Unsupervised Neural Network Verification – Benedikt Böing

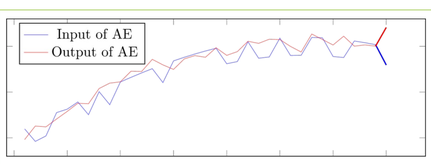
Worst-Case-Error Verification of Autoencoders

Challenges:

- Define unsupervised verification problem
- Encode problems for SMT solvers

Contributions:

- Base verification problem on loss function
- Define unsupervised adversarial attacks based on autoencoder loss function
- Show use-cases of verifying autoencoders



Let $f : \mathbb{R}^n \to \mathbb{R}^n, A \subset \mathbb{R}^n, dist : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}_+$

- $x \in \mathbb{R}^n$ is ε-adversarial attack if $dist(x, f(x)) > \varepsilon$
- $wce(f, A) := \sup\{dist(x, f(x)) : x \in A\}$

Figure 1: Definition of Unsupervised Adversarial Attacks

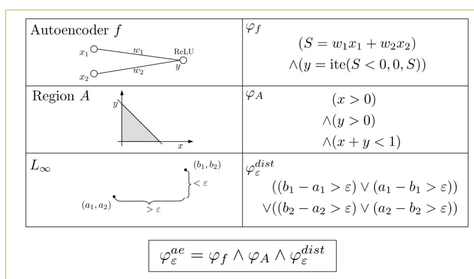


Figure 2: Encoding of Autoencoder for SMT Solvers

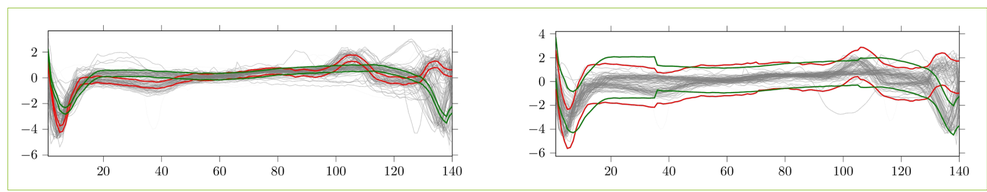


Figure 3: Bounded Image Spaces of Autoencoder

Böing et al. Quality Guarantees for Autoencoders via Unsupervised Adversarial Attacks [ECML 2020]

Training Autoencoders for Robustness and Verification Scalability

Challenges:

- Verification of neural networks is slow
- Autoencoders show non-robust behaviour

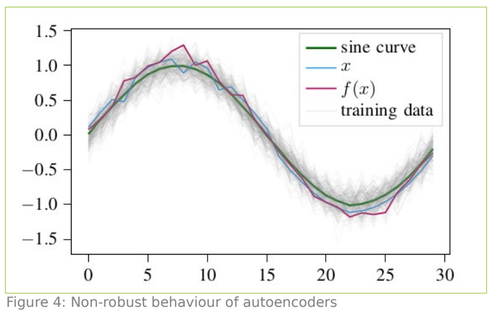


Figure 4: Non-robust behaviour of autoencoders

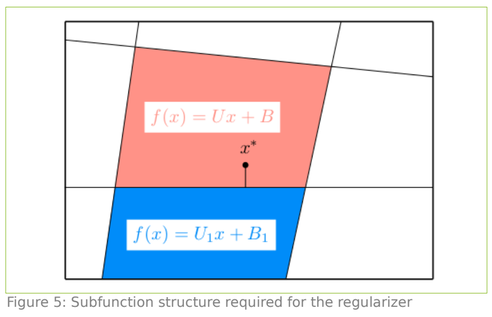


Figure 5: Subfunction structure required for the regularizer

Contributions:

- New regularizer for autoencoder training
- Decrease of affine subfunctions leading to faster verification and more robustness

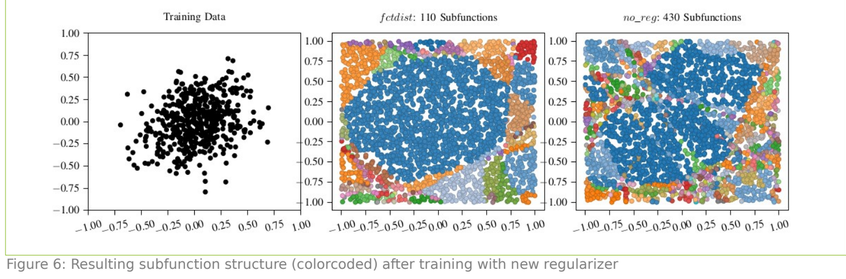


Figure 6: Resulting subfunction structure (colorcoded) after training with new regularizer

Böing et al. On Training and Verifying Robust Autoencoders [DSAA 2022]

Post-Robustifying a given Anomaly Detection Ensemble

Challenges:

- Adapt existing model for robustness
- Scalability of neural network verification

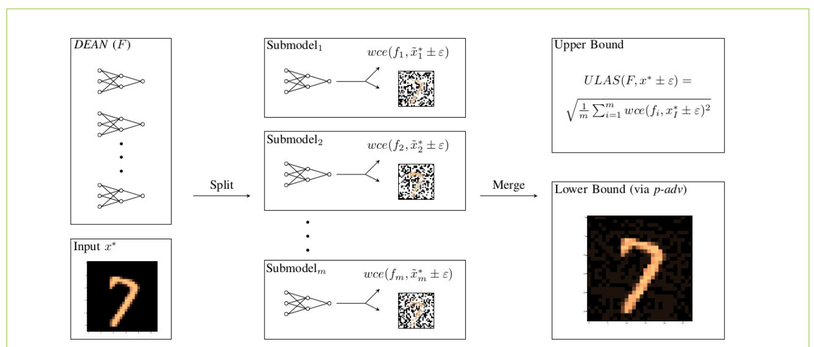


Figure 7: Approach for verification of anomaly detection ensemble *DEAN*

Contributions:

- Robustify a given ensemble method as post-processing step by model selection
- Divide-and-Conquer approach exploiting ensemble properties for verification

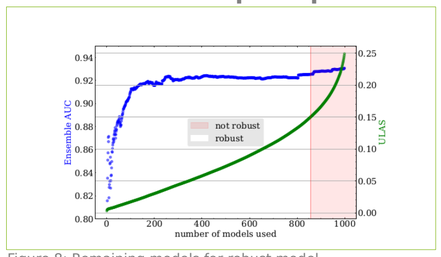


Figure 8: Remaining models for robust model

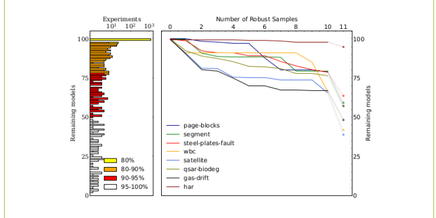


Figure 9: Remaining models for multiple samples

Böing et al. Post-Robustifying Deep Anomaly Detection Ensembles by Model Selection [ICDM 2022]

Literature:

- Ehlers. Formal Verification of Piece-Wise Linear Feed-Forward Neural Networks [ATVA 2017]
- Szegedy et al. Intriguing Properties of Neural Networks [ICLR 2014]
- Katz et al. Reluplex: An Efficient SMT Solver for Verifying Deep Neural Networks
- Albarghouthi. Introduction to Neural Network Verification

Technical University of Dortmund
Chair of Data Science and Data Engineering
Prof. Dr. Emmanuel Müller

E-Mail: emmanuel.mueller@cs.tu-dortmund.de
Tel.: +49 231 755 7722
Web: https://ls9-www.cs.tu-dortmund.de/

technische universität dortmund

Chair of Data Science and Data Engineering
Prof. Dr. Emmanuel Müller