Project for Team 2

Project Title: Drainage Crossing Classification using Convolutional Neural Networks (CNNs)

Objective: To design a ResNet-based Convolutional Neural Network (CNN) for efficient and accurate classification of drainage crossing locations

Description: Remotely sensed images provide an advantageous perspective for geographical feature detection and classification tasks, including drainage crossings. Students will undertake a project to design a suitable ResNet-based CNN model optimized for deployment on a resource-constrained GPU. The project will be divided into three key stages:

1) Model Design and Neural Architecture Search (NAS): The first stage requires students to design a ResNet-based model and use Neural Network Intelligence (NNI) to perform Neural Architecture Search (NAS) on key hyperparameters like the number of layers, the size of kernels, and filters. Here, the constraint on memory used by the model should be less than 100 MB.
2) Performance Prediction and Pareto Front Analysis: Subsequently, students will predict the inference time using nn-Meter, with the aim of keeping it below 150 ms. This predicted inference time and the model accuracy will form a multidimensional optimization problem. Using these factors, students will construct a Pareto front to visualize the trade-off between model accuracy and inference time.
3) Resource Usage and Computational Cost Analysis: In the final stage, students will further analyze the solutions on the Pareto front using Nsight Systems/Compute. They will compare memory usage, computational cost, and other performance metrics. Following this analysis, the optimal CNNs for the given task will be selected.

Expected Outcome: Students will understand how to design, optimize, and implement machine learning models under specific constraints. They will learn how to balance accuracy and efficiency, which is crucial in real-world applications like geographical feature detection.

Reference

[1] Quanlu Zhang, Zhenhua Han, Fan Yang, Yuge Zhang, Zhe Liu, Mao Yang, and Lidong Zhou. Retiarii: A Deep Learning Exploratory-Training Framework. USENIX

Symposium on Operating Systems Design and Implementation (OSDI), pp.919-936, Nov. 2020
[2] https://github.com/microsoft/nni
[3] https://nni.readthedocs.io/en/stable/
[4] Li Lyna Zhang, Shihao Han, Jianyu Wei, Ningxin Zheng, Ting Cao, Yuqing Yang, Yunxin Liu. nn-Meter: towards accurate latency prediction of deep-learning model inference on diverse edge devices. Proceedings of the 19th Annual International Conference on Mobile Systems, Applications, and Services (MobiSys), pp. 81-93, Jun. 2021.
[5] https://github.com/microsoft/nn-Meter
[6] DL-GPU workshop tutorial on “CUDA Kernel Profiler and Performance Analysis Tool”

Project Presentation: 3:00 PM (CST) on Friday, August 4th, 2023
30 minutes for each team
Paper Submission: an 8-page paper to SHDA 2023 in conjunction with SC 2023
https://shda-workshop.github.io