SWE-699/CS-690: AI Safety & Assurance

Fall '25

Meetings:	Wed 4:30PM – 7:10PM	**Place:**	PLANET 122
Instructor:	ThanhVu Nguyen	**Email:**	tvn@gmu.edu (Canvas is preferred)
Office Hr:	(email to set appointment)	**Place:**	Zoom or ENGR 4430
GTA:	Linhan Li	**Email:**	lli34@gmu.edu
Office Hr:	10AM–12PM	**Place:**	in-person or Zoom (email for info)

1 Description

Course Overview This special topic course is a **research seminar** on *AI Verification and Assurance*. AI, in particular Deep Neural Networks (DNNs), have emerged as an effective approach for solving challenging real-world problems. 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.

Learning Outcomes By the end of the course, students will gain a solid understanding of the principles of AI verification and assurance, particularly techniques for verifying the safety and robustness of DNNs.

1.1 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.

1.2 Prerequisite

- No prerequisite courses. However, basic knowledge in linear algebra and AI/ML, e.g., CS 580, is strongly recommended
- Programming knowledge (Python): we will use Python for assignments and projects.
- *Important*: This class does *not* assume prior knowledge of any AI or deep learning topics, e.g., how NNs are created and trained, as all background material will be provided. It also does not teach these topics and instead focuses on formal analysis of given NNs.

1.3 Course Materials

- T. Nguyen and H. Duong, *Engineering a Verifier for Deep Neural Networks* (**Required**, free)
- A. Albarghouthi. *Introduction to Neural Network Verification*. 2021. (**Recommended**, free)
- Other useful resources:
  - Neural Network Verification Tutorial
  - Advanced Topics in ML and Formal Methods (Grad level course at UIUC)

1.4 Assignment Submission and Communication

We will use **Canvas** for communication and submitting assignments, and to keep track of grades (§2). It's the student's responsibility to ensure that your grade records are correct.

When submitting to Canvas, you can either submit a PDF, Word, or text (code) file. If you manually write your answers, take a picture and submit it. **DO NOT** submit link (e.g., to Google Docs or some other services); you will receive a 0 for the assignment if you submit a link.

2 Grading

Assignments	**Percentage**
Participation	10%
Homework Assignments	25%
Quizzes	20%
Programming Assignments	20%
Project	25%
Total	100%

Scales

A+	$\geq 97\%$	A	$\geq 93\%$	A-	$\geq 90\%$
B+	$\geq 87\%$	B	$\geq 83\%$	B-	$\geq 80\%$
C	$\geq 70\%$	D	$\geq 60\%$	F	$< 60\%$

Groups For homework, PAs, and the project, you can team up with another student (a group of 2). You will need to submit a *single group* solution for each assignment. You also need to include a statement of contributions from each group member. Once you form a group, you cannot change it and will work with your group on all PAs and the project.

2.1 Participation

I place great emphasis on peer learning and interactive engagement. The class is structured to leverage interactions to the largest extent possible. Often, there will be in-class exercises for every class. The in-class exercises will be closely related to an upcoming homework assignment.

2.2 Homework Assignments

There are weekly *group homework assignments*, which are given in **Appendix A** of the DNN verification book. Homework assignments are due **before class**, i.e., before 4:30 PM on the day we meet. Late submissions are *not* accepted except in truly exceptional circumstances.

- Each group should be prepared to present their homework solution in class.
- There are **no make-ups**.
- Everyone in the group gets the same credit.

2.3 Quizzes

We will have a short quiz every week. Each quiz is worth 10 pts. The quiz will be based on the material covered in the previous weeks. Each quiz happens during the last part of class. *You must be present to take the quiz.*

Quiz Make-up Policy You will have the opportunity to make up a quiz if you miss it. The grading and make-up policy is as follows:

- Contact the GTA and schedule a make-up quiz (likely will be offered during TA office hours). The make-up can be different from the quiz given in class, but focuses on the same topics.

- The make-up must be taken promptly and within *a week of the quiz*.
- All quizzes count towards the final grade. Each quiz is scored on a 10 point scale. Missed quizzes score 0/10.
- The maximum possible score on the make-up is 8/10.

2.4 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.

2.5 Project

You will work on a group project that involves implementing and evaluating a DNN analysis technique. Each group will choose specific technique to focus on, and you will be responsible for understanding it in depth, implementing a prototype, and evaluating it. The project will culminate in a report and a presentation where you will summarize your findings and demonstrate your understanding of the assigned technique.

Project topics, format of report, and presentation guidelines will be provided in a separate document.

3 GMU Policies

3.1 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.

3.2 Learning Disabilities

Disability Services at George Mason University is committed to providing equitable access to learning opportunities for all students by upholding the laws that ensure equal treatment of people with disabilities. If you are seeking accommodations for this class, please first visit `https://ds.gmu.edu/` for detailed information about the Disability Services registration process. Then please discuss your approved accommodations with me. Disability Services is located in Student Union Building I (SUB I), Suite 2500. Email: ods@gmu.edu — Phone: (703) 993-2474