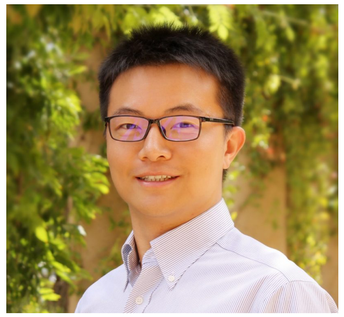
MIT HAN LAB

EfficientML.ai Course Summary



Song Han

Associate Professor, MIT
Distinguished Scientist, NVIDIA

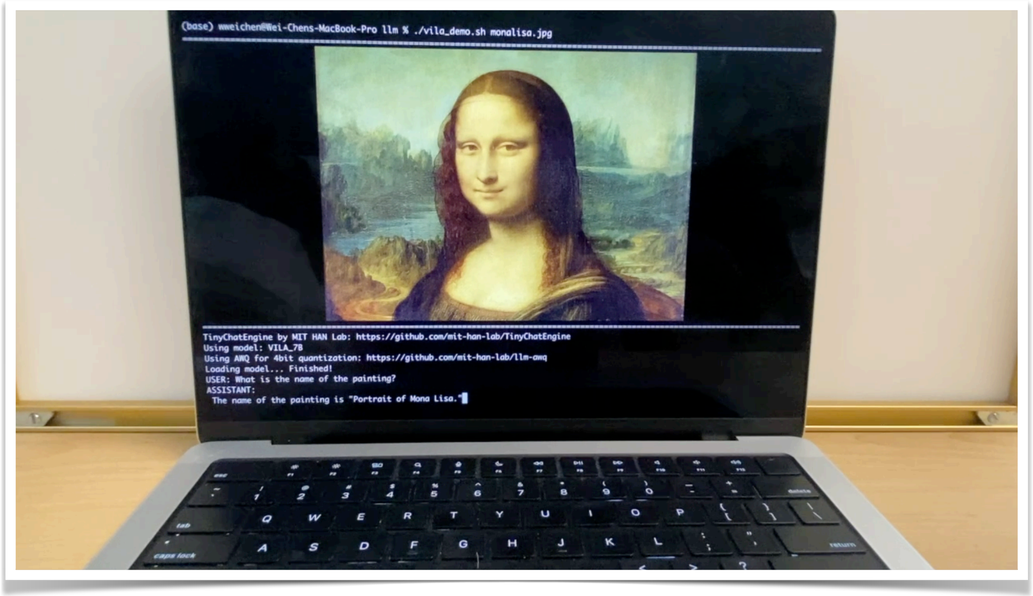
@SongHan_MIT

Course Summary

TinyML and Efficient Deep Learning Computing

This course introduces efficient AI computing techniques that enable **powerful deep learning applications** on **resource-constrained devices**.

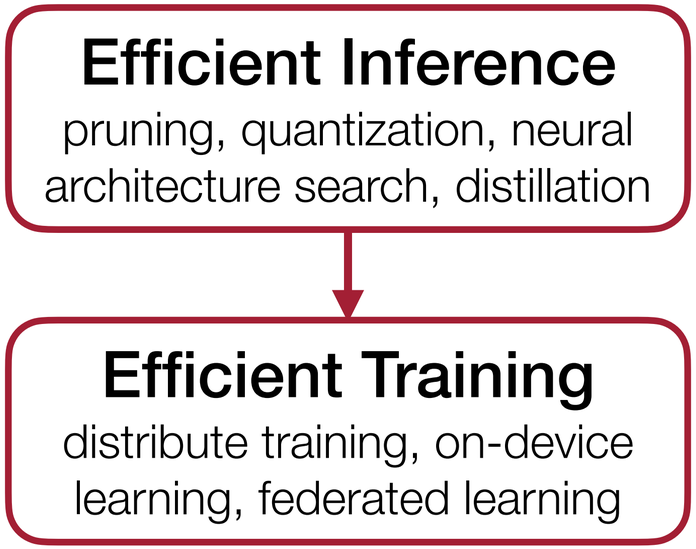
Students get **hands-on experience** implementing model compression techniques and efficient LLM inference library to **deploy large language models (Llama2-7B) on a laptop**.



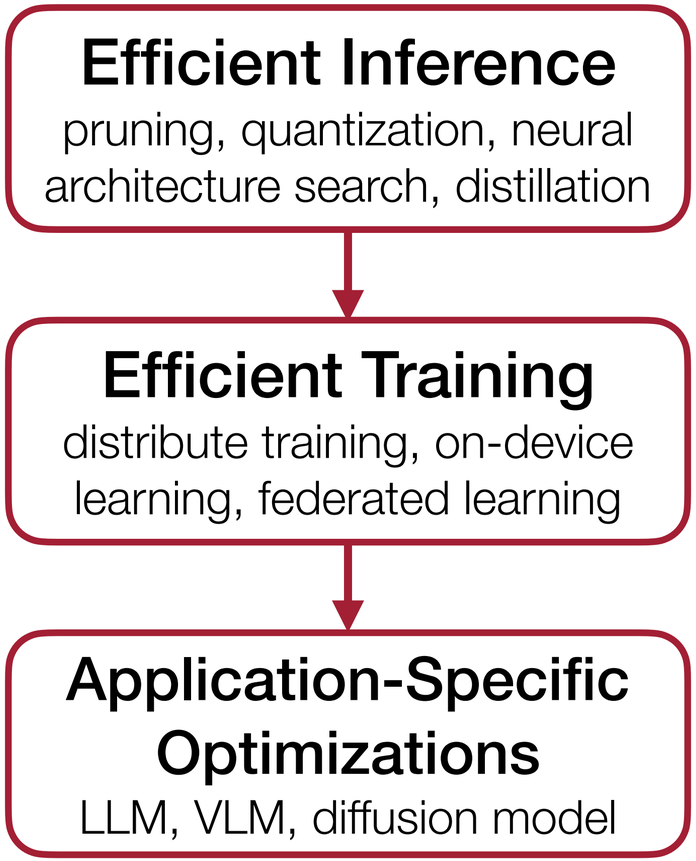
Course Overview

Efficient Inference
pruning, quantization, neural architecture search, distillation

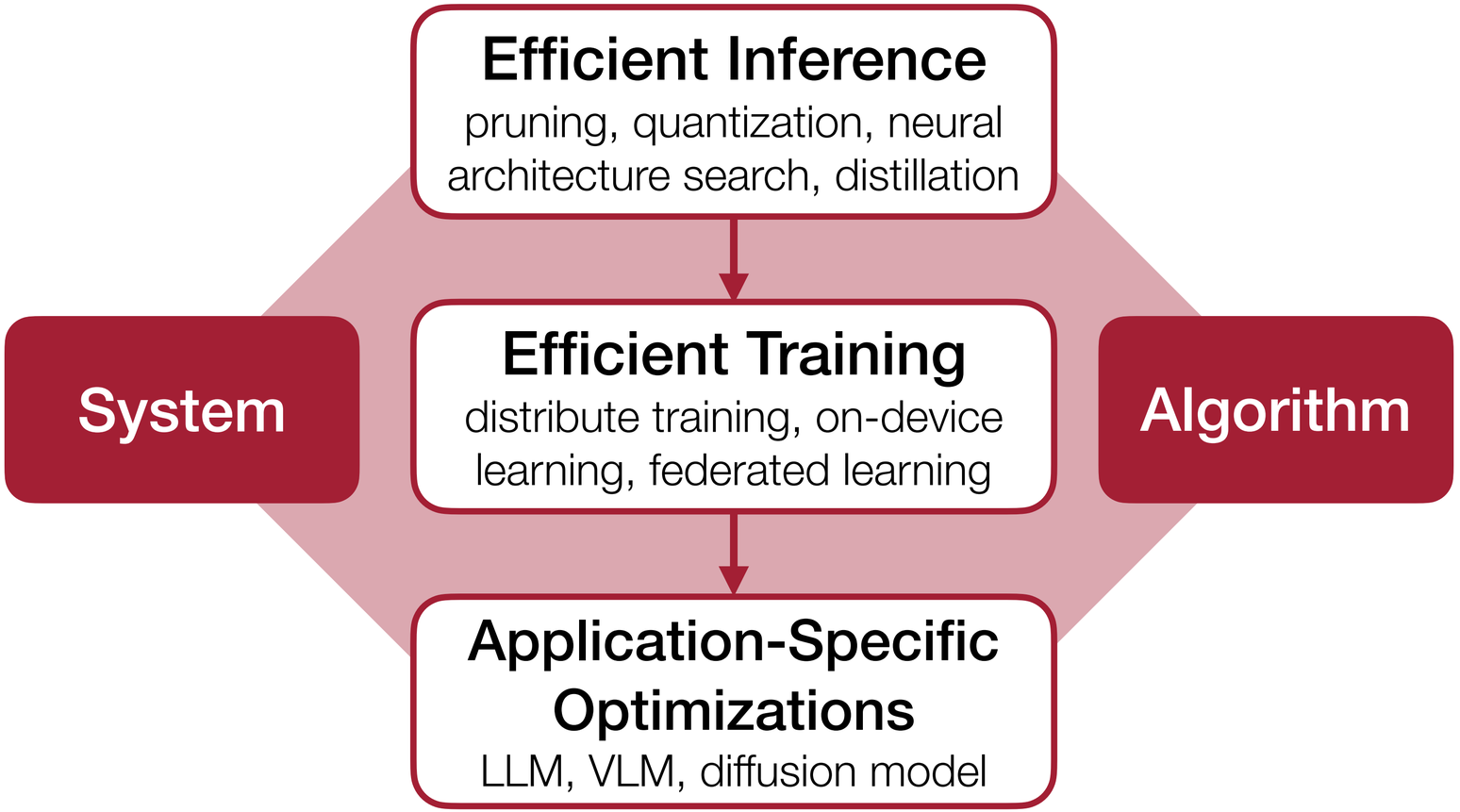
Course Overview



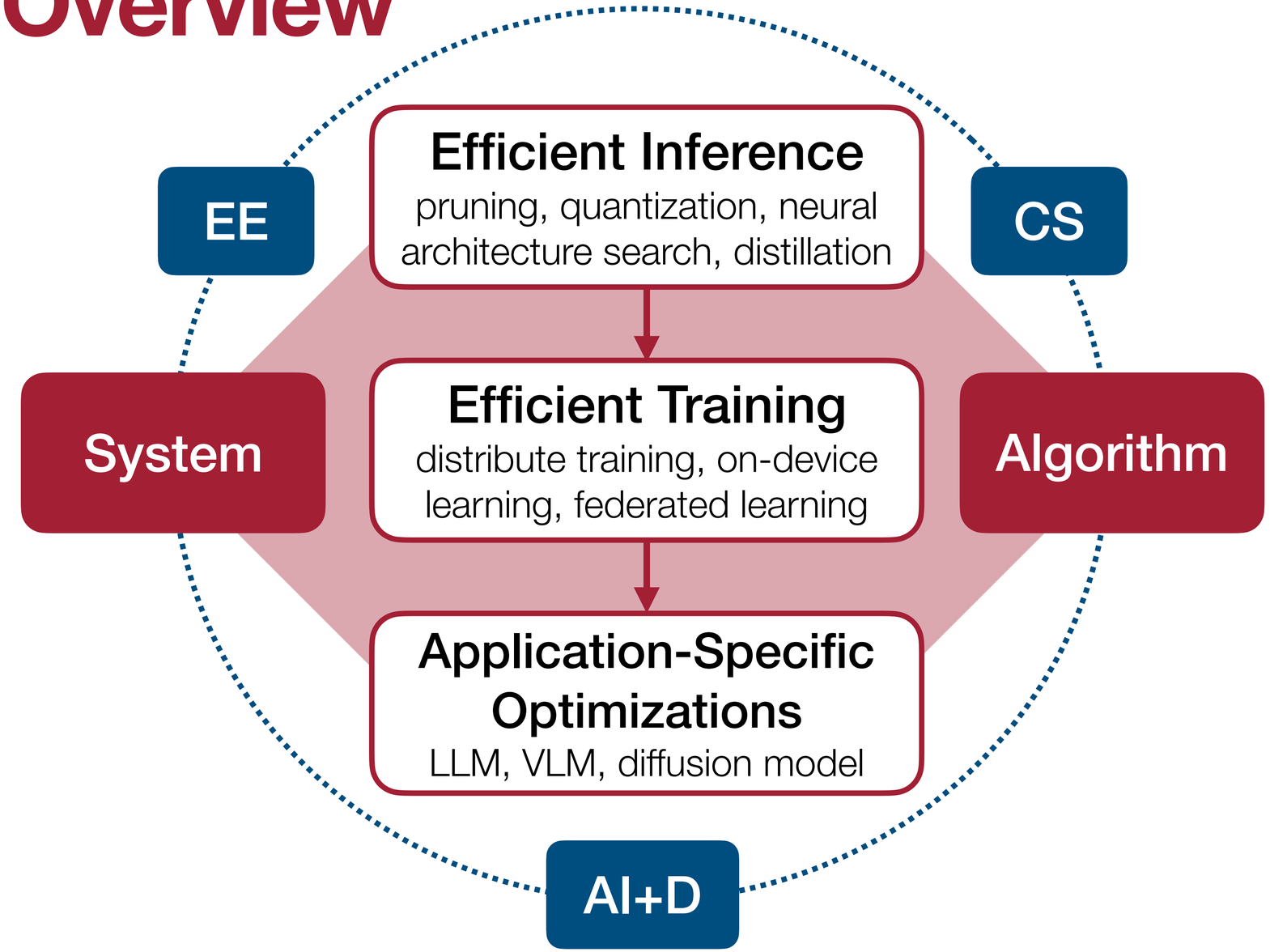
Course Overview



Course Overview



Course Overview

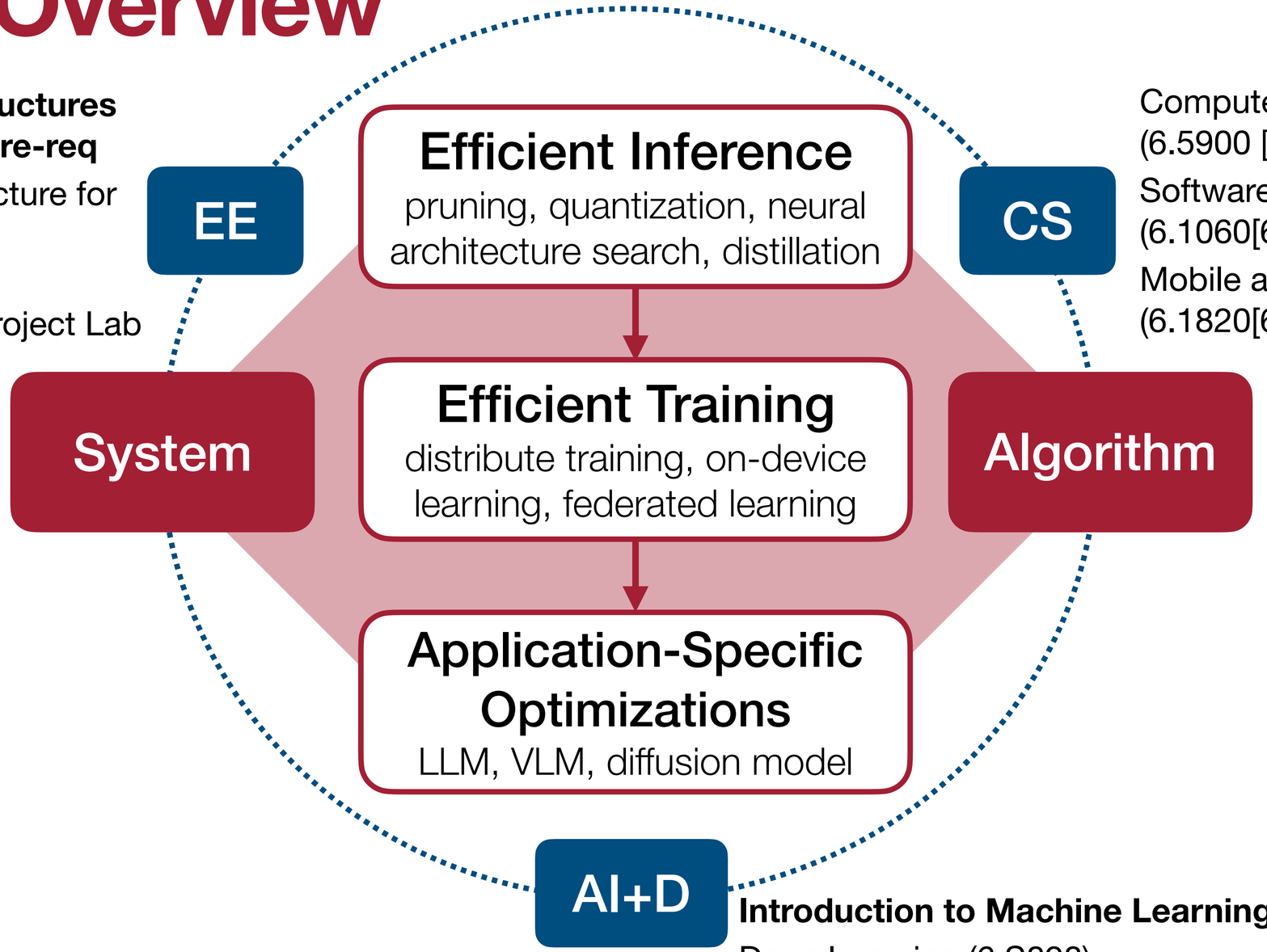


Course Overview

Computation Structures (6.1910[6.004]), pre-req

Hardware Architecture for Deep Learning (6.5930[6.825])

Microcomputer Project Lab (6.2060[6.115])



Computer System Architecture (6.5900 [6.823])

Software Performance Engineering (6.1060[6.172])

Mobile and Sensor Computing (6.1820[6.808])

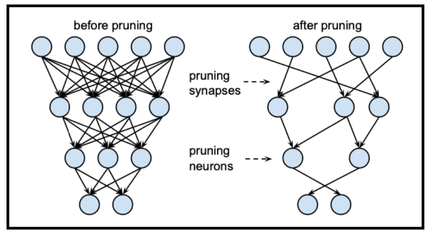
Introduction to Machine Learning (6.3900[6.036]), pre-req

Deep Learning (6.S898)

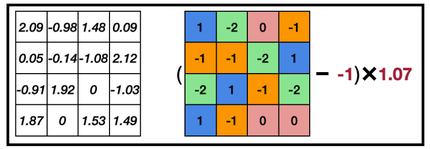
Advances in Computer Vision (6.8300)

Lecture Structure

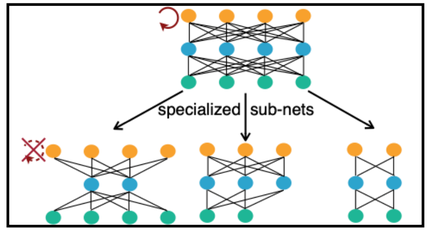
Efficient Inference



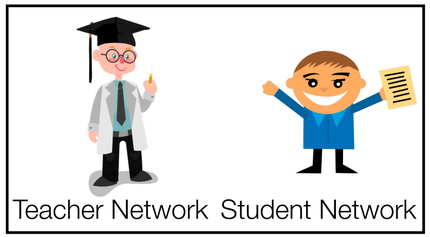
Pruning



Quantization

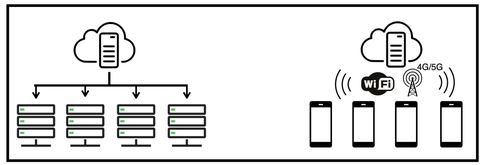


Neural Architecture Search

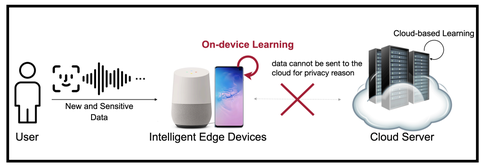


Knowledge Distillation

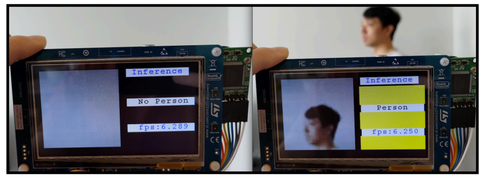
Efficient Training



Distributed Training

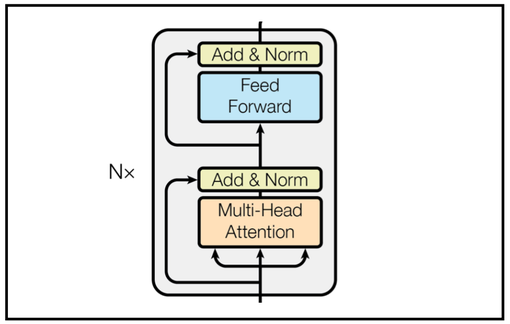


On-Device Learning

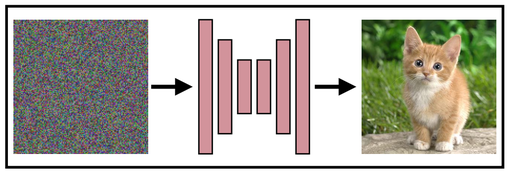


TinyEngine on MCU

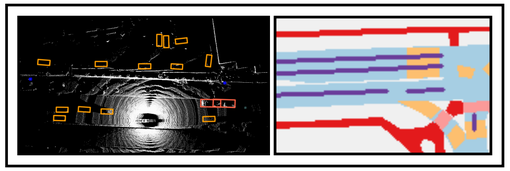
Domain-Specific Optimization



Large Language Models



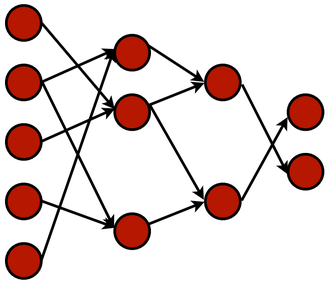
Diffusion Models



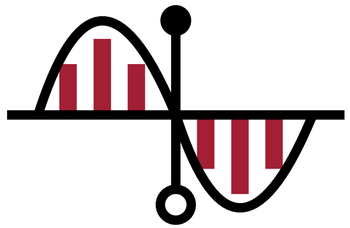
Autonomous Driving

Labs

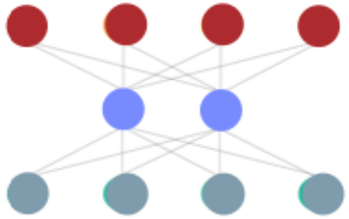
Lab 0 — Getting Started with PyTorch



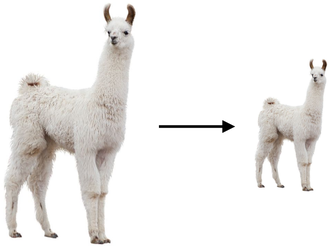
Lab 1 — Pruning



Lab 2 — Quantization



Lab 3 — Neural Architecture Search



Lab 4 — LLM Compression

Lab 5 — LLM Deployment on Laptop

Logistics - Final Project

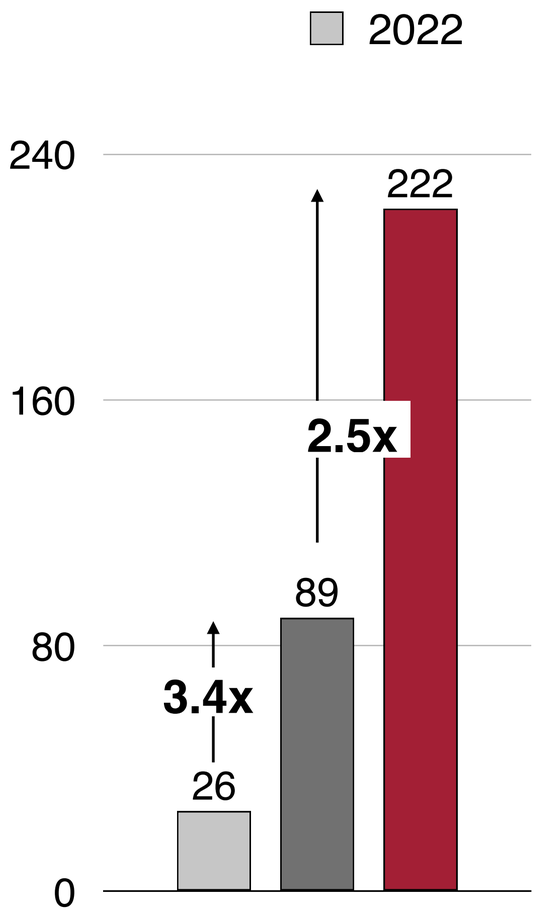
- Poster presentation will be on **Dec 3/Dec 5/Dec 10**
  - Demo is encouraged!

- Please **submit the poster** and **written report** by **Dec 14, 11:59 pm ET.**
  - PDF format, **4 pages at least**.
  - Use NeurIPS template: https://neurips.cc/Conferences/2023/PaperInformation/StyleFiles,
  - With Github link to open source the code.
  - We highly encourage adding a demo video link in the report.
  - Each team only needs to submit one copy of the poster and report, with everyone's name on it.

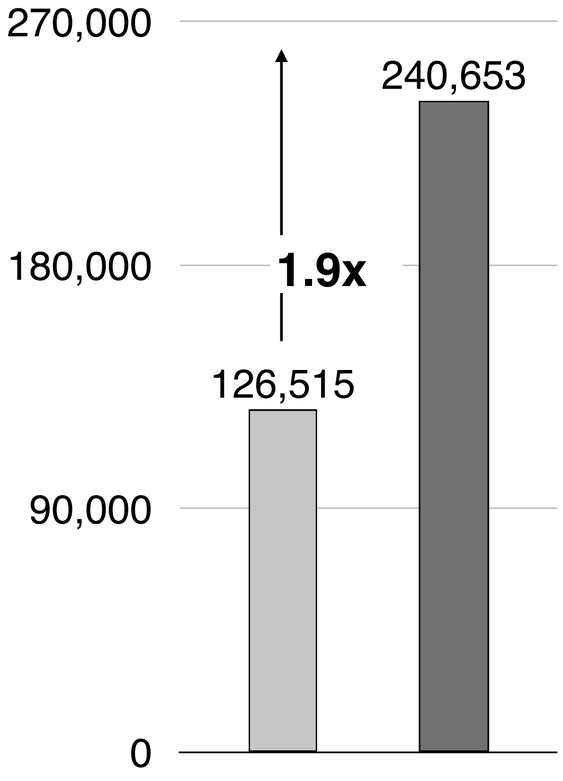
- Rubric:

Motivation	Technical Soundness	Novelty	Evaluation	Poster Presentation	Written Report	Open Source
10	10	10	10	10	10	10

We are growing!



Total Students



 # Views on Youtube

Subject Evaluation

- Complete the end-of-term subject evaluation **December 2** - **December 15**.
  - We'll give 4 participation bonus points for those who submits the evaluation.
  - Please upload the submission confirmation as a proof on Canvas by **Dec 16**. Your support is appreciated.
  - http://registrar.mit.edu/subjectevaluation