Towards Certification of Deep Learning: Falsification, Explanation, Verification, Enhancement, and Reliability

Xiaowei Huang
University of Liverpool, UK

ICFEM2022, 26th October, 2022

Outline

Safety Properties

Certification Framework – F.E.V.E.R.

- Falsification
- Explanation
- Verification
- Enhancement
- Reliability

Conclusions

Safety Properties

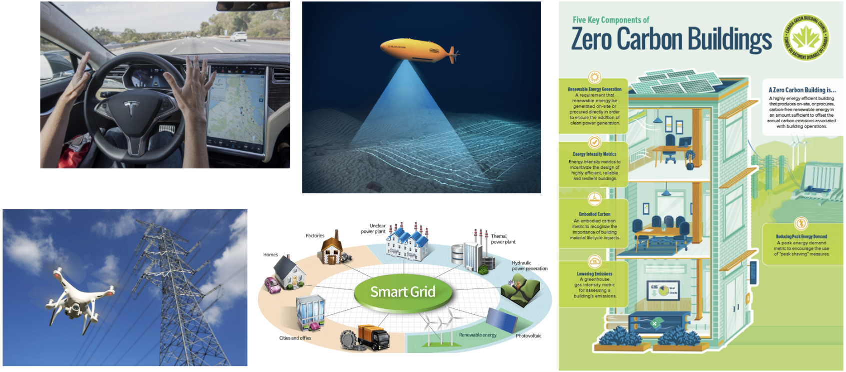


Figure: Driverless Car [18], Autonomous Underwater Vehicles [19], Drone for inspection [17], Smart Grid [2], Net-zero buildling [1], etc.

nature

Explore content ⌄ About the journal ⌄ Publish with us ⌄

nature > articles > article

Article | Open Access | Published: 15 July 2021

Highly accurate protein structure prediction with AlphaFold

John Jumper, Richard Evans, ... Demis Hassabis + Show authors

Nature **596**, 583–589 (2021) | Cite this article

577k Accesses | **1061** Citations | **2997** Altmetric | Metrics

nature International weekly journal of science

Home | News | Research | Careers & Jobs | Current Issue | Archive | Audio & Video | For Authors

Archive > Volume 542 > Issue 7639 > Letters > Article > Article metrics > News

Article metrics for:

Dermatologist-level classification of skin cancer with deep neural networks

Andre Esteva, Brett Kuprel, Roberto A. Novoa, Justin Ko, Susan M. Swetter, Helen M. Blau & Sebastian Thrun

- Drug Discovery and Development
- Automatic Medical Diagnosis

nature reviews drug discovery

Review Article | Published: 11 April 2019

Applications of machine learning in drug discovery and development

Jessica Vamathevan, Dominic Clark, Paul Czodrowski, Ian Dunham, Edgardo Ferran,

Figure: Original detected tracks

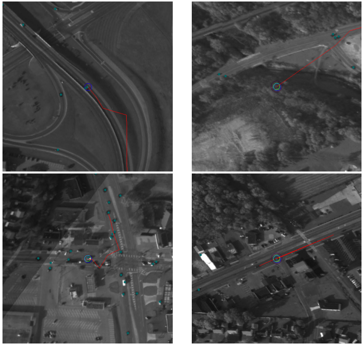
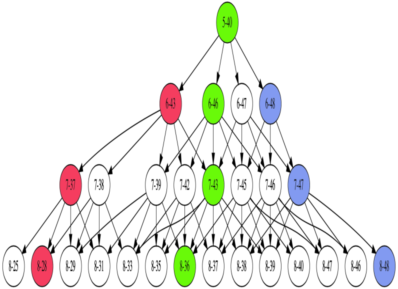


Figure: Distorted tracks

[25] Reliability Validation of Learning Enabled Vehicle Tracking. ICRA2020

(a) Heuristic search

(b) Verification



(c) Enumeration of all possible Tracks

[9] *Practical Verification of Neural Network Enabled State Estimation System for Robotics. IROS2020.*

Trustworthiness = Certification + Explanation

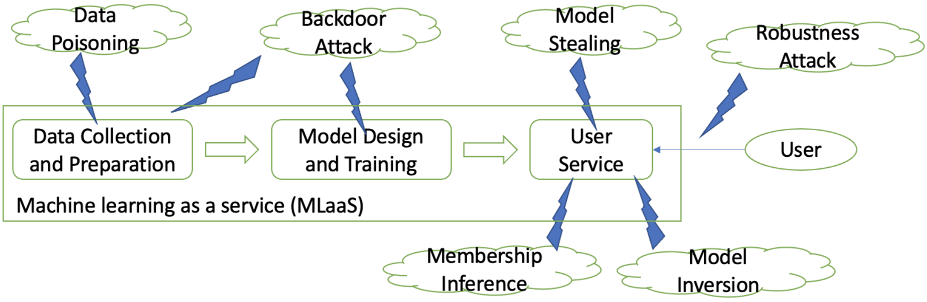
- Certification can be property-based, considering safety and security properties.

[11]: A Survey of Safety and Trustworthiness of Deep Neural Networks: Verification, Testing, Adversarial Attack and Defence, and Interpretability, Computer Science Review. 37 (2020): 100270.

Properties

1. Generalisation
2. Uncertainty
3. Robustness
4. Data Poisoning
5. Backdoor
6. Model Stealing
7. Membership Inference
8. Model Inversion
9. etc

Attack in ML Development Cycle



[10] Machine Learning Safety. Springer, 2022.

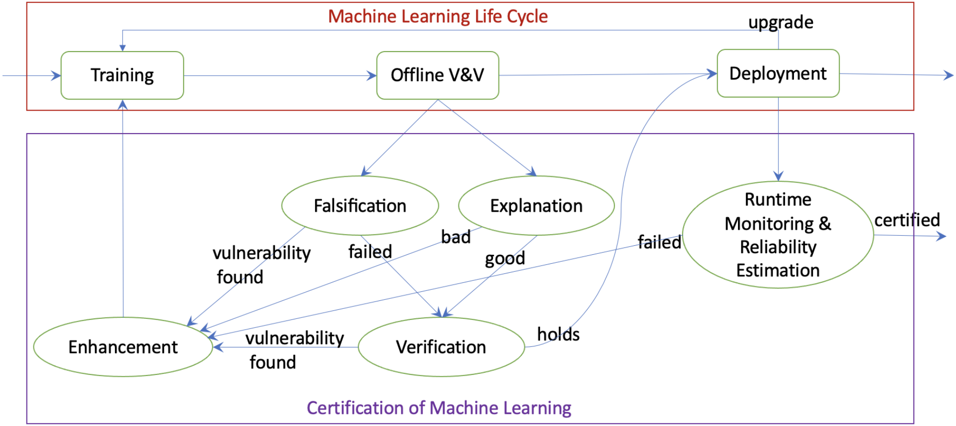
Certification Framework – F.E.V.E.R.

Assurance is a description of what high-quality software *development processes* should be put in-place to create (safety-critical) software that performs its desired function.

If *life cycle evidence* can be produced to demonstrate that these processes have been correctly and appropriately implemented, then such software should be assured.

leads to software standards such as

- DO-178B/C, Software Considerations in Airborne Systems and Equipment Certification
- ISO 26262: standards for the functional safety of road vehicles



[11] A survey of safety and trustworthiness of deep neural networks: Verification, testing, adversarial attack and defence, and interpretability. Computer Science Survey, 2020

Trend of relevant research

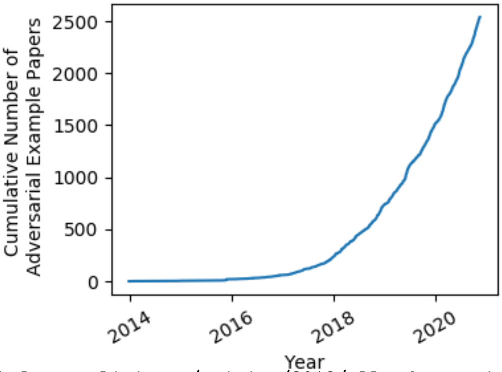


Figure: `https://nicholas.carlini.com/writing/2019/all-adversarial-example-papers.html`

Comprehensive ones: *1. A Survey of Safety and Trustworthiness of Deep Neural Networks: Verification, Testing, Adversarial Attack and Defence, and Interpretability, Computer Science Review. 37 (2020): 100270. [11]*; *2. Machine Learning Safety. Springer, 2022.* [10]

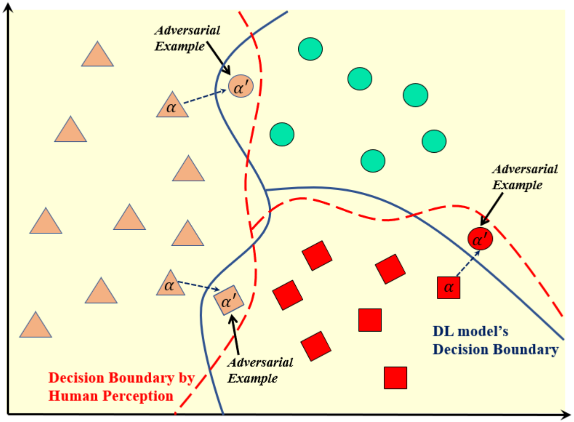
UNIVERSITY OF LIVERPOOL

Falsification

Falsification aims to find evidence to demonstrate the weaknesses of a trained machine learning model or a machine learning training process. Popular techniques include

- adversarial attack
- testing
- Monte Carlo sampling based methods,
- genetic algorithm based methods,
- etc

Robustness Error



DL model: classifies α and α' **differently**

Human: should remain the **same**

Limited-Memory BFGS Attack (L-BFGS)

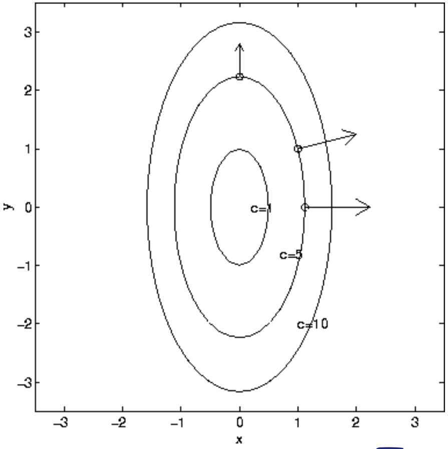
For robustness, one of earliest adversarial attack : optimization based formulation with L_2-norm metric

- Model $f : R^{s_1} \to \{1 \dots s_K\}$ with s_K labels
- $x \in R^{s_1} = [0,1]^{s_1}$ is an input
- $t \in \{1 \dots s_K\}$ is a target misclassification label

Find the adversarial perturbation r via

$$\begin{aligned} & \min ||r||_2 && \text{assure human-decision unchanged} \\ \textit{s.t.} \quad & \arg\max_l f_l(x+r) = t && \text{assure misclassification} \\ & x + r \in R^{s_1} && \text{assure perturbed image feasible} \end{aligned} \tag{1}$$

The gradient vector $\nabla f(x, y)$ points in the direction of greatest rate of increase of $f(x, y)$



FGSM Attack

Fast Gradient Sign Method is able to find adversarial perturbations with a fixed L_∞-norm constraint **very efficiently**

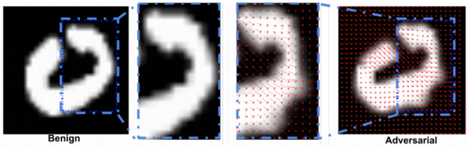
- θ: the model parameters,
- x, y: the input and the label
- $J(\theta, x, y)$: the loss function

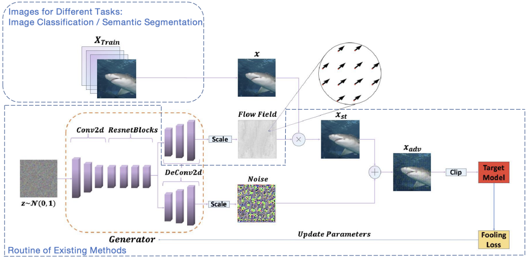
Find adversarial perturbation r by linearizing the loss function around the current value of θ,

$$r = \epsilon \, \text{sign} \left(\nabla_x J(\theta, x, y) \right) \tag{2}$$

- A **one-step modification** to all pixel values to increase the loss function with a L_∞-norm constraint ϵ

Universal Attack on Both Additive and Nonaddictive Noise



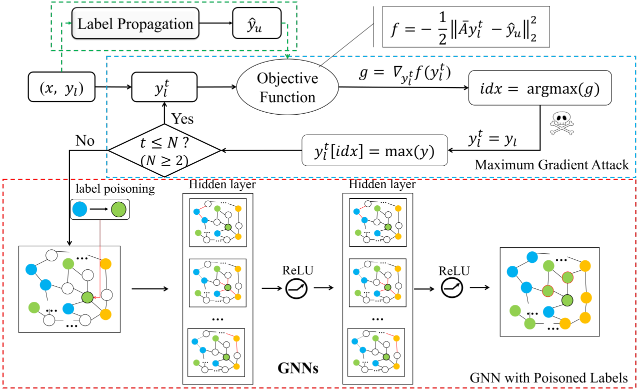


- Instead of perturbing the pixel values, adversarial attacks can be achieved by **spatial transformation** – on MNIST: digit "0" is misclassified as "2" (left figure)
- Different metric is required to measure pixel's **spatial displacement**
- Perturb spatial location and values of pixels simultaneously on **a set of images**?

[24] Generalizing Universal Adversarial Perturbations for DNNs. ICDM2020

Label Poisoning Attack on Graph Neural Networks

1. label propagation to generate predictive labels
2. maximum gradient attack to poison data labels
3. GNN training with poisoned labels



---

[16] Adversarial Label Poisoning Attack on Graph Neural Networks via Label Propagation. ECCV2022

- Well established in many industrial standard for software used in safety critical systems, such as ISO26262 for automotive systems and DO 178B/C for avionic systems.

- Coverage-guided testing
  - (step 1) generate as many as possible the test cases according to the structural information of the model, and
  - (step 2) use the test cases to evaluate if the model performs well with respect to certain properties

- Coverage Metrics
  - Structural Coverage, e.g., MC/DC coverage metrics [23] (Core idea: not only the presence of a feature needs to be tested but also the causal effects of less complex features on a more complex feature must be tested.)
  - Scenario Coverage
- Test Case Generation Methods
  - Fuzzing
  - Symbolic/Concolic execution [24], etc

- check DeepConcolic: `https://github.com/TrustAI/DeepConcolic`

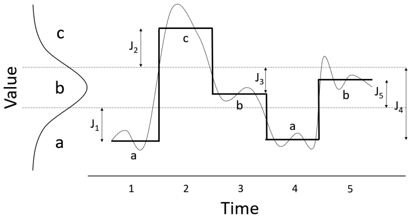
---

[23] Structural Test Coverage Criteria for Deep Neural Networks. ICSE2019

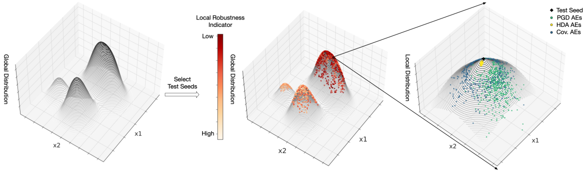
[24] Concolic Testing for Deep Neural Networks. ASE2018

More Advanced Testing

Coverage-Guided Testing for Recurrent Neural Networks [6]



Hierarchical Distribution-Aware Testing of Deep Learning [7]



[6] Coverage-Guided Testing for Recurrent Neural Networks. IEEE trans. on Reliability, 2021

[7] Hierarchical Distribution-Aware Testing of Deep Learning. ArXiv, 2022

Explanation

The black-box nature of deep neural networks (DNNs) makes it impossible to understand why a particular output is produced, creating demand for “Explainable AI”.

(a) ‘cowboy hat’ (b) ‘dog’ (c) ‘numbfish’ (d) ‘sheep’

Figure: Input images and explanations from for Xception (red labels highlight misclassification or counter-intuitive explanations) [22]

For certification, we need not only correct classification but also correct explanation.

[22] Explaining Image Classifiers using Statistical Fault Localization. ECCV2020

Adopting the definition of explanations by Halpern and Pearl, which is based on their definition of actual causality. What we required:

1. an explanation is a *sufficient* cause of the outcome;
2. an explanation is a *minimal* such cause (that is, it does not contain irrelevant or redundant elements);
3. an explanation is *not obvious*; in other words, before being given the explanation, the user could conceivably imagine other explanations for the outcome.

What we propose:

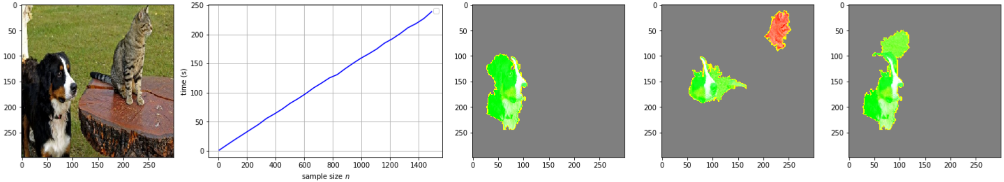
- SFL (stochastic fault localisation) measures to rank the set of pixels of x by slightly abusing the notions of passing and failing tests

---

[22] Explaining Image Classifiers using Statistical Fault Localization. ECCV2020

Utilising Bayesian variant to deal with

- consistency in repeated explanations of a single prediction (as shown below, with LIME, different explanations can be generated for the same prediction)



(A) (B) (C) (D) (E)

- explanation fidelity
- robustness to kernel settings

[31] BayLIME: Bayesian Local Interpretable Model-Agnostic Explanations. UAI2021

SAFARI: Robustness ∧ Interpretability

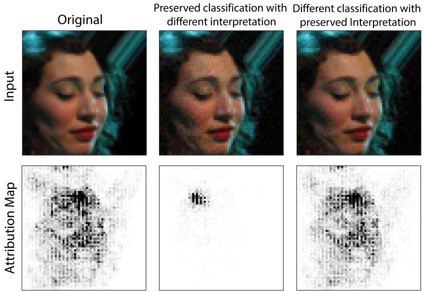


Figure: Two types of misinterpretations after perturbation

Novel black-box evaluation methods:

- based on Genetic Algorithm
- for both *worst-case* and *overall* robustness of explanations
- new interpretation Discrepancy Metrics

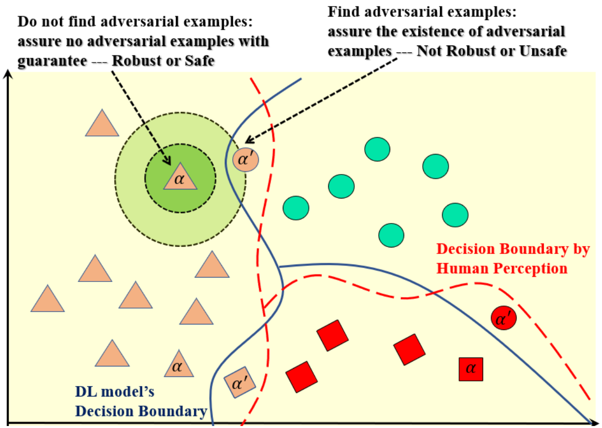
---

[8] SAFARI: Versatile and Efficient Evaluations for Robustness of Interpretability. ArXiv, 2022.

Verification

Verification aims to determine if a model satisfies certain properties. Popular techniques include

- reduction to constraint solving
- over-approximation
- global optimisation based methods
- statistical evaluation
- coverage-guided testing
- etc



(Robustness) Verification: verify if a certain input area can exclude misclasssification with **guarantees**

Verification by Reduction to Constraint Solving

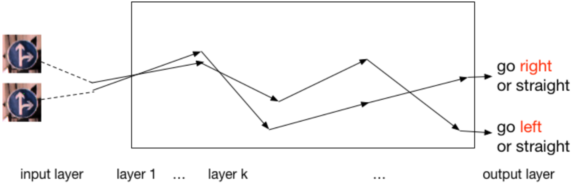
- (step 1) encode the entire network
- (step 2) encode the robustness constraint over the input
- (step 3) compute the result by solving the constraints

- encode the network
- Let $\vec{t}_{i+1}$ have value 0 or 1 in its entries and have the same dimension as $\vec{v}_{i+1}$, and M be a very large constant number that can be treated as ∞.
- we have the following MILP constraints for every layer $i = 1..K-2$

$$
\begin{aligned}
\vec{v}_{i+1} &\geq \mathbf{W}_i\vec{v}_i + \vec{b}_i, \\
\vec{v}_{i+1} &\leq \mathbf{W}_i\vec{v}_i + \vec{b}_i + M\vec{t}_{i+1}, \\
\vec{v}_{i+1} &\geq \mathbf{0}, \\
\vec{v}_{i+1} &\leq M(1 - \vec{t}_{i+1}),
\end{aligned}
\tag{3}
$$

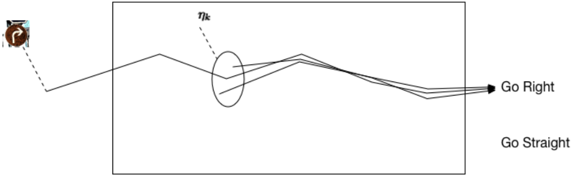
Verification by Region Propagation

How does neural network process (two very similar) inputs?



How does verification work?

A layer-by-layer explicit search with SMT solver



[12] Safety verification of deep neural networks. CAV2017

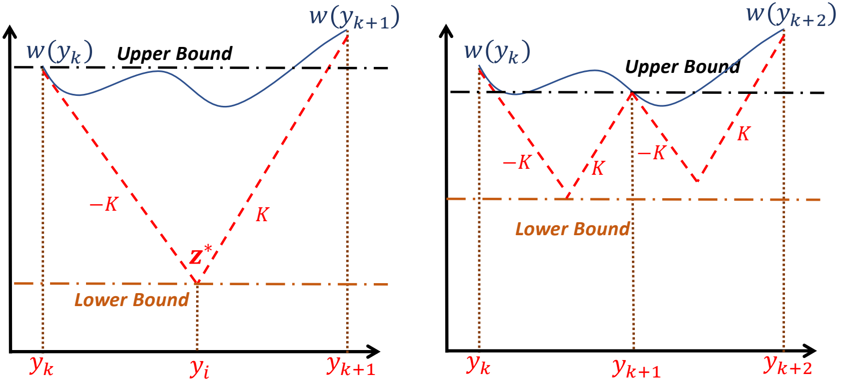


Figure: A lower-bound function designed via Lipschitz constant

[20] Reachability Analysis of Deep Neural Networks with Provable Guarantees. IJCAI2018

Robustness Verification by Other Methods

- Reduction to Monte-Carlo Tree Based Search
- Reduction to Other Global Optimisation Method
- Reduction to Two-player Game

---

[26] Feature-guided black-box safety testing of deep neural networks. TACAS2018.

[21] Global robustness evaluation of deep neural networks with provable guarantees for the Hamming distance. IJCAI2019

[27] A game-based approximate verification of deep neural networks with provable guarantees. Theoretical Computer Science, 2020.

- Scalability
- Mostly work with Robustness
- Can only deal with deterministic variables/neurons, but machine learning problems are mostly statistical ...

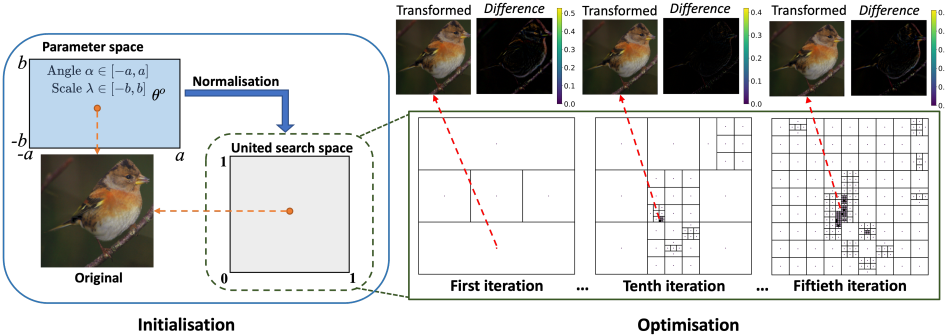


Figure: After normalising the parameter space to a unit search space, GeoRobust performs a sequence of space divisions to find the global worst-case transformation.

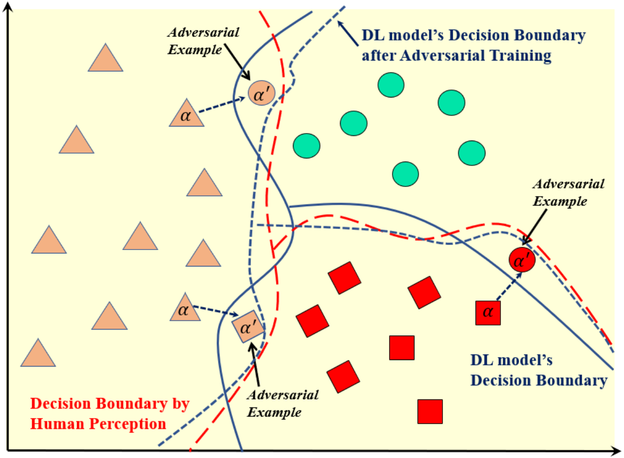
[3] Towards Verifying the Geometric Robustness of Large-scale Neural Networks. ArXiv. 2022

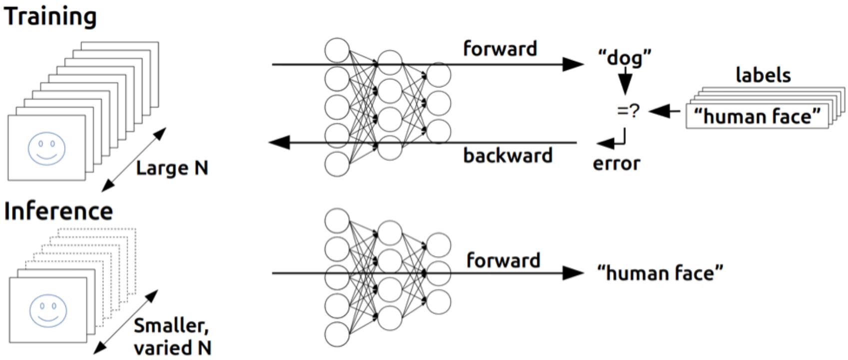
Enhancement through Rectification

Rectification aims to enhance the machine learning training process or the trained machine learning model, so that the resulting machine learning model performs better with respect to the properties. Popular techniques include

- adversarial training
- regularisation
- outlier detection
- randomisation (based on differential privacy)
- etc

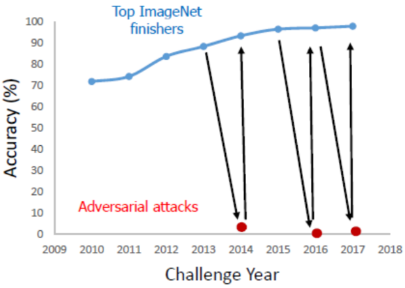
Model Improvement for Robustness




Training
forward
"dog"
labels
=?
"human face"
backward
error
Large N
Inference
forward
"human face"
Smaller,
varied N

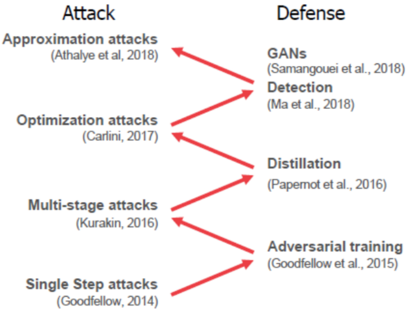
Attack vs. Defence: An Endless Game

Adversarial attacks cause a catastrophic reduction in ML capability



ImageNet Classification

Many defenses have been tried and failed to generalize to new attacks



Attack / Defense Cycle

@ DARPA's GARD programme

Consider weight correlation during the training

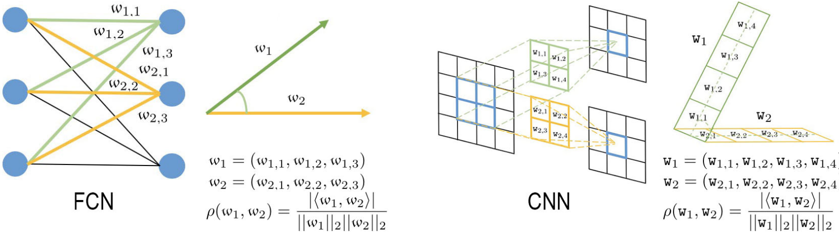


Figure: For fully connected networks, the weight correlation of any two neurons is the cosine similarity of the associated weight vectors. For convolutional neural networks, the weight correlation of any two filters is the cosine similarity of the reshaped filter matrices.

[15] How does Weight Correlation Affect Generalisation Ability of DNNs? NeurIPS2020

PAC-Bayes Bound

(McAllester, 1999) considers a generalization bound on the parameters

Posteriori distribution Q on parameters Θ

Priori distribution P on parameters Θ

$$\mathbb{E}_{\Theta\sim Q}[\mathcal{L}_D(f_\Theta)] \leq \mathbb{E}_{\Theta\sim Q}[\mathcal{L}_S(f_\Theta)] + \sqrt{\frac{\mathrm{KL}(Q||P) + \log\frac{m}{\delta}}{2(m-1)}}$$

Likelihood δ

Expected loss on input space D

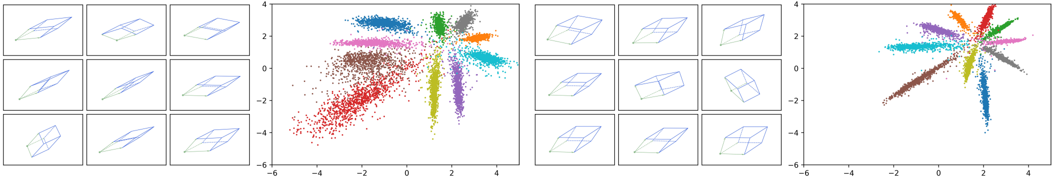
Expected loss on samples S from D

Number of samples

KL divergence plays a key role in the generalization bound

- a small KL term will help tighten the bound
- a larger KL term will loose the bound

[14] How does Weight Correlation Affect Generalisation Ability of DNNs? NeurIPS2020



(a) Normal CNN weight volume: 0.1592 (b) Normal CNN feature visualization (c) Dropout CNN weight volume: 0.3239 (d) Dropout CNN feature visualization

Figure: Visualization of weight volume and features of the last layer in a CNN on MNIST, with and without dropout during training

---

[15] Weight Expansion: A New Perspective on Dropout and Generalization. Transactions on Machine Learning Research. 2022

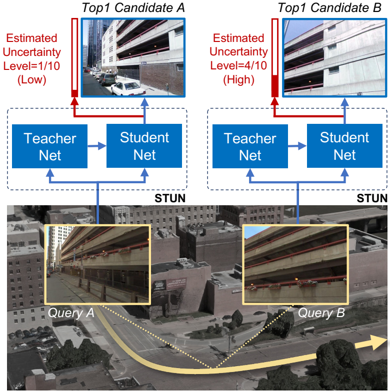
- treating model weights as random variables allows for enhancing adversarial training through **S**econd-Order **S**tatistics **O**ptimization (S^2O) with respect to the weights
- derive an improved PAC-Bayesian adversarial generalization bound, which suggests that optimizing second-order statistics of weights can effectively tighten the bound.
- thhrough experiments, we show that S^2O not only improves the robustness and generalization of the trained neural networks when used in isolation, but also integrates easily in state-of-the-art adversarial training techniques like TRADES, AWP, MART, and AVMixup, leading to a measurable improvement of these techniques.

---

[13] Enhancing Adversarial Training with Second-Order Statistics of Weights. CVPR2022

1. train a teacher net
2. supervised by the pretrained teacher net, a student net with an additional variance branch is trained
3. During the online inference phase, we only use the student net to generate both a place prediction and the uncertainty

This can not only generate uncertainty for each prediction but also improve the accuracy (i.e., generalisation).



[4] *STUN: Self-Teaching Uncertainty Estimation for Place Recognition. IROS2022*

Safety assurance via a reliability assessment model.

- Safety assurance: processes that function systematically to ensure the performance and effectiveness of safety risk controls and that the organization meets or exceeds its safety objectives through the collection, analysis, and assessment of information
- Software reliability: the probability of failure-free software operation for a specified period of time in a specified environment

Approach: a reliability assessment model to construct probabilistic safety argument by deriving reliability requirements from low-level ML functionalities

A RAM built upon statistical testing evidence, while inspired by conventional partition-based testing and operational profile (OP)-based testing

$$\textbf{Reliability} = \textbf{Generalisation} \times \textbf{Local Robustness/Safety/Security/...} \tag{4}$$

Specifically,

$$\lambda := \int_{x \in \mathbb{R}^{s_1}} I_{\{x \text{ causes a misclassification}\}}(x) \mathrm{Op}(x) \, \mathrm{d}x \,, \tag{5}$$

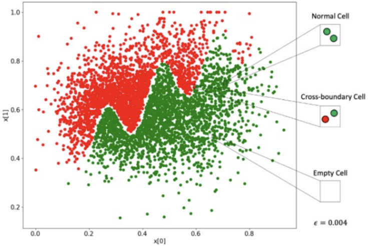
where x is an input in the input domain $\mathbb{R}^{s_1}$, and $I_{\mathtt{S}}(x)$ is an indicator function—it is equal to 1 when $\mathtt{S}$ is true and equal to 0 otherwise. The function $\mathrm{Op}(x)$ returns the probability that x is the next random input.

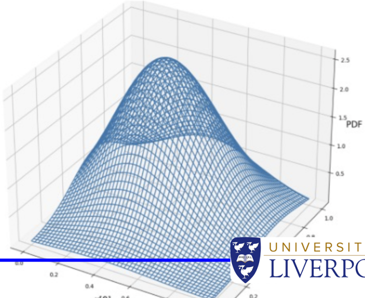
[28] A safety framework for critical systems utilising deep neural networks. SafeCOMP2020.
[29] Assessing Reliability of Deep Learning Through Robustness Evaluation and Operational Testing. AISafety2021

- Partition the input space into "cells", with the guidance of r-separation
- Approximation the operational profile OP
- Cell robustness evaluation
- "Assemble" cell-wise estimates for reliability

$$\lambda = \sum_{i=1}^{m} Op_i \lambda_i \tag{6}$$

Then we can have the mean and variance of λ



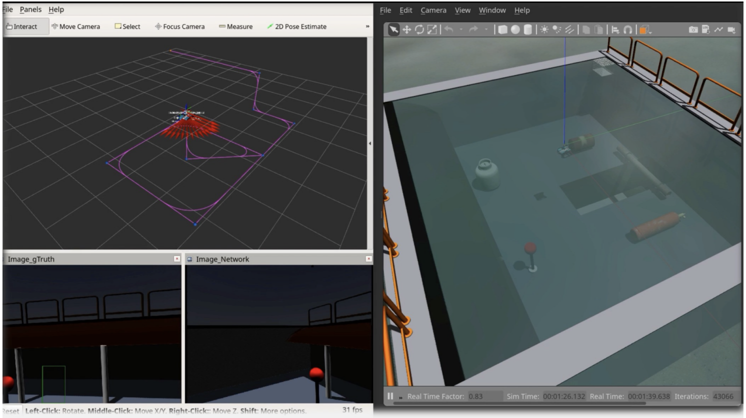


---

[30] Reliability Assessment and Safety Arguments for Machine Learning Components in Assuring Learning-Enabled Autonomous Systems. ACM Trans. Embedded Computing Systems, 2022.

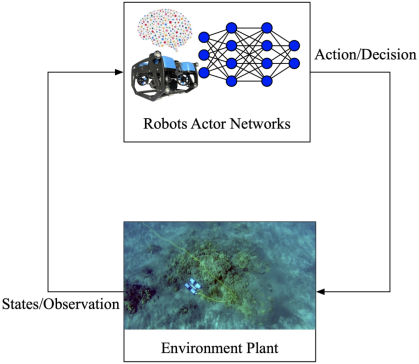
Autonomous Underwater Vehicle (AUV) Case Study

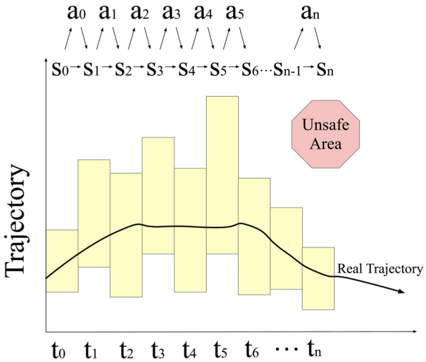
- An autonomous inspection/survey mission with several waypoints and docking
- 6 simulated objects per mission: pipe, barrel, dock-cage, etc
- the mission is subject to dynamic noise factors



[30] Reliability Assessment and Safety Arguments for Machine Learning Components in System Assurance. ACM Trans. Embedded Computing Systems, 2022.

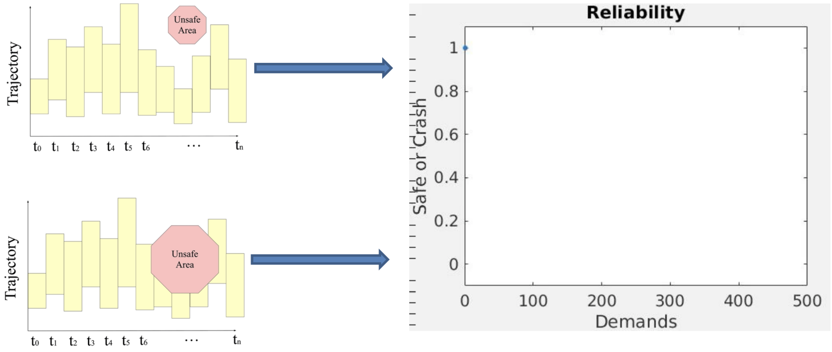
RAM for Motion Planning





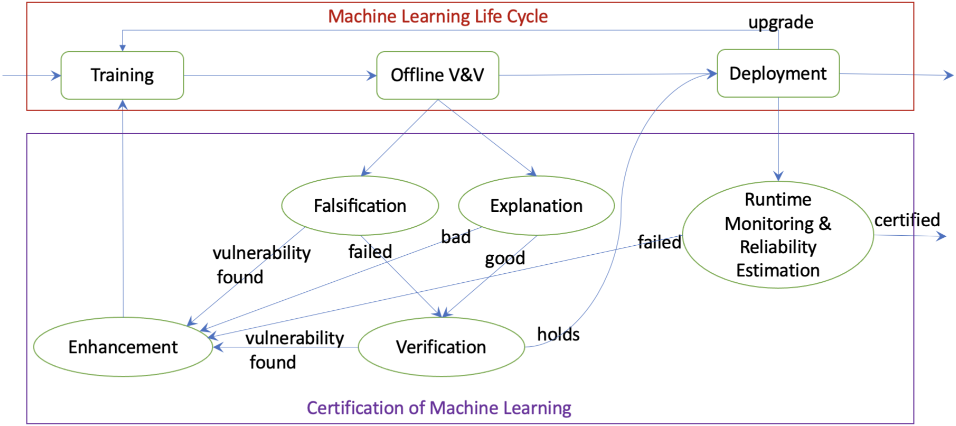
[5] Dependability Analysis of Deep Reinforcement Learning based Robotics and Autonomous Systems. IROS2022.

RAM for Motion Planning



---

[5] Dependability Analysis of Deep Reinforcement Learning based Robotics and Autonomous Systems. IROS2022.



Machine Learning Life Cycle
upgrade
Training
Offline V&V
Deployment
Falsification
Explanation
Runtime Monitoring & Reliability Estimation
certified
vulnerability found
failed
bad
good
failed
Enhancement
vulnerability found
Verification
holds
Certification of Machine Learning

Conclusions

- There is no single tool/method that can work for the certification of deep learning
- None of the F.E.V.E.R. has been sophisticated – many to be done for not only each analysis technique but also the interfacing between them
- More than one properties to work with – probably an expressive formal language will help.

FOCETA

Thank you

http://www.foceta-project.eu/

https://www.linkedin.com/company/foceta-project

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 956123.

UNIVERSITY OF LIVERPOOL

91.67%

References I

Getting serious about net zero buildings.

Smart grid.

Towards verifying the geometric robustness of large-scale neural networks.
In *ArXiv*, 2022.

K. Cai, C. X. Lu, and X. Huang.
Stun: Self-teaching uncertainty estimation for place recognition.
In *IROS2022*, 2022.

Y. Dong, X. Zhao, and X. Huang.
Dependability analysis of deep reinforcement learning based robotics and autonomous systems.
In *IROS2022*, 2022.

W. Huang, Y. Sun, X. Zhao, J. Sharp, W. Ruan, J. Meng, and X. Huang.
Coverage-guided testing for recurrent neural networks.
IEEE Transactions on Reliability, pages 1–16, 2021.

W. Huang, X. Zhao, A. Banks, V. Cox, and X. Huang.
Hierarchical distribution-aware testing of deep learning, 2022.

W. Huang, X. Zhao, G. Jin, and X. Huang.
Safari: Versatile and efficient evaluations for robustness of interpretability, 2022.

References II

W. Huang, Y. Zhou, Y. Sun, J. Sharp, S. Maskell, and X. Huang.
Practical verification of neural network enabled state estimation system for robotics.
In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 7336–7343, 2020.

X. Huang, G. Jin, and W. Ruann.
Machine Learning Safety.
Springer, 2022.

X. Huang, D. Kroening, W. Ruan, J. Sharp, Y. Sun, E. Thamo, M. Wu, and X. Yi.
A survey of safety and trustworthiness of deep neural networks: Verification, testing, adversarial attack and defence, and interpretability.
Computer Science Review, 37:100270, 2020.

X. Huang, M. Kwiatkowska, S. Wang, and M. Wu.
Safety verification of deep neural networks.
In *International Conference on Computer Aided Verification*, pages 3–29. Springer, 2017.

G. Jin, X. Y. annd Wei Huang, S. Schewe, and X. Huang.
Enhancing adversarial training with second-order statistics of weights.
In *CVPR2022*, 2022.

G. Jin, X. Yi, P. Yang, L. Zhang, S. Schewe, and X. Huang.
Weight expansion: A new perspective on dropout and generalization.
Transactions on Machine Learning Research, 2022.

G. Jin, X. Yi, L. Zhang, L. Zhang, S. Schewe, and X. Huang.
How does weight correlation affect the generalisation ability of deep neural networks.
In *NeurIPS'20*, 2020.

G. Liu, X. Yi, and X. Huang.
Adversarial label poisoning attack on graph neural networks via label propagation.
In *ECCV2022*, 2022.

RBR.
Neurala, avisight team up for ai-powered drone inspections, 2020.

Driverless cars: everything you need to know about autonomous car revolution.
https://www.autoexpress.co.uk/car-tech/85183/driverless-cars-everything-you-need-to-know-about-autonomous-car-revolution.
[Online; accessed 11-April-2013].

Geoscience and auv surveys.
https://www.oceaneering.com/survey-and-mapping/geoscience-and-auv-surveys/.
[Online; accessed 11-April-2013].

W. Ruan, X. Huang, and M. Kwiatkowska.
Reachability analysis of deep neural networks with provable guarantees.
In *IJCAI*, pages 2651–2659, 2018.

W. Ruan, M. Wu, Y. Sun, X. Huang, D. Kroening, and M. Kwiatkowska.
Global robustness evaluation of deep neural networks with provable guarantees for the hamming distance.
pages 5944–5952. International Joint Conferences on Artificial Intelligence Organization, 2019.

Y. Sun, H. Chockler, X. Huang, and D. Kroening.
Explaining image classifiers using statistical fault localization.
In *Computer Vision – ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXVIII*, page 391–406, Berlin, Heidelberg, 2020. Springer-Verlag.

Y. Sun, X. Huang, D. Kroening, J. Shap, M. Hill, and R. Ashmore.
Structural test coverage criteria for deep neural networks.
In *ICSE2019*, 2019.

Y. Sun, M. Wu, W. Ruan, X. Huang, M. Kwiatkowska, and D. Kroening.
Concolic testing for deep neural networks.
In *Automated Software Engineering (ASE), 33rd IEEE/ACM International Conference on*, 2018.

Y. Sun, Y. Zhou, S. Maskell, J. Sharp, and X. Huang.
Reliability validation of learning enabled vehicle tracking.
In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 9390–9396, 2020.

M. Wicker, X. Huang, and M. Kwiatkowska.
Feature-guided black-box safety testing of deep neural networks.
In *International Conference on Tools and Algorithms for the Construction and Analysis of Systems*, pages 408–426. Springer, 2018.

M. Wu, M. Wicker, W. Ruan, X. Huang, and M. Kwiatkowska.
A game-based approximate verification of deep neural networks with provable guarantees.
Theoretical Computer Science, 2020.

X. Zhao, A. Banks, J. Sharp, V. Robu, D. Flynn, M. Fisher, and X. Huang.
A safety framework for critical systems utilising deep neural networks.
In *SafeComp2020*, pages 244–259, 2020.

X. Zhao, W. Huang, A. Banks, V. Cox, D. Flynn, S. Schewe, and X. Huang.
Assessing reliability of deep learning through robustness evaluation and operational testing.
In *SafeComp2021*, 2021.

X. Zhao, W. Huang, V. Bharti, Y. Dong, V. Cox, A. Banks, S. Wang, S. Schewe, and X. Huang.
Reliability assessment and safety arguments for machine learning components in assuring learning-enabled autonomous systems.
ACM Transactions on Embedded Computing Systems, 2022.

X. Zhao, W. Huang, X. Huang, V. Robu, and D. Flynn.
Baylime: Bayesian local interpretable model-agnostic explanations.
pages 887–896, 2021.
37th Conference on Uncertainty in Artificial Intelligence 2021, UAI 2021 ; Conference date: 27-07-2021 Through 30-07-2021.