IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol.-7, Issue-5, May 2019 E-ISSN: 2347-2693

# Strategies to architect AI Safety: Defense to guard AI from Adversaries

**Rajagopal. A[1], Nirmala. V[2*]**

[1] Interdisciplinary research, Indian Institute of Technology, Madras, India
[2] PG and Research Dept. of Physics, Queen Mary's College, Chennai, India

[*]*Corresponding Author: gvan.nirmala@gmail.com, Tel.: +91 9445412134*

DOI: https://doi.org/10.26438/ijcse/v7i5.451456 | Available online at: www.ijcseonline.org

Accepted: 19/May/2019, Published: 31/May/2019

***Abstract*—** The impact of designing for safety of AI is critical for humanity in the AI era. With humans increasingly becoming dependent of AI, there is a need for neural networks that work reliably, inspite of Adversarial attacks. Attacks can be one of 3 types: I) Similar looking adversarial images that aim to deceive both human and computer intelligence, II) Adversarial attacks such as evasion and exploratory attacks, III) Hacker introduced occlusions/perturbations to misguide AI. The vision for Safe and secure AI for popular use is achievable. To achieve safety of AI, this paper contributes both a strategy and a novel deep learning architecture. To guard AI from adversaries, paper proposes 3 strategies: 1) Introduce randomness at inference time to hide the representation learning from adversaries/attackers, 2) Detect presence of adversaries by analyzing the input sequence to AI, 3) Exploit visual similarity against adversarial perturbations. To realize these strategies, this paper proposes a novel architecture, Dynamic Neural Defense (DND). This defense has 3 deep learning architectural features: I) By hiding the way a neural network learns from exploratory attacks using a random computation graph, DND evades attack. II) By analyzing input sequence to cloud AI inference engine with CNN-LSTM, DND detects fast gradient sign attack sequence. III) By inferring with visual similar inputs generated by VAE, any AI defended by DND approach doesn't succumb to hackers. Thus, a roadmap to develop reliable, safe & secure AI is presented.

***Keywords*—** AI, Deep Learning, AI Safety, AI Security, Neural Networks, Adversarial Attacks and Defences, autonomous AI.

## *I.* INTRODUCTION: WHY AI SAFETY

### *A. The importance of safety & security of AI*

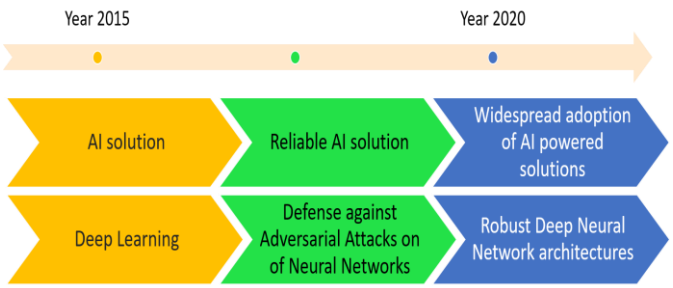


Figure 1. Why safe AI is critical for adoption of AI

The field of Artificial Intelligence is witnessing amazing progress thanks to extraordinary work in research in academia and open source Deep Learning frameworks. But the practical widespread adoption of Deep Learning based solutions is still yet to happen. One of the key challenges is that they are vulnerable to Adversarial Attacks [1]. Research on AI security will open doors for widespread adoption of AI as per Fig 1.

### *B. Purpose of the Contribution: Safety & Security of AI*

The goal of this paper is to improve safety of AI by proposing a novel approach that learns to become resilient inspite of:

1. Adversarial attacks such as evasion and exploratory attacks on CNNs;
2. Adversarial attacks that introduce perturbations to create visually similar looking adversarial images that aim to deceive both human and AI; and
3. Hacker introduced occlusions or alterations in physical environment to misguide AI decisions.

### *C. Why this paper? Is it safe to depend upon AI?*

The enormous potential of Deep Learning can be realized for masses only when it is depend-able. But researchers are beginning to ask the question: Is it depend-able? Can mankind depend on decisions made by AI. When mankind begins to depend upon a new technology, it shouldn't be vulnerable. A reliable technology needs to work irrespective of the circumstances.

 

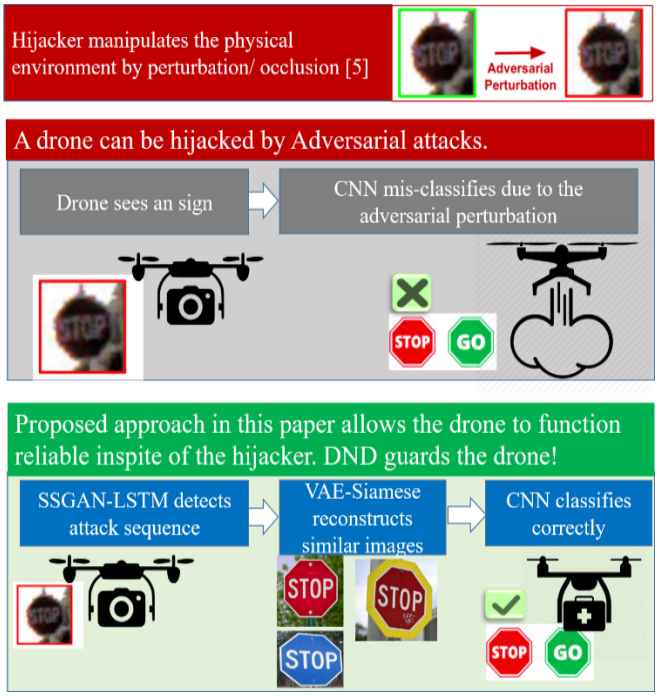


Figure 2. With and without Safe AI.

With and without safe AI, the outcome can be totally different as illustrated in Fig 2. Is it safe to depend upon AI?

1. Can an autonomous self-flying drone be hijacked by rival introduced perturbations?
2. Can a drone take down a terrorist walking on the road?

*D. Contribution of the paper*

Towards achieve the goal of guarding AI, this paper contributes on two aspects

1) Strategies to design safety into AI; and
2) Novel neural network architectural framework to realize the strategies.

The above two contributions are summarized in Fig 3. The proposed strategy presented in this paper offers a significant potential to improve resilience of AI systems against adversaries, setting a forward vision for safe AI in mainstream deployments. In addition to this contribution, the paper also contributes by exploring a novel approach in architecting neural network to realize the strategies.

The paper is organized as per the content flow shown in Fig 3.

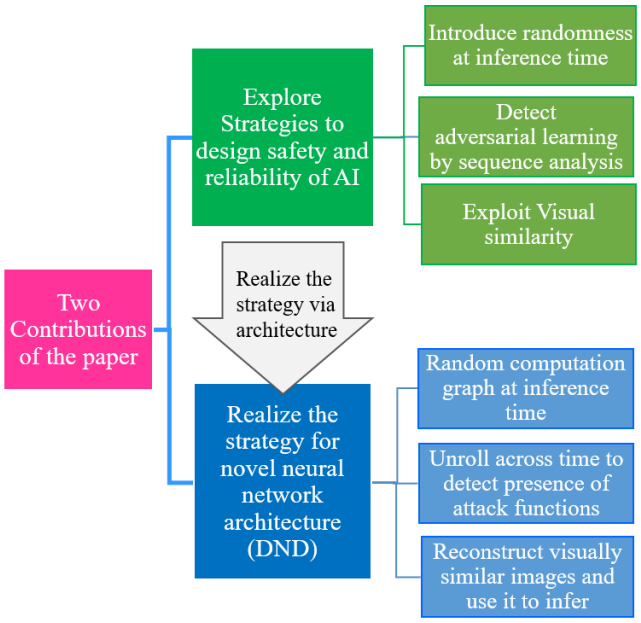


Figure 3. Two contributions. 1) Strategy 2) AI Architecture

## II. CONTRIBUTION #1: STRATEGIES FOR AI SAFETY

This paper proposes strategies to defend AI against adversaries. To explore the insights, Table 2 discusses it. The 3 strategies are

1. Introduce randomness to hide the learning function from adversaries:
   a. Through a random computation graph at inference time, this defense strategy introduces randomness to hide the learning function from an attacker.
   b. Neural architecture search automatically identifies a set of network graphs that optimizes for minimizing transferability of adversarial examples.
2. Detect adversaries by analyzing the sequence of inputs:
   a. As adversaries can attack only after a sequence of inference requests to the cloud served model, the defense strategy is to analyze by unrolling across time to detect presence of attack functions.
   b. Further, a decoy neural network is used to early detect such adversaries
3. Exploit visual similarity against adversarial perturbations:
   a. Like humans, this approach looks for visual similarity to improve robustness inspite of adversarial perturbations. It reconstructs visually similar images and infers from a random reconstruction, thus insulating the AI for any attacking systems that may try to learn about the model. This idea is used in DND of Fig 5.
   b. In addition, this mechanism improves robustness as inference is based on a visually similar images rather than the adversarial input.

 

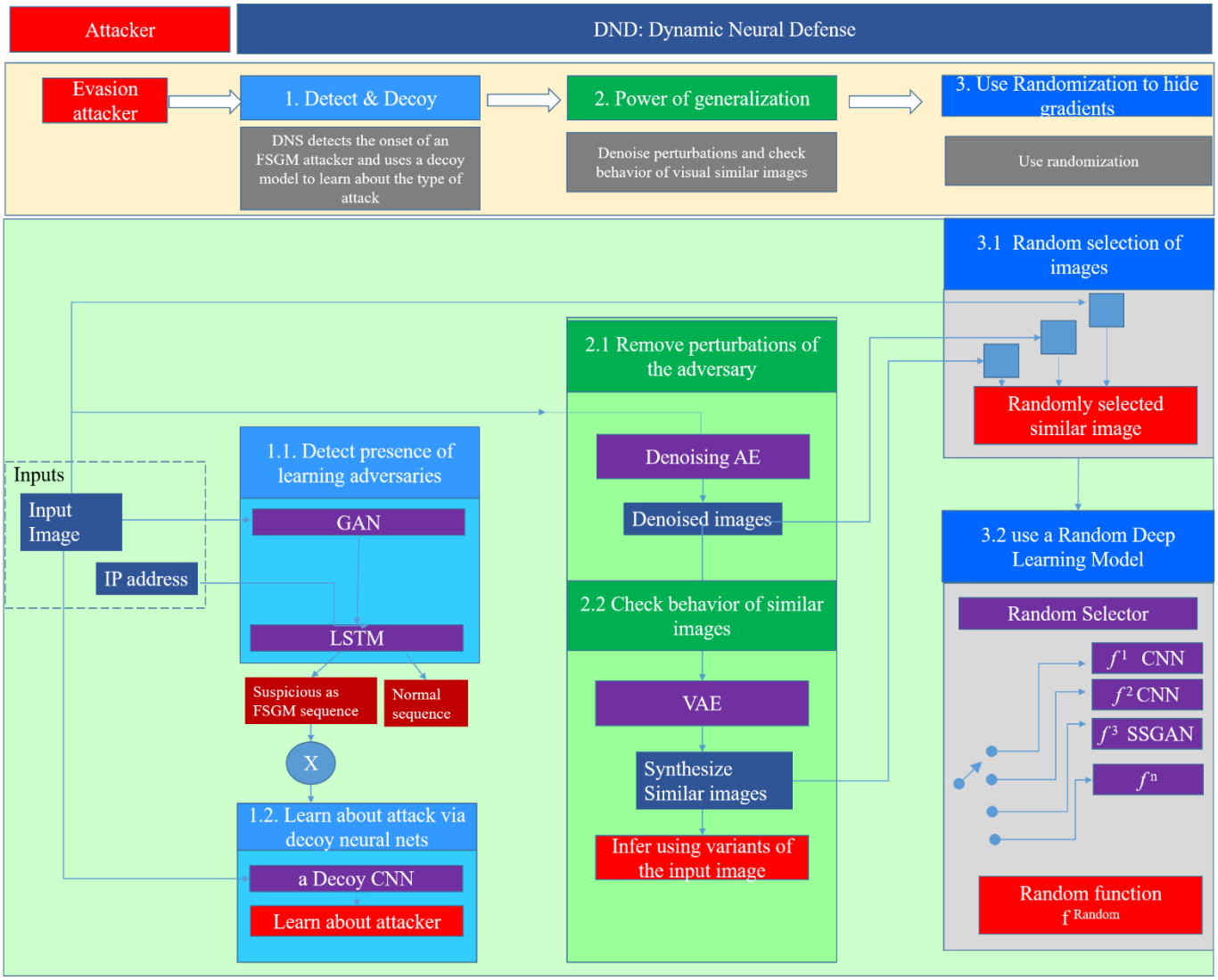


Figure 4. Proposed DND architecture to enable Safe AI.

## III. CONTRIBUTION #2: NEURAL DEFENSE FOR SAFE AI

To realize this strategy, the paper also presents a novel architectural approach as illustrated in Fig 4. The strategies are realized as a novel architecture, called Dynamic Neural Defence (DND). DND, a Neural network approach has three major deep learning based architectural features:

1. By hiding the way a neural network learns from attacks, DND evades attack. DND dynamically infer by a random choice of neural network (or a randomly selected ensemble) at run-time or a dynamically choosen computation graph. This stochastic computation hides the learning algorithm from an attacker. Thus, a three layers of "**AI firewall**" is illustrated in Fig 5. One may add 4th layer with SHIELD[16]. Further, a Neural Architecture Search is employed to design a set of CNN models in such a way, so that transferability of adversarial examples is minimized. Refer module 3.2 of DND architecture presented in Fig 4.
2. DND can learn to adapt dynamically by learning about occlusion/attack patterns from a sequence of attacks via SSGAN-LSTM detector that detects the presence of a fast gradient sign attack sequence. Refer block 1.1 in Fig 4.
3. DND not just only denoises using AE, but also infers using additional set of visual similar inputs generated by VAE with small variance, thus improving reliability of classification by inferring through other examples that are visual similar to the adversarial input. Please refer block 2.1 and 2.2 in the architecture presented in Fig 4.

 

Table 1: Nomenclature used in this paper

| | |
|---|---|
| DNN | Deep neural networks |
| *Safe*AI | Safety of AI systems also Robustness of DNN |
| DND | A proposed architecture in this paper called Dynamic Neural Defense |
| $\boldsymbol{f}$ | Deep Learning Model |
| **(X,y)** | Input pair in the dataset of unsupervised Deep Learning |
| $\mathbf{W} \leftarrow \text{train}(\boldsymbol{f}, X,y)$ | Set of Weights matrix of DNN learnt after training the model $\boldsymbol{f}$ on **(X,y)**. |
| $\underline{\mathbf{y}^{\mathbf{pred}}} = \boldsymbol{f}(\mathbf{W}*\mathbf{X})$ | Predicted label when a Deep Learning Model, $\boldsymbol{f}$ infers a input,X using the learnt Weights,**w** |
| NAS | Neural Architecture Search Algorithm |
| $\boldsymbol{f}^{\mathbf{1}}, \boldsymbol{f}^{\mathbf{2}}$ | Two different Deep Learning Models created by NAS |
| $\boldsymbol{f}^{\mathbf{n}}$ | A $n^{th}$ Model $\boldsymbol{f}$ created by NAS |
| *Attacker* | Adversarial Attacks on Deep Neural Networks [1] |
| *Defender* | Defender with a goal of *Safe*AI and applies Defences against *Attacker* [1] |
| $\boldsymbol{f}^{\mathbf{[Random\ i]}}$ | A *Defender* selects a random $i^{th}$ Model $\boldsymbol{f}^{\mathbf{i}}$ at inference time and uses it to classify |
| ATM | Adversarial Threat Model [1] |
| WBA | White Box Attacks [1] |
| FGSM | Fast Gradient Sign Method based Attack [3] |
| ***Attacked*** $\boldsymbol{f}^{\mathbf{1}}$ | *Attacker* learns about the vulnerability/gradients of $1^{st}$ Model $\boldsymbol{f}^{\mathbf{1}}$ about using FSGM |
| CNN | Convolutional Neural Network |
| LSTM | Long Short Term Memory Recurrent Neural Network |
| AE | AutoEncoder |
| VAE | Variational AutoEncoder |
| SSGAN | Semi Supervised Learning using Generative Adversarial Network |

## IV. IS TODAY'S AI VULNERABLE ?

### A. *Are all DNNs vulnerable to Adversarial Attacks?*

How easy is it today to compromise a state of art Deep Learning solution?

Due to intriguing properties of DNN[2], even an DNN based image classifier performing at super-human level is vulnerable to such attacks. Even today, even state-of-art Deep Learning Neural Networks to vulnerable to adversarial examples [1]. This is because of the neural network tries to generalize the decision boundaries by learning from a set of trainable examples, but an attacker can always learns to leverage the neural network architecture and representations learnt $\boldsymbol{f}$ **(W)**, by a particular Deep Learning Model $\boldsymbol{f}$ , and then generate Adversarial examples which can be misclassified by $\boldsymbol{f}$ **(***Adversarial example* $\mathbf{X}^{\mathbf{perturbed}}$ **)**, such that the predictions made by $\underline{\mathbf{y}^{\mathbf{pred}}} = \boldsymbol{f}(\mathbf{W}* \mathbf{X}^{\mathbf{perturbed}})$ are misclassified. Thus an attacker can force a DNN to mis-classify by gaining knowledge about network architecture $\boldsymbol{f}$ and the weights **W** that has been learnt by training $\boldsymbol{f}$ over dataset **(X,y).** The **attacker** gains knowledge about the learning function $\boldsymbol{f}$ modelled by the neural network and the learnt weights **W** on a training distribution. This type of attack is called "White Box Attacks".

### B. *What is the most common type of attack?*

The most common type of attack is when an attacker generates malicious inputs so as to force the learn model $\boldsymbol{f}$ to make a mis-classification. This is called "*Evasion Attacks*" [4]. This knowledge about the trained neural network model $\mathbf{W} \leftarrow \text{train}(\boldsymbol{f}, X,y)$ is utilized by the attacker to identify feature spaces for which the model will misclassify, such that the MSE loss is maximized $\|\mathbf{y} - \mathbf{y}^{\mathbf{pred}}\|^2$ where **y** is predicted class $\mathbf{y}^{\mathbf{pred}} = \boldsymbol{f}(\mathbf{W}* \mathbf{X}^{\mathbf{perturbed}})$ for an manipulated legitimate input, $\mathbf{X}^{\mathbf{perturbed}}$ . Refer Fig 5 for evasion attack.

### C. *Who wins today? Adversarial Attackers vs Defenses?*

Anirban's survey paper on Adversarial Attacks and Defenses[1] shows the state of art on as 2018, demonstrating the wide range of possibilities for attack.

Recently Princeton[5] printed malicious signboards and deceived autonomous vehicles as illustrated in Fig 4. In fact there exists an open source library of adversarial attack approach such as CleverHans [3].

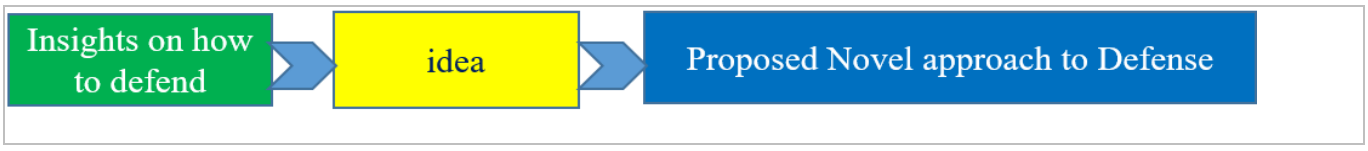


 

Table 2: Summary of the key ideas in this paper

| | | |
|---|---|---|
| 1. Make it difficult for Evasion Attacker to gain knowledge about how DNN behaves | **Introduce randomness at inference time** | The inference engine uses a random Deep Learning Model $f^{[Random\ i]}(W)$ for each inference, thus masking the gradients from attackers. 3 layers of **AI firewalls** block flow of gradient information to any attack, as illustrated in Fig 5. |
| 2. Remove the perturbations of the Adversary | **DNN learns to a remove perturbations** | De-noising Autoencoders remove perturbations/noise as shown in block 2 in Fig 4. Presence of other attacker's objects in the scene are identified with YOLO, and hence deliberately introduced occlusions are detected & even removed, improving resilience of AI. |
| 3. Predict using a visually similar images such as an input | **Generate visually similar images & classify using them** | Inference engine employs Variational autoencoders with minor variance to generate more similar looking variants of the input, and then infers from them. With varience, the DNS shields the gradient from the adversary as Fig 5. |
| 4. Detect presence of Attacker by analysing sequence of inference | **Detect presence of functions such as FGSM** | Watch a sequence of images from an IP address, and detect the presence of FGSM algorithms. If so, deploy a decoy to learn about the attack algorithm as shown in block 1 in Fig 4. |
| 5. Block transferability of adversarial examples by learning to identify them | **Generate Adversarial examples & learn to detect them ahead of deployment.** | DND can anticipate how a typically classifier can be evaded using adversarial examples, and uses this understanding to avoid taking actions under such adversarial situations. An over fitted classifier which has memorized to detect adversarial examples, along with a semi supervised GAN to senses adversarial examples. |

## V. RELATED WORK

### A. *The advances in defences against Adversariral attackers*

As per Anirban's recent survey [1], approaches to defences are

1) Augment training dataset with adversarial examples [1]
2) Hide the model's gradient from FGSM [3]
3) Defensive distillation [13]
4) Blocking the transferability by labelling adversarial examples as n+1 class [7]
5) Minimize reconstruction error using DefenseGAN [1]
6) Denoising [10]

The following gap exists:

1) Need for a comprehensive architecture that combines all the current insights. [1]
2) Design of novel approaches based on intriguing properties of neural networks [2]

As per Anirban's recent survey [1], there are only a few defenses that can handle all types of attack scenarios.

## VI. RESULTS AND DISCUSSION

To enable safety & security of AI, this paper's contributed strategic ideas and novel architecture framework. The key ideas are summarized in Fig 3.

To advance the security of AI, two key results are

1) Insights to enable AI Safety are discussed in Table 2.
2) Novel Deep Learning architecture to enable AI Safety is conceptualized & visualized in Fig 4.

## VII. CONCLUSION AND FUTURE SCOPE

### A. *Why contribute to Security & Safety of AI ?*

The imperative to design for security in early stages of deep neural networks and not an after-thought especially with the risk of hacker exploiting powerful autonomous AI. The black box nature of neural networks make it difficult to even detect that a neural network has been hacked into, and even pose a great danger for human existence.

 

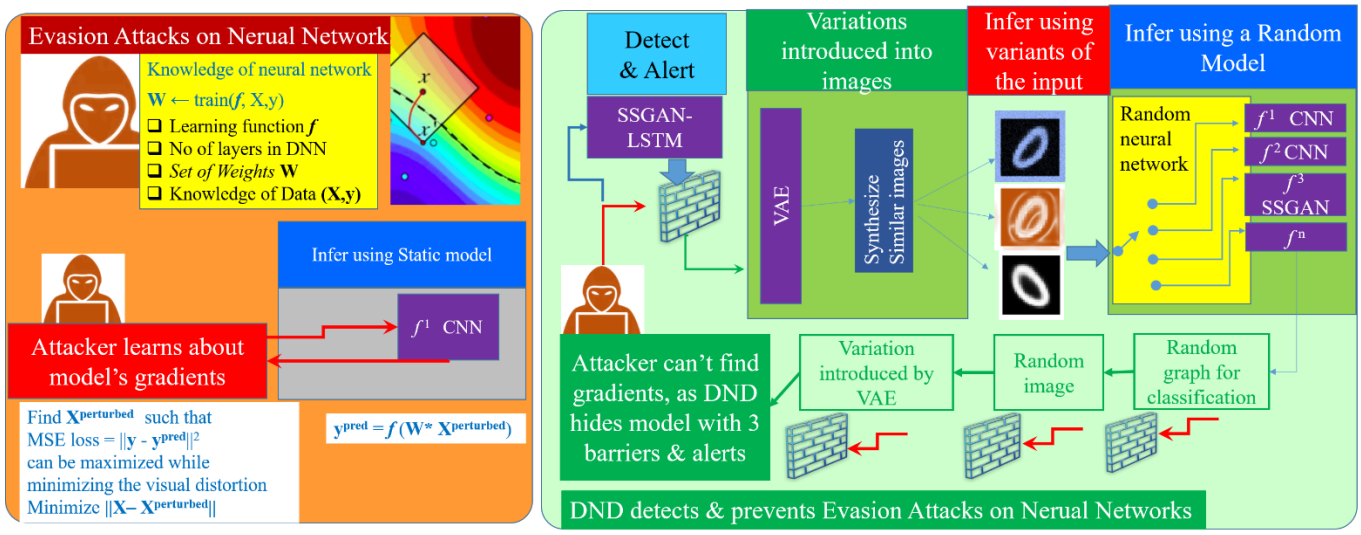


Figure 5. How DND detects and hides from an attacker?

### *B. Contributions: Roadmap, Strategies & novel Neural Net*

This paper offers a significant potential to improve power of resilience of AI systems against adversaries, setting a forward roadmap for safe usage of AI by one and all. Together with insights, ideas and neural architecture principles, the roadmap to architect for reliable AI is offered by this paper.

Summary of key contributions

1) The paper contributions are summarized in Fig 3.
2) A roadmap for Safe AI is presented in Table 2
3) Neural defence is conceptualized in Fig 4 & Fig 5.

### *C. Future directions: Defense of the future*

Safety of AI is an emerging topic for AI researchers. Research on ideas discussed in Table 2 can lead to breakthrough.

## REFERENCES


[1] Chakraborty, A., Alam, M., Dey, V., Chattopadhyay, A., & Mukhopadhyay, D, *"Adversarial Attacks and Defences: A Survey"*, CoRR, arXiv:**1810.00069, 2018**.

[2] Szegedy, C at el., *"Intriguing properties of neural networks"*, arXiv:**1312.6199, 2013**.

[3] Papernot, N., Goodfellow, I., Sheatsley, R., Feinman, R. and McDaniel, P., *"cleverhans v1. 0.0: an adversarial machine learning library"*, arXiv:**1610.00768**, **2016**.

[4] Biggio, B at el, *"Evasion attacks against machine learning at test time"*, Joint European conference on machine learning and knowledge discovery in databases, Springer, pp. **387-402**, **2013.**

[5] Sitawarin, C., Bhagoji, A.N., Mosenia, A., Chiang, M., Mittal, P., *"Darts: Deceiving autonomous cars with toxic signs"*, arXiv:**1802.06430**, **2018**.

[6] Kurakin, Alexey, I. Goodfellow, and S. Bengio. *"Adversarial machine learning at scale."* arXiv:**1611.01236**, **2016**

[7] Yuan, Xiaoyong, Pan He, Qile Zhu, Xiaolin Li., *"Adversarial examples: Attacks and defenses for deep learning."* IEEE transactions on neural networks and learning systems, **2019**.

[8] Amodei, Dario, Chris O, Jacob S, Paul C, John S, Dan M. *"Concrete problems in AI safety"*, arXiv:**1606.06565**, **2016**.

[9] Liu, G, Issa K, Abdallah K. *"GanDef"*, arXiv:**1903.02585**, **2019**.

[10] Carlini, Nicholas. *"Is AmI Robust to Adversarial Examples?."*, arXiv:**1902.02322**, **2019**.

[11] Carlini, Nicholas, David W. *"Defensive distillation is not robust to adversarial examples."*, arXiv:**1607.04311**, **2016**

[12] Mahmood S, Sruti B, Lujo B, and Michael K. *"Accessorize to a Crime: Real and Stealthy Attacks on State-of-the-Art Face Recognition"*, Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security, pp. **1528-1540**, **2016**.

[13] Tramer, Florian, A Kurakin, N Papernot, I Goodfellow, D Boneh, P McDaniel. *"Ensemble adversarial training: Attacks and defenses"* arXiv:**1705.07204** , **2017**

[14] U.Kaur, Mahajan, Singh, *"Trust Models in Cloud Computing"*, International Journal of Scientific Research in Network Security and Communication, Vol.**6**, Issue.**2**, pp.**19-23**, **2018**

[15] Arora, Sharma, *"Synthesis of Cryptography and Security Attacks"*, International Journal of Scientific Research in Network Security and Communication, Vol.**5**, Issue.**5**, pp.**1-5**, **2017**

[16] Das at el., *"Shield: Fast, practical defense & vaccination for deep learning"*, 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, ACM, pp. **196-204**, **2018**


## Authors Profile


*Rajagopal A* , BTech from IIT Madras, MS from IISc, is currently research scholar at IIT Madras.

*Nirmala V* works as an Assitant Professor at PG & Research Dept of Physcis at Queen Mary's College, Chennai.


 