x' + \epsilon$
$v_{post} \equiv \forall j \in \{0,1,2,\ldots,9\} - \{8\}: y'_8 \ge y'_j$
$H \equiv \forall x: v_{pre} \to v_{post} = \sim(\exists x: v_{pre} \wedge \sim v_{post})$

Limitations

- Specification Expressiveness
  Temporal Logics, Nested Conjunctions
- Architecture Support
  e.g. RNN, LSTM, Transformer

Future Work

- Apply these verification tools to other NNs deployed at CERN
- Discover more techniques to express safety properties as specifications
- Explore verification techniques for more complex network structures such as RNNs, transformers, etc.

References

[1] Weng, Lily, Huan Zhang, Hongge Chen, Zhao Song, Cho-Jui Hsieh, Luca Daniel, Duane Boning, and Inderjit Dhillon. "Towards fast computation of certified robustness for relu networks." In *International Conference on Machine Learning*, pp. 5276-5285. PMLR, 2018.

[2] Müller, Mark Niklas, Christopher Brix, Stanley Bak, Changliu Liu, and Taylor T. Johnson. "The third international verification of neural networks competition (VNN-COMP 2022): summary and results." *arXiv preprint arXiv:2212.10376* (2022).

CERN Beams Department
Industrial Controls and Safety Systems Group (ICS)

CERN Summer Student Poster Session 2023
25 July 2023