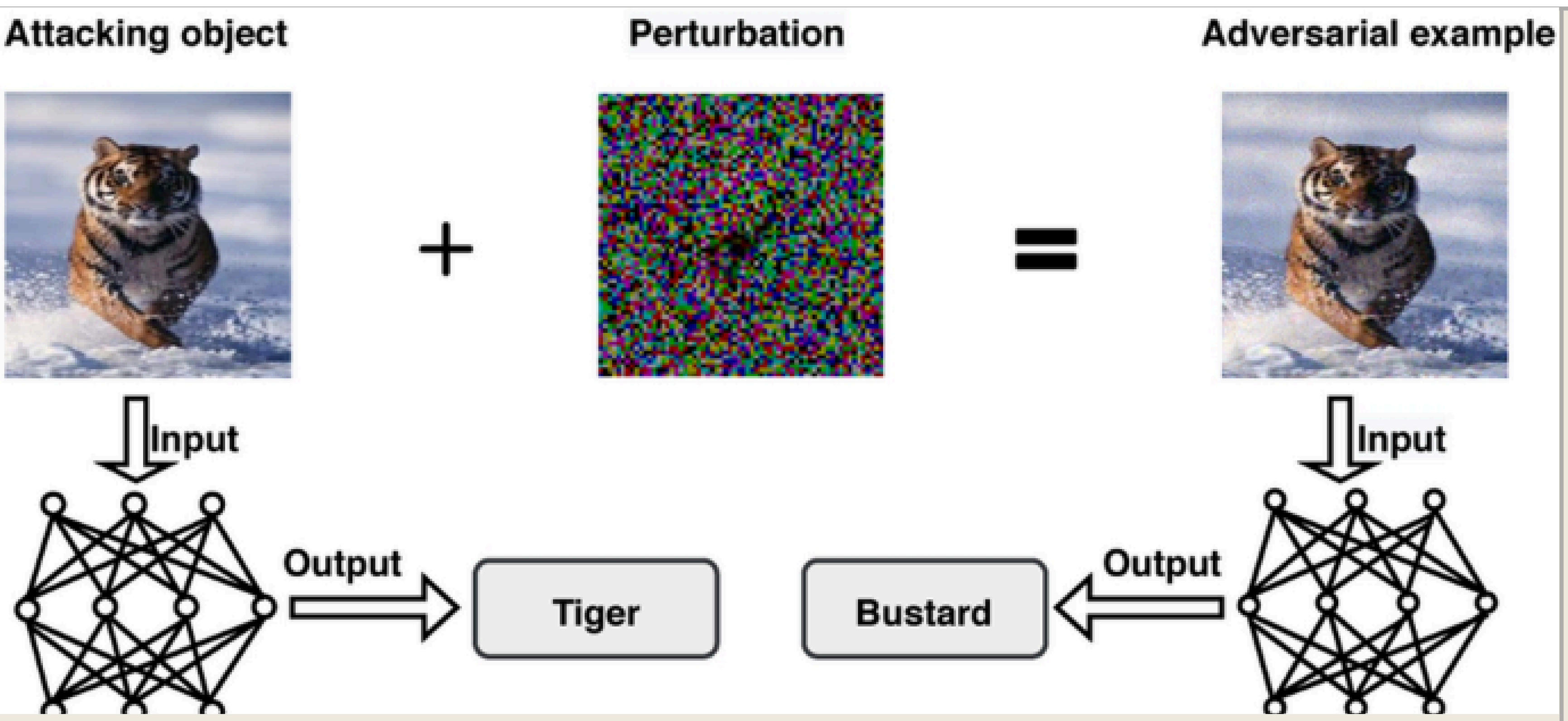
Adversarial Attacks and Robustness

Are the machine learning models we use intrinsically flawed?

By: Aashi Soni RA2411026010894

01. Introduction

Adversarial attacks involve subtly altering the input data in a way that confuses a machine learning model into making incorrect predictions, without the changes being obvious to humans. These attacks exploit weaknesses in the model, and they can have serious consequences. Adversarial Attacks and Robustness is an exciting and critical area in machine learning research. It focuses on understanding how ML models, particularly deep learning models, can be vulnerable to intentional manipulation of input data (known as adversarial attacks), and how to make these models more robust to such attacks. In critical systems like facial recognition, self-driving cars, medical diagnosis systems, or financial fraud detection, adversarial attacks can cause catastrophic failures. For example, in a self-driving car, an adversarially altered stop sign (with only a few changes) might be misclassified as a yield sign, leading to accidents.



02. Objectives

The objective of our research is to:

1. **Develop and Evaluate Robust Defense Mechanisms:** Explore, implement, and test various defense techniques, such as adversarial training and defensive distillation, to improve model robustness against adversarial attacks.
2. **Analyse the Trade-offs Between Robustness and Performance:** how enhancing robustness impacts model performance on clean data, and explore ways to balance robustness with generalization to unseen data.
3. **Address Ethical and Security Implications:** Study the ethical concerns surrounding adversarial attacks and propose guidelines to ensure the secure and fair deployment of machine learning models
4. **Identify Vulnerabilities in Machine Learning Models:** Understand how adversarial attacks exploit weaknesses in machine learning models
5. **Impact of Adversarial Attacks in Real-World Applications:** how adversarial attacks affect machine learning systems in critical areas like healthcare, finance, and autonomous driving, and assess the potential risks.
6. **Investigate Novel Adversarial Attack Methods:** Explore and develop new types of adversarial attacks

03. Methodology

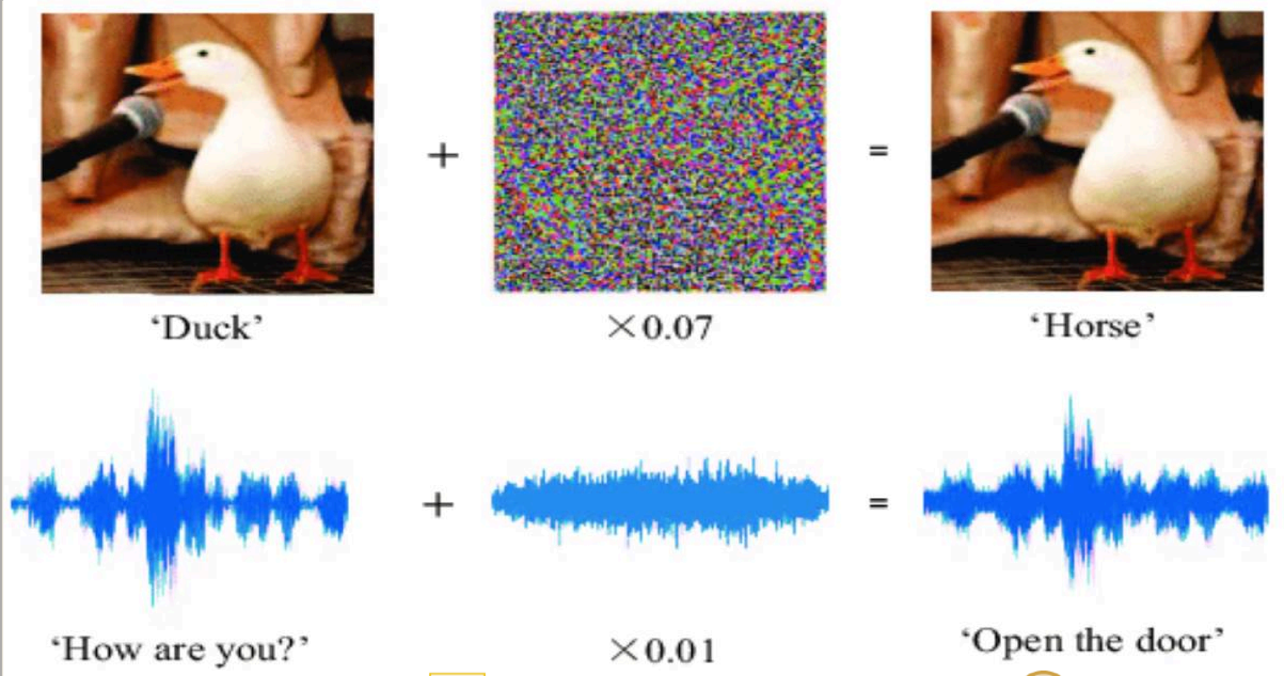
The approach and methods that will be used to achieve the project's objectives are:

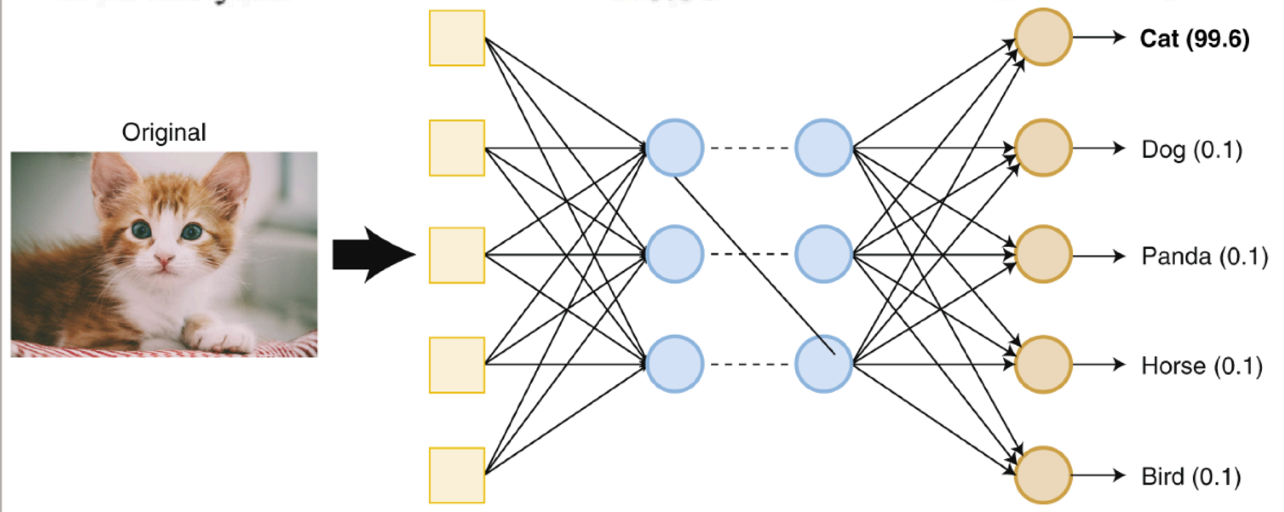
1. **Research and Review Papers :** academic papers and articles on attack methods like FGSM, PGD, and defences like adversarial training and distillation to identify key gaps and trends.
2. **Use standard datasets** (e.g., MNIST, CIFAR-10) and **machine learning models** (CNNs for images, RNNs for text) **for testing.** Train models and prepare them for attack simulations by splitting datasets into training, validation, and testing sets.
3. **Implement adversarial attacks** like FGSM and PGD on trained models to assess their vulnerabilities and measure performance degradation.
4. **Test different defences to improve robustness.** Apply defence strategies such as adversarial training and noise injection. Evaluate their effectiveness by reapplying attacks and measuring improvements.
5. **Compare model accuracy** on clean and adversarially perturbed data to evaluate trade-offs between robustness and generalization.
6. **Analyse potential misuse in real-world applications** and propose ethical guidelines for deploying robust, fair AI systems.

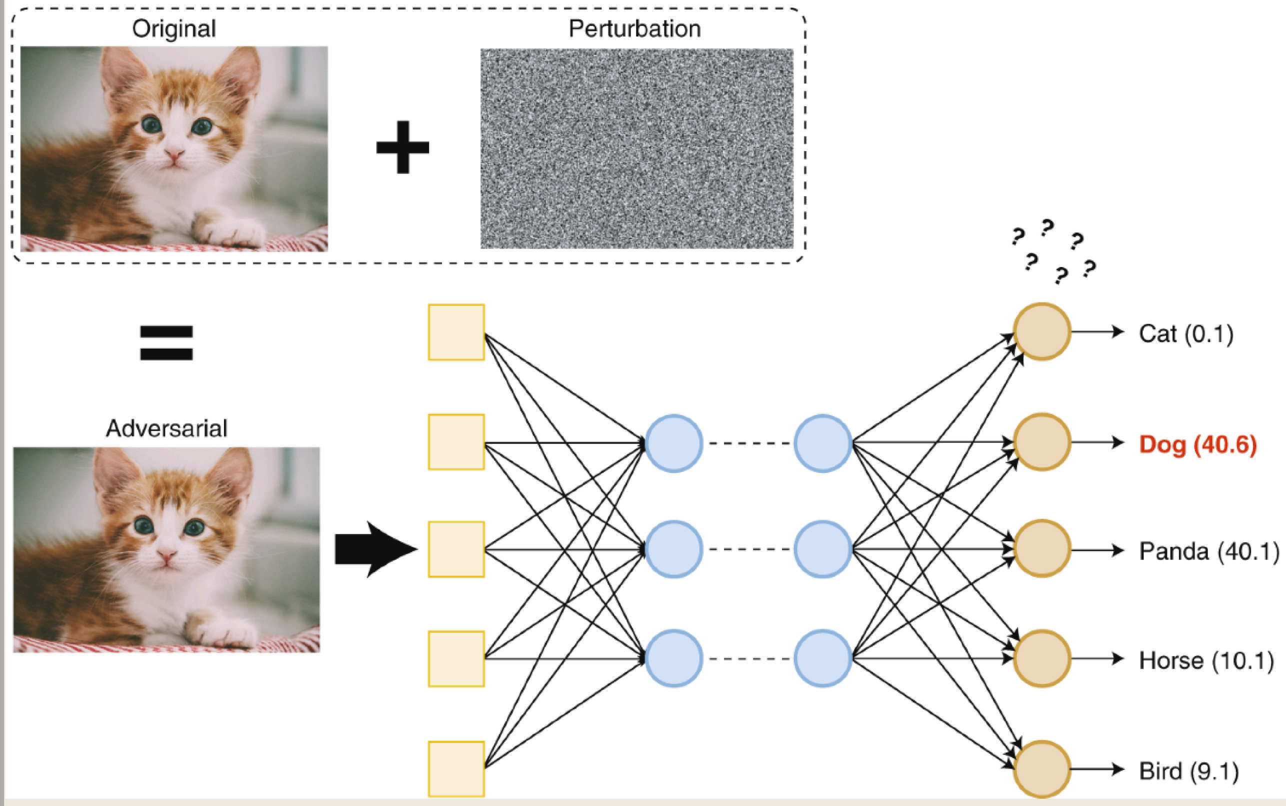
04. Expected Outcomes

The expected outcomes of the research are :

1. **Understanding of Vulnerabilities :** A detailed analysis of how different machine learning models are vulnerable to various types of adversarial attacks, highlighting specific weaknesses and common attack strategies.
2. **Performance Metrics and Evaluation Framework :** A set of performance metrics and evaluation frameworks to measure the effectiveness of both models and defence mechanisms, allowing for better comparison and assessment of adversarial robustness.
3. **Increased Public Awareness :** Raising awareness about the implications of adversarial attacks on AI systems
4. **Long-Term Model Performance Insights :** how machine learning models evolve in terms of robustness and performance over time, particularly as they are exposed to more adversarial examples or new data.
5. **Improved Domain Robustness:** improved robustness for machine learning models







05. References

- https://towardsdatascience.com/breaking-neural-networks-with-adversarial-attacks-f4290a9a45aa
- https://viso.ai/deep-learning/adversarial-machine-learning/
- https://openai.com/index/attacking-machine-learning-with-adversarial-examples/