Course Code	Course Title
COMP 4901Y	**Special Topics in Computer Science**

Course Description

Selected topics of current interest to the Department not covered by existing courses. Offerings are announced each semester. May be graded by letter, P/F, or DI/PA/F for different offerings.

Topic to be covered:

In recent years, foundation models have fundamentally revolutionized the state-of-the-art of artificial intelligence. Thus, the computation in the training or inference of the foundation model could be one of the most important workflows running on top of modern computer systems. This course unravels the secrets of the efficient deployment of such workflows from the system perspective. Specifically, we will i) explain how a modern machine learning system (i.e., PyTorch) works; ii) understand the performance bottleneck of machine learning computation over modern hardware (e.g., Nvidia GPUs); iii) discuss four main parallel strategies in foundation model training (data-, pipeline-, tensor model-, optimizer- parallelism); and iv) real-world deployment of foundation model including efficient inference and fine-tuning.

List of Topics

Date	Topic
W1 - 01/31	Introduction and Logistics
W2 - 02/05, 02/07	Machine Learning Preliminary & PyTorch Tensors
W3 - 02/14	Stochastic Gradient Descent
W4 - 02/19, 02/21	Auto-Differentiation & PyTorch Autograd
W5 - 02/26, 02/28	Nvidia GPU Performance & Collective Communication Library
W6 - 03/04, 03/06	Transformer Architecture & Large Scale Pretrain Overview
W7 - 03/11, 03/13	Data Parallel Training & Pipeline Parallel Training
W8 - 03/18, 03/20	Tensor Model Parallel Training & Optimizer Parallel Training
W9 – 03/25, 03/27	Mid-Term Review & Mid-Term Exam
W10 - 04/08, 04/10	Generative Inference Workflow & Hugging Face Library
W11 - 04/15, 04/17	Generative Inference Optimization & Speculative Decoding
W12 - 04/22, 04/24	Prompt Engineering Overview & Practices
W13 - 04/29	Parameter Efficient Fine-tuning (LoRA)
W14 - 05/06, 05/08	Guest Speech (TBD) & Final Exam Review

Textbooks (Optional)

N/A

Reference books

N/A

Grading Scheme

4 Homework (4 X 5%)	20%
Mid-term exam	30%
Final exam	50%
Total	100%

Course Intended Learning Outcomes

N/A

Assessment Rubrics
N/A