Project for Team 2

Title

Exploration of TPU Architectures for Optimized Transformer Performance in Image Detection of Drainage Crossings

Introduction

As high-resolution digital elevation models (HRDEMs) continue to improve the mapping of hydrographic features, the challenge of accurately identifying under-road drainage structures like culverts remains significant. Utilizing the Transformer architecture, renowned for its ability to handle complex visual data through self-attention mechanisms, this project proposes to refine image analysis techniques for environmental science by optimizing the performance of Transformer models through targeted TPU architecture exploration using the Scalesim tool [3].

Objective

This project will explore various TPU architectures using the Scalesim tool, with the goal of identifying configurations that significantly reduce the latency of the Transformer model, thereby enhancing its processing efficiency for HRDEMs.

Methodology

The deep learning model will be trained using a predeveloped dataset from four different watersheds in different U.S. states. The Transformer architecture, known for its effectiveness in processing sequential data and its adaptability to computer vision tasks, will be utilized to develop an object detection model. The model will focus on identifying culvert locations by enhancing CNN architectures with Transformer capabilities.

Training and evaluation of the models will be performed on the compiled dataset, with a split of training, validation, and test sets to ensure generalizability. The models will be evaluated using metrics such as accuracy, precision, recall, and F1 score to assess their performance in classifying and identifying culverts.

Students will conduct simulations with Scalesim [4] (https://github.com/scalesim-project/scale-sim-v2) to evaluate different TPU architectures and their effects on the performance of a Transformer model trained to detect drainage crossings. The methodology includes:

1. Configuring Scalesim for simulating a range of TPU architectures
2. Assessing the impact of each TPU configuration on model latency and computational efficiency.
3. Discussing the performance bottlenecks identified in the Transformer architecture and proposing solutions to mitigate these issues.

Expected Outcomes

- Development of accurate and scalable models based on the Transformer architecture for enhanced identification of drainage crossings, serving as a benchmark for future research in this area.

- Comprehensive analysis detailing the influence of various TPU architectures on the latency and performance of Transformer models in environmental image processing.
- Identification of optimal TPU configurations that effectively balance performance with computational resource demands, potentially guiding similar future applications

Reference

[1] Ananda Samajdar, Jan Moritz Joseph, Yuhao Zhuz, Paul Whatmoughx, Matthew Mattinax, and Tushar Krishna. A Systematic Methodology for Characterizing Scalability of DNN Accelerators using SCALE-Sim. IEEE International Symposium on Performance Analysis of Systems and Software (ISPASS), August 2020.
[2] Ananda Samajdar, Yuhao Zhu, Paul Whatmough, Matthew Mattina, and Tushar Krishna. SCALE-Sim: Systolic CNN Accelerator Simulator. https://arxiv.org/abs/1811.02883 , February 2019
[3] https://scalesim-project.github.io
[4] https://github.com/scalesim-project/scale-sim-v2

Project presentation: 30-minute team presentation on Friday, June 7th, 2024
Paper draft completion deadline: August 2nd, 2024

Paper draft: a 10-page paper including references in the IEEE 2-column format
Paper format should follow IEEE Computer Society Proceedings Manuscript Formatting Guidelines (see the link to "formatting instructions" below).
https://www.ieee.org/conferences/publishing/templates.html