HARVARD
John A. Paulson
School of Engineering
and Applied Sciences

Course Offering: Fall 2020
Instructor: Prof. Vijay Janapa Reddi
Teaching Fellow: Colby, Radhika, Srivatsan, and Brian
Office: Maxwell Dworkin 147
Email: vj@eecs.harvard.edu

CS249r: Tiny Machine Learning

About the Course

Tiny Machine Learning (TinyML) is an introductory course at the intersection of Machine Learning and Embedded IoT 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 modern computer science and electrical engineering students.

As such we have designed an introductory course on Applied AI at the intersection of Machine Learning and Embedded IoT Devices. We provide background on both topics and then dive into the unique challenges faced at that intersection point with hands-on assignments using TensorFlow, Google Colab, and Arduino.

HARVARD
John A. Paulson
School of Engineering
and Applied Sciences

Course Offering: Fall 2020
Instructor: Prof. Vijay Janapa Reddi
Teaching Fellow: Colby, Radhika, Srivatsan, and Brian
Office: Maxwell Dworkin 147
Email: vj@eecs.harvard.edu

Course Topics

The course will cover all things related to machine learning on embedded devices. The topics range from tinyML applications and algorithms to the design of the frameworks built for running ML on embedded systems to the microcontroller hardware designed, optimized, and built for ultra-low power (milliwatts) computing:

- tinyML Applications and Usecases
- tinyML algorithms machine learning algorithms and optimizations
- tinyML frameworks, tools, and techniques
- Ultra-low-power system design

Meeting Time

Mondays and Wednesdays form 10:30am-11:45am via Zoom.

Teaching Assistants

Colby Banbury <cbanbury@g.harvard.edu>
Radhikah Ghosal <rghosal@g.harvard.edu>
Srivatsan Krishnan <srivatsan@g.harvard.edu>
Brian Plancher <brian_plancher@g.harvard.edu>

Office Hours

Sri (M 3-4 PM EST)
Pete (T 12-1 PM EST)
Radhika (T 3-4 PM EST)
VJ (W 3-4 PM EST)
Colby (F 12-1 PM EST)
Brian (Special OHs will be announced)

Webpage

General overview: https://sites.google.com/g.harvard.edu/tinyml/home
Details about assignments, projects etc: https://canvas.harvard.edu/courses/75295

Textbook

There is no real textbook yet on TinyML. However, the course requires you to use the following book: TinyML which will serve as a reference material for the lab assignments.

We can make an online version available to you via Canvas that cannot be distributed outside of class. But we encourage you to get the book so that it is easier to follow the material.

HARVARD
John A. Paulson
School of Engineering
and Applied Sciences

Course Offering: Fall 2020
Instructor: Prof. Vijay Janapa Reddi
Teaching Fellow: Colby, Radhika, Srivatsan, and Brian
Office: Maxwell Dworkin 147
Email: vj@eecs.harvard.edu

Prerequisites

Not all are required, but the following are recommended.

1. CS 51/61/161 and/or a basic systems programming experience
2. CS 181 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 assume few students will have full exposure to the full breadth of topics we will cover. As such, we intend to provide some background on all of the topics. That said, students may find it helpful if they also have some background in some of 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.

Course Schedule

- The course will consist of lectures that are interspersed with paper readings
- The instructor will teach some classes, but the course will also consist of guest lectures 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
- Note that the project proposal is also due in the middle of the term and the project report and presentation will be due/presented during the exam block.
- The most up-to-date schedule will be posted to the course website, and you can find it here: https://sites.google.com/g.harvard.edu/tinyml/home#h.ummto0wwhsk8

Grade Formula

- Paper Reviews – 10%
    - Students will be expected to read the assigned papers or articles before each class and submit a short response. Each Discussion response will be graded on a two point scale. 1 point is awarded if a response is posted before the deadline and it is clear that the student completed the reading. The other point is awarded if the response is thoughtful and conveys some new information or opinion.
- Paper Presentation – 10%
    - Students will each be expected to present at least one paper (chosen from the provided list) to the class and lead the class through an in-class discussion
- Class Participation – 10%
    - Students will be expected to attend class and participate in discussions about the assigned paper for that class

HARVARD
John A. Paulson
School of Engineering
and Applied Sciences

Course Offering: Fall 2020
Instructor: Prof. Vijay Janapa Reddi
Teaching Fellow: Colby, Radhika, Srivatsan, and Brian
Office: Maxwell Dworkin 147
Email: vj@eecs.harvard.edu

- Programming Assignments – 35%
    - Students will do a series of programming assignments as the course progresses. These assignments will require you to use different programming languages (mostly Python and C/C++) and work within IDE frameworks like Visual Studio and Arduino. You will also learn to use TF Lite for Microcontrollers to do your tiny machine learning assignments.
- Project (including final project presentation) – 35%
    - Students will undertake an in-depth course project on a topic of their choosing. The project will introduce students to research in the field of computer architecture and robotics. Project proposals will be due about midway through the class and regular meetings with the teaching staff will help refine the goals and milestones. Final output of the project is a presentation to the class and a project write up in the form of a research style paper.

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.

Assignment Descriptions

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.
- Develop the skill of quickly taking away the big picture ideas from a paper to decide whether to invest time in reading and understanding it in detail.
- Crowdsource a best practices guide on writing papers for use in the final project

To that end students are required to:

1. Log onto the course canvas site and submit a "review" for all of the pre-reads and assigned papers being presented in each class **36 HOURS BEFORE THE CLASS**. These "reviews" will ask students to submit a couple of sentences on what they thought the key takeaways were from the paper and any questions/comments they have on the content AS WELL AS one thing they liked and disliked about the writing style.

Paper Presentation – 10%

Purpose:

- Develop the skill of understanding a paper in detail
- Develop the skill of presenting a (conference) paper to an audience

HARVARD
John A. Paulson
School of Engineering
and Applied Sciences

Course Offering: Fall 2020
Instructor: Prof. Vijay Janapa Reddi
Teaching Fellow: Colby, Radhika, Srivatsan, and Brian
Office: Maxwell Dworkin 147
Email: vj@eecs.harvard.edu

- Develop the skills of teaching a concept to a class
- Get feedback on presentation skills both in terms of delivery and slides

To that end students will be required to:

1. Give at least one (depending on class size) **18 minute** presentation on one of the papers at the end of this document (organized by topic -- see schedule for dates for each topic)
    a. ~5 minutes of setup (What is the problem? Why is it important? What are the key challenges in solving it?)
    b. ~5 minutes of contribution (What did the author(s) do? Why was it novel?)
    c. ~8 minutes of context (How did it compare to other work? What work did this build on? What are the relative strengths and weaknesses?)
2. The presentation should be supported by slides. Two guides for good slide design can be found here and also here (and the course staff will upload some templates to the course website). Also, here is a guide on effective presentations. Finally, international students who want additional help can contact Sarah Emory or Pamela Pollock
3. Each presentation will be followed by **10 minutes of Q&A / Discussion**
4. To help students develop their presentation, students are required to **meet with the course staff for 15 minutes one week prior** (ideally right before or after class) to their presentation to discuss their approach and the key ideas they want to get across. Ideally with ideas / outlines for slides. *The farther along the presentation is, the better the feedback the course staff can give (we do not expect finished presentations).*

Class Participation – 10%

Purpose:

- Ensure that all students are able to end each paper presentation understanding the high level takeaways from a given paper
- To give feedback to presenters

To that end, students are required to:

1. Submit anonymous feedback on the provided forms on the presentation
2. Ask questions of the presenter both about their presentation and about things they chose not to present that they either found confusing or interesting

Programming Assignments – 35%

Purpose:

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

HARVARD
John A. Paulson
School of Engineering
and Applied Sciences

Course Offering: Fall 2020
Instructor: Prof. Vijay Janapa Reddi
Teaching Fellow: Colby, Radhika, Srivatsan, and Brian
Office: Maxwell Dworkin 147
Email: vj@eecs.harvard.edu

To that end, students are required to:

- Read the assignments that are given to them
- Read the associated pre assignment material from the TinyML book
- Understand what is required of them for completing the homework assignment
- Complete the assignment and demonstrate knowledge of what was accomplished
- Submit the completed assignment to the canvas website.

Final Project – 35%

Purpose:

- Provide an opportunity for students to apply, extend, and integrate the foundational concepts learned in the course toward a topic of interest in the fields of computer architecture and/or robotics
- 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 anything you do will require demonstrating on an embedded system and so we encourage you to think carefully.
2. Submit a Project Proposal midway through the semester (**see website for dates**):
    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 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 (**see website for dates**). 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 (**see website for dates**). That is, teams are given **12 minutes** (subject to change based on class size) to present their paper in the style of the paper presentations.

HARVARD
John A. Paulson
School of Engineering
and Applied Sciences

Course Offering: Fall 2020
Instructor: Prof. Vijay Janapa Reddi
Teaching Fellow: Colby, Radhika, Srivatsan, and Brian
Office: Maxwell Dworkin 147
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 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.

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.

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.