Project for Team 1

Project Title: Drainage Crossing Object Detection using Advanced Convolutional Neural Networks (CNNs)

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

Motivation: Traditionally, the Faster R-CNN is used for object detection to draw bounding boxes. The Faster R-CNN requires the pixel values to be compressed into the 0-255 range (due to the 8-bit requirement), which compromises the data's spatial details. Recently, SPPNet is proposed to be an alternative approach, which allows 32-bit data input.

Description: Remotely sensed images provide an advantageous perspective for geographical feature and object detection tasks, including drainage crossings. Students will undertake a project to design a suitable SPPNet-based CNN model optimized for deployment on a powerful 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 set of SPPNet-based models for different input image sizes and use Neural Network Intelligence (NNI) to perform Neural Architecture Search (NAS) on their network structures and hyperparameters. Here, the memory size used by each model should be less than 24GB.
2) Inference Latency Reduction and Pareto Front Analysis: Subsequently, students will minimize the inference latency of each multi-branch model using the Inter-Operator Scheduler (IOS). This reduced inference latency and the model accuracy will form a multi-objective optimization problem. Using these factors, students will construct a Pareto front to visualize the trade-off between model accuracy and inference latency.
3) Resource Usage and Computational Cost Analysis: In the final stage, students will further analyze the models on the Pareto front using Nsight Systems/Compute. They will compare memory usage, computational cost, and other performance metrics. From the observation, hardware utilization is analyzed for different Pareto-optimal models.

Expected Outcome: Students will understand how to design, optimize, and implement deep learning models targeting a specific platform. They will learn how to balance accuracy and efficiency, which is crucial in real-world applications like geographical feature detection. Also, they will have a deep insight into the resource consumption of different models.

Reference

[1] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 37, no. 9, pp. 1904-1916, Sep. 2015.
[2] https://github.com/bahl24/SpatialPyramidPooling_Keras
[3] 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. pp.919-936, Nov. 2020
[4] https://github.com/microsoft/nni
[5] https://nni.readthedocs.io/en/stable/
[6] Yaoyao Ding, Ligeng Zhu, Zhihao Jia, Gennady Pekhimenko, Song Han. IOS: Inter-Operator Scheduler for CNN Acceleration. Proceedings of Machine Learning and Systems (MLSys), pp. 1-14, Apr. 2021
[7] https://github.com/mit-han-lab/inter-operator-scheduler
[8] 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