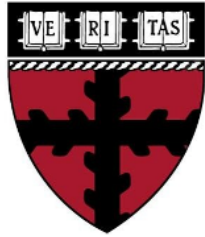# CS249r: Tiny Machine Learning

sites.google.com/g.harvard.edu/cs249-tinyml-2023

CS249R: TINYML
THE FUTURE OF ML IS TINY AND BRIGHT

Course Overview

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 and 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.

PrevioUs Years

Fall 2022

Fall 2020

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

🔗
What You'll Learn & Do in the Course

At the end of the course, you would have been exposed to the following:

- Brief introduction to ML and IoT

- Industry talks about real-world deployments

- Discussion of bleeding edge academic research

- Practical experience through hands-on project assignments

At the end of the course, you would have achieved the following:

- Gained familiarity with cutting edge literature in the field of tinyML

- Learnt to train and deploy models on microcontrollers with TF-micro

- Conceived and developed a (novel) TinyML application running on a MCU

🔗 Instructor

🔗 Teaching Assistants

🔗

Ikechukwu Uchendu

Computer Science, PhD Student, Harvard University

🔗

Jason Jabbour

Computer Science, PhD Student, Harvard University

🔗

Jessica Quaye

Computer Science, PhD Student, Harvard University

🔗

Class Information

🔗

Schedule

When: Every Monday

- Time: 12:45 PM - 3:30 PM EST

- Where: SEC LL2.229

🔗

General Information

🔗

Resources

🔗

LECTURE NOTES

This year, we are doing an experiment! We, the students, staff, and the instructor of this course, have embarked on a collaborative journey to assemble our class notes into an Introduction to Machine Learning Systems for TinyML book.

Our collective effort is driven by the shared vision of creating a reference material that encompasses the vast expanse of knowledge and insights gained throughout our time studying Machine Learning Systems.

This compilation is not just an aggregation of individual notes but a testament to our dedication, curiosity, and passion for the subject. Each page reflects the collective wisdom of our class, interwoven with individual experiences and understandings.

Please note that this book is a work in progress. As the field of Machine Learning Systems is ever-evolving, so too are our notes. We continually strive to update and enhance the content, ensuring its relevance and accuracy. We hope this book serves as a valuable resource for anyone venturing into the world of Machine Learning Systems, whether you're a novice, a practitioner, or an expert.

🔗 OTHER Reference BookS

You may use the TinyML O'Reilly book as a reference material when needed. We will be drawing some of the content from the textbook titled "AI at the Edge" as well, which contains more up-to-date methods and examples.

🔗 Coding Assignments

To get everyone familiar with coding on embedded systems with ML, we will be using the examples provided in this book as a starting point. Each assignment will build on the examples provided.

🔗 Projects

The course will culminate with project demos! You will have an opportunity to showcase what you have learned by incorporating your experience into a hands-on project of your liking. Alternatively, we will provide a list of suggested projects that will allow you to start from the class assignments.

🔗 Prerequisites

You must be confident and comfortable with the following topics:

- Proficiency in C/C++ and Python. All class assignments will be involve one or both languages.

- Familiarity with command line tools in Mac, Windows or Linux. Assignments will require you to operate a terminal.

You will get more out of the class if you have familiarity with:

- College Calculus, Linear Algebra, Basic Probability, and Statistics: Many ML topics revolve around taking derivatives, understanding matrix vector operations and notation, and the concepts of gaussian distributions, means, standard deviations, etc.

- Basic Machine Learning / Aritificial Intellifence: The class will provide some introductory material but/and any previous knowledge you have will allow you to dive deeper into the material.
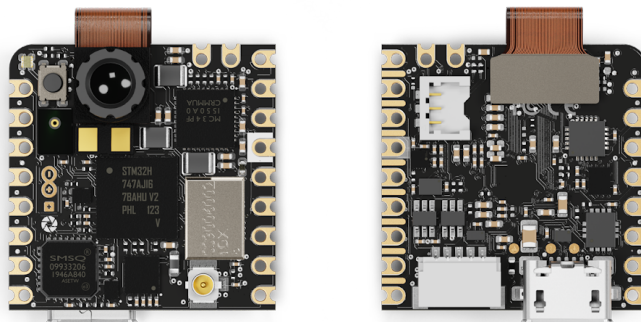
🔗 Grading

- Paper Reviews: 10%

- Paper Presentation: 10%

- Class Participation: 10%

- Programming assignments: 25%

- Final Project: 45%

🔗 Development Platforms

🔗 EMbedded system

Deploy TinyML models on a STM3-H747AII6 Dual ARM® Cortex® M7/M4 IC processor with a 2MP color camera and a smart 6-axis motion sensor, integrated microphone and distance sensor.

🔗
embedded ML framework

You will use TF Lite (Micro) to deploy your ML models,
which is offered free of cost by Google.

🔗
Active Learning Labs Staff

🔗
Leonardo
Gomez

🔗
Benjamin Brown