CS249r: Data Engineering

Sep. 18, 2023

Goals for today

- 1. Course logistics
- 2. Lecture
- 3. Paper discussions
- 4. Guest speaker

Course Logistics

Class Schedule

		Start	~End
Lecture		12:45:00 PM	1:40:00 PM
Break		1:40:00 PM	1:45:00 PM
Paper Discussions	Paper 1	1:45:00 PM	2:05:00 PM
	Paper 2	2:05:00 PM	2:25:00 PM
Break		2:25:00 PM	2:30:00 PM
Guest Lecture		2:30:00 PM	3:30:00 PM

Office Hours



Matthew Stewart

Postdoctoral Researcher, Harvard University

Office Hours: 5-6 PM Thursday (SEC 1.412)

<u>lkechukwu Uchendu</u>

Computer Science, PhD Student, Harvard University

Office Hours: 9-10 AM Thursday (SEC 1.412)



<u>Jason Jabbour</u>

Computer Science, PhD Student, Harvard University

Office Hours: 5-6 PM Tuesday



Jessica Quaye

Computer Science, PhD Student, Harvard University

Office Hours: 1-2 PM Tuesday (SEC 5.403)

Paper Discussion Sign-up Sheet

- 1. Please take a few minutes to sign up for next week's paper discussion:
- 2. Submit to Canvas:
 - a. Paper Reading Group 1 for your first time
 - b. Paper Reading Group 2 for your second time



Kook

Goal:

- Open source "living" textbook
- Putting in "boiler plate" material
- Welcome to read through the first few chapters of background
 - Improve it! \bigcirc

Embedded Al: Principles, Algor × + ☆ (i) localhost:3902 $\leftarrow \rightarrow C$ Q ₫ ☆) 🙀 🧮 🗟 🔹 🐺 🕼 🔱 🙁 🖯 🖈 🔲 🕵 E 🗎 TinyML 🗎 Harvard 🗎 Funding 🗎 MLC 📄 Meta 📄 Nora 📄 LLMs 🔅 🔯 🗛 🛧 🛆 LLMx 🛆 CS249r Other Bookmarks

Embedded AI: Principles. Algorithms, and Applications 00.1.2 1

FRONT MATTER

Acknowledgements

About This Book

1 Introduction

2 Embedded Systems

3 Deep Learning Primer 4 Embedded ML

Preface

Dedication

Copyright

WELCOME

DEEP DIVE

5 ML Workflow

6 Data Engineering

7 Pre-processing

D Case Studies

Embedded AI: Principles, Algorithms, and Applications

Preface

In "Embedded AI: Principles, Algorithms, and Applications", we will embark on a critical exploration of the rapidly evolving field of artificial intelligence in the context of embedded systems, originally nurtured from the foundational course, tinvML from CS249r.

The goal of this book is to bring about a collaborative endeavor with insights and contributions from students, practitioners and the wider community, blossoming into a comprehensive guide that delves into the principles governing embedded AI and its myriad applications.

"If you want to go fast, go alone, if you want to go far, go together."-African Proverb

As a living document, this open-source textbook aims to bridge gaps and foster innovation by being globally accessible and continually updated, addressing the pressing need for a centralized resource in this dynamic field. With a rich tapestry of knowledge woven from various expert perspectives, readers can anticipate a guided journey that unveils the intricate dance between cutting-edge algorithms and the principles that ground them, paving the way for the next wave of technological transformation.

The Philosophy Behind the Book

We live in a world where technology perpetually reshapes itself, fostering an ecosystem of open collaboration and knowledge sharing stands as the cornerstone of innovation. This philosophy fuels the creation of "Embedded Al: Principles, Algorithms, and Applications." This is a venture that transcends conventional textbook paradigms to foster a living repository of knowledge. Anchoring its content on principles, algorithms, and applications, the book aims to cultivate a deep-rooted understanding that empowers individuals to navigate the fluid landscape of embedded AI with agility and foresight. By embracing an open approach, we not only democratize learning but also pave avenues for fresh perspectives and iterative enhancements, thus fostering a community where knowledge is not confined but is nurtured to grow, adapt, and illuminate the path of progress in embedded AI technologies globally.

Conventions Used in This How to Contact Us How to Contribute Contributors O Edit this page

Report an issue View source

Table of contents

The Philosophy Behind the

Preface

Book

8 ML Frameworks 9 Model Training 10 Efficient Al 11 Optimizations 12 Deployment 13 On-Device Learning 14 Hardware Acceleration 15 MLOps 16 Privacy and Security 17 AI Sustainability 18 Responsible AI 19 Generative Al References Appendices A Tools B Resources C Communities

Book Chapter Contributions

- Link to sign-up sheet
- 3-4 people sign up
- Produce at least 10 pages of content
 - Brainstorm sections within a chapter
 - Meet with Matthew Stewart and Prof. VJ to discuss sections
 - Work collaboratively to improve the content/sections
- Grading
 - Content quality
 - Topic coverage
 - Figures/References
 - Peer Review

Assignment 1: Setup Arduino Nicla Vision + Data Engineering for Person Detection

- TinyML workflow
- Data Engineering
- Edge Impulse
- Transfer Learning

Assignment 1: Setup Arduino Nicla Vision + Data Engineering for Person Detection

- Part 1:
 - What: You will set up the Arduino Nicla Vision and submit an image of yourself taken with the Nicla. Your image will be included in the testing dataset for Part 2.
 - **How:** You will use the Edge Impulse pipeline to bootstrap your system.

Assignment 1: Setup Arduino Nicla Vision



Assignment 1: Setup Arduino Nicla Vision + Data Engineering for Person Detection

• Part 1:

- What: You will set up the Arduino Nicla Vision and submit an image of yourself taken with the Nicla. Your image will be included in the testing dataset for Part 2.
- **How:** You will use the Edge Impulse pipeline to bootstrap your system.
- Part 2:
 - What: You will gather training data for a person detection model.
 - **How:** You will use Edge Impulse + OpenMV tools to get your training data.

Assignment 1: Collect Training Data for Person Detection Model









Keyword Spotting v. General Speech Recognition

- Keyword spotting is one of the most successful examples of TinyML
 - Low-power, continuous, on-device, machine learning
 - Common Voice SWTS expands keyword spotting to more languages

Keyword Spotting v. General Speech Recognition

- Keyword spotting is one of the most successful examples of TinyML
 - Low-power, continuous, on-device
 - Common Voice SWTS expands keyword spotting to more languages
- General ASR still requires larger, power-hungry models
 - But it can run on mobile devices (offline dictation on smartphones)



Now Playing by Google







Imagine a world where every device has a voice assistant in *your* favorite language.



你好 HALLO 안녕 HOLA नमस्ते CIAO ΗΕLLΟ こんにちは привет BONJOUR LA







• Speech commands for the whole planet?



- Speech commands for the **whole planet**?
- For **more than** just voice assistants





People





People

How can we democratize ML?

Data Engineering

Requirements

- Problem definition
- Permissions & rights
- Machine & human usable format

Data Engineering

Requirements

Gathering

- Problem definition
- Permissions & rights
- Machine & human usable format
- People
- Collection
- Labeling
- Data sources

Data Engineering

Requirements

Gathering

Refinement

- Problem definition
- Permissions & rights
- Machine & human usable format
- People
- Collection
- Labeling
- Data sources

- Processing
- Validation
- Augmentation
Data Engineering

Requirements

Gathering

Refinement

Sustainment

- Problem definition
- Permissions & rights
- Machine & human usable format
- People
- Collection
- Labeling
- Data sources

- Processing
- Validation
- Augmentation

- Storage
- Security
- Errors
- Versioning

Data Engineering

Requirements

Gathering

Refinement

Sustainment

- Problem definition
- Permissions & rights
- Machine & human usable format
- People
- Collection
- Labeling
- Data sources

- Processing
- Validation
- Augmentation

- Storage
- Security
- Errors
- Versioning

Data engineering requires significant capital & effort.

Classifying Images



(<u>Deng et al., 2009</u>)

Detecting Objects

• Common Objects in Context (COCO)—2.5M+ segmented images



(Lin et al., 2014)

Datasets require *significant effort*

- Waymo—1,950 20-second driving segments (cameras, LIDAR, labels)
- KITTI 360-73KM+ of annotated driving data



Datasets require *significant effort*

These massive machine learning datasets are constructed by hand

- Common Voice—5000+ hours of spoken audio
- Common Objects in Context (COCO)—2.5M+ labeled images
- ImageNet-4M+ labeled images
- Waymo—1,950 20-second driving segments
- **KITTI 360–73KM+** of annotated driving data

Data Engineering is costly and tedious.

Public vs. Non-public datasets





(<u>Reddi et al., 2021</u>)

How do we build 1000 words for 1000 languages?

Cost Model v. Community Model?





Cost Model v. Community Model?







Social Good

https://commonvoice.mozilla.org

(Ardila et al., 2019)

• Crowdsourcing platform









- Crowdsourcing platform
- Over 50,000 volunteers

Common Voice is Mozilla's initiative to help teach machines how real people speak.

Voice is natural, voice is human. That's why we're excited about creating usable voice technology for our machines. But to create voice systems, developers need an extremely lage amount of voice data.

Most of the data used by large companies isn't available to the majority of people. We think that stifles innovation. So we've launched Common Voice, a project to help make voice recognition open and accessible to everyone.

READ MORE



- Crowdsourcing platform
- Over 50,000 volunteers
- 54 different languages

We're building

an open source, multi-language dataset of voices that anyone ca use to train speech-enabled applications.

We believe that large, publicly available voice datasets will foster innovation and healthy commercial competition in machine-learning based speech technology.

Common Voice's multi-language dataset is already the largest publicly available voice dataset of its kind, but it's not the only one.

Look to this page as a reference hub for other open source voice datasets and, as Common Voice continues to grow, a home for our release updates.

Validated Hours **5,671**

Recorded Hours

Language -German French Welsh Breton Chuvash Turkish Tatar Kyrgyz Irish Kabyle Catalan Chinese (Taiwan) Slovenian Italian Dutch Hakha Chin Esperanto Estonian Persian Basque Spanish Chinese (China) Mongolian Sakha Dhivehi Kinvarwanda Swedish Russian Indonesian Arabic Tamil Interlingua Portuguese Latvian Japanese Votic Abkhaz Chinese (Hong Kong) Romansh Sursilvan Sorbian, Upper Romanian Frisian Czech Greek Romansh Vallader Polish Assamese Ukrainian Maltese Georgian Puniabi Odia Vietnamese

What's inside the Common Voice dataset? 06-2

- **Crowdsourcing** platform
- Over 50,000 volunteers
- 54 different languages
- Goal: speech recognition for all languages on the planet



ψ

Recorded Hours

7 226

Display a menu

What's inside the **Common Voice** dataset?

V

















How do we democratize the data engineering pipeline?

Specify Wanted Keyword
1. Up 2. Down 3. Yes 4. No 5 6 265









(Mazumder et al., 2021)





(Mazumder et al., 2021)

Nine-Language Embedding Model



Nine-Language Embedding Model



Generalizing to Any Language


Generalizing to Any Language











How do we widen access to applied machine learning?





••• <>		O 🗎 tinymlx.org						Ċ	0 1 4 1	Ø
	Syllabus Timesheet OpenReview D	iscourse edX Insights Disc1 Disc2 Disc3 FixMe Frame.io g141 c1	41 gTinyML-EdX	NAE ASCR_Xstack_F	FY21 tinyML paper	ogle Sheets Apr2-Culture & Be	longing MLCommons 2	Nomination A Chat with A Al -	YouTube Discussion TinyML3 edX	
	C 1000 words in 1000 languages - 3.29.2021 - Google Slides	🐣 Data Engineering for Everyone SIGARCH			G dataset icon - Google Search				Welcome to the Tiny Machine Learning Open Education Initiative (TinyMLx) - TinyMLx	+
		TinyMLx The Tiny Machine Learning Open Education Initiative	Home	Courses	Discuss	TinyML4STEM	TinyML4D	Team		

Welcome to the Tiny Machine Learning Open Education Initiative (TinyMLx)

TinyML is a cutting-edge field that brings the transformative power of machine learning (ML) to the performance- and powerconstrained domain of embedded systems. Successful deployment in this field requires intimate knowledge of applications, algorithms, hardware, and software.

Take an edX Course to Lean More Build and Teach your own TinyML Course

Take one of our exciting courses!



Foundations of TinyML

Focusing on the basics of machine learning and embedded systems, such as smartphones, this course will introduce you to the "language" of TinyML.



Applications Of TinyML

Get the opportunity to see TinyML in practice. You will see examples of TinyML applications, and learn first-hand how to train these models for tiny applications such as keyword soottino. visual wake words. and gesture



Deploying TinyML

Learn to program in TensorFlow Lite for microcontrollers so that you can write the code, and deploy your model to your very own tiny microcontroller. Before you know it, you'll be implementing an entire TinVML application.

An Applied Machine Learning Journey







Topics to Scribe

- Data Engineering Challenges
- Data Engineering Collection Methods
- Democratizing the Data Engineering Pipeline

Paper Discussions

Paper 1: Model Cards for Model Reporting

<u>Link</u>

Paper 2: Multilingual Spoken Words Corpus

Link

Guest Speaker

Guest Speaker – Kasia Chmielinski

Kasia is the Co-Founder of the <u>Data Nutrition Project</u> and will dive into the importance of understanding datasets beyond numbers. We will learn why nutritional labels are crucial for responsible Al and how we can interpret them effectively.

With hands-on experience mitigating bias in AI and building tools for responsible data systems, Kasia brings a unique and crucial perspective to our exploration of tiny machine learning.

