Reliable and Trustworthy Artificial Intelligence

Lecture 1 [Part I]: Introduction, topics, organization

Martin Vechev

ETH Zurich

Fall 2022

SRILAB

http://www.sri.inf.ethz.ch

ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

whoami

Professor of Computer Science at ETH since January 2012

Sofia, Bulgaria

SFU, Canada, B.Sc.

Cambridge, England, PhD

Researcher @
IBM T.J. Watson Research Center, New York, USA

Professor at ETH Zurich,
Lead SRI: http://www.sri.inf.ethz.ch

Architect of INSAIT
https://insait.ai/

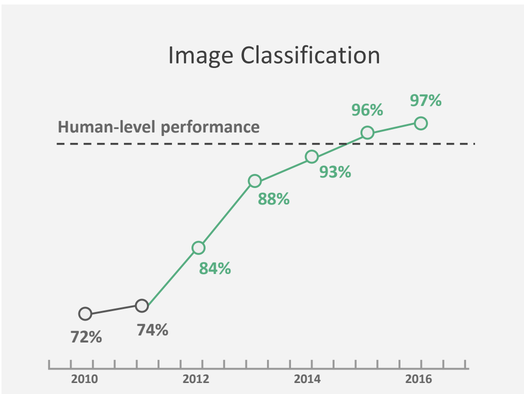
DEEPCODE — AI for Code

CHAINSECURITY — Security

LatticeFlow — Robust AI

Startup co-founder

Motivation



Focus on improving standard accuracy

Real-world AI requires guarantees beyond standard accuracy

Reliable and Trustworthy AI (this course)

This course: A glimpse into latest research

Course Breakdown: by areas

Robustness

attacks and defenses, certification (relaxations, branch and bound, certified training, smoothing), logic + deep learning

Privacy

attacks, differential privacy, secure synthetic data, data minimization, federated learning vulnerabilities

Fairness/Bias

individual fairness, group fairness, methods for building fair systems for tabular, NLP and visual data

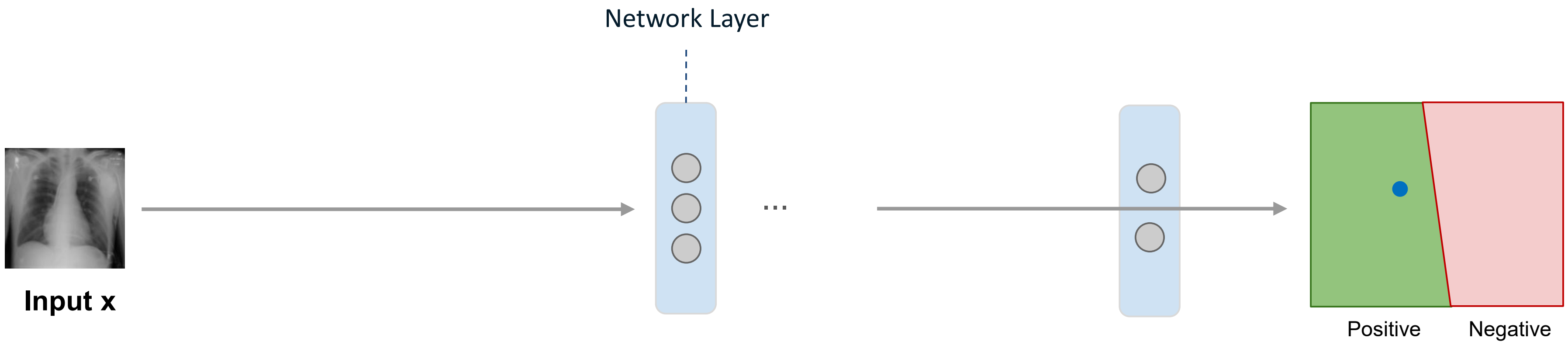
Common theme: provable mathematical guarantees for all of the above

Course Breakdown: by areas

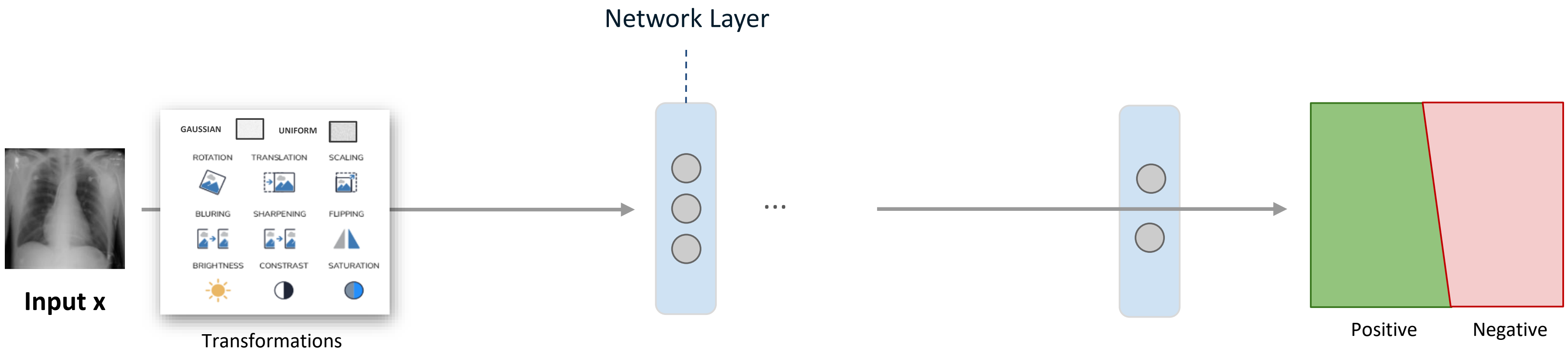
Robustness

attacks and defenses, certification (relaxations, branch and bound, certified training, smoothing), logic + deep learning

Why is it hard to certify robustness of AI models?

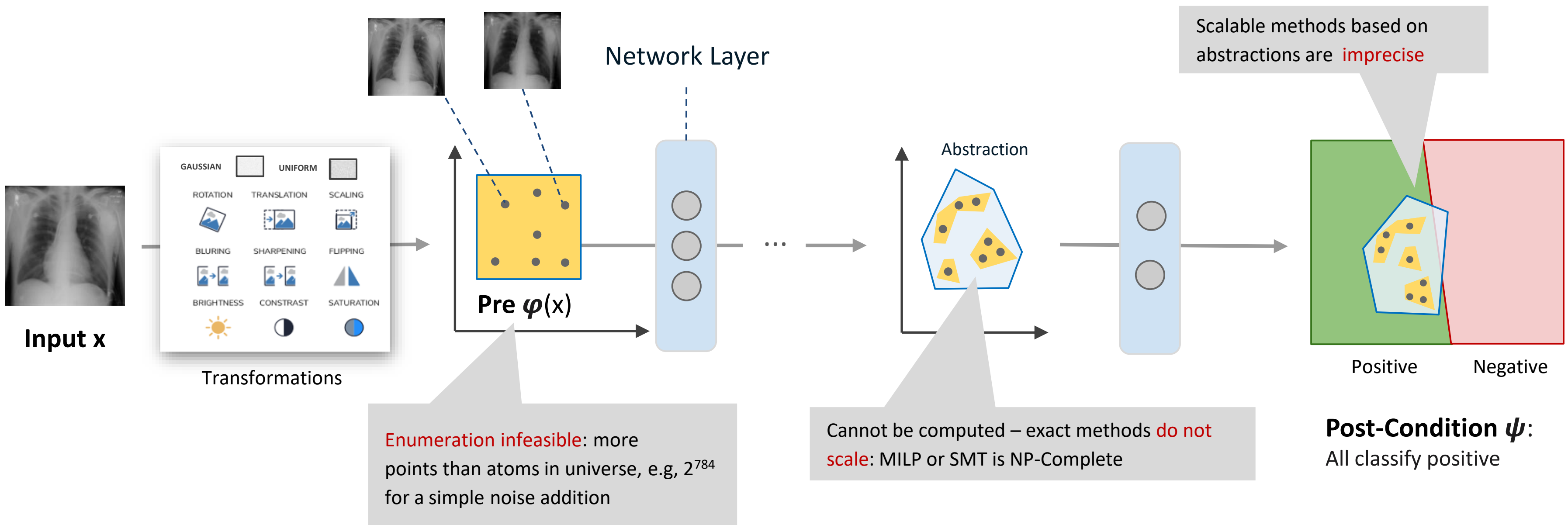


Why is it hard to certify robustness of AI models?



Goal: prove that image transformations do not change the classification

Why is it hard to certify robustness of AI models?



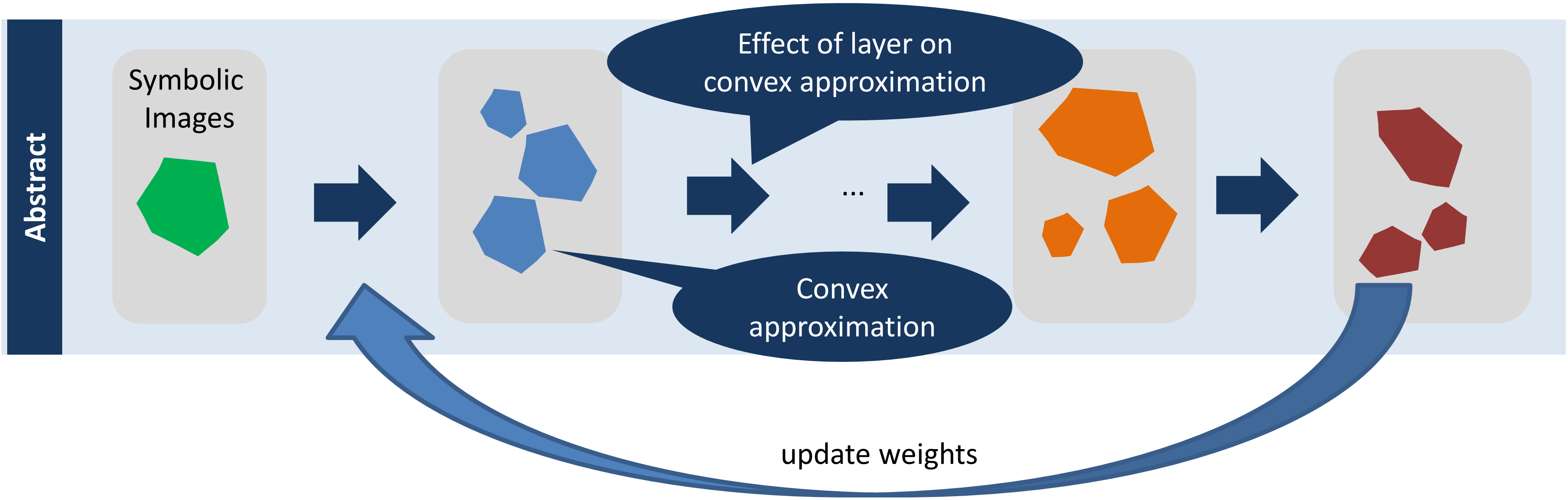
Beyond verification: Provable Defenses of Deep Models

However, an observation here is that if a network is not trained to be provably defended, then it can be difficult to prove properties about it.

Question: can we train the network to be more provable? How?

Provable Defenses of Deep Models: The Idea

Do propagation of convex shapes in the forward pass…must be fast!

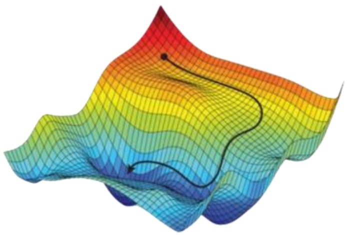


Do back propagation using the symbolic information 

Requires a new loss, which one? What happens to the standard loss?

Many technical parts needed to make this work well (e.g., annealing).

Why is it hard to train certified models?



Standard Training

$$\min_{\theta} E[loss(\theta, x, y)]$$

Training with Specification ($\boldsymbol{\varphi}$, $\boldsymbol{\psi}$)

$$\min_{\theta} E[\max_{x' \in \varphi(x)} loss_{(\varphi, \psi)}(\theta, x', y)]$$

- Intractable to compute exactly
- Current training methods unable to provably enforce specification (φ, ψ)

Training with specifications requires more complex optimization problems

Course Breakdown: by areas

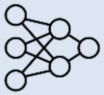
Privacy

attacks, differential privacy, secure synthetic data, data minimization, federated learning vulnerabilities

Example: ML Privacy Attacks

Model Stealing

- Given black-box access to model **f**, extract its weights
- Direct IP theft in case of proprietary models
- + Having a faithful copy of **f** allows to mount further attacks that require white-box access (e.g., variants of training data extraction, finding adversarial examples)

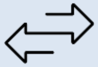


Membership Inference

- Given a target data record **x** and black-box access to model **f**, determine if **x** was used in model training
- Presence of a person may leak sensitive information
- + Useful as a measure of model leakiness, i.e., the risk a person incurs if they allow their data to be used

Model Inversion

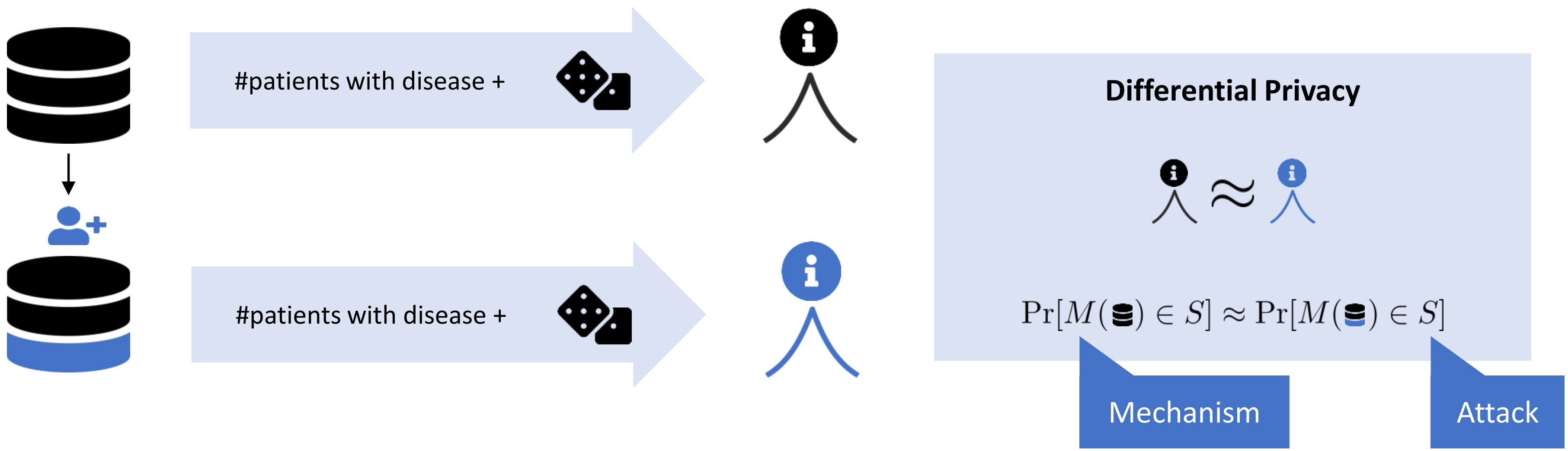
- Given black-box/white-box access to model **f**, extract representative inputs for a particular class
- Has direct privacy implications (if 1 class is only about 1 person)
- Still, does not leak actual training samples



Training Data Extraction

- For various threat models reconstruct actual samples **x** from the training dataset of **f**
- Strong, completely breaks the privacy
- Example: Large language model GPT-2 leaks training text!

Example: Differential Privacy



We cover: applications of DP in ML (DP with SGD, PATE, DP with Federated Learning, DP with Synthetic Data Generation)

We do not cover: cryptographic constructions to protect data in ML (FHE, MPC); see our recent paper:

Private and Reliable Neural Network Inference, ACM CCS'22

Example (DP-SGD): Here, *M* is the training process (DP-SGD) and *S* is a set of possible weights. With DP we guarantee that the probability model weights are in S is close to the probability trained on similar data => we cannot recover membership of a data point from the output weights.

Course Breakdown: by areas

Fairness/Bias

individual fairness, group fairness, methods for building fair systems for tabular, NLP and visual data

Why fairness and bias?

ML makes decisions that **impact people:**

- Should person get a loan?
- Is person likely to commit a crime?
- Should person get hired?

The European Commission is creating regulations with a goal that AI systems "do not create or reproduce bias“.

A.I. Could Worsen Health Disparities

In a health system riddled with inequity, we risk making dangerous biases automated and invisible.

The never-ending quest to predict crime using AI

The practice has a long history of skewing police toward communities of color. But that hasn't stopped researchers from building crime-predicting tools.

SCIENCEINSIDER | EUROPE

Europe plans to strictly regulate high-risk AI technology

How AI Is Deciding Who Gets Hired

Employee advocates say hiring software is making discrimination worse. But some applicants are hacking the system.

Tabular data

Age	Salary	Loan
37	85K	True
26	60K	False
52	100K	True

Vision

NLP

The first is a training problem. A.I. must learn to diagnose disease on large data sets, and if that data doesn't include enough patients from a particular background, it won't be as reliable for them. Evidence from other fields suggests this isn't just a theoretical concern. A recent study found that some facial recognition programs incorrectly classify less than 1 percent of light-skinned men but more than one-third of dark-skinned women. What happens when we rely on such algorithms to diagnose melanoma on light versus dark skin?

Medicine has long struggled to include enough women and minorities in research, despite knowing they have different risk

Key challenges we study:

How to **define** fairness?

How to **enforce** fairness?

How to **prove** fairness?

Example: Defining Fairness

Individual fairness

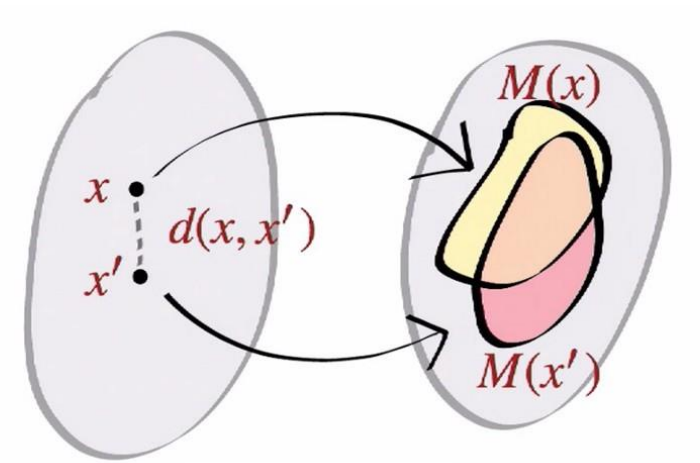
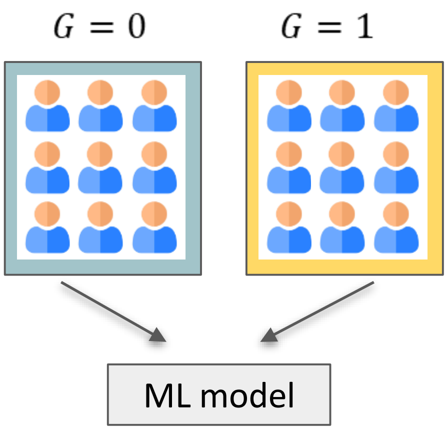


Image source: Moritz Hardt

Group fairness



$$P(Y = 1|G = 0) = P(Y = 1|G = 1)$$

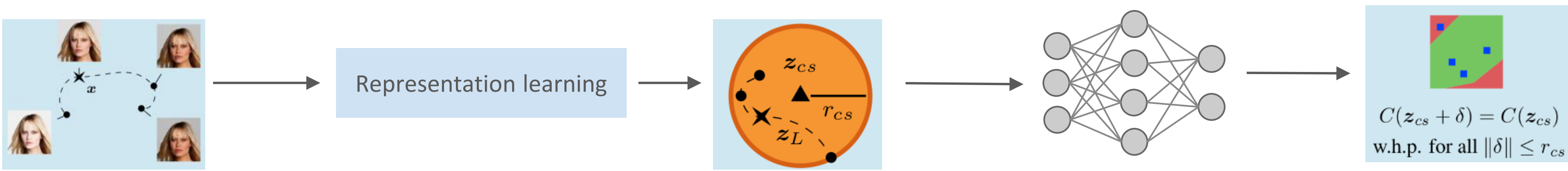
Requires that if two individuals x and x' are similar (according to some similarity notion), decisions of ML model $M(x)$ and $M(x')$ should be similar for these two individuals.

Key challenge: How to find a suitable similarity metric d (e.g. some norm in feature space)?

Requires the probability an ML model assigns a label to different groups is the same (e.g. groups can be different races). Variants of group fairness differ in the way groups are formed: demographic parity, etc.

Key challenge: How to define groups?

Example: Enforcing individual fairness



1. Use generative model to capture the set of images similar to **x**

2. Use smoothing to guarantee that representations of similar individuals **get mapped to similar representations** *with high probability*

3. Use smoothing to guarantee that similar representations **get classified the same** *with high probability*

This course should help you create new science

Ultimate goal: create new science.

The material in the course is structured in a way where after each lecture, you can think creatively about new ideas in the space, and doing your own research. We want to teach you the key ideas and concepts but with the purpose that you think **creatively** and **critically**.

You do not have to know all the material to be creative!

Try to be creative as soon as you learn one part.

Student in Course → New Science

M.Sc. Thesis, Research in CS, Research in Data Science Students who took course

Sample list from last 2 years, we list 1 paper per person but some have more than 1:

- Claudio Ferrari: Complete Verification via Multi-Neuron Relaxation Guided Branch-and-Bound, ICLR’22
- Robin Staab: Bayesian Framework for Gradient Leakage, ICLR’22
- Nikola Jovanovic: Certified Defenses: Why Tighter Relaxations May Hurt Training, TMLR’22
- Christian Sprecher: Shared Certificates for Neural Network Verification, CAV’22
- Miklos Horvath: Boosting Randomized Smoothing with Variance Reduced Classifiers, ICLR’22 (Spotlight)
- Anian Ruoss: Latent Space Smoothing for Individually Fair Representations, ECCV’22
- Chengyuan Yao: Automated Discovery of Adaptive Attacks on Adversarial Defenses, NeurIPS’21
- Alexander Hägele: Robustness Certification with Generative Models, PLDI’21
- Gregory Bonaert: Fast and Precise Certification of Transformers, PLDI’21.
- Mark N. Mueller: Boosting Certified Robustness with Compositional Architectures, ICLR’21
- Tobias Lorenz: Robustness Certification for Point Cloud Models, ICCV’21
- Wonryong Ryou: Scalable Polyhedral Verification of Recurrent Neural Networks, CAV’21
- …

Let me know if interested to do research internship, project, thesis, etc. in this space

Course Project

- The course project will be about verification of neural networks.
- The project be advertised by the end of October in a special lecture.
- The project will be done in Python in groups of 2
- The project will be automatically graded.
- 2 TA's are going to be involved with the project.

What this course aims to do

- Introduce you to some of the latest and most important research in A.I. as related to safety and reliability

- Convey core and general concepts, with a focus on applying the concepts in a system building project

- Introduce open research problems in the area and enable you to contribute, be creative and formulate new tasks

- Many students who took the course and did follow-up research (e.g., M.Sc. Thesis, Research in CS, Research in Data Science) ended up with top publications (e.g., in ICLR, ICML, PLDI, ICCV, ECCV, NeurIPS, etc.).

What this course is not

- It does not cover how to design neural nets to solve vision or robotics tasks (though we look at such networks). There are already such courses at ETH.

- This is not a course on gradient-based optimization algorithms. Such a course already exists at ETH.

- It is not an introductory course to Deep Learning or Python.

Course Organization

Grading

- 70% final written exam (make sure you do the homework)
- 30% course project (groups of two)

Course web site:

https://www.sri.inf.ethz.ch/teaching/rtai22

All information posted there: lectures notes, exercises, Q&A, etc.

Exercises

- Every week, we will publish an exercise sheet and its solutions on the course webpage.

- The exercise session will consist of a discussion of selected exercises (typically not all exercises). On demand, the teaching assistant can also discuss questions on specific exercises brought up by students.

- Some exercise sessions will also discuss prerequisites for the course.

- We strongly recommend to solve the exercises before next week's exercise session, and before looking at the solutions. The style of the exam will be similar to the exercises, so first-hand experience solving exercises is critical.

Course in 2022

- This is the sixth installment of the course
- Updates in 2022
  - New lectures: privacy, fairness, bias, federated, regulations, deep-tech perspective from industry, also lecture from DeepMind
  - Course explicitly structured into some of the latest trends: certification, privacy, fairness.

We aim to keep the course up-to-date, which can be very challenging ☺