Embedded AI: Principles, Algorithms, and Applications

Table of contents

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, tinyML 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.

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 AI: 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.

Prerequisites

Venturing into “Embedded AI: Principles, Algorithms, and Applications” does not mandate you to be a maestro in machine learning from the outset. At its core, this resource seeks to nurture learners who bear a fundamental understanding of systems and harbor a curiosity to explore the confluence of disparate, yet interconnected domains: embedded hardware, artificial intelligence, and software. This confluence forms a vibrant nexus where innovations and new knowledge streams emerge, making a basic grounding in system operations a pivotal tool in navigating this dynamic space.

Moreover, the goal of this book is to delve into the synergies created at the intersection of these fields, fostering a learning environment where the boundaries of traditional disciplines blur to give way to a holistic, integrative approach to modern technological innovations. Your interest in unraveling embedded AI technologies and low-level software mechanics would be guiding you through a rich learning experience.

Conventions Used in This Book

Please follow the conventions listed in Conventions

How to Contact Us

Please contact *vj@eecs.harvard.edu*

How to Contribute

Please see instructions at here.

Contributors

Please see Credits.

Credits

coming soon.

1 About Us

1.1 Who's This Book For

1.2 Course Structure

1.3 Course Requirements

1.4 Course Materials

1.5 What You'll Learn

2 Dedication

3 Introduction

3.1 AI for Beginners

3.2 Machine Learning and Deep Learning

3.3 Machine Learning

3.4 Deep Learning

3.5 Applications of Deep Learning

3.6 Quiz

4 Deep Learning

4.1 What are Neural Networks

4.2 What is Deep Learning Training

4.3 What is Deep Learning Inference

5 Embedded Systems

5.1 Sensors

5.2 Power

coming soon.

6 Embedded ML

6.1 CloudML

6.2 EdgeML

6.3 TinyML

6.3.1 TinyML for IoT Systems

6.3.2 How does TinyML Work

6.3.3 Resources are Limited, but so is the Competition

6.4 Exercises

7 ML Workflow

7.1 Data Collection

7.2 Pre-Processing

7.3 Training

7.4 Optimization

7.5 Deployment

7.6 Evaluation

7.7 Quiz

8 Data Engineering

8.1 Data Sources

8.2 Training Data

8.3 Training Data Splits

8.4 Data Labeling

8.5 Types of Data

9 Pre-processing

9.1 What is Data Pre-processing?

9.2 What's Involved with Data Pre-processing?

9.3 What's The Importance Of Data Pre-Processing?

10 ML Frameworks

coming soon.

11 Model Training

11.1 Selecting a Training Dataset

11.2 Neural Network Architectures

11.2.1 Multilayer Perceptron (MLP)

11.2.2 Convolutional Neural Networks

11.2.3 Recurrent Neural Networks

11.2.4 Transformers

11.3 Back Propagation

11.4 Convergence

11.5 Overfitting and Underfitting

11.6 Hyperparameters

11.6.1 Epochs

11.6.2 Learning Rate

11.7 Transfer Learning

11.7.1 Optimizer

11.8 Summary

11.9 Quiz

12 Efficient AI

coming soon.

13 Optimizations

13.1 Software Optimizations

13.1.1 Compression

13.1.2 Quantization

13.1.3 Weight Pruning

13.1.4 Knowledge Distillation

13.2 Hardware Optimizations

13.2.1 GPUs

13.2.2 TPUs

13.2.3 NPUs

14 Deployment

15 On-Device Learning

15.1 Federated Learning

15.2 On-Device Training

coming soon.

16 Hardware Acceleration

coming soon.

17 MLOps

18 AI Sustainability

coming soon.

19 Responsible AI

coming soon.

20 Generative AI

coming soon.

References

Acknowledgements

coming soon.