ECE 554: Machine Learning for Embedded Systems

Instructor: Dr. Weiwen Jiang
ECE Department, College of Engineering and Computing
Email: wjiang8@gmu.edu

Semester and Year: Fall 2025

Class Meeting Day(s) and Time(s): Thursday, 4:30pm to 7:10pm
Modality: Face-to-Face
Class Location: Room 2413, Peterson Hall

Course Materials: Course materials will be posted before or after the class. No formal textbook is required. The book from http://www.deeplearningbook.org/ will be referred to on the course.

Course Description: Machine learning (ML) has gradually become the core component of wide applications in different computing scenarios, ranging from edge computing to cloud computing. This course focuses on resource-constrained edge computing, in particular the embedded systems, and introduces techniques for developing energy/time efficient ML algorithms and models for the embedded systems. Topics that are covered include (i) commonly used ML algorithms, (ii) ML model compression techniques, (iii) hardware-aware machine learning, (iv) hardware and neural architecture co-design. The course also provides a comprehensive team-based research and development experience through projects and presentations. Offered by Electrical & Comp. Engineering. May not be repeated for credit.

Prerequisites: The course topics are self-contained so that a background in machine learning is not required. Students should be familiar with programming and embedded systems to complete the course projects.

Course Schedule: there are three sections of the course, including:

- Section I: Introduction of Machine Learning and Deep Neural Networks
- Section II: Optimization of ML/DNN Models for Resource-Constrained Devices
- Section III: Co-Optimization of ML/DNN and Hardware Designs

SECTION I: Introduction of Machine Learning and Deep Neural Networks

Date	Topic
Week 1	Course Information & Introduction to Machine Learning
Week 2	MLP Programming and Implementation with Pytorch
Week 3	Train Neural Networks
Week 4	Deep Convolutional Neural Networks (CNN) - Part 1
Week 5	Deep Convolutional Neural Networks (CNN) - Part 2

SECTION II: Optimization of ML/DNN Models for Resource-Constrained Devices

Date	Topic
Week 6	Transformer
Week 7	Model Compression - Part 1
Week 8	Model Compression - Part 2
Week 9	Mid-Term Exam

SECTION III: Co-Optimization of ML/DNN and Hardware Design

Date	Topic
Week 10	ML System Implementation and Optimization
Week 11	Reinforcement Learning
Week 12	Neural Architecture Search
Week 13	Hardware-Aware Neural Architecture Search
Week 14	HW/SW Co-Design with Neural Architecture Search

* The schedule might change during the semester depending on the progress of the class.

Goals and Outcomes:

- Understand the basic training and inference techniques of a neural network. (Section I)
- Get familiar with commonly used neural networks, such as CNN (Section I)
- Be able to implement neural networks using machine learning tools (Section I)
- Be able to optimize neural networks for resource constrained hardware (Section II)
- Be able to apply compression techniques, i.e., pruning and quantization (Section II)
- Be able to design the customized neural network (Section III)
- Be able to co-design neural networks and hardware accelerators (Section III)

Homework Labs: There are 2-4 take-home labs in Section I assigned to the students to practice basic skills of machine learning implementation using TensorFlow or Pytorch. A total of 7 days after the due dates are permitted for all assignments. However, after the due date, each assignment will be deducted 10 points for each day late. Submissions that exceed the 7 days after the due dates will not be considered for get points.

Presentation: The project-oriented research articles will be assigned to the teams of students at the end of Section I. Students then need to prepare a presentation on the assigned research articles in Section II or III.

Project: Students will form teams to implement several open-topic projects in terms of different hardware platforms: mobile devices or FPGAs. According to different projects, open-source or free-version software would be involved, e.g., the TensorFlow Lite for mobile devices and Vivado for FPGA synthesis. Each team will be assigned one project. At the end of the course, students will give a demonstration of the completed projects and deliver a comprehensive project/technical report.

Grading:

- Homework assignments (2-4 take-home labs) 20%
- Quizzes (3-4 quizzes in the class) 10%
- Research paper presentation (Presentation and Q&A) 20%
- Project final review/report 30%
- Midterm exam 20%

Grading Schema:

A	95-100	C	70-77.9
A-	90-94.9	F	0-69.9
B+	86-89.9		
B	82-85.9		

Grading-related Policies:

- There are no make-ups except in a situation of extended illness or long-term personal circumstances.
- Work submitted after grading is complete will not be accepted.

AI (Artificial Intelligence) Tools Policy:

The use of AI-based tools is permitted for purposes of learning, exploring ideas, and identifying credible references. Students may use such tools to clarify concepts, brainstorm topics, or locate scholarly sources. However, AI tools must not be used to generate complete solutions to assignments, assessments, or projects, nor may students present AI-generated text, code, or other output as their own original work. Copying, paraphrasing, or otherwise incorporating AI-generated materials without attribution constitutes academic dishonesty and will be treated as plagiarism under the University's Academic Standards. Students are responsible for critically evaluating and verifying any information obtained through AI tools, ensuring that their submissions reflect their own understanding, analysis, and synthesis of course material.

University and Course-Specific Policies:

Common policies affecting all courses at George Mason University, including

- Academic Standards
- Accommodations for Students with Disabilities
- FERPA and Use of GMU Email Addresses for Course Communication
- Title IX Resources and Required Reporting,

are available at

https://stearnscenter.gmu.edu/home/gmu-common-course-policies

You are strongly encouraged to get familiar with this additional information.

The course has the following specific policies:

- Email Communications: Students must use their MasonLive email account to receive important University information, including messages related to this class. See http://masonlive.gmu.edu for more information. Homework assignments and other course material will be emailed to your MasonLive email account. Also, when you send me an email, please write ECE554 on the subject line.

- Avoid Reposting Course Material: It is not allowed to reposting course material. The course materials (lecture notes, homework, projects, exams, solutions, and anything else posted on the course website) are copyrighted. You may not upload them to any other website or share them with any on-line or off-line test bank.
- Honesty and Integrity: Mason expects students to pursue their academic work with honesty and integrity. Students should feel free to work in groups to discuss lecture material and homework assignments; however, under no circumstance should a student represent another's work as his or her own. Copying solutions for assigned homework problems, from any source, constitutes a violation of the university honor code. Any form of cheating may cause penalties, from getting an F in this course to academic actions in accordance with university policy.
- Office of disability services: Mason provides accommodations through the Office of Disability Services (ODS) http://ods.gmu.edu. If you are a student with a disability and you need academic accommodations, please see me and contact ODS at 993-2474.