CS-695/SWE-699 (Fall '23): AI Safety and Assurance

Course Page https://nguyenthanhvuh.github.io/class-verification

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

1 Description

This special topic course is a **research seminar** on *AI Verification and Analysis*. AI, in particular Deep Neural Networks (DNNs), have emerged as an effective approach for solving challenging real-world problems. Among many others, they have been used for image recognition, autonomous driving, airplane collision control, power grid control, fake news detection, drug synthesis and discovery, and COVID-19 detection and diagnosis.

However, just like traditional software, DNNs can have "bugs", e.g., producing unexpected results on inputs that are different from those in training data, and be attacked, e.g., small perturbations to the inputs by a malicious adversary or even sensorial imperfections result in misclassification. These issues, which have been observed in many DNNs and demonstrated in the real world, naturally raise the question of how DNNs should be tested, validated, and ultimately *verified* to meet the requirements of relevant robustness or safety standards.

In this class, we will learn various techniques and tools to verify DNNs. We will cover topics including the applications of verification, testing, analysis, constraint solving, and abstraction techniques to DNNS such as Feedforward Neural Networks (FNNs), Residual Networks (ResNet), and Convolutional Neural Networks (CNNs). We will focus on scalable and precise techniques that can deal with large, real-world DNNS.

The course will focus on active research areas in formal AI and DNN reasoning, but the specific topics will be largely determined by a combination of instructor fiat and the interests of the students.

1.1 Prerequisite

- No prerequisite courses. However, basic knowledge in linear algebra and AI/ML, e.g., CS 580, is strongly recommended
- Programming knowledge (Python)

1.2 Learning Outcomes

- **Understanding of AI Verification Techniques and Tools**: Students will gain a deep understanding of AI verification concepts, techniques, and tools. They will learn how to apply these techniques to various types of neural networks, including FNNs, REsNets, CNNs, and RNNs. The student will also learn how to use existing tools to analyze and verify DNNs.
- **In-Depth Understanding**: For the final project, students will delve deeply into a specific DNN analysis technique. They will not only understand the theory but also be able to provide concrete examples and *implement* the technique themselves, gaining a comprehensive understanding of the chosen topic.

- **Strengthen knowledge in Linear Algebra, AI/ML, and Programming**: Students will develop or strengthen their foundational knowledge in linear algebra and AI/ML, making them well-prepared to tackle advanced topics and real-world problems in AI safety. Programming assignments will require students to implement AI analysis techniques in Python. This will improve their programming skills and their ability to apply theoretical concepts to practical problems.

- **Critical Reading and Evaluation**: Through weekly reading assignments, students will learn to critically evaluate both book chapters and research papers related to AI verification and analysis. They will be able to identify the problem addressed, assess proposed solutions, analyze the strengths and weaknesses of different approaches, and evaluate and compare related techniques.

- **Presentation and Discussion Skills**: Students will have the opportunity to lead group discussions and presentations on assigned readings. This will enhance their presentation and communication skills, as well as their ability to facilitate meaningful discussions among peers.

2 Grading

You will be evaluated based on

1. Participation: weekly reading assignments, discussion, and participation (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. *Only one member of the group needs to submit the 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.

Unless specified otherwise, all assignments are submitted through *Blackboard*. All communication will be through *Piazza*.

2.1 Participation

On average, we will have **two reading assignments** each week (about 45 mins for each). You are responsible for reading the assignments in advance for any given discussion. Typically there are two types of reading assignments: (i) **chapters** from a neural networks textbook and (ii) from **research papers**.

2.1.1 What to think about for a reading assignment

Book Chapter When reading from a book, you should read the assigned chapters carefully, try the examples by hand (e.g., on a piece of paper), and answer the following:

1. **What is the problem?** (what is the problem the chapter is devoted to? why is it interesting?)

2. **What is the solution?** (what is the proposed solution? what are its strengths and weaknesses?)

3. Provide **your thoughts** on the reading (e.g., what you like, what you don't like, is it clear?)

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

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). Typically a paper has an illustrating example, you *should understand it in detail*.

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.)

6. **Tools**: Usually each research paper has a free implementation tool. I will give bonus points if you try out the tool, i.e., downloading and installing the tool on your computer and run it. Then 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.

2.1.2 Lead Groups

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

- You will present **in depth** the assigned reading to the class (about 30 mins). You should use slides or presentation and whiteboard for additional illustration.
  - Your presentation should answer the questions given in Section 2.1.1.
  - In addition, if the reading has examples (e.g., book chapters have various examples and research papers often have working examples or small code illustration), then your presentation must include and explain those in detail.

- You will guide the discussion (about 10 mins), e.g., ask questions to the class, engage others to ask questions and participate in discussion.

2.1.3 Reading Summaries

If your group is *not assigned* to lead a reading assignment, then your group will write a 1-page summary of the assignment (so 2 pages for 2 reading assignments). If your group *is assigned* to lead a reading assignment, then your group will not need to write the summary of this assignment. However you will still need to write the summary of the other assignments that you were not assigned to lead. You will submit it to me 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 structure your writing by answering 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

Your group will be assigned an analysis technique and you will need to understand it in depth (i.e., *own* it). You will demonstrate your understanding by providing a full example illustration and description of the algorithm and summarize its results. **The NeuralSAT paper** `https://arxiv.org/pdf/2307.10266.pdf` **provides a concrete example to do this.**

2.3.1 Example Illustration

You will describe a full concrete example illustrating the DNN technique assigned to you. This writing *must consist of a complete step by step illustration* on how the technique works on a given example (e.g., starting with a concrete DNN and example (e.g., the DNN and property given in PA1), how Reluplex works step by step to derive its result). You can write this in **Markdown** and include graphics and other files as needed.

IMPORTANT This part *is* the main focus of the project as it demonstrates your understanding of the technique. You will need to do a thorough work on this. **For a concrete illustration example, see *Section 3.1 Illustration of the NeuralSAT paper*.**

2.3.2 Writing

You will describe the DNN analysis algorithm and experimental results.

Algorithm description You will first write *pseudocode* describing how the technique works. Then you will create subsection describing important components, giving small examples to illustrate the ideas, etc. This kind of description is often used in many papers you have read throughout the course.

You are expected to do additional research on the technique you were assigned to. Reading just one paper is likely not enough for you to master that technique, e.g., it might require additional backgrounds. For example, the PLANET technique relies on many SAT solving concepts (e.g., conflict analysis, learning clause, unit propagations) that we have seen many times from the DNN book and in class. All of these concepts are popular and can also be found online, e.g., Wikpedia.

For a concrete algorithm description, see *Section 4 of the NeuralSAT paper.* Also look at Figures 4, 6, 7 for pseudocode examples.

Evaluation Here you will summarize the results of the technique. These results are from the paper you were given (i.e., you don't need to use external source for this). Here you will talk about

- what benchmarks were used ? e.g., MNIST, CIFAR-10, etc benchmarks, their sizes (number of neurons, parameters) and the types of properties (e.g., safety for ACAS models or robustness or something else)
- Briefly describe the tools/techniques that were used for comparison
- Main evaluation results. For example, for MNIST it performs X for CIFAR it performs Y. Compared to tool Z it is faster/slower, etc.

Due 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.