VERIFICATION OF NEURAL NETWORKS

MOHAMED AZOUZ MRAD

Software Verification and Validation

Slides based on :

Testing and verification of neural-network-based safety-critical control software: A systematic literature review

DeepSafe: A Data-driven Approach for Checking Adversarial Robustness in Neural Networks

2020/2021

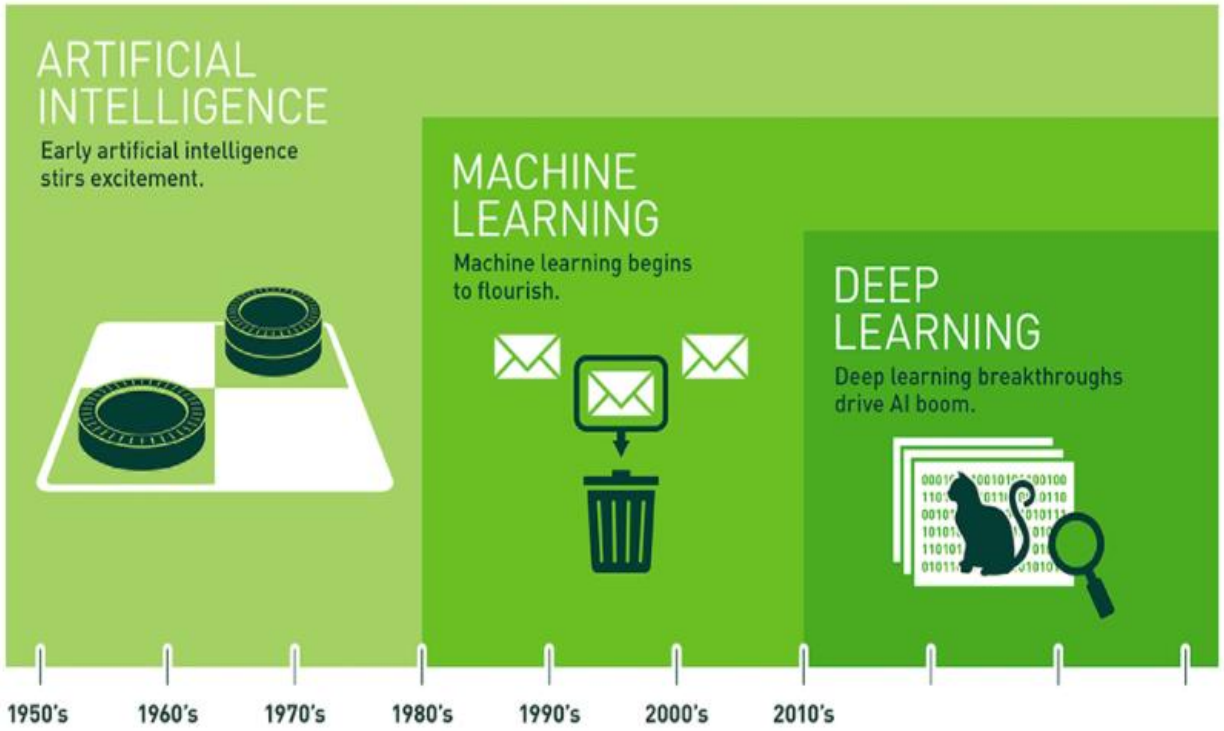
DNN overview

- Machine learning and deep learning
- Neural networks
- Deep neural network
- Application domain

1. DNN overview

Machine learning and deep learning

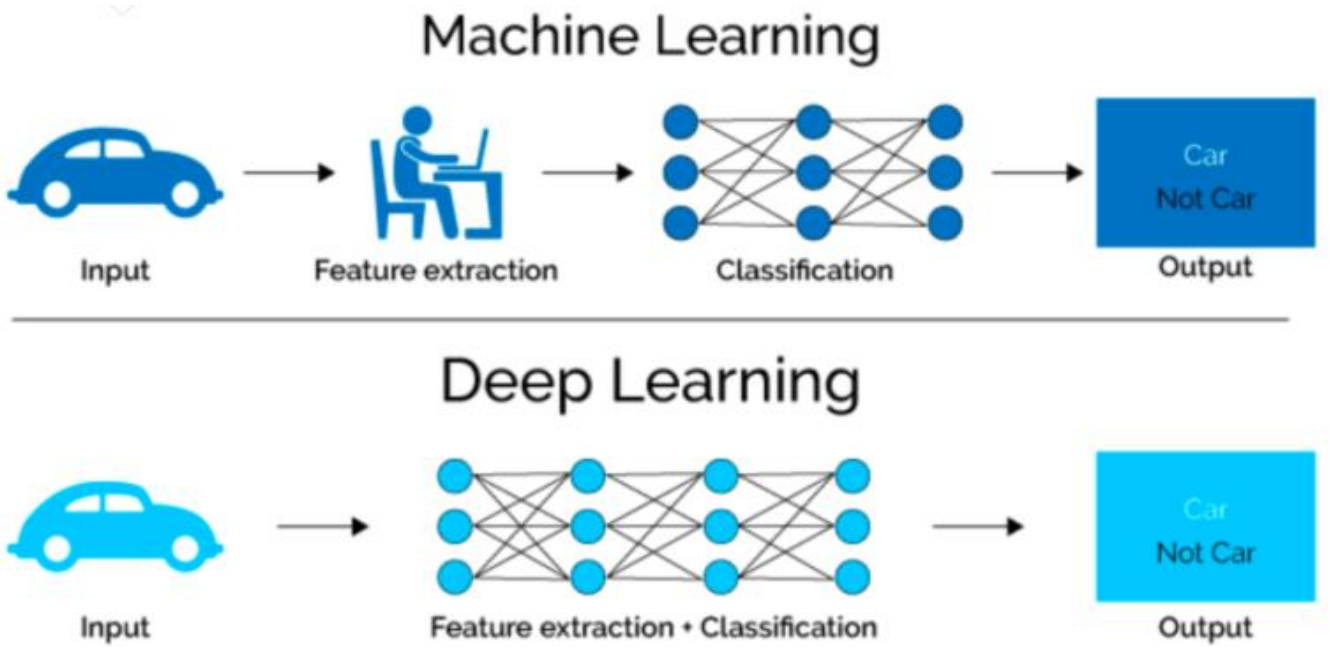
- *Machine learning is a subset of AI. It is the tendency of machines to learn from data analysis and achieve Artificial Intelligence.*
- *Deep learning is a subset of machine learning where neural networks, algorithms, learn from large amounts of data.*
- *Similarly, to how we learn from experience, the deep learning algorithm would perform a task repeatedly, each time tweaking it a little to improve the outcome.*



1. DNN overview

Machine learning and deep learning

- *Machine learning needs some guidance for performing a task, whereas deep learning the model will do it on its own without the interference of programmer.*



Source: https://thedatascientist.com/what-deep-learning-is-and-isnt/

1. DNN overview

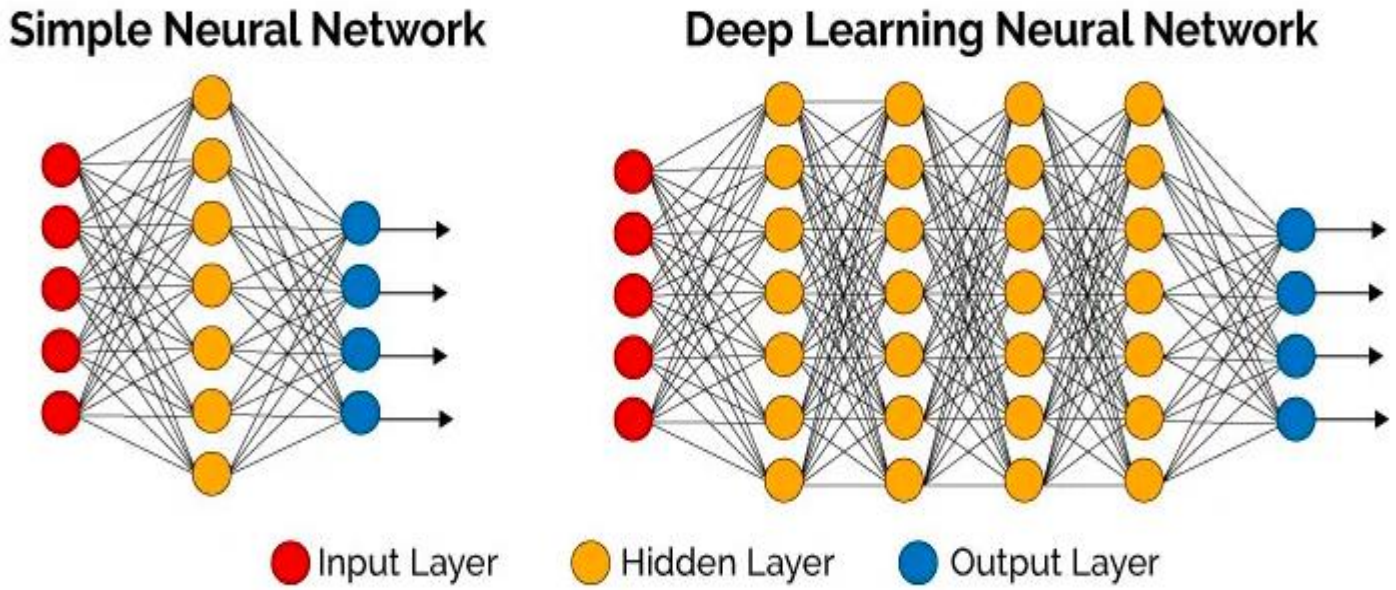
Neural networks

- *A neural network is a simulation of the human brain using machine learning. In fact, it is a mathematical model that uses learning algorithms.*
- *It represents a network of neurons used to analyze and process the data.*
- *A neural network is built up by organizing layers of neurons in a network architecture.*

1. DNN overview

Deep neural network

- *A deep neural network is simply a neural network with many layers.*
- *The extra layers provide a huge increase in computational power, which have allowed deep neural networks to reach amazing performance in multiple tasks.*



Source: https://thedatascientist.com/what-deep-learning-is-and-isnt/

1. DNN overview

Application domain

- *Deep learning is used in many fields such as:*

  - Autonomous vehicles.
  - Fraud Detection.
  - Natural Language Processing (NLP), etc.…

- *Systems such self driving cars are **safety critical domains**, that is why we need some verification approaches and testing tools to ensure the safety and robustness of those systems.*

DNN Verification

- Verification approaches
- Example of a verification approach in the robustness evaluation of DNN theme

2. DNN verification

Verification approaches

- *Neural networks could be trained offline or online depending on the nature of the system.*
- *Being trained offline means the neural network only learns during development and it will act deterministically. Therefore, static verification methods could be possible.*
- *Online training gives the neural network the ability to learn during operation, which needs run-time verification methods.*
- *In some systems, in order to meet the requirements, we need both online and offline training strategies, that is why, both verification methods are required.*
- *IEC 61508 and ISO 26262 are two standards highly relevant to the test and verification of safety critical control systems.*

2. DNN verification

Verification approaches

- *IEC 61508 defines 4 Safety Integrity Levels (SIL) and more time and effort for verification are needed for higher the SIL level a system requires.*
- *Formal methods are highly recommended techniques for verifying high SIL systems..*

SIL Levels According IEC 61508 / IEC 61511

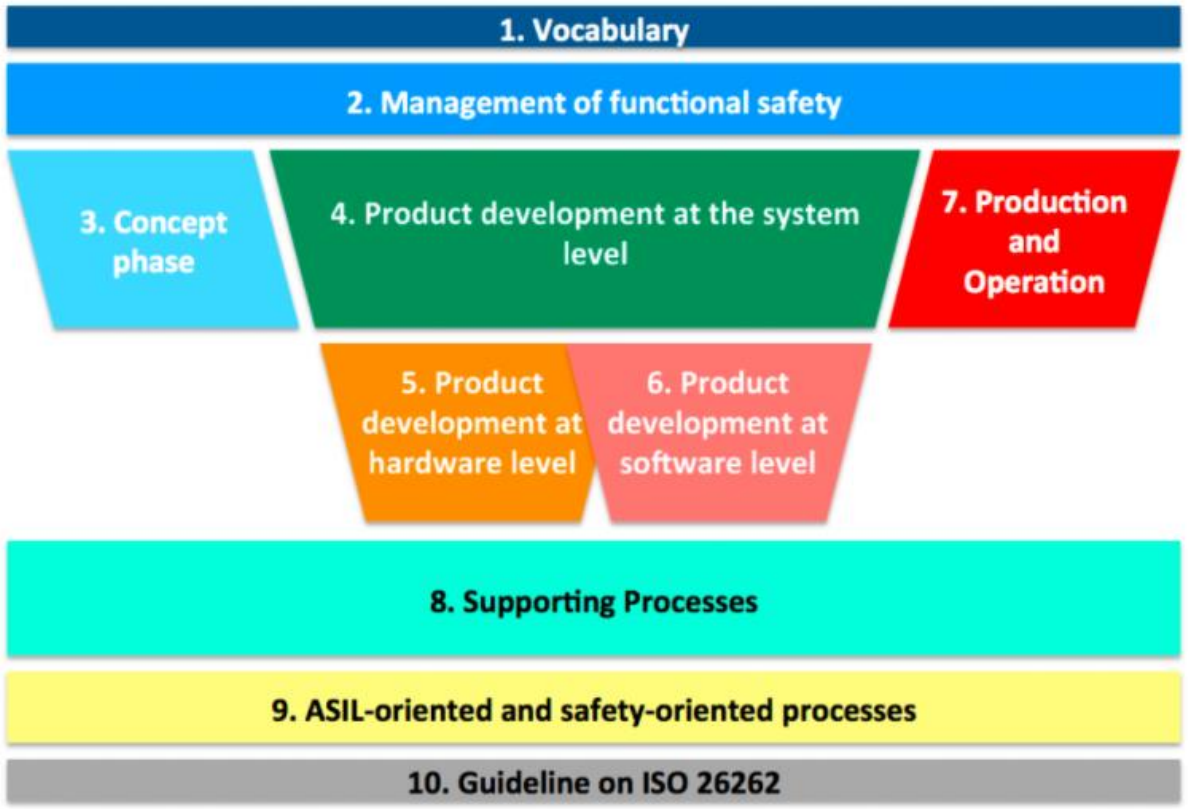
SIL Safety Integrity Level	PFDavg Average probability of failure on demand per year (low demand mode)	RRF Risk Reduction Factor	PFDavg Average probability of failure on demand per hour (high demand or continuous mode)
SIL 4	$\geq 10^{-5}$ and $< 10^{-4}$	100000 to 10000	$\geq 10^{-9}$ and $< 10^{-8}$
SIL 3	$\geq 10^{-4}$ and $< 10^{-3}$	10000 to 1000	$\geq 10^{-8}$ and $< 10^{-7}$
SIL 2	$\geq 10^{-3}$ and $< 10^{-2}$	1000 to 100	$\geq 10^{-7}$ and $< 10^{-6}$
SIL 1	$\geq 10^{-2}$ and $< 10^{-1}$	100 to 10	$\geq 10^{-6}$ and $< 10^{-5}$

Source: https://instrumentationtools.com/understanding-safety-integrity-level-iec-61511/

2. DNN verification

Verification approaches

- *ISO 26262, titled Road vehicles functional safety, is an international standard for the functional safety of electrical and/or electronic systems in production automobiles*
- *ISO 26262 explicitly states that the production of a safety case is mandated to assure system safety.*
- *It defines a safety case as an argument that the safety requirements for an item are complete and satisfied by evidence compiled from work products of the safety activities during development.*



Source: https://icomod.com/ressources/aux-and-goodies/iso26262-flow-of-workproducts-visualized/

2. DNN verification

Verification approaches

- *The development of suitable approaches, which can verify the system behavior and misbehavior is always challenging and the architecture of NNs and especially DNNs makes it harder to decipher how the algorithmic decisions were made.*

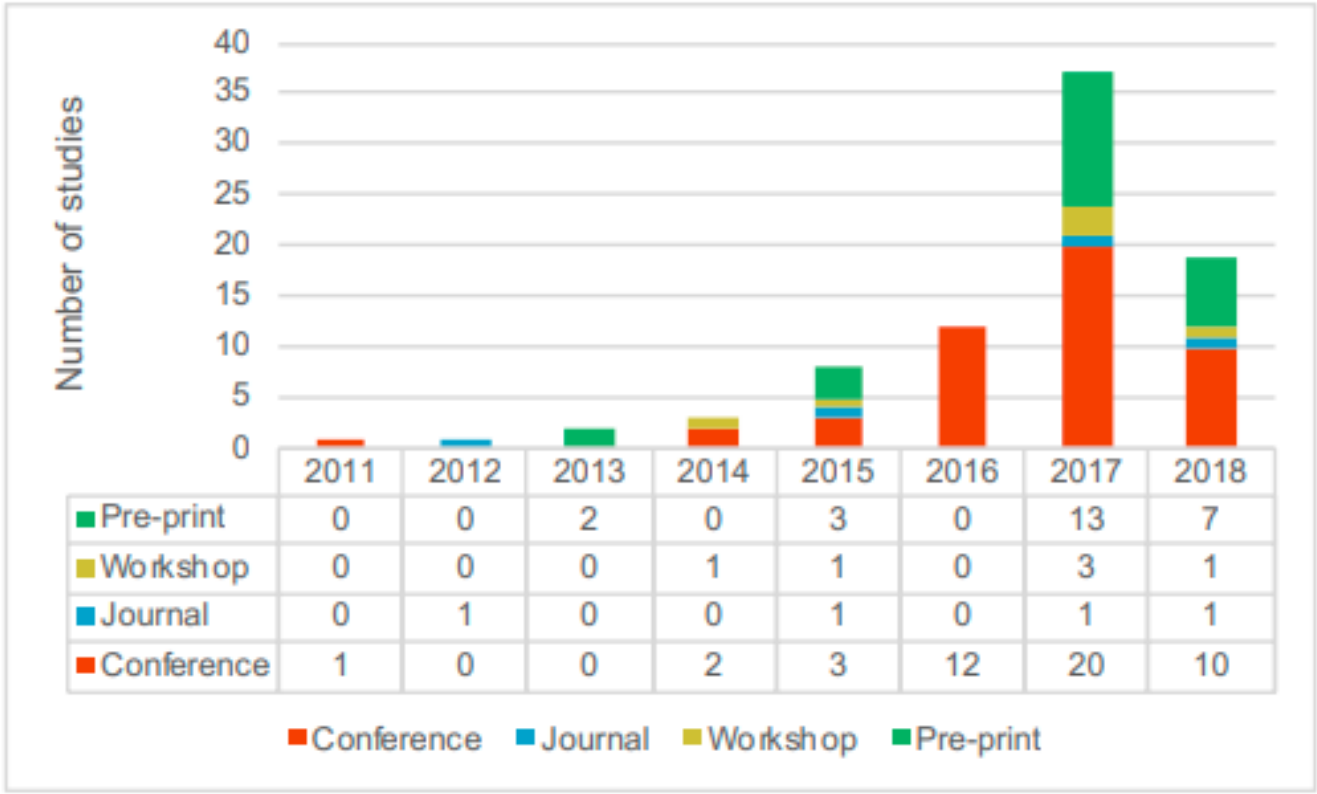
- *The current version of IEC 61508 is not applicable for the verification of NN because AI technologies are not recommended there.*

- *The latest version of ISO 26262 and its extension, ISO/PAS 21448, which is also known as safety of the intended functionality (SOTIF), are also not ready for the verification of NN-based autonomous vehicles.*

2. DNN verification

Verification approaches

- *Since the traditional verification approaches can not be applied to NN verification, researchers have focused on the topic.*
- *There have been 68 papers (81.9%) published between 2016 and 2018 indicating that researchers are paying more attention to the test and verification of NN for safety critical systems.*



	2011	2012	2013	2014	2015	2016	2017	2018
Pre-print	0	0	2	0	3	0	13	7
Workshop	0	0	0	1	1	0	3	1
Journal	0	1	0	0	1	0	1	1
Conference	1	0	0	2	3	12	20	10

Source: https://arxiv.org/pdf/1910.06715.pdf

2. DNN verification

Verification approaches

- *From all the approaches defined in papers and publications, five were identified.*
- *The next table shows the different approaches with their references found in the article: Testing and verification of neural-network-based safety-critical control software: A systematic literature review.*

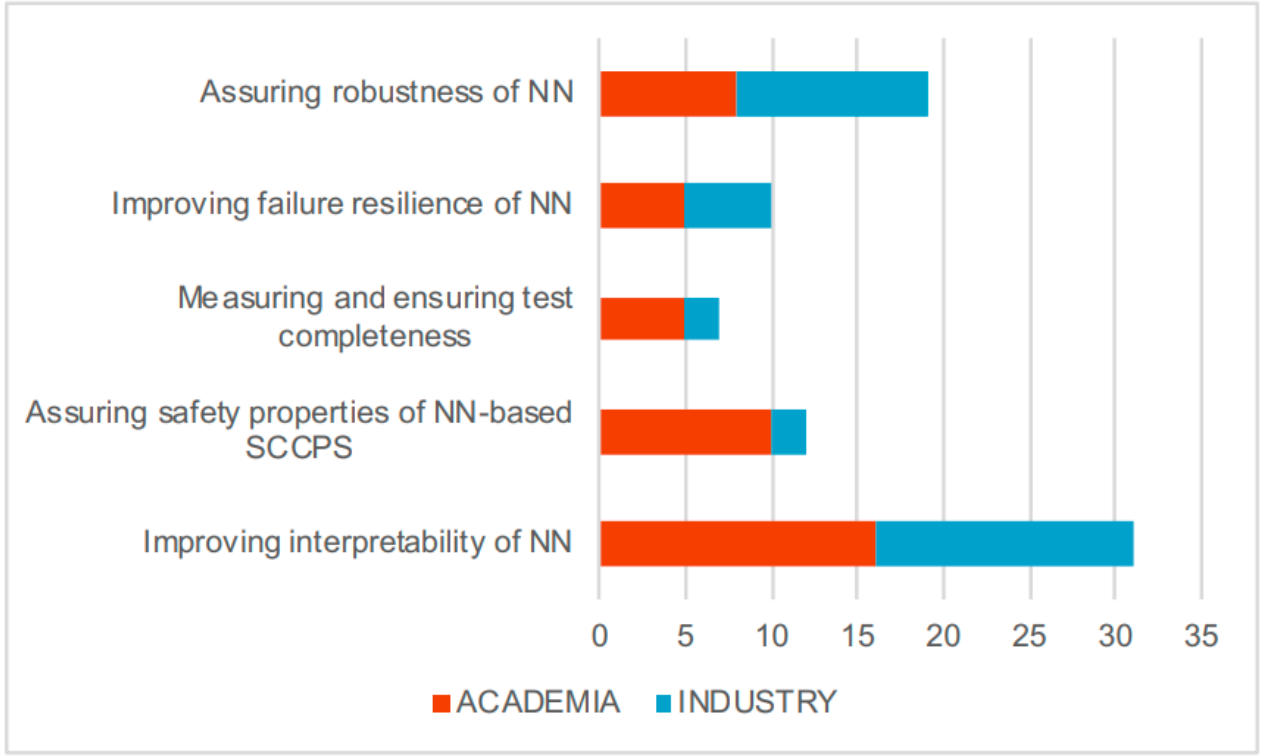
Themes	Sub-themes	Papers	#
Assuring robustness of NNs	Understanding the characteristics and impacts of adversarial examples	[56],[57],[58],[59],[60], [61], [62]	17
	Detect adversarial examples	[63],[64], [65], [66], [67], [68]	
	Mitigate impact of adversarial examples	[69], [70]	
	Improving robustness of NNs through using adversarial examples	[71], [72]	
Improving failure resilience of NNs		[73],[74],[75],[76],[77],[78],[79],[80],[81],[82],[83]	11
Measuring and ensuring test completeness		[84],[85],[86],[87],[88],[89],[90]	7
Assuring safety properties of NN-based CPSs		[91],[92],[93],[94],[95],[96],[97],[98],[99], [100],[101],[102],[103]	13
Improving interpretability of NNs	Understand how a specific decision is made	[104],[105],[106],[107],[108],[109],[110],[111], [112],[113],[114],[115],[116],[117],[118], [119], [120],[121],[122]	31
	Facilitate understanding of the internal logic of NNs	[123],[124],[125],[126],[127],[128]	
	Visualizing internal layers of NNs to help identify errors in NNs	[129],[130],[131],[132],[133],[134]	

Source: https://arxiv.org/pdf/1910.06715.pdf

2. DNN verification

Verification approaches

- *Since Deep learning is the future. Both industry and academic research are showing interest to NN verification.*
- *The next figure compares the interest's difference of academia and industry for the five identified themes.*



Source: https://arxiv.org/pdf/1910.06715.pdf

2. DNN verification

Deep Safe: an example of a verification approach in the robustness evaluation of DNN theme

- *Robustness of a NN is its ability to cope with erroneous inputs.*
- *The erroneous inputs can be an adversarial example (i.e., an input that adds small perturbation intentionally to mislead classification of an NN), or benign but wrong input data.*
- *A proposed approach called **DeepSafe** deals with this theme.*
- *This approach can automatically identify safe regions of the input space, within which the network is robust against adversarial perturbations.*
- *The approach is data-guided, relying on clustering to identify well-defined geometric regions as candidate safe regions. Then verification techniques are used to confirm that these regions are safe using counter-examples as a proof*

2. DNN verification

DeepSafe Steps

- *The DeepSafe approach is composed of 4 steps :*

  - Clustering of training inputs.
  - Cluster analysis.
  - Cluster verification.
  - Processing of possible adversarial examples.

2. DNN verification

DeepSafe: Clustering of training inputs

- *A modified **kMeans** clustering algorithm is used to perform clustering over the training inputs, using not only the similarity of the data-points but also using their labels.*
- *This produces, small and dense clusters of consistently-labeled inputs.*
- *Such clusters are assumed to be **Safe regions**, in which all inputs should be labeled consistently.*
- *The main hypothesis (H1) behind the clustering algorithm is that for a given cluster C, with centroid **cen** and radius **r**, any input x within distance r from cen has the same true label l as that of the cluster*

$$\|x - cen\| \leq r \quad \Rightarrow \quad label(x) = l$$

```
function rep = next_fun_acas(p)
movefile('count.csv','countin.csv');
nlin = importdata("countin.csv");
m = size(nlin);
rep = 0;
for sn = 1 : m(1)
    num = nlin(sn,1);
    lo = "cluster"+ num2str(num)
    +".csv";
    ln = "clusterin"+num2str(num)
    +".csv";
    loc = char(lo);
    lnc = char(ln);
    movefile(loc,lnc);
end
cnt = 0;
for yn = 1 : m(1)
    X = importdata("clusterin"+
          num2str(nlin(yn,1))+".csv");
    [idx,C] = kmeans(X,nlin(yn,2),
          'Distance','sqeuclidean');
    Xn = [X,idx];
    Yn = sortrows(Xn,7);
    Un = unique(Yn(:,7));
    nn = size(Un);
for xn = 1 : nn(1)
  Zn = Yn(Yn(:,7) == Un(xn,1),1:6);
  Bn = unique(Zn(:,6));
  n0n = size(Bn);
  wn = xn + cnt;
  if (n0n(1,1) > 1)
    rep = 1;
    csvwrite("cluster"+num2str(wn)
        + ".csv",Zn);
    nln(1,1) = wn;
    nln(1,2) = n0n(1,1);
    dlmwrite("count.csv",nln,
        '-append','delimiter',',');
  else
    sn = Un(xn,1);
    csvwrite("clusterFinal"+num2str(wn)
        + "_" + num2str(p)+".csv",C(sn,:));
    dlmwrite("clusterFinal"+num2str(wn)
        + "_" + num2str(p)+".csv",Zn,
        '-append','delimiter',',');
  end
end
cnt = wn;
end end
```

Source:
https://arxiv.org/pdf/1710.00486.pdf

2. DNN verification

DeepSafe: Cluster Analysis

- *The clusters obtained characterize the behavior of the network over large chunks of the input space.*
- ***Cluster Analysis** can provide useful insights regarding the behavior and accuracy of the network. The **density** and the **centroid** behavior are the cluster properties that have been used.*
- *Hypothesis (H1), is assumed to hold in clusters with high density, and for the purpose of robustness the clusters with low density are disregarded in order to increase the chance of examining only valid **Safe regions** and of detecting only valid **adversarial perturbations.***
- *The centroid of a cluster can be also a representative of the cluster behavior, since targeted adversarial perturbations are more likely to exist for the labels whose level of confidence of the NN is higher than others. The label scores assigned by the NN are used to look for targeted safe regions.*

2. DNN verification

DeepSafe: Cluster Verification/ Processing of possible adversarial examples

- *Having identified and analyzed the clusters, the Reluplex tool, which is an efficient SMT solver for verifying deep neural networks, is used to verify a formula representing the negation of the previous hypothesis (H1).*

$$\exists x. \quad \|x - \text{cen}\|_{L_1} \leq r \quad \wedge \quad \text{score}(x, l') \geq \text{score}(x, l) \quad \text{(Eq 2)}$$

- *If the negated hypothesis is shown not to hold, the region is indeed safe with respect to that label (targeted safe region); and otherwise, Reluplex provides a satisfying assignment, which constitutes a valid adversarial perturbation.*

- *a satisfied solution to Eq. 2 for any target label (l') indicates the presence of an input within the region for which the network assigns a higher score to label l' than to l.*

- *The check for the validity of the adversarial example needs to be done by the user/domain expert.*

2. DNN verification

DeepSafe: Case studies

- *On the ACAS Xu dataset, DeepSafe was able to identify 125 regions which are completely safe, 52 targeted safe between 210 clusters.*
- *On the Mnist Image dataset, DeepSafe was able to identify 7 regions which are completely safe, 63 targeted safe between 80 clusters.*
- *Preliminary experiments on the ACAS Xu and MNIST datasets highlight the potential of the approach in providing formal guarantees about the robustness of neural networks in a scalable manner.*

Conclusion

3. Conclusion

- *The research on verification of neural networks is gaining interest and attention from researchers.*
- *The approaches can be classified into five high-order themes, namely, assuring robustness of NNs, improving failure resilience of NNs, measuring and ensuring test completeness, assuring safety properties of NN-based SCCPSs, and improving interpretability of NNs.*
- *DeepSafe is a data guided technique that search for adversarial perturbations or prove they cannot happen, by doing so, the approach identifies and provides proof for regions of safety in the input space within which the network is robust with respect to target labels.*
- *Preliminary experiments on the ACAS Xu and MNIST datasets highlight the potential of the DeepSafe in providing formal guarantees about the robustness of neural networks.*

References

- Neural network: https://www.sciencedirect.com/topics/neuroscience/neural-networks
- Difference between artificial intelligence, machine learning and deep learning: https://hackernoon.com/difference-between-artificial-intelligence-machine-learning-and-deep-learning-1pcv3zeg
- Testing and verification of neural-network-based safety-critical control software: https://arxiv.org/pdf/1910.06715.pdf
- DeepSafe: A Data-driven Approach for Checking Adversarial Robustness in Neural Networks: https://arxiv.org/pdf/1710.00486.pdf

THANK YOU FOR YOUR ATTENTION