CSCI-GA.3033 Special Topics: Efficient AI Computing: Algorithm and Implementation

Fall 2025

Instructor: Sai Qian Zhang **Office Hour:** Fri 1:30 - 2:30PM (link)
Email: sai.zhang@nyu.edu **Office:** Rm 1003, 370 Jay Street

Course Pages: `https://www.saiqianzhang.com/COURSE/`

Course email: `efficientaiaccelerator@gmail.com`

Lecture Time: Wed 7:10 - 9:10PM

Main References: Course slides and Goodfellow, Ian. "Deep learning." (2016).

Description: The course will focus on recent advancements in the design of efficient neural networks, specifically on how to create and optimize AI models for improved performance, and resource efficiency. Students will explore essential techniques such as pruning, quantization, and model distillation across different model architectures like CNNs, RNNs, Transformers, and LLMs. These techniques aim to reduce computational complexity while preserving accuracy. Additionally, the course will cover efficient training and inference methods, including distributed computing, parallelism, and low-precision computation, essential for deploying AI on resource-constrained platforms. Lastly, students will gain a foundational understanding of computer architectures and learn how to deploy AI algorithms on actual edge devices.

Grading Policy: Coding assignments (30%), In-class quiz (15%), Midterm (25%), Final project (30%).

Course Structure: The course will involve 13 lectures, 3 coding assignments 1 final project, in-class quiz and 1 midterm exam.

Course Schedule (Tentative):

Lecture #	Date	Topic	Description
Lecture 1	Sep 3	Intro to Basic topics of DNN	• Deep Neural Networks Basics • Introduction on Efficient AI
Lecture 2	Sep 10	Intro to Convolutional Neural Networks	• Basics of convolutional operations • Batch normalization, layer normalizations, RMS norm, ReLU, GeLU • Popular CNN architectures: MobileNet, DenseNet, SqueezeNet.

Lecture 3	Sep 17	Intro to Transformer and Large Model	• Transformer Basics, vision transformer basics • LLM Basics, RLHF, KV cache • Vision-language model
Lecture 4	Sep 24	Neural Network Pruning	• Different pruning techniques • Sparse matrix encoding for efficiency storage • CNN pruning, transformer pruning
Lecture 5	Oct 1	Neural Network Quantization	• Different types of DNN quantization • Quantization-aware training • Post-training quantization
Lecture 6	Oct 8	Distillation, Low Rank Decomposition and NAS	• Low-rank factorization • Reparameterization • Neural architecture search (NAS)
Lecture 7	Oct 15	Efficient Algorithm for Large Model	• Data distribution of large model • Large model pruning • Large model quantization
Lecture 8	Oct 22	Efficient DNN Training	• Efficient training of DNNs • Parameter efficient finetuning • Federated Learning
No lecture	Oct 29	Midterm	No class
Lecture 9	Nov 5	Distributed System for DNN Training and Inference	• Federated Learning Continue • Distributed DNN Training • Distributed DNN Inference
Lecture 10	Nov 12	Machine Learning System for Large Model	• Speculative Decoding • Flash Attention & Flash Decoding • System and Algorithm Codesign

Lecture 11	Nov 19	AI Accelerator Introduction and CNN Accelerators	• Convolutional operation conversion to Matmul • Hardware architecture of CNN accelerator • Systolic array-based CNN accelerator
No Lecture	Nov 26	Legislative Friday	Classes meet according to a Friday schedule.
Lecture 12	Dec 3	Transformer & LLM Accelerators	• Hardware design for nonlinear blocks, system optimization of LLMs • Popular transformer accelerator design
Lecture 13	Dec 10	New playground for Efficient AI: AR/VR	Invited talk
No lecture	Dec 17	Final Presentation	No class