Project for Team 1

Title

Accurate and Time-Efficient Transformer Inference on TPU Architecture for Drought Classification

Introduction

The U.S. Drought Monitor (USDM) plays a vital role in tracking and assessing drought conditions across the United States, utilizing a variety of meteorological and environmental observations. Despite its effectiveness, the dynamic nature of drought phenomena presents challenges in timely and accurate classification. Leveraging the Transformer architecture, renowned for its ability to handle dynamic and complex data through self-attention mechanisms, presents a promising opportunity to enhance the predictive accuracy and operational efficiency of the USDM. This project aims to develop a Transformer-based deep learning model to accurately predict drought classifications and optimize the performance of Transformer models through targeted TPU architecture exploration using the Scalesim tool [5].

Objectives

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 USDMs.

Methodology

The methodology for this project involves several key steps:

- Data Collection: Assemble a comprehensive dataset including historical meteorological data, satellite imagery, and environmental observations relevant to drought conditions, sourced from NOAA, NASA, and other databases.
- Model Development: Utilize the Transformer architecture, known for its effectiveness in handling sequential data and its adaptability to complex pattern recognition tasks, to develop a model capable of synthesizing and predicting drought conditions.
- Model Training and Validation: Train the model on a designated dataset divided into training, validation, and test sets to ensure rigorous learning and generalizability. Utilize cross-validation techniques to optimize model parameters.
- Performance Evaluation: Assess the model using key performance metrics such as accuracy, precision, recall, and the F1 score. Special attention will be given to the model's ability to differentiate between various drought severity levels across different geographical regions.
- TPU Architecture Exploration: Conduct simulations with Scalesim [6] (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 specific tasks include:

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 a robust and scalable model for accurate prediction of drought classifications.
- 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] U.S. Drought Monitor, https://droughtmonitor.unl.edu
[2] Jia-Li Zhang, Xiao-Meng Huang, and Yu-Ze Sun. Multiscale spatiotemporal meteorological drought prediction: A deep learning approach. Advances in Climate Change Research, vol. 15, no. 2, pp. 211-221, April 2024.
[3] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, Illia Polosukhin. Attention is All You Need. Conference on Neural Information Processing Systems (NIPS), Long Beach, CA, USA, December 2017.
[4] 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.
[5] 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
[6] https://scalesim-project.github.io
[7] https://github.com/scalesim-project/scale-sim-v2
[8] Norman P. Jouppi, Cliff Young, Nishant Patil, David Patterson, Gaurav Agrawal, Raminder Bajwa, Sarah Bates, Suresh Bhatia, Nan Boden, Al Borchers, Rick Boyle, Pierre-luc Cantin, Clifford Chao, Chris Clark, Jeremy Coriell, Mike Daley, Matt Dau, Jeffrey Dean, Ben Gelb, Tara Vazir Ghaemmaghami, Rajendra Gottipati, William Gulland, Robert Hagmann, C. Richard Ho, Doug Hogberg, John Hu, Robert Hundt, Dan Hurt, Julian Ibarz, Aaron Jaffey, Alek Jaworski, Alexander Kaplan, Harshit Khaitan, Daniel Killebrew, Andy Koch, Naveen Kumar, Steve Lacy, James Laudon, James Law, Diemthu Le, Chris Leary, Zhuyuan Liu, Kyle Lucke, Alan Lundin, Gordon MacKean, Adriana Maggiore, Maire Mahony, Kieran Miller, Rahul Nagarajan, Ravi Narayanaswami, Ray Ni, Kathy Nix, Thomas Norrie, Mark Omernick, Narayana Penukonda, Andy Phelps, Jonathan Ross, Matt Ross, Amir Salek, Emad Samadiani, Chris Severn, Gregory Sizikov, Matthew Snelham, Jed Souter, Dan Steinberg, Andy Swing, Mercedes Tan, Gregory Thorson, Bo Tian, Horia Toma, Erick Tuttle, Vijay Vasudevan, Richard Walter, Walter Wang, Eric Wilcox, and Doe Hyun Yoon. In-Datacenter Performance Analysis of a Tensor Processing

Unit. Proceedings of the 44th Annual International Symposium on Computer Architecture (ISCA), pp. 1-12, June 2017.

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

Paper submission target: IEEE Big Data 2024
https://www3.cs.stonybrook.edu/~ieeebigdata2024/CallPapers.html