/ADVERSARIAL TRAINING IS ALL YOU NEED

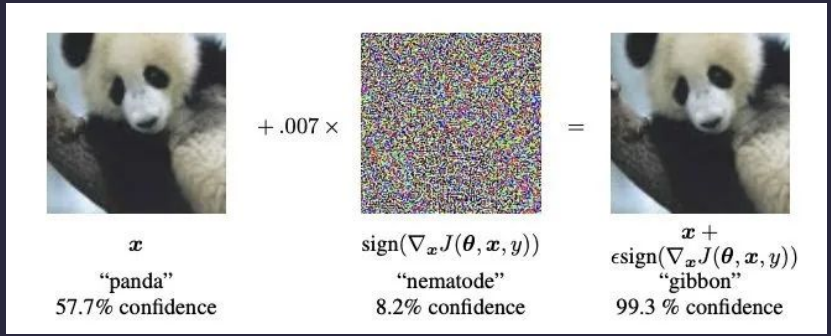
Group No. 1
Group Name: Four of a Kind

Group Members: Antreev Singh Brar, Anubhav Kalyani, Gurbaaz Singh Nandra, Pramodh Gopalan

Instructor: Prof. Priyanka Bagade

/Adversarial Attacks

- Adversarial image is an image that is intentionally manipulated by adding adversarial **perturbation** to a natural image, **unchanged to the human eye.**

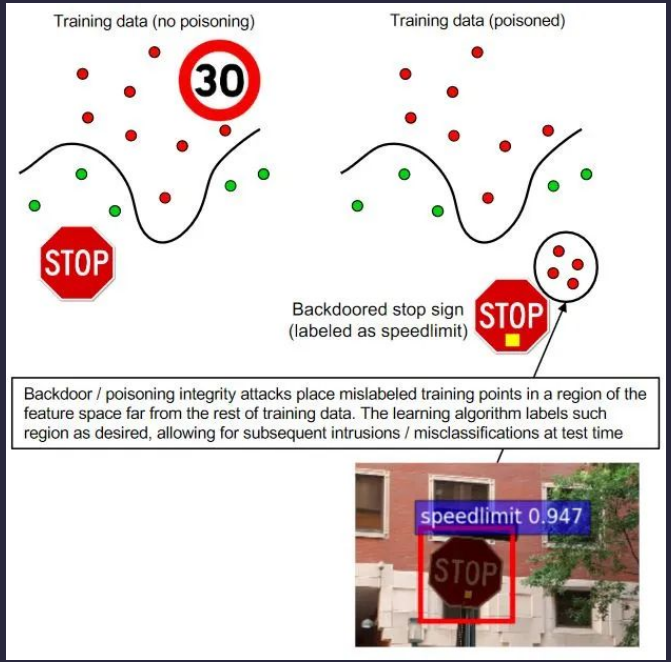


https://viso.ai/deep-learning/adversarial-machine-learning/

- Most extensive studies of adversarial machine learning have been conducted in the area of image recognition, where modifications are performed on images, causing the classifier to produce **incorrect predictions.**

/Poisoning Attacks

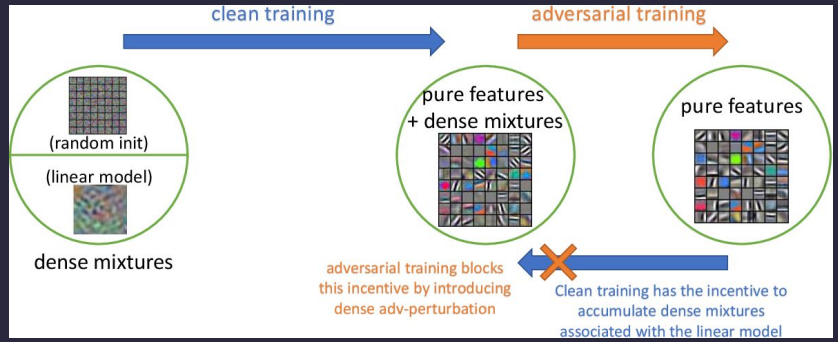
- This type of adversarial attack occurs during the training time itself.
- Typically involves manipulating and **contaminating the train data** with malicious train samples.
- Model trained on such poisoned dataset **learns false features**, which can later be exploited by the attacker during deployment.



https://viso.ai/deep-learning/adversarial-machine-learning/

/Adversarial Training

- A defense method used to increase adversarial robustness by **retraining the model** on adversarial examples.
- Adversarial examples are generated at each iteration based on current state of the model, and are used to retrain the model.
- Considered to be a **reliable defense** against adversarial attacks, as it cannot be broken by adaptive attacks.



https://arxiv.org/pdf/2005.10190.pdf

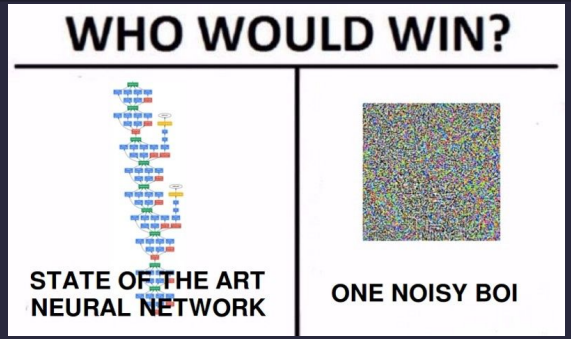
/Importance

/Security of Deep Learning

With machine learning rapidly becoming core to organizations economy, the need to protect them is growing fast.

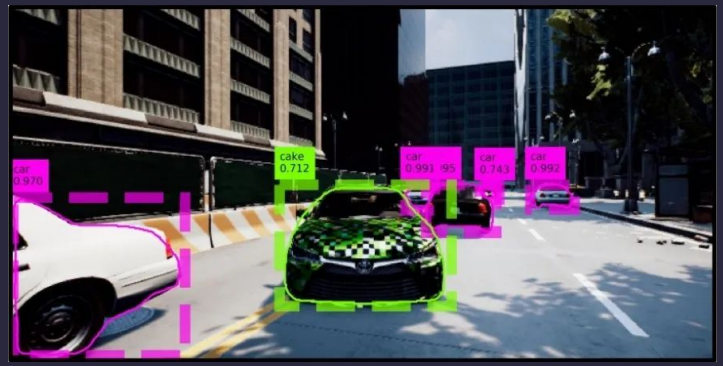
/Rapidly Evolving Space

Cat and mouse game: some propose defenses, others break them



/Impact on real-life situations

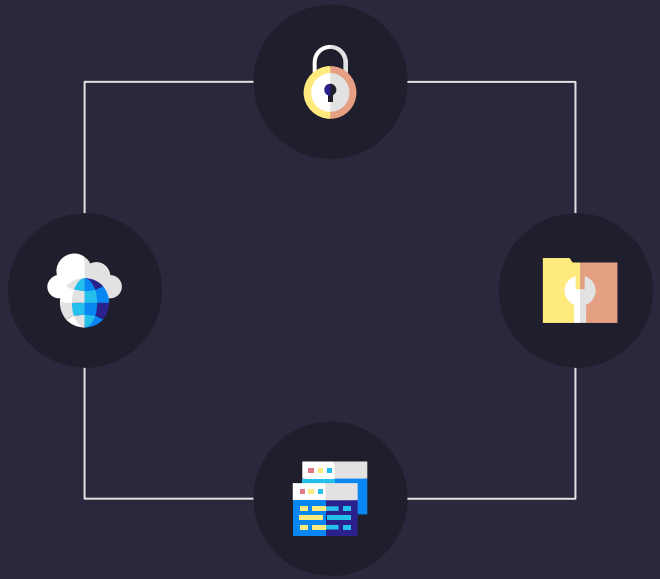
Domains like self-driving cars and AI-driven healthcare are most prone to hazard and has human lives at stake



/Can Adversarial training defend against Poisoning attacks?

Hence we arrive at the **problem statement** we aim to answer.

- Adversarially perturbed points have been shown to work as strong poisons.
- Whereas, adversarial training utilise adversarial samples generated from strong attacks to make the model more robust.
- Interesting to observe the results when one is tested against the other. We aim to do this in a **computationally less expensive** approach than existing research.



/THANKS!

/We are open to questions and suggestions, if any.