Verification and Validation of Deep Learning

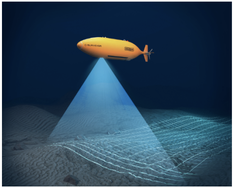
Xiaowei Huang
University of Liverpool, UK

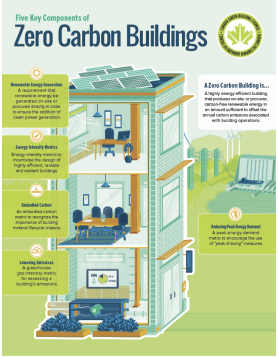
UDRC Workshop, November, 2021

Outline

UNIVERSITY OF LIVERPOOL

Introduction





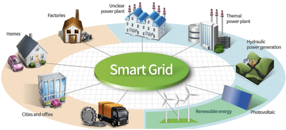


Figure: Driverless Car [6], Autonomous Underwater Vehicles [7], Drone for inspection [5], Smart Grid [2], Net-zero buildling [1], etc.

Can We Trust its Decisions?

Question: Can we really *trust* the decisions made by deep learning models, especially on safety-critical applications?

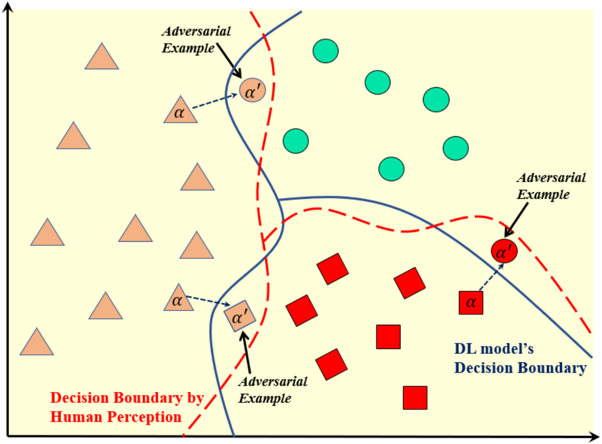
This question can be broken down into a number of more concrete questions, such as

- How does a deep learning model make a decision?
- Does deep learning always make a correct decision?
- Under what circumstances a deep learning model will make a wrong/correct decision?
- ... ...

Vulnerability

- Robustness – local (input-level) safety
  - wrt input perturbation, weight perturbation, etc
- Generalisation – global (model-level) safety
  - wrt different operational environment
- Security
  - wrt data corruption & poisoning, data privacy, etc
- Explainability

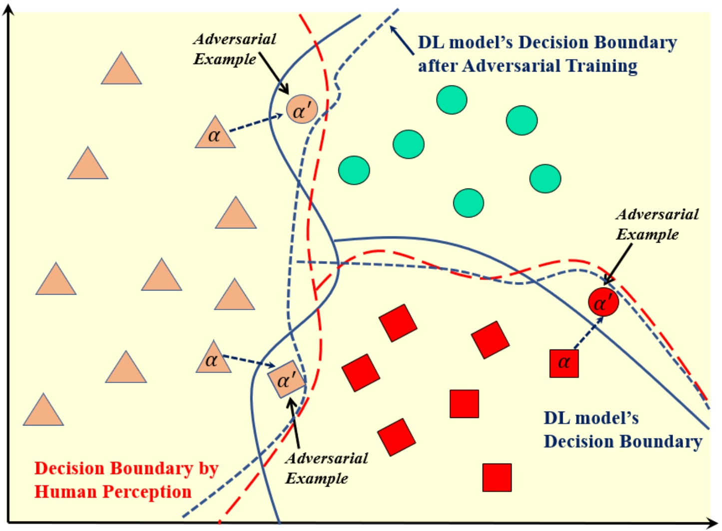
Robustness Error



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

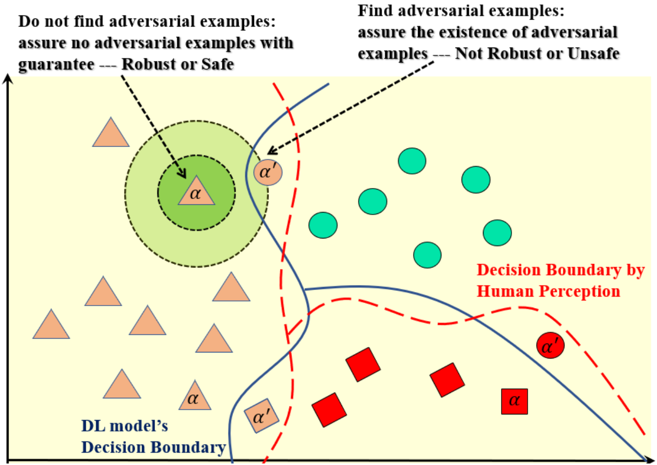
Human: should remain the **same**

Model Improvement for Robustness



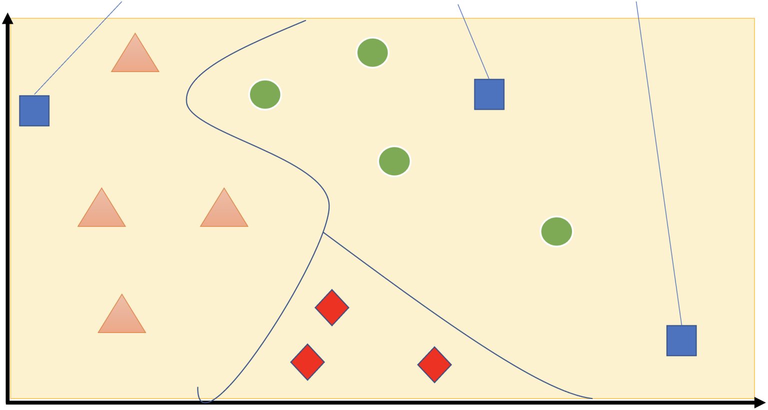
UNIVERSITY OF LIVERPOOL

Verification



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

How about these unseen datapoints which are far away from known data?



Research Dealing with AI Safety

- Direction 1: identify the errors
  - adversarial attack, security attack, etc
- Direction 2: determine if it is without error
  - verification
- Direction 3: reduce errors by improving models
  - adversarial training, adversarial defence, etc
- Direction 4: quantify the errors
  - statistical evaluation, software testing, etc
- Direction 5: demonstrate the safety for development cycle
  - safety assurance, reliability estimation, etc

Trend of relevant research

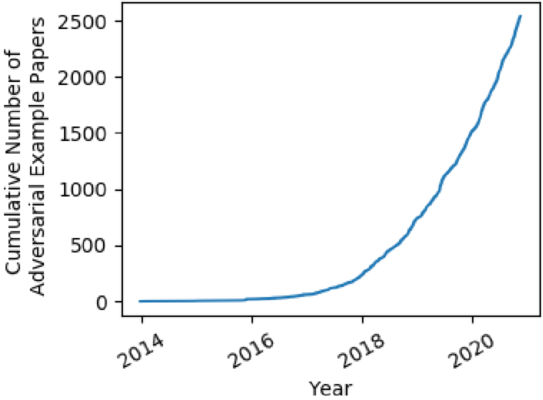
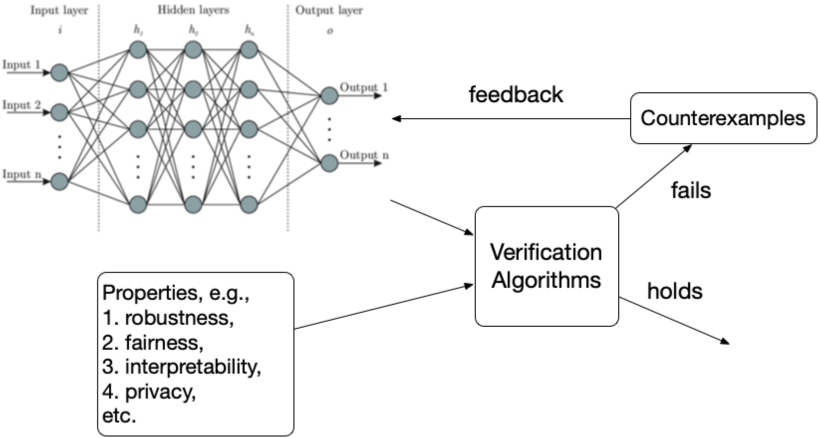


Figure: https://nicholas.carlini.com/writing/2019/all-adversarial-example-papers.html

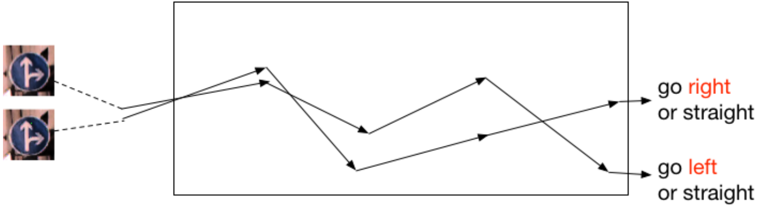
Comprehensive one: *A Survey of Safety and Trustworthiness of Deep Neural Networks: Verification, Testing, Adversarial Attack and Defence, and Interpretability, Computer Science Review. 37 (2020): 100270.* [3]

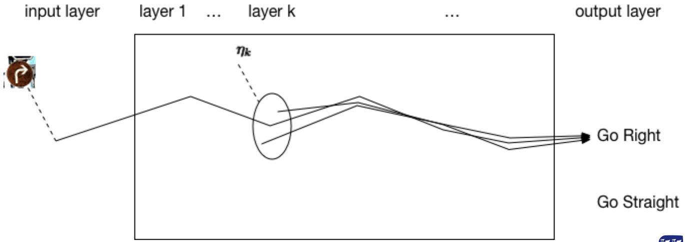
Formal Verification

What is Verification?



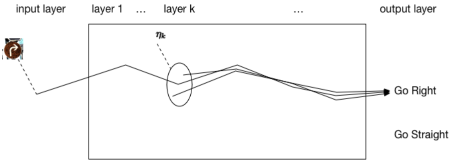
Robustness Verification





Robustness Verification for Neural Network

- Verification Methods
  - Constraint Solving
  - Over-approximation
  - Global optimisation



- What is the difficulty?
  - Scalability
  - Only deal with robustness

Statistical Evaluation

State-of-the-art

- Sampling-based methods
- Software testing methods
- Deep learning theory based methods

May work with **both robustness and generalisation**.

Sampling-based Methods

- Sampling & fitting distributions
- Guarantee from some theories e.g., minimum adversarial distortion follows extreme value distributions [9].

- Enhanced Monte Carlo sampling [8]
- Guarantee from statistics theory

- 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 Metrics
  - structural coverage
  - scenario coverage
- Test Case Generation Methods
  - fuzzing
  - symbolic execution, etc

- to determine if the generated test cases include bugs.

- Industrial standards need to be upgraded

- A few Coverage Metrics
  - DeepXplore: Automated Whitebox Testing of Deep Learning Systems. SOSP 2017
  - Structural Test Coverage Criteria for Deep Neural Networks. EMSOFT 2019
- A few Test Case Generation Methods
  - Concolic Testing for Deep Neural Networks. ASE 2018

- Use a set of generated test cases to either finding bugs or evaluating the performance of a neural network

- How to statistically predict the generalisation capability of a network, according to some structural information such as weights, architectures, etc?

Methods	Generation of input samples?	Utilisation of structural information?
Sampling	Yes	No
Software testing	Yes	Maybe
DL theory	No	Yes

Table: Comparison between statistical evaluation methods

Which DL theory?

- PAC Bayesian Theory, to upper bound the gap between expected loss on input space and expected loss on training samples, by taking into consideration the change of weights before and after the training.
- Vapnik–Chervonenkis (VC) dimension, to measure of the capacity (complexity, expressive power, richness, or flexibility) of a set of functions that can be learned by a statistical binary classification algorithm.
- etc.

PAC Bayes + Weight Correlation [4]

Relax the i.i.d. assumption on the posterior distribution, and define quantities such as weight correlation (WC) based on structural information

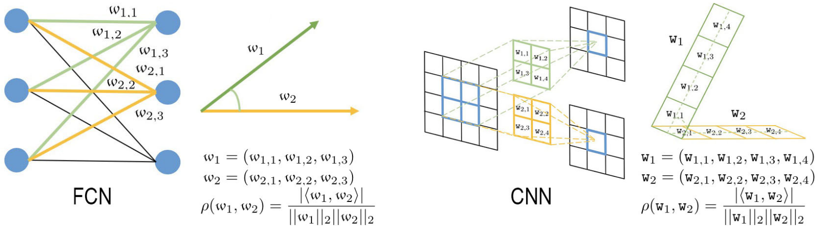


Figure: **(FCN)** The WC of any two neurons is the cosine similarity of the associated weight vectors. **(CNN)** The WC of any two filters is the cosine similarity of the reshaped filter matrices.

PAC Bayes + Weight Correlation

Table 1: Complexity Measures (Measured Quantities)

Generalisation Error (GE)	$\mathcal{L}_D(f_{\theta^F}) - \mathcal{L}_S(f_{\theta^F})$
Product of Frobenius Norms (PFN)	$\prod_\ell \|\theta_\ell^F\|_{\mathrm{Fr}}$
Product of Spectral Norms (PSN)	$\prod_\ell \|\theta_\ell^F\|_2$
Number of Parameters (NoP)	Total number of parameters in the network
Sum of Spectral Norms (SoSP)	Total number of parameters $\times \sum_\ell \|\theta_\ell^0 - \theta_\ell^F\|_2$
Weight Correlation (WC)	$\frac{1}{\ell}\sum_\ell \rho(w_\ell)$
PAC Bayes (PB)	$\sum_\ell \|\theta_\ell^0 - \theta_\ell^F\|_{\mathrm{Fr}}^2 / 2\sigma_\ell^2$
PAC Bayes & Correlation (PBC)	$\sum_\ell (\|\theta_\ell^0 - \theta_\ell^F\|_{\mathrm{Fr}}^2 / 2\sigma_\ell^2 + \mathrm{g}(w_\ell))$

New measure

Table 2: Complexity measures for CIFAR-10

Network	PFN	PSN	NoP	SoSP	PB	PBC	WC	GE
FCN1	8.1e7	1.4e4	3.7e7	1.6e9	1.1e4	1.14e5	0.297	2.056
FCN2	3.3e7	8.5e3	4.2e7	1.61e9	8.8e3	1.24e5	0.296	2.354
VGG11	8.5e10	1.4e5	9.7e6	2.4e8	2.0e3	3.41e4	0.273	0.929
VGG16	5.1e15	1.3e7	1.5e7	5.2e8	2.6e3	3.73e4	0.275	0.553
VGG19	1.1e19	2.9e8	2.1e7	8.1e8	3.3e3	4.26e4	0.274	0.678
ResNet18	2.5e22	1.1e12	1.1e7	8.4e8	4.7e3	1.34e5	0.732	2.681
ResNet34	9.9e34	4.9e16	2.1e7	3.1e9	1.0e4	1.30e5	0.733	2.552
ResNet50	1.4e76	7.5e46	2.3e7	6.1e9	1.6e7	1.62e7	0.278	2.807
DenseNet121	5.9e176	1.4e151	6.8e6	1.5e10	1.0e9	1.04e9	0.357	1.437
Concordant Pairs	21	21	22	26	24	**29**	24	-
Discordant Pairs	15	15	14	10	12	**7**	12	-
Kendall's τ	0.16	0.16	0.22	0.44	0.33	**0.61**	0.33	-

Safety Assurance

Play video demo :
https://www.youtube.com/watch?v=akY8f5sSFpY&t=1s

Conclusions

Conclusions

- Most efforts are taken on finding errors on pre-trained models;
- Adversarial examples are inherent to deep learning, i.e., errors cannot be eliminated;

What shall we do?

- focus on acceptable level of safety;
- consider how the deep learning models are used;
- consider safety assurance on not only the pre-trained models but also the development cycle.

References I

Getting serious about net zero buildings.

Smart grid.

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.

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.

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].

S. Webb, T. Rainforth, Y. W. Teh, and M. P. Kumar.
A statistical approach to assessing neural network robustness.
In *ICLR2019*, 2019.

References II

T.-W. Weng, H. Zhang, P.-Y. Chen, J. Yi, D. Su, Y. Gao, C.-J. Hsieh, and L. Daniel.
Evaluating the Robustness of Neural Networks: An Extreme Value Theory Approach.
In *ICLR2018*, 2018.