HARVARD
UNIVERSITY

Course Offering: Fall 2023
Instructor: Prof. Vijay Janapa Reddi
Office: SEC 5.305
Email: vj@eecs.harvard.edu

CS249r: Tiny Machine Learning

About the Course

Tiny machine learning (TinyML) is defined as a fast-growing field of machine learning technologies and applications including hardware (dedicated integrated circuits), algorithms and software capable of performing on-device sensor (vision, audio, IMU, biomedical, etc.) data analytics at extremely low power, typically in the mW range and below, and hence enabling a variety of always-on use-cases and targeting battery-operated devices. The pervasiveness of ultra-low-power embedded devices, coupled with the introduction of embedded machine learning frameworks like TensorFlow Lite for Microcontrollers, will enable the mass proliferation of AI-powered IoT devices. The explosive growth in machine learning and the ease of use of platforms like TensorFlow (TF) make it an indispensable topic of study for computer science and electrical engineering students.

HARVARD
UNIVERSITY

Course Offering: Fall 2023
Instructor: Prof. Vijay Janapa Reddi
Office: SEC 5.305
Email: vj@eecs.harvard.edu

Course Topics

The course provides a sweeping overview of machine learning systems, from foundational concepts like the stages of machine learning to advanced topics such as hardware acceleration and on-edge generative AI. This includes a journey through data engineering, optimized model frameworks, and sustainability dimensions of ML, all tailored to embedded environments.

- Overview and Introduction to Embedded Machine Learning
- Data Engineering
- Embedded Machine Learning Frameworks
- Efficient Model Representation and Compression
- Performance Metrics and Benchmarking of ML Systems
- Learning on the Edge
- Hardware Acceleration for Edge ML: GPUs, TPUs and FPGAs
- Embedded MLOps
- Secure and Privacy-Preserving On-Device ML
- Responsible AI
- Sustainability at the Edge
- Generative AI at the Edge

Meeting Time

Mondays from 12:45pm - 3:30pm in SEC LL2.229

Teaching Assistants

Ikechukwu Uchendu <iuchendu@g.harvard.edu>

Jason Jabbour <jasonjabbour@g.harvard.edu>

Jessica Quaye <jquaye@g.harvard.edu>

HARVARD
UNIVERSITY

Course Offering: Fall 2023
Instructor: Prof. Vijay Janapa Reddi
Office: SEC 5.305
Email: vj@eecs.harvard.edu

Office Hours

Ikechukwu Uchendu: Wednesday 9 - 10 AM (SEC 5.401)

Jason Jabbour: Friday 11 AM - 12 PM (SEC LL2.225)

Jessica Quaye: Tuesday 1 - 2 PM (SEC 5.403)

Matthew Stewart: Thursday 5 - 6 PM (SEC 1.412 Seminar Room)

Vijay Janapa Reddi: Monday 3:30 - 4:30 PM (after class)

Webpage

Canvas Site: https://canvas.harvard.edu/courses/122580

Textbook

The field of embedded machine learning systems is rapidly evolving. While there is a TinyML textbook available, it has become somewhat dated. We'll primarily rely on the latest academic publications for our studies. You may use the TinyML textbook as a reference material when needed. We will be drawing some of the content from the textbook titled "*AI at the Edge*," which contains more up-to-date methods and examples. We can make an online version available to you via Canvas that cannot be distributed outside of class. However, we encourage students to get their own copy so that it is easier to follow and work through the material.

Prerequisites

Not all are required, but the following prerequisites are recommended:

1. CS 51/61/161 and/or a basic systems programming experience
2. CS 181/CS 182 or something to that effect
3. CS 141 or something that exposes you to an embedded system

We hope to have a diverse class and we intend to provide some background on embedded computing. That said, we recommend students have prior experience in the algorithms employed in machine learning from classes such as CS 181/182. Please contact the instructor or teaching fellow if you are interested in taking the course but are unsure about whether the background you have is suitable.

Note that registering for this course is lottery based! All students who wish to register must complete the survey form on Canvas: https://forms.gle/NbpqQxy5m6f3cHAP6

HARVARD
UNIVERSITY

Course Offering: Fall 2023
Instructor: Prof. Vijay Janapa Reddi
Office: SEC 5.305
Email: vj@eecs.harvard.edu

Course Schedule

- The course will consist of instructor and guest lectures, student-led paper discussions, and research project presentations to expose students to a variety of topics at the intersection of machine learning and embedded systems.
- There are programming assignments that are interspersed that you are required to complete, and there is a capstone project at the end of the semester
- The course schedule is subject to change. Please refer to the following spreadsheet for a more up-to-date version.

Date	Topic
Sep 11	Overview, Introduction to Machine Learning, and Introduction to Embedded Systems
Sep 18	Data Engineering
Sep 25	Machine Learning Frameworks
Oct 2	Efficient Model Representation and Compression
Oct 9	*Columbus Day*
Oct 16	Learning on the Edge
Oct 23	Hardware Acceleration for ML: GPUs, TPUs and FPGAs
Oct 30	MLOps
Nov 6	Secure and Privacy-Preserving On-Device ML
Nov 13	Responsible AI
Nov 20	Sustainability at the Edge
Nov 27	Generative AI at the Edge
Dec 5	Presentations

Late Policy

Late assignments will receive a max of 50% credit if submitted 1 day after the deadline, a max of 20% credit if submitted 2 days after the deadline, and no credit more than 2 days after the deadline. If you anticipate that your assignment will be late, please contact the TFs.

HARVARD
UNIVERSITY

Course Offering: Fall 2023
Instructor: Prof. Vijay Janapa Reddi
Office: SEC 5.305
Email: vj@eecs.harvard.edu

Assignment Descriptions and Grading

Paper Reviews – 10%

Purpose:

- Develop the skill of reading papers (especially those outside of one's main discipline) through practice. There is no one correct way to read a paper but here is a helpful guide.

To that end, students are required to:

- Read the assigned papers or articles before each class.
- Lead a group discussion about one of the week's assigned papers.

Class Participation – 10%

Purpose:

- Engage with the presenter by posing questions about the presented material and related subjects.

To that end, students are required to:

- Students will be expected to attend class and engage with other students by asking questions, providing feedback, and actively participating in the group discussion.

HARVARD UNIVERSITY

Course Offering: Fall 2023
Instructor: Prof. Vijay Janapa Reddi
Office: SEC 5.305
Email: vj@eecs.harvard.edu

Paper Presentation – 10%

Purpose:

- Develop the skill of understanding a paper in detail.
- Develop the skill of presenting a (conference) paper to an audience.
- Develop the skills of teaching a concept to peers.
- Receive feedback on presentation skills, focusing on content delivery and depth of topic comprehension.

To that end, each student will be assigned a paper to present in their small group during every class session. Students are expected to be familiar with their assigned paper and prepared to answer inquiries about its content and implications. Specifically, students will be required to:

1. Deliver comprehensive paper presentations throughout the semester.
2. Each paper presentation should consist of the following:
    a. **Setup:** What is the problem? Why is it important? What are the key challenges in solving it?
    b. **Contribution:** What did the author(s) do? Why was it novel?
    c. **Context:** How did it compare to other work? What work did this build on? What are the relative strengths and weaknesses?
3. Engage in a Q&A session after each presentation, fostering active group discussions.

Programming Assignments – 25%

Purpose:

- Students will gain hands-on experience with machine learning systems.
- The course will have a series of programming assignments that require students to demonstrate technical skills on training machine learning models, optimizing the models, and deploying the models onto embedded systems.

To that end, students are required to:

- Read the assignments that are given to them.
- Read any associated pre-assignment material from related publications.
- Understand what is required of them to complete the homework assignment.
- Complete the assignment and demonstrate knowledge of what was accomplished.
- Submit the completed assignment to the canvas website.

HARVARD UNIVERSITY

Course Offering: Fall 2023
Instructor: Prof. Vijay Janapa Reddi
Office: SEC 5.305
Email: vj@eecs.harvard.edu

Final Project – 45%

Purpose:

- Provide an opportunity for students to apply, extend, and integrate the foundational concepts learned in the course.
- Practice conducting a formal (conference) research paper.
- Practice collaborating with others on research.
- Practice writing conference papers in appropriate Latex templates.
- Practice getting feedback from advisers on research ideas.

To that end, students are required to:

1. Work by themselves or in teams of 2-3 students. Note that we expect all students to demonstrate a ~equal amount of work, so teams of 3 should be sure to tackle appropriately sized problems. Please remember that we expect most projects to include demonstrating on an embedded system and so we encourage you to think carefully.
2. Submit a Project Proposal midway through the semester
    a. A brief discussion of the problem and algorithms you intend to investigate and the system you intend to build in doing so.
    b. Examples of expected behavior of the system or the types of problems the algorithms you investigate are intended to handle.
    c. A list of papers or other resources you intend to use to inform your project effort. This list will form the core of your project report reference list. If your project includes anything unusual (such as having significant systems demands), please state this as well.
3. Submit a 6-8 page two column Project Report in Latex. The course staff suggests using the overleaf online editor which is free when students sign up with a Harvard email and the course staff will provide the required Latex template for the project. Your paper should contain (at a minimum): an abstract, introduction, related work, algorithm/system description, experiments, conclusion, and references. Strong papers will have descriptive figures that reveal good experiment design and execution.
4. Present their project in a simulated conference environment (anticipated date: Monday, December 5).

HARVARD
UNIVERSITY

Course Offering: Fall 2023
Instructor: Prof. Vijay Janapa Reddi
Office: SEC 5.305
Email: vj@eecs.harvard.edu

Diversity and Inclusion

In an ideal world, science would be objective. However, much of science is subjective and is historically built on a small subset of privileged voices. We acknowledge that it is possible that there may be both overt and covert biases in the material due to the lens with which it was written, even though the material is primarily of a scientific nature. Since integrating a diverse set of experiences is important for a more comprehensive understanding of science, please contact the course staff (in person or electronically) or submit anonymous feedback if you have any suggestions to improve the quality of the course materials.

We would like to create a learning environment that supports a diversity of thoughts, perspectives, and experiences, and honors your identities. If you have a name and/or set of pronouns that differ from those that appear in your official records, please let us know! If you feel like your performance in the class is being impacted by your experiences outside of class, please don't hesitate to contact us. If you prefer to speak with someone outside of the course, the SEAS Director of Diversity, Inclusion, and Belonging is an excellent resource.

So that the course staff has enough time to implement accommodations, students needing academic adjustments or accommodations because of a documented disability must present their Faculty Letter from the Accessible Education Office (AEO) and speak with the course staff by the end of the second week of the term. All discussions will remain confidential, although the course staff may contact the AEO to discuss appropriate implementation.

HARVARD
UNIVERSITY

Course Offering: Fall 2023
Instructor: Prof. Vijay Janapa Reddi
Office: SEC 5.305
Email: vj@eecs.harvard.edu

Academic Integrity

This course's policy on academic honesty is best stated as "be reasonable." We recognize that interactions with classmates and others can facilitate mastery of the course's material. However, there remains a line between asking for help and submitting someone else's work. Please see the University Guidelines on the use of language models and other generative AI tools within the class. If any questions arise, please reach out to the class instructor.

For Paper Reviews and Paper Presentations, students are permitted to ask classmates and others for conceptual help so long as that help does not reduce to another doing your work for you (e.g., writing your response or making your slides). Collaboration on the course's final project is permitted to the extent prescribed by its specification.

If in doubt as to whether some act is reasonable, do not commit it until you solicit and receive approval in writing from the course staff. Acts considered not reasonable are handled harshly. If the course refers some matter to the Administrative Board and the outcome is Admonish, Probation, Requirement to Withdraw, or Recommendation to Dismiss, the course reserves the right to impose local sanctions on top of that outcome. If you commit some act that is not reasonable but bring it to the attention of the course staff within 48 hours, the course may impose local sanctions, but the course will not refer the matter to the Administrative Board except in cases of repeated acts.