Trustworthy Aritificial Intelligence

Presented by

Dr. Narinder Singh Punn,

Assistant Professor,

Dept. of CSE,

ABV-IIITM Gwalior

विश्वजीवनामृतं ज्ञानम्

About Me

Born in Delhi

Solan, Himachal Pradesh

B.Tech (2015), CSE, NIT Hamirpur

intellect™
A Polaris Group Company

Software Developer (2017), Chennai

M.Tech + PhD. (2022), IIIT Allahabad, Prayagraj

Postdoc (2023), Mayo Clinic, Arizona

Assistant Professor, ABV-IIITM Gwalior

Outline

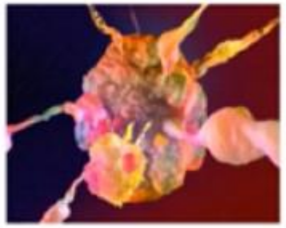
- Motivation
- What Will We Learn
- Marking Scheme

Ubiquitous AI

Autonomous Driving

Medicine

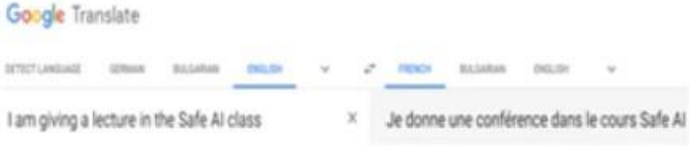
Algorithms Can Now Identify Cancerous Cells Better Than Humans



Game Playing

Google's A.I. Program Rattles Chinese Go Master as It Wins Match

Natural Language Understanding



I am giving a lecture in the Safe AI class

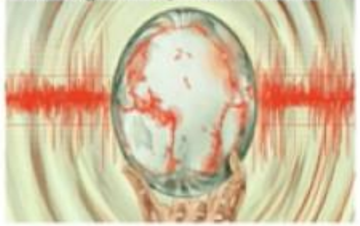
Je donne une conférence dans le cours Safe AI

Fraud Prevention

How AI is transforming the fight against money laundering

Earthquake prediction

A.I. Is Helping Scientists Predict When and Where the Next Big Earthquake Will Be



Ubiquitous AI

Autonomous Driving

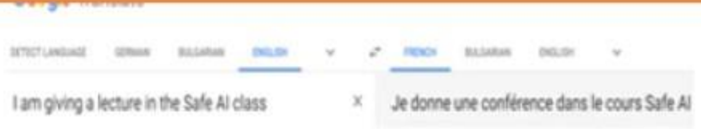
Medicine

Algorithms Can Now Identify Cancerous Cells Better Than Humans

Game Playing

Google's A.I. Program Rattles Chinese Go Master as It Wins Match

But there are problems…

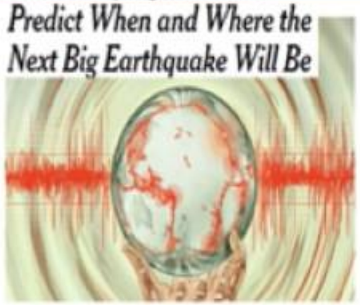


I am giving a lecture in the Safe AI class

Je donne une conférence dans le cours Safe AI

money laundering

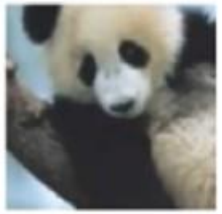
A.I. Is Helping Scientists Predict When and Where the Next Big Earthquake Will Be



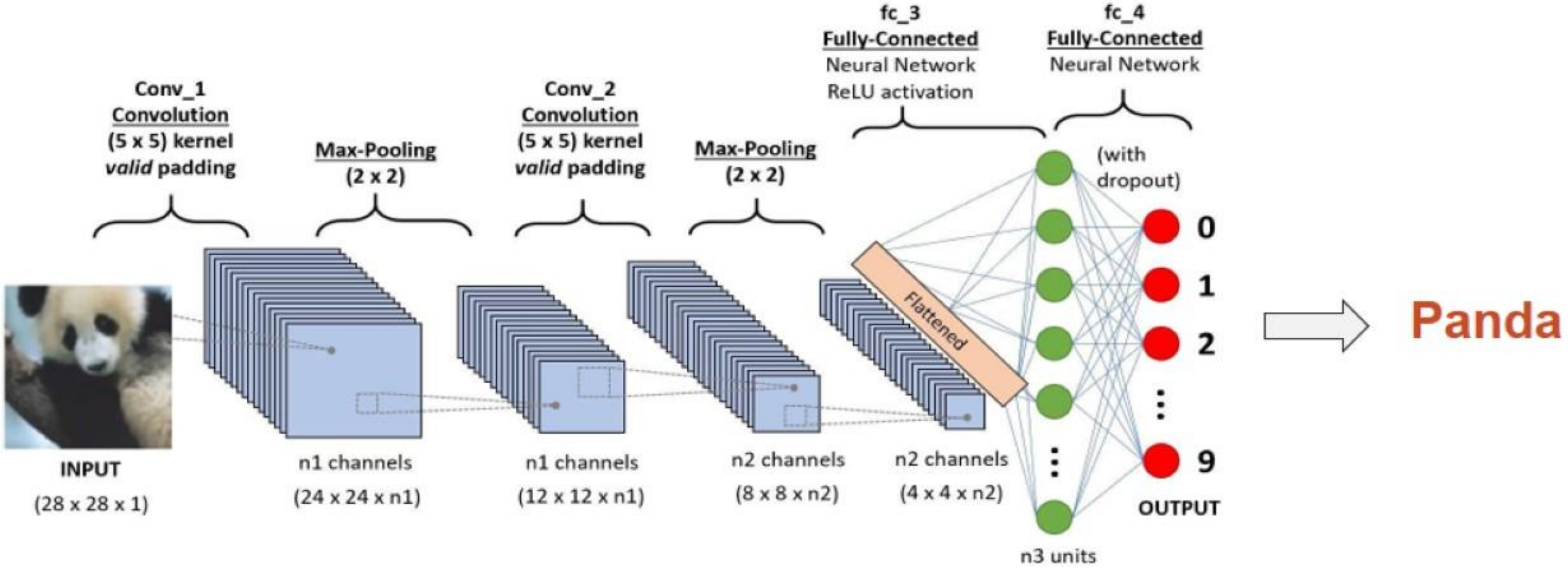
Building Trustworthy AI System is Hard

An example:

What animal do you see in the following picture?

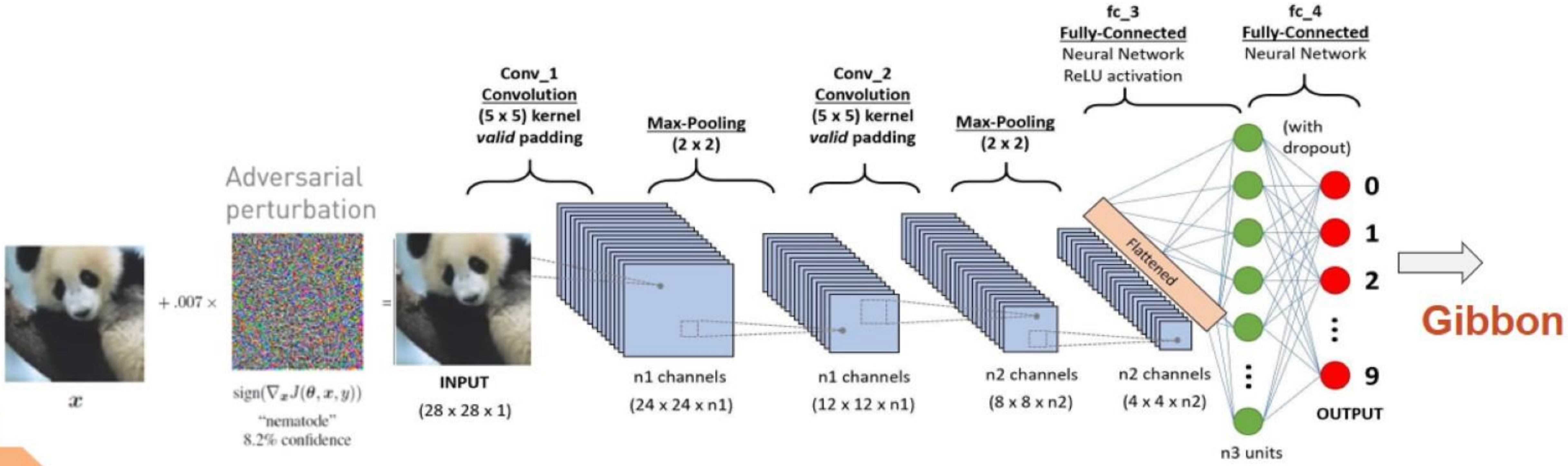


Building Trustworthy AI System is Hard



CNN Model correctly predicts

Building Trustworthy AI System is Hard

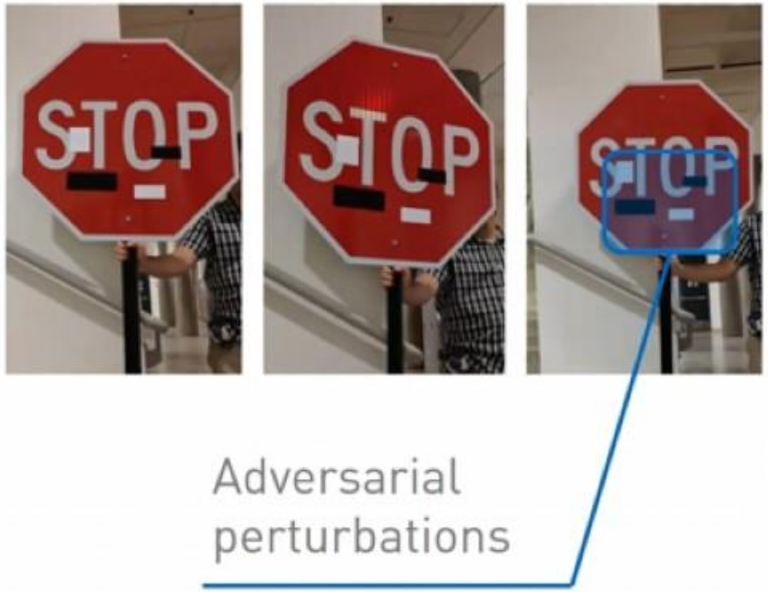


CNN Model incorrectly predicts

Goodfellow, I. J., Shlens, J., & Szegedy, C. (2014). Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572.*

Building Trustworthy AI System is Hard

1 Attacker modifies signs



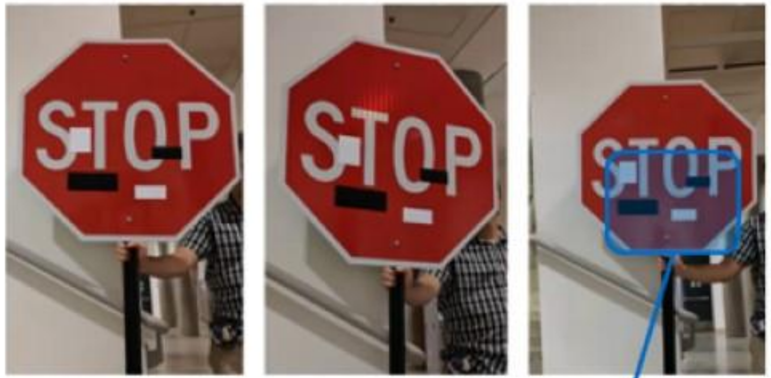
Adversarial perturbations

What sign do **you** see?

Eykholt, Kevin, Ivan Evtimov, Earlence Fernandes, Bo Li, Amir Rahmati, Chaowei Xiao, Atul Prakash, Tadayoshi Kohno, and Dawn Song. "Robust physical-world attacks on deep learning visual classification." In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1625-1634. 2018.

Building Trustworthy AI System is Hard

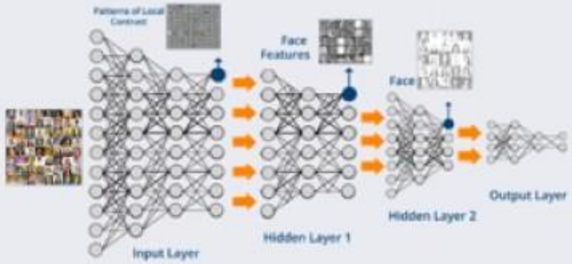
1 Attacker modifies signs



Adversarial perturbations

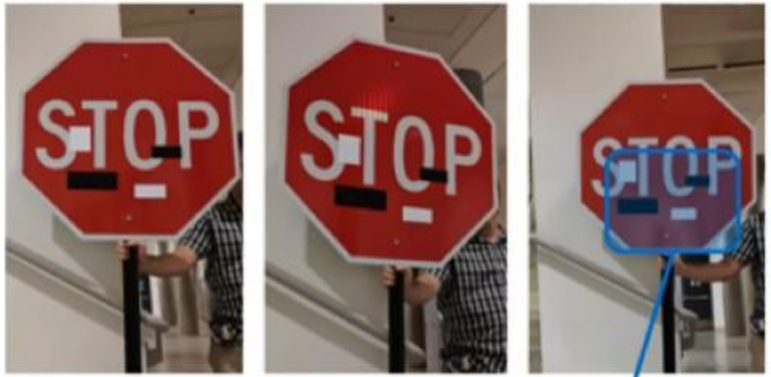
2 **AI vision system** incorrectly classifies to

LIMITED TO
45
MPH



Building Trustworthy AI System is Hard

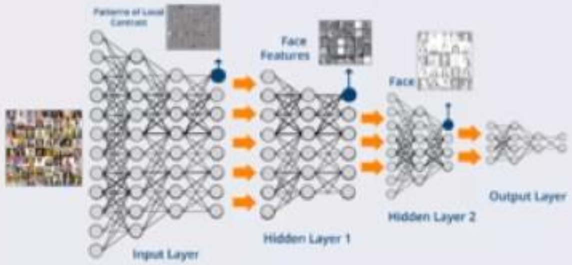
1 Attacker modifies signs



Adversarial perturbations

2 **AI vision system** incorrectly classifies to

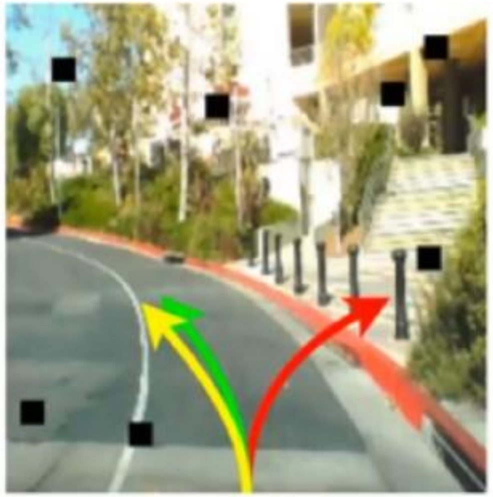
LIMITED TO
45
MPH



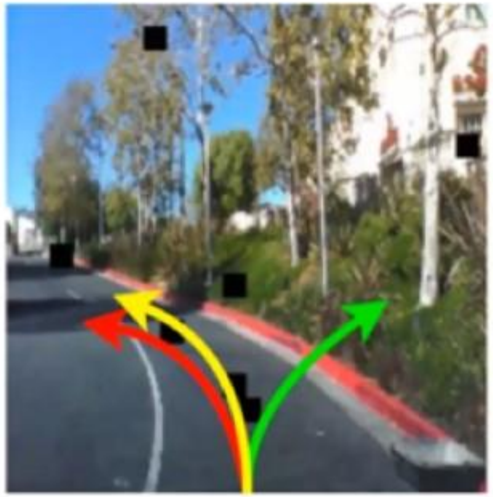
3 Car crash

Building Trustworthy AI System is Hard

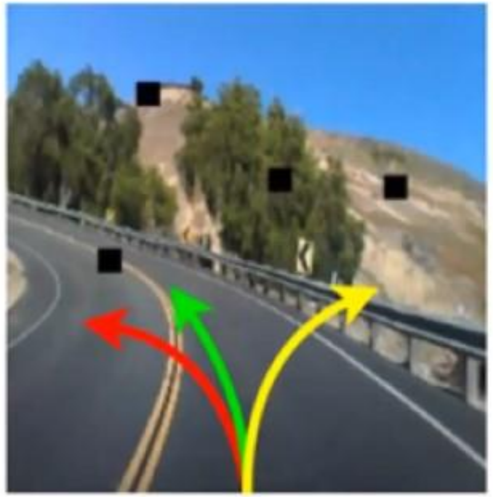
Self-driving car: in each picture one of the 3 networks makes a mistake...



DRV_C1: right



DRV_C2: right



DRV_C3: right

Pei, Kexin, Yinzhi Cao, Junfeng Yang, and Suman Jana. "Deepxplore: Automated whitebox testing of deep learning systems." In *proceedings of the 26th Symposium on Operating Systems Principles*, pp. 1-18. 2017.

Government Actions

EU: Ethics Guidelines for Trustworthy AI

https://ec.europa.eu/futurium/en/ai-alliance-consultation

"AI systems need to be reliable, secure enough to be resilient against both overt attacks and more subtle attempts to manipulate data"

"Explainability of the algorithmic decision-making process, adapted to the persons involved, should be provided to the extent possible"

Independent

High-Level Expert Group on Artificial Intelligence

set up by the European Commission

A Definition of AI: Main Capabilities and Disciplines

[illegible]

Apr 8, 2019

DARPA: Guaranteeing AI Robustness against Deception (GARD)

https://www.darpa.mil/attachments/GARD_ProposersDay.pdf

Develop theoretical foundations for AI robustness

Develop principled defenses

Guaranteeing AI Robustness against Deception (GARD)

Dr. Hava Siegelmann

I2O

Proposers' Day Brief

6 February 2019

Feb 6, 2019

Three Waves of AI

- **First Wave (up to early 2000's):**
  - Systems based on rules, deduction, typically handcrafted exact rules, based on logic, deduction and symbolic reasoning. Can explain why decision was taken (causality). Does not deal well with noise or uncertainty.

Expert system

Three Waves of AI

- **Second Wave (mid 2000's to now):**
  - Systems based on data and statistical learning, search, no human effort required, deals well with uncertainty. Hard time explaining their decisions, hard to ensure reliability and safety, limited logic.

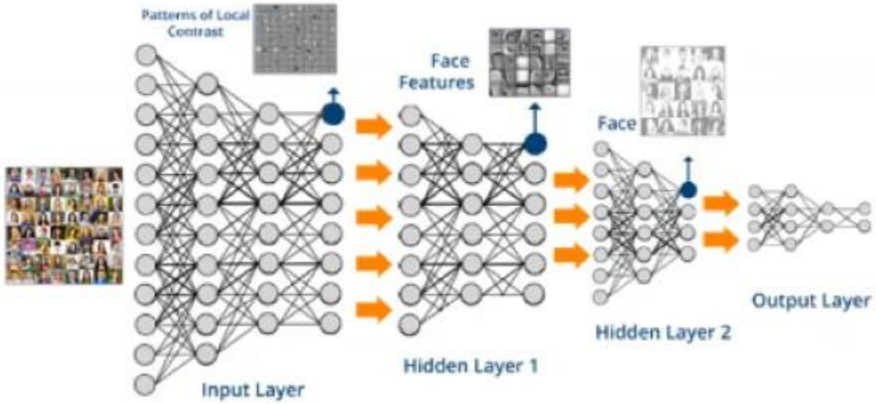


Image classification

Three Waves of AI

- **Third Wave of AI (today-?):**
    - Systems combine strengths of both approaches...can deal well with uncertainty, can explain decisions via logic, yet safe.

I decided to turn right because...?

Three Waves of AI

- **Third Wave of AI (today-?):**
  - Systems combine strengths of both approaches...can deal well with uncertainty, can explain decisions via logic, yet safe.

I decided to turn right because...?

This course aligns with this wave!!!

About the Course

- **AI Fundamentals:** Foundation and concepts of AI.
- **Adversarial Robustness:** Understanding and mastering the latest advancements in attacking and defending deep neural networks against adversarial examples.
- **Safety and Fairness:** Developing methods to automatically ensure and prove that AI models are safe, fair, and robust.
- **Interpreting Neural Network Behaviors:** Gaining insights into how neural networks operate and interpret their behaviors.

Marking Scheme

- 05 Marks - Attendance (75% attendance is must)
- 10 Marks - Test
- 15 Marks - Group Assignment
- 30 Marks - Minor
- 40 Marks - Major
- Total 100 Mark

Resources

- Christopher M. Bishop, Pattern Recognition and Machine Learning - Springer, 2nd edition
- Ian Goodfellow, Yoshoua Bengio, and Aaron Courville, Deep Learning MIT Press Ltd, Illustrated edition
- Wojciech Samek et al. Explainable AI: Interpreting, Explaining and Visualizing Deep Learning" by 2019 edition
- Research Papers

THANK YOU