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

Fall 2026

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

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

Course Email: efficientaiaccelerator@gmail.com

Course Prerequisite: Basic knowledge on deep learning

Lecture Time: Wed 7:10 - 9:10 PM (link)

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, 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.

Course Objectives: By successfully completing this course, students will be able to:

- Understand core principles of efficient AI algorithms.
- Design and optimize scalable AI models for real-world applications.
- Critically analyze and present recent research in efficient AI.

Discussion board: Discussion board is available in Brightspace.

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 2	Intro to Basic topics of DNN	• Deep Neural Networks Basics • Introduction on Efficient AI

Lecture 2	Sep 9	Intro to Convolutional Neural Networks	• Basics of convolutional operations • Batch normalization, layer normalizations, RMS norm, ReLU, GeLU • Popular CNN architectures
Lecture 3	Sep 16	Intro to Transformer and Large Model	• Transformer Basics, vision transformer basics • LLM Basics, RLHF, KV cache • Vision-language model
Lecture 4	Sep 23	Neural Network Pruning	• Different pruning techniques • Sparse matrix encoding for efficiency storage • CNN pruning, transformer pruning
Lecture 5	Sep 30	Neural Network Quantization	• Different types of DNN quantization • Quantization-aware training • Post-training quantization
Lecture 6	Oct 7	Distillation, Low Rank Decomposition and NAS	• Low-rank factorization • Reparameterization • Neural architecture search (NAS)
No lecture	Oct 14	Legislative Monday	No lecture - Classes meet according to a Monday schedule
Lecture 7	Oct 21	Efficient Algorithm for Large Model	• Data distribution of large model • Large model pruning • Large model quantization
Lecture 8	Oct 28	Efficient DNN Training	• Efficient training of DNNs • Parameter efficient finetuning • Federated Learning
No lecture	Nov 4	Midterm	No class

Lecture 9	Nov 11	Distributed System for DNN Training and Inference	• Federated Learning Continue • Distributed DNN Training • Distributed DNN Inference
Lecture 10	Nov 18	Machine Learning System for Large Model	• Speculative Decoding • Flash Attention & Flash Decoding • System and Algorithm Codesign
Lecture 11	Nov 25	AI Accelerator Introduction and CNN Accelerators	• Convolutional operation conversion to Matmul • Hardware architecture of CNN accelerator • Systolic array-based CNN accelerator
Lecture 12	Dec 2	Transformer & LLM Accelerators	• Hardware design for nonlinear blocks, system optimization of LLMs • Popular transformer accelerator design
Lecture 13	Dec 9	Guest Lecture	Invited talk on efficient AI

Grading Policy: Coding assignments (30%), In-class quiz (10%), Midterm (30%), Final project (30%).

Generative Artificial Intelligence Use: Generative AI tools may be used to support learning (e.g., clarifying course materials, improving project report writing), but they may not be used to generate assignment answers.