Reliable and Interpretable Artificial Intelligence

Martin Vechev

ETH Zurich

Fall 2020

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

DEEPCODE AI for Code

CHAINSECURITY Security

LatticeFlow Safe AI

Co-founder

Today

Motivation for material

What will we learn

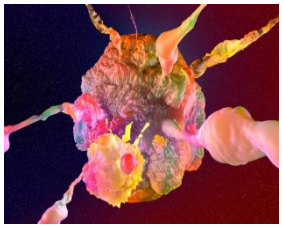
Course organization

AI Disruptions

Autonomous Driving

Medicine

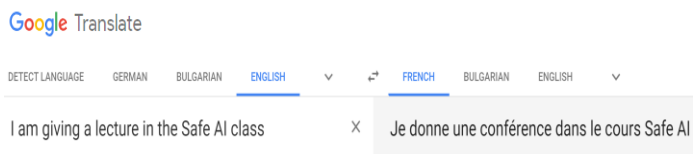
Algorithms Can Now Identify Cancerous Cells Better Than Humans



Game Playing

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

Natural Language Understanding

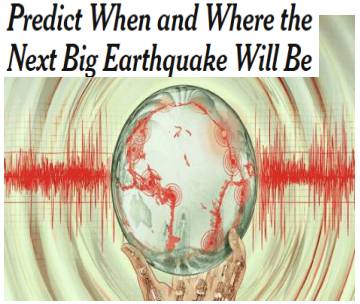


Fraud Prevention

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



AI Disruptions

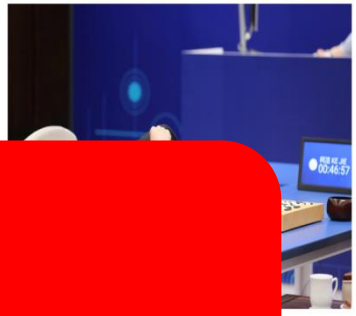
Autonomous Driving

Medicine

Algorithms Can Now Identify Cancerous Cells Better Than Humans

Game Playing

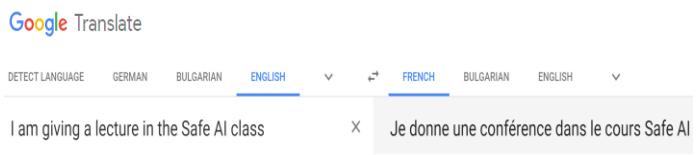
Google's A.I. Program Rattles Chinese Go Master as It Wins Match



But there are problems…

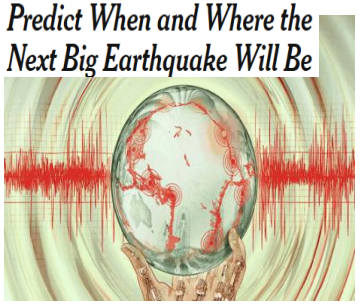
Natural Langu

diction



How AI is transforming the fight against money laundering

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

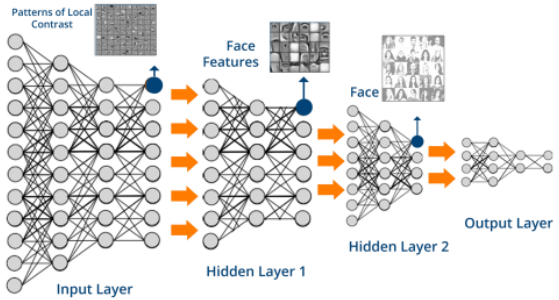


Building Reliable AI Systems is Hard

What animal do **you** see?

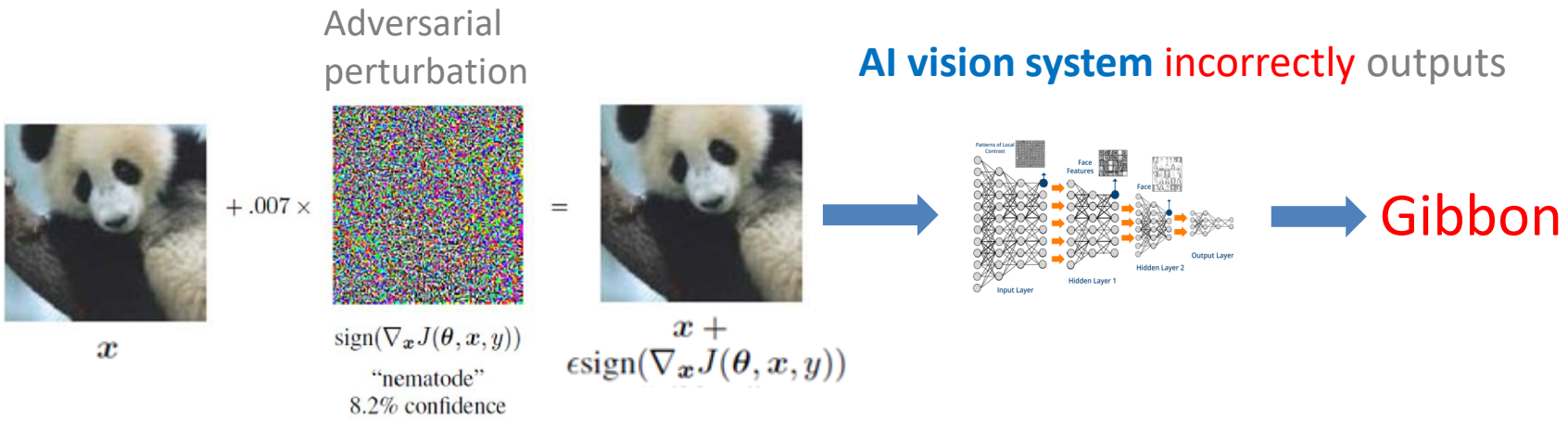
Building Reliable AI Systems is Hard

AI vision system correctly outputs



Panda

Building Reliable AI Systems is Hard



Explaining and Harnessing Adversarial Examples, ICLR '15

Building Reliable AI Systems is Hard

1 Attacker modifies signs

What sign do **you** see?

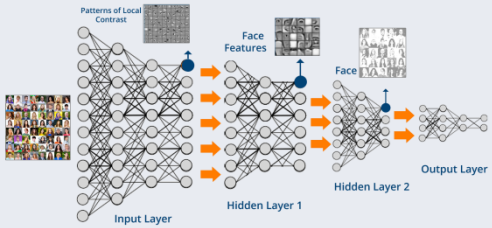
Robust Physical-World Attacks on Deep Learning Visual Classification, CVPR'18

Building Reliable AI Systems is Hard

1 Attacker modifies signs

Adversarial perturbations

2 **AI vision system** incorrectly classifies to



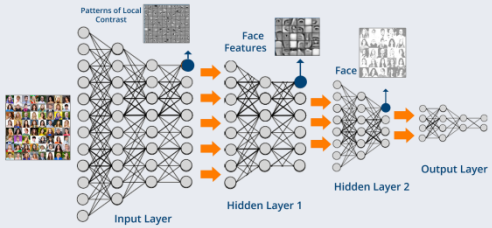
Robust Physical-World Attacks on Deep Learning Visual Classification, CVPR'18

Building Reliable AI Systems is Hard

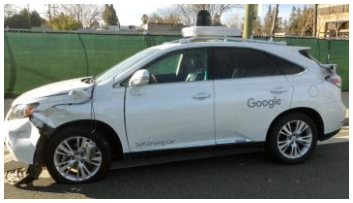
1 Attacker modifies signs

Adversarial perturbations

2 **AI vision system** incorrectly classifies to



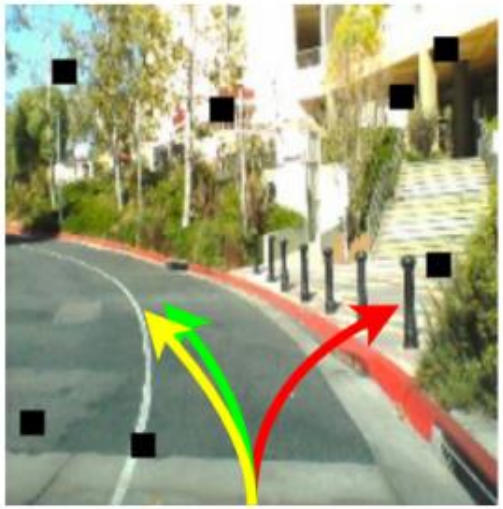
3 **Car crash**



Robust Physical-World Attacks on Deep Learning Visual Classification, CVPR'18

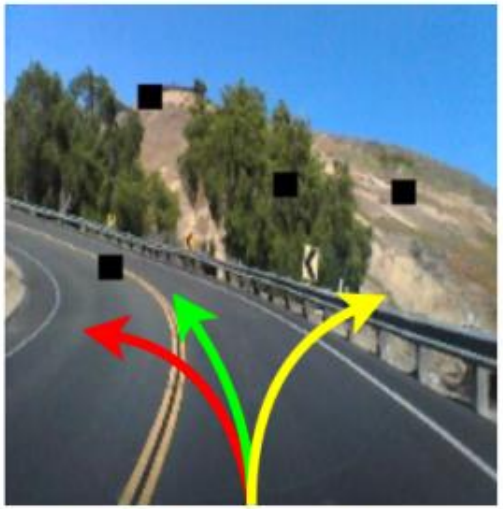
Building Reliable AI Systems is Hard

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



DRV_C1: right

DRV_C2: right



DRV_C3: right

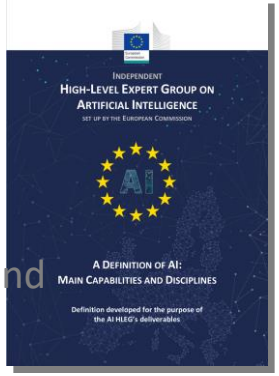
DeepXplore: Automated Whitebox Testing of Deep Learning Systems, SOSP'17

Government Action

First steps by the US and the EU towards regulation of AI systems

EU: Ethics Guidelines for Trustworthy AI

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



Apr 8, 2019

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

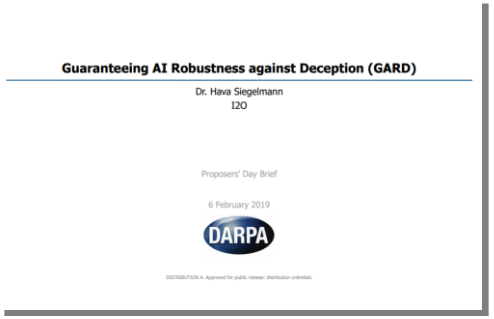
"Explainability of the algorithmic decision-making process, adapted to the persons involved, should be provided to the extent possible"

DARPA: Guaranteeing AI Robustness against Deception (GARD)

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

Develop theoretical foundations for AI robustness

Develop principled defenses



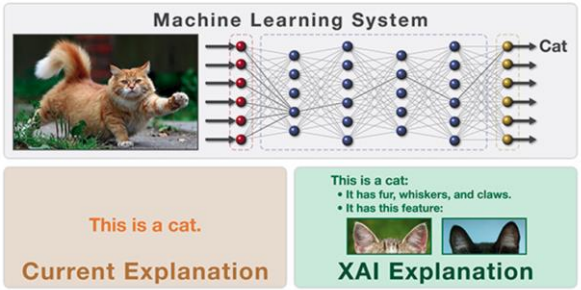
Feb 6, 2019

DARPA DEFENSE ADVANCED RESEARCH PROJECTS AGENCY ABOUT US / OUR RESEARCH /

Defense Advanced Research Projects Agency › Program Information

Explainable Artificial Intelligence (XAI)

Mr. David Gunning



2,067 views | Sep 7, 2018, 07:10pm

DARPA Plans To Spend $2 Billion Developing New AI Technologies

Intelligent Machines

The Dark Secret at the Heart of AI

No one really knows how the most advanced algorithms do what they do. That could be a problem.

How well can we get along with machines that are unpredictable and inscrutable?

European Union regulations on algorithmic decision-making and a "right to explanation"

Bryce Goodman, Seth Flaxman

(Submitted on 28 Jun 2016 (v1), last revised 31 Aug 2016 (this version, v3))

We summarize the potential impact that the European Union's new General Data Protection Regulation will have on the routine use of machine learning algorit effect as law across the EU in 2018, it will restrict automated individual decision-making (that is, algorithms that make decisions based on user-level predictors affect" users. The law will also effectively create a "right to explanation," whereby a user can ask for an explanation of an algorithmic decision that was made al that while this law will pose large challenges for industry, it highlights opportunities for computer scientists to take the lead in designing algorithms and evaluati avoid discrimination and enable explanation.

Comments: presented at 2016 ICML Workshop on Human Interpretability in Machine Learning (WHI 2016), New York, NY
Subjects: **Machine Learning (stat.ML)**; Computers and Society (cs.CY); Learning (cs.LG)
Cite as: arXiv:1606.08813 [stat.ML]

DARPA Is Funding Research Into AI That Can Explain What It's "Thinking"

Artificial intelligence pioneer says we need to start over

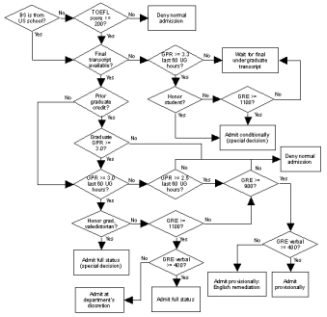
"My view is throw it all away and start again"

"I don't think it's how the brain works," he said. "We clearly **don't need all the labeled data**."

"The future depends on **some graduate student** who is deeply suspicious of everything I have said."

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

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

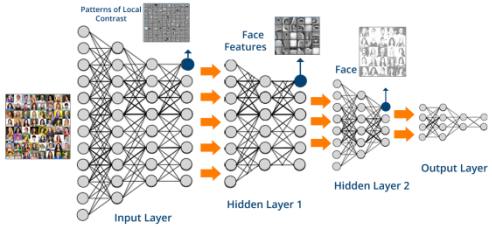


Image classification

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…?

A DARPA Perspective on Artificial Intelligence: https://www.youtube.com/watch?v=-O01G3tSYpU

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.

Symbolic Reasoning

- Logic
- Deduction
- Modularity
- Abstraction
- Compositionality

$$\frac{\Gamma \vdash e_1 \qquad \Gamma \vdash e_2}{\Gamma \vdash e_1 + e_2} \qquad \frac{}{\Gamma \vdash x = e}$$

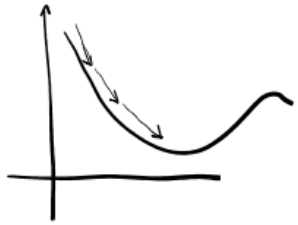
fold(_, _, _)

map(_, _, _) cons(_,_)

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

Machine Learning

- Optimization
- Probability
- Data Driven



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

Symbolic + Probabilistic

Many of the developments here are still in its infancy…the course material is an instance of this direction.

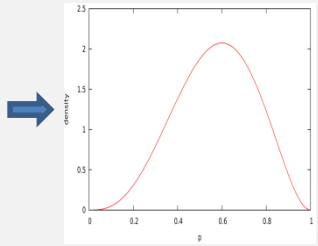
3rd Wave of AI in our lab (sample)

Probabilistic Programming [psisolver.org]

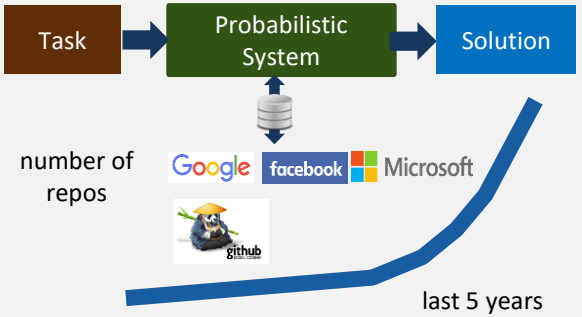
Probabilistic Program

```
def main() {
 p := Uniform(0,1);
 r := [1,1,0,1,0];
 for  i in [0..r.len]
  observe(Bern(p) == r[i]);
 return p;
}
```

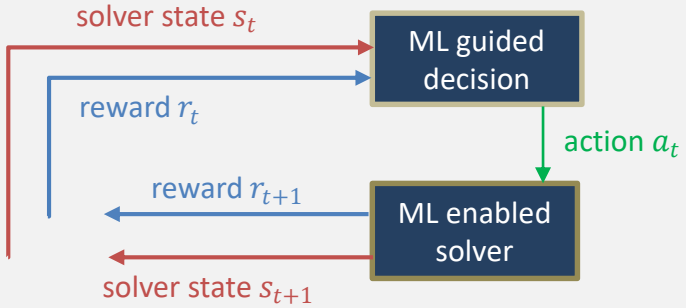
Probability Density



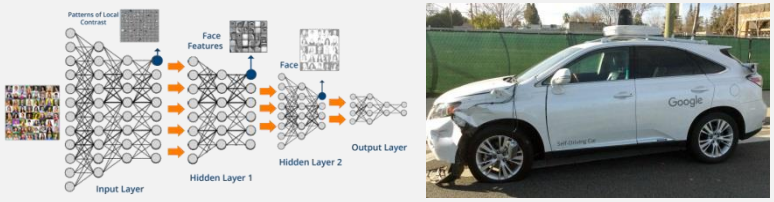
ML for Big Code [deepcode.ai]



ML-guided Solvers [fastsmt.ethz.ch]



Trusted Artificial Intelligence [safeai.ethz.ch]



The RIAI course covers this direction

This course: A glimpse into latest AI research…

The focus of the course will be on understanding and mastering the latest advances in attacking and defending deep neural networks against adversarial examples, ways to automatically ensure and prove deep learning is safe, fair and robust, combining symbolic logic and neural networks for querying and training in order to incorporate background knowledge, as well as interpreting behaviors of neural networks.

Let us go over (at a very high level) the sequence of topics as they will be presented in the lectures and their motivation…

We will start with attacks on deep neural networks

Adversarial Attacks on Deep Models

x

"panda"
57.7% confidence

$+ .007 \times$

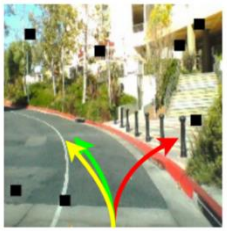
$\text{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, x, y))$

"nematode"
8.2% confidence

$=$

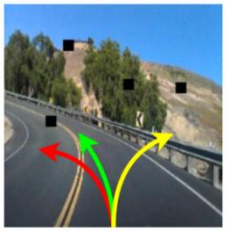
$x + \epsilon\text{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, x, y))$

"gibbon"
99.3 % confidence



DRV_C1: right

DRV_C2: right



DRV_C3: right

Adversarial Attacks on Deep Models

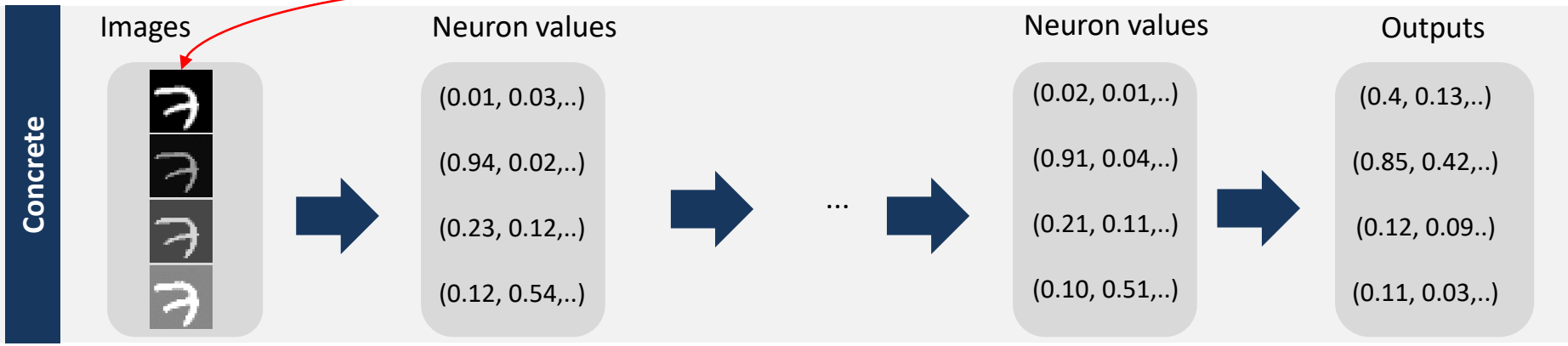
Here, we will see various examples of adversarial attacks, including geometric perturbations of images, attacks on NLP models, audio processing, and others.

Technically, we will study how to perform these attacks, in particular, using gradient based methods, as well as study how to defend the neural network by training it to be less susceptible to attacks. In the process, we will discuss fundamental trade-offs between accuracy and experimental robustness.

Experimental robustness is desirable, but can we mathematically prove once and for all that deep models cannot be attacked?

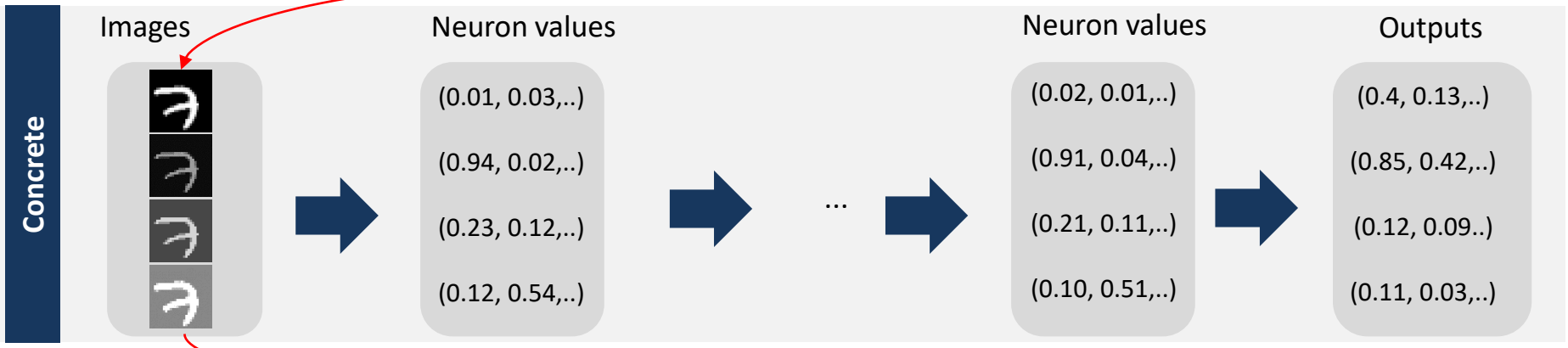
Automated Verification of Deep Models

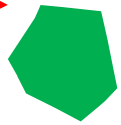
An attack can generate an infinite number of images…we cannot enumerate…



Automated Verification of Deep Models

An attack can generate an infinite number of images…we cannot enumerate…

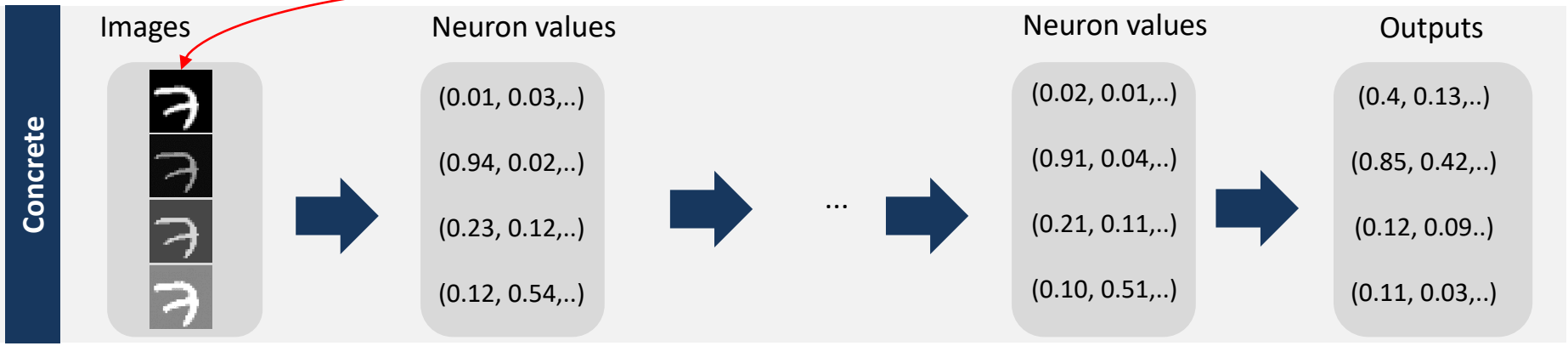


Can we summarize these somehow into some symbolic, finite region 

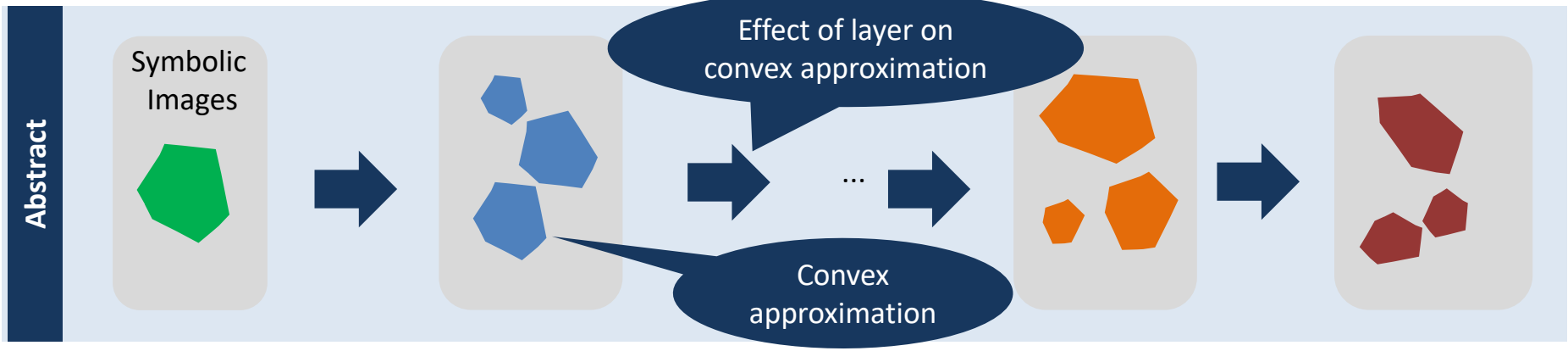
and work with that summary  instead?

Automated Verification of Deep Models

An attack can generate an infinite number of images…we cannot enumerate…



Automatically reason about an infinite set of all possible images at once!



Automated Verification of Deep Models

Verifying properties of deep models is non-trivial and requires carefully designed convex approximations. A key question here is scalability vs. approximation: too much approximation means the verifier will fail to prove valid properties, too little means it won't scale to large networks.

We will study several state-of-the-art convex approximations (introduced during the last year) to verify deep networks. We will also study how to combine these with MILP (Mixed-Integer Linear Solvers) and how to verify end-to-end robustness against audio, geometric and norm-based attacks.

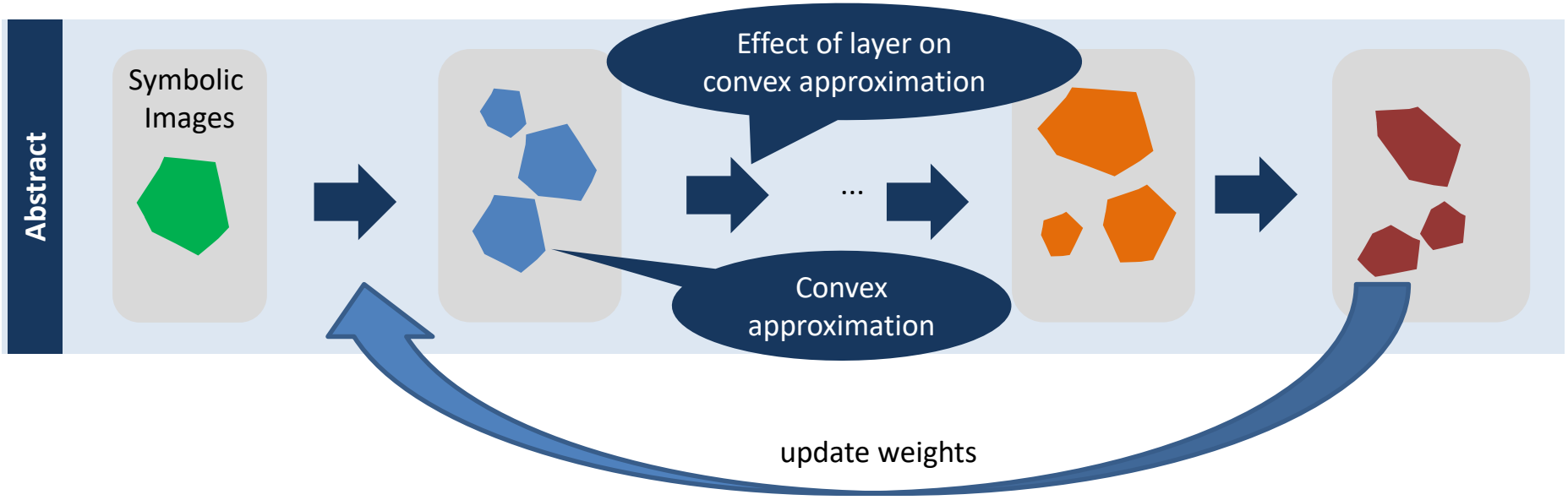
Beyond verification:
Provable Defenses of Deep Models

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).

Provable Defenses of Deep Models: (few) fundamental Questions

Current provable defenses improve provability with special training, but also negatively affect accuracy.

Open hard problems:

Can we train a model that is both accurate and provable? Can we prove it is always possible to find such a network? What is the connection between the convex approximation, accuracy and provability?

Course Project

- The course project will be about verification of neural networks.
- The project be advertised at the end of October.
- We will have a 45 min lecture introducing the project.
- 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.

We will be exploring topics beyond attacks, defenses and provability of deep neural networks.

In particular, we will be exploring the connection between learning and logic as well as the connection between differential privacy and statistical robustness guarantees.

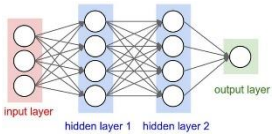
Combining Logic and Deep Learning

As Deep Learning makes more and more decisions (e.g: bank credit, job applications, university admissions, political elections), it becomes critical to be to understand how these decisions can be influenced and understood.

Query

What should the applicant change to receive a bank credit?

Neural Network NN



Rejected for credit by NN

- income
- residence
- conditions
- job
- …

Deep Learning Query Engine

To be in the > 84% probability of receiving credit, increase income by at least 5K and be employed for at least 3 more months…

Combining Logic and Deep Learning

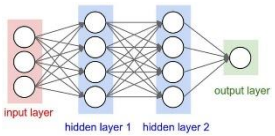
Adversarial examples are in fact just a special case of a query…

Query

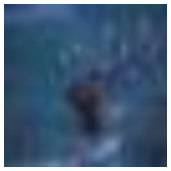
```
find i[32, 32, 3]
where i in [0, 255],
    class(NN(i)) = 9,
    ||i - deer||∞ < 25,
    ||i - deer||∞ > 5
```

Find an image i which gets classified to 9 (truck) where the image i is within some distance of the image deer.

Neural Network NN

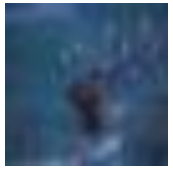


deer



Deep Learning Query Engine

image i



classified as truck by NN!

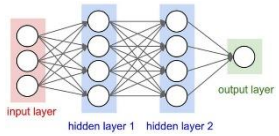
Combining Logic and Deep Learning

Comparing neural networks

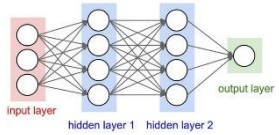
```
find i[28, 28]
where i in [0, 1],
    i[0:9,:] = nine[0:9,:],
    class(NN1(i)) = 8,
    class(NN2(i)) = 9
```

Find an image i which gets classified to 8 with network 1 and to 9 with network 2, such that pixels in row 0:9 of image i are the same as image nine

Network NN1



Network NN2



Deep Learning Query Engine

image i

Image nine

Combining Logic and Deep Learning

Adversarial examples are in fact just a special case of a query…

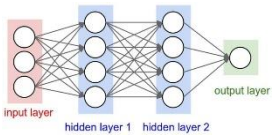
Hmm, but this involves logical constraints and a neural network!

How can we unify these?

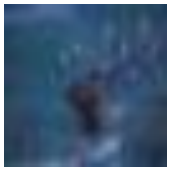
Query

```
find i[32, 32, 3]
where i in [0, 255],
    class(NN(i)) = 9,
    ||i - deer||∞ < 25,
    ||i - deer||∞ > 5
```

Neural Network NN

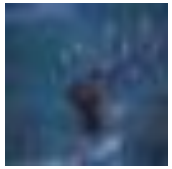


deer



Deep Learning Query Engine

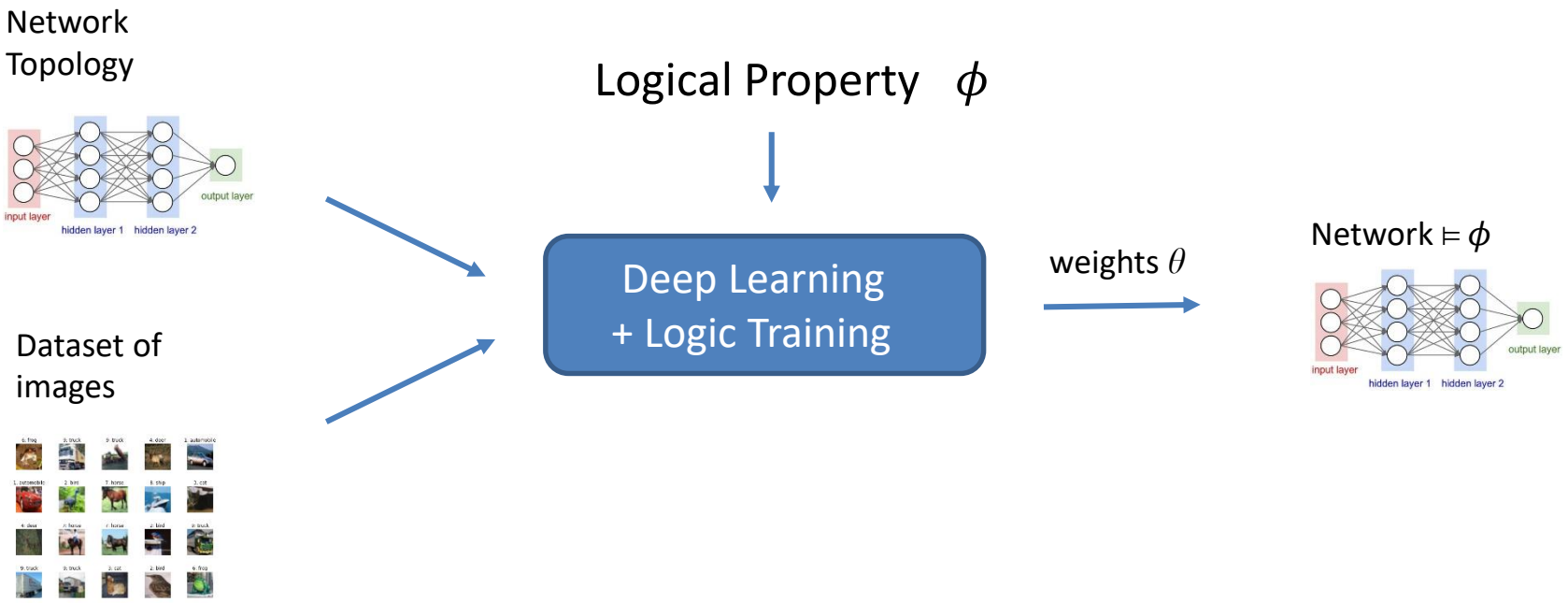
image i



classified as truck by NN!

Combining Logic and Deep Learning

We can also train neural networks to satisfy a logical property



In fact, this can help accuracy as we can label part of the data and specify properties on the remaining, unlabeled data.

Key Questions to Study

How do we reconcile logic with differentiable gradient-based optimization? What are the guarantees?

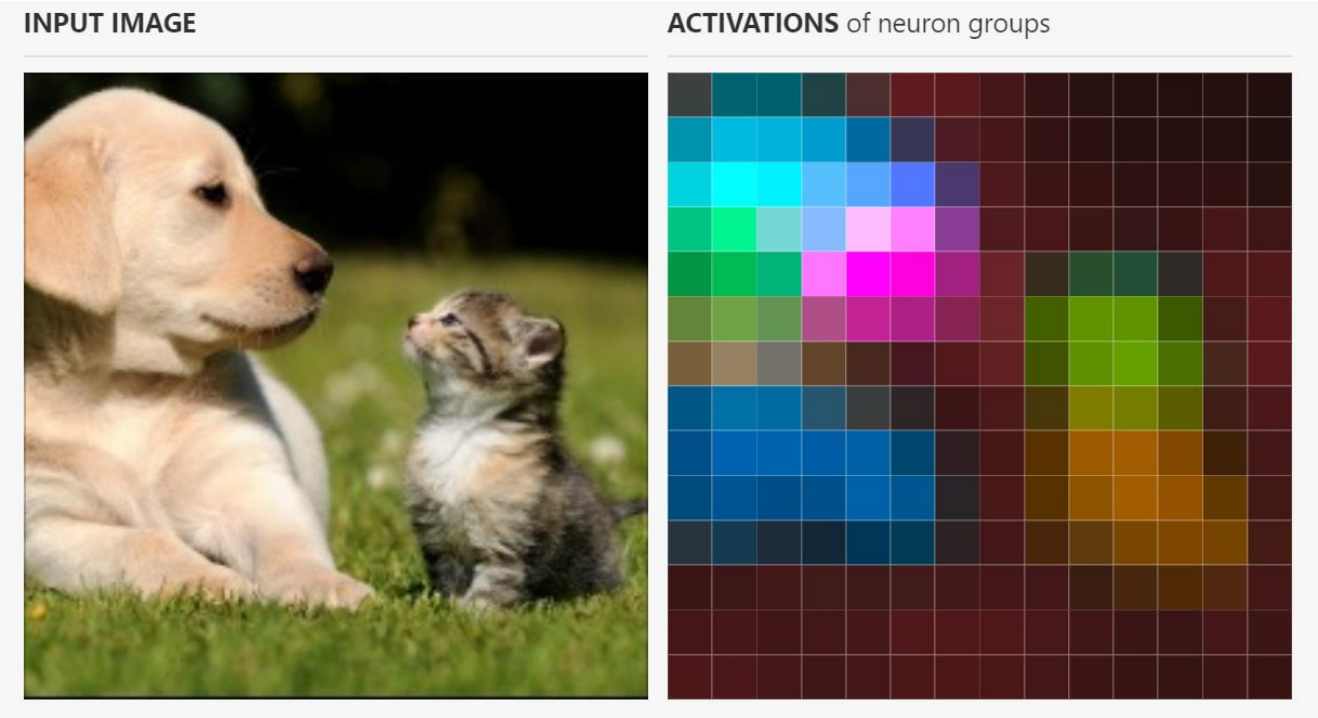
What kind of properties can one enforce? How is enforcement done?

How are counter-example inputs to the property found?

What is the connection between the optimization method and the logic fragment? What is the asymptotic complexity of the projection?

Understanding Deep Learning

In addition to interacting and training the network with logic, we may want to more deeply examine how individual network neurons work.



In particular, we will look at the Distill line of work from Google Brain/DeepMind

So far: (very) high level preview of technical material

Next: Course Details

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 and formulate new tasks

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/riai2020

All information posted there: lectures notes, exercises, etc.

Exercises

- Come with questions/solutions
  - The TA will go over the homework

- Exercises are not graded, but if you do them, you will have an advantage on the final exam

Feedback from RIAI 2018 and 2019

”…The course is great, I don't see any major thing that should be changed…The course content is amazing…I really love this course… This course has been different from the other courses I have studied so far. The fact that the course dealt with topics that are pretty recent made it very interesting…I would generally keep most of it unchanged…The lectures are very interesting…”

Topics are fun…but…the course had ~30% failure rate, so think before taking it ☺

RIAI 2020

- This is the fourth installment of the course (2020, 2019, 2018, 2017)

- Improvements in 2020:
  - Course used to be 5 credits in 2019, now its 6 to reflect the amount of content, project and work.
  - Additional lectures including fairness, bias and provability.
  - Exam time is now 120 min and not 90 min (main complaint in 2019)

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

Sample Research from ETH RIAI students

M.Sc. Thesis, Research in CS, Research in Data Science

- *Anian Ruoss*: neural verifier for vector fields transformations. Submission in progress.
- *Joel Oskarsson*: provably robust training for geometric perturbations. Semester project.
- *Wonryong Ryou*: verifier for audio perturbations. M.Sc. Thesis.
- *Jiayu Chen*: verifier for sigmoid networks (with ABB Research). M.Sc. Thesis.
- *Rupanshu Ganvir*: new k-ReLU convex approximations. Research in CS.
- *Mark Muller:* neural architectures for provability. M.Sc. Thesis.
- *Christian Sprecher*: proof transfer for certified networks. M.Sc. Thesis.
- *Nicola Jovanovic:* theoretical considerations of provable training. Research in CS.
- *Samriddhi Jain:* interpretable reinforcement learning. Research in CS.
- *Shubhangi Ghosh:* Relational geometric verification. Research in Data Science.
- *Chengyuan Yao:* learning adversarial attacks. MSc. Thesis
- More not listed here…

The course enables students to participate and develop new research.

If you are interested, let me know early.

Next lecture: Adversarial attacks and defenses