The Hong Kong University of Science and Technology

UG Course Syllabus

[Course Title] Large-Scale Machine Learning for Foundation Models

[Course Code] COMP4901Y

[No. of Credits] 3

[Any pre-/co-requisites] COMP 3211

Name: [Instructor(s) Name] Binhang Yuan

Email: [Your Email Address] biyuan@ust.hk

Course Description

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, etc.); iv) real-world deployment of foundation model including efficient inference and fine-tuning.

Assessments:

[List specific assessed tasks, exams, quizzes, their weightage]

Assessment Task	Contribution to Overall Course grade (%)
Mid-Term	30%
Homework	20%
Final examination	50%

Required Texts and Materials

N/A.

[Optional] Additional Resources

N/A.