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

Fall 2023

Meetings:	Tues 4:30PM – 7:10PM	**Place:**	Innovation Hall 215G
Instructor:	ThanhVu Nguyen	**Email:**	tvn@gmu.edu
Office Hr:	Tues 11:00AM – 12:00AM (email to confirm)	**Place:**	ENGR 4430
TA:	Dhiman Goswami	**Email:**	dgoswam@gmu.edu
Office Hr:	Fri 10:00AM – 11:00AM	**Place:**	ENGR 4456

1 Description

This special topic course is a research seminar on **AI Verification and Analysis**. We will learn various AI verification 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).

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 (Python)

2 Grading

You will be evaluated based on

1. Participation: weekly reading assignments, participation, and summarization (40%),
2. Programming assignments: 3–4 PA's (40%),
3. Project: 1 final project (20%)

Group and Submission

For your assignments, you can work in **groups** of 2–3 students. You can also work by yourself. One member turns in one solution for each assignment.

Students on a group are expected to participate equally in the effort and to be thoroughly familiar with all aspects of the joint work. All members bear full responsibility for the completion of assignments. Each member receives the same grade for the assignment. You may change group for different assignments but groups may not be dissolved in the middle of an assignment.

2.1 Participation

On average, we will have **two reading assignments** each week (about 45 mins for each reading). You are responsible for reading the assigned papers in advance for any given discussion.

2.1.1 Reading a paper

When reading a paper, you should focus on the following questions:

1. the **problem** (what is the problem we are trying to solve? why is it interesting?)
2. the limitations of the **state of the art** (what are existing approaches? what are their limitations?)
3. the proposed **approach** (its novelty, strength, and how it addresses the weaknesses of existing work)
4. the **evaluation** and comparison with other approaches (what are the results of the work, how was it evaluated and compared to others?)
5. provide **your thoughts** on the paper (e.g., what you like, what you don't like, what you think is interesting, etc.)

2.1.2 Lead Groups

Each reading will be assigned to a group. That group will lead the class discussion on that paper. The group will be responsible for the following:

- You will present **in depth** the paper you read to the class (about 25 mins). You can use slides or whiteboard.
  - Your presentation should answer the questions given in Section 2.1.1.
  - In addition, you will include providing concrete examples illustrating how the technique (usually given in the paper).
- After your presentation, you will guide the discussion (15 mins), e.g., ask questions to the class, have the class ask questions and participate in discussion.
- **Tools**: Usually each research paper has a free implementation tool. I will give bonus points 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.

2.1.3 Summarization

If your group is not assigned to lead the paper, your group will write a 1-page summary of the paper. You will submit it to me and the TA on Piazza **before** the day of class that we will discuss the paper (i.e., by 4:29 PM Tuesday). For the summary, you should answer the questions given in Section 2.1.1.

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.

2.2 Programming Assignments (PA's)

This course consists of several Programming Assignments (PA's) in Python. These PAs are designed for you to gain fundamental knowledge of state of the art AI analysis. All assignments have similar grading weights.

Your submissions will be evaluated for correctness, organization, and documentation. We will not attempt to fix broken submissions that fail to execute properly; only limited partial credit will be given in such situations. Assignments are due at **11:59pm** on the due date.

2.3 Project

You (your group) will understand in depth (i.e., *own*) an analysis technique. This technique can be from one of the papers you read or lead. Your project consists of two parts:

2.3.1 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 discussed.
- 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).

2.3.2 Implementation

You will **implement** the DNN analysis technique using Python. In your blog entry above, you will **write** how you apply your implementation to the example illustration you had. The project is **due on the last day class** (Sat, Dec 2nd).

3 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.

4 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.