CS-695/SWE-699: AI Safety and Assurance

Fall 2023

Instructor:	ThanhVu Nguyen	**Meetings:**	T 4:30PM – 7:10PM
Email:	tvn@gmu.edu	**Place:**	Innovation Hall 215G

Description

This special topics course is a research seminar on **AI Verification and Analysis**. The course will focus on active research areas in formal AI reasoning, but the specific topics will be largely determined by a combination of instructor fiat and the interests of the students.

Prerequisite

- No prerequisite courses (but knowledge in linear algebra is strongly recommended)
- Programming knowledge (preferably Python): capable of developing a project of roughly about 1000 lines of code.

Grading

You will be evaluated based on

1. weekly reading assignments and discussions,
2. a programming project

Reading and Discussion (Reading Assignments)

We will read papers covering various topics including the applications of verification, testing, analysis, constraint solving, and abstraction techniques to Deep Neural Networks such as Feedforward Neural Networks (FNNs), Residual Networks (ResNet), Convolutional Neural Networks (CNNs), and Recurrent Neural Networks (RNNs).

On average, we will discuss **two papers** a week (about an hr for each paper). You are responsible for reading the assigned papers in advance for any given discussion.

Each student will be assigned to **lead a paper** that talks about a technique

- You will present **in depth** the paper you read
  - You will need to do a Powerpoint presentation style
  - This include providing concrete examples illustrating how the technique
- You will guide the discussion
- **Tools**: Usually each research paper has a free implementation tool. I will take into account when evaluating if you try out the tool and discuss some interesting things that it can or cannot do (e.g., try the the tool on some small but nontrivial examples). This will help you understand the readings better and give you ideas on how to use existing tools in your own work.

At the beginning of **each paper** discussion I will choose at random up to **three students** who is **not** the paper.

- Each student will give a **five-minute** presentation that at least talk about

1. the problem (what is it? why is it interesting?)
2. existing approaches (what are they? what are their limitations?)
3. the proposed technical approach (also talk about the strengths of approach and how the approach addresses the weaknesses of existing works)
4. limitations of the proposed approach and lists ways in which it might be improved.

- The goals of this approach are to encourage all participants to read the material thoroughly in advance, to provide jumping-off points for detailed discussions, and to allow me to evaluate participation.

Project

Each student will understand in depth (or own) a DNN analysis technique. This technique is from one of the papers you lead. Your project consists of 2 parts

Example Illustration

- You will write a "blog" describing full concrete example illustrating the DNN technique assigned to you. This example will be due **2 weeks** after the paper is read.
- Format
  - Written in **Markdown**
  - Posted on the class's wiki
  - Consist of a full illustration on how the technique works on a given example (e.g., how Reluplex works on a simple DNN).

Programming Project (Project Info)

- You will implement the DNN analysis technique using Python.
  - You will be given some example code in Python.
  - If you prefer a different language, talk to me first.
- In your blog entry above, you will write how you apply your implementation to the example illustration you had.

Honor Code

As with all GMU courses, this class governed by the GMU Honor Code. In this course, all assignments carry with them an implicit statement that it is the sole work of the author.

Learning Disabilities

Students with learning disabilities (or other conditions documented with GMU Office of Disability Services) who need academic accommodations should see me and contact the Disability Resource Center (DRC) at (703) 993-2474. I am more than happy to assist you, but all academic accommodations must be arranged through the DRC.

Links

- Schedule and Assignments
- Project Info

Relation courses

- Eth Zurich Reliable and Trustworthy Artificial Intelligence