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

Fall 2025


**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 3 | Intro to Basic topics of DNN | • Deep Neural Networks Basics<br>• Introduction on Efficient AI |

| | | | |
|---|---|---|---|
| Lecture 2 | Sep 10 | Intro to Convolutional Neural Networks | • Basics of convolutional operations<br>• Batch normalization, layer normalizations, RMS norm, ReLU, GeLU<br>• Popular CNN architectures: MobileNet, DenseNet, SqueezeNet. |
| Lecture 3 | Sep 17 | Intro to Transformer and Large Model | • Transformer Basics, vision transformer basics<br>• LLM Basics, RLHF, KV cache<br>• Vision-language model |
| Lecture 4 | Sep 24 | Neural Network Pruning | • Different pruning techniques<br>• Sparse matrix encoding for efficiency storage<br>• CNN pruning, transformer pruning |
| Lecture 5 | Oct 1 | Neural Network Quantization | • Different types of DNN quantization<br>• Quantization-aware training<br>• Post-training quantization |
| Lecture 6 | Oct 8 | Distillation, Low Rank Decomposition and NAS | • Low-rank factorization<br>• Reparameterization<br>• Neural architecture search (NAS) |
| Lecture 7 | Oct 15 | Efficient Algorithm for Large Model | • Data distribution of large model<br>• Large model pruning<br>• Large model quantization |
| Lecture 8 | Oct 22 | Efficient DNN Training | • Efficient training of DNNs<br>• Parameter efficient finetuning<br>• Federated Learning |
| No lecture | Oct 29 | Midterm | No class |
| Lecture 9 | Nov 5 | Distributed System for DNN Training and Inference | • Federated Learning Continue<br>• Distributed DNN Training<br>• Distributed DNN Inference |

| | | | |
|---|---|---|---|
| Lecture 10 | Nov 12 | Machine Learning System for Large Model | • Speculative Decoding<br>• Flash Attention & Flash Decoding<br>• System and Algorithm Codesign |
| Lecture 11 | Nov 19 | AI Accelerator Introduction and CNN Accelerators | • Convolutional operation conversion to Matmul<br>• Hardware architecture of CNN accelerator<br>• 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<br>• 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 |

**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.

**Moses Center Statement of Disability:** If you are a student with a disability who is requesting accommodations, please contact New York University's Moses Center for Student Accessibility (CSA) at 212-998-4980 or mosescsa@nyu.edu. You must be registered with CSA to receive accommodations. Information about the Moses Center can be found at www.nyu.edu/csa. The Moses Center is located at 726 Broadway on the 2nd floor.