DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING, DEPARTMENT OF EVOLUTION, ECOLOGY AND ORGANISMAL BIOLOGY, OHIO STATE UNIVERSITY, OAK RIDGE NATIONAL LABORATORY

Building an AI powered Assistant for Computational Scientists

OSU: Swathi Vallabhajosyula; Bryan Carstens; Jian Chen; Rajiv Ramnath ORNL: Matthew Wolf
Award: OAC/1945347

Introduction

The availability of high-performance computing (HPC) cyberinfrastructures (CI) like Ohio Supercomputer (OSC)[1] or Titan[2] enables scientists to perform computationally intensive experiments (reaching Exascale levels).

These resources are expensive and hard to optimize:

a) shared between simple/complex tasks – long/short run times and

b) billed per hour based on processing and memory requirements.

We studied how AI injected into existing CIs could provide resource recommendations from application and system configurations, prior workflows and resulting interactions (CI traces, logs).

Directions:

1. Creating a knowledge base from data sources and knowledge sources (e.g. publications)
2. Synthetic data generation to create training data for machine learning models

Publications:

"Towards Practical, Generalizable Machine-Learning Training Pipelines to build Regression Models for Predicting Application Resource Needs on HPC Systems", published in PEARC '22, which explores the potential for cost-effectively developing generalizable and scalable machine-learning-based regression models for predicting the approximate execution time of an HPC application given its input data and parameters. This work examines:

(a) To what extent models trained on scaled-down datasets on commodity environments adapt to production environments,

(b) to what extent models built for specific applications can generalize to other applications within a family, and

(c) how the most appropriate model may change based on the type of data and its mix.

We also describe and show the use of an automatable pipeline for generating the necessary training data and building the model.

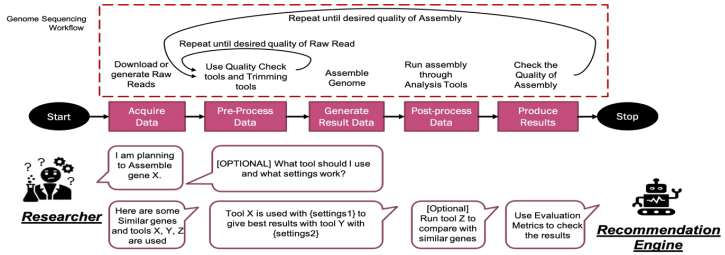

Figure 1: Aligning the Data Analysis phase of Genome Sequencing with a generic Identified workflow pipeline.

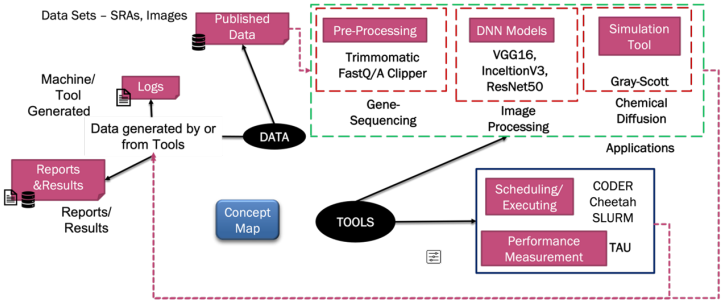

Figure 2: Concept map showing applications and tools in the pipeline

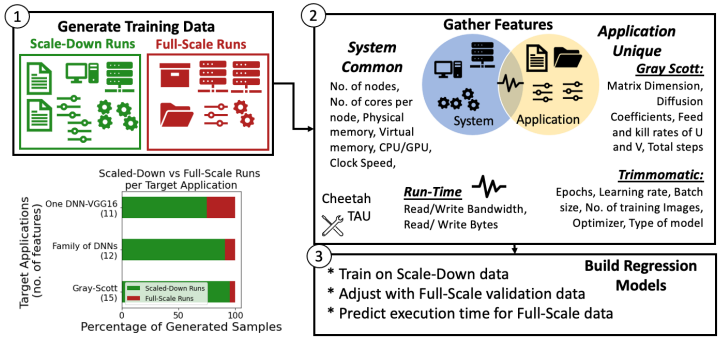

Figure 3: Visualizing the Automated Data Collection Pipeline and Model Generation along with system-specific and application specific features used in our experiments

Training Data and Model:

1. Training Data: By providing scaled-down input to target application we generate scaled down runs with several syste, configurations and application settings. We Generate Full-scale data by running the application on complete input with limited settings and configurations.

2. Model Building: We train a one hidden-layer neural network model based on scaled-down runs, scale the prediction to fit the actual application execution by adjusting the execution time based on full-scale validation data. We gradually add a few full-scale runs to training data to see the model accuracies.

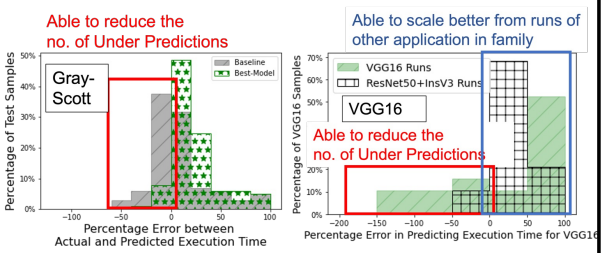

Figure 4: Percentage error between actual and predicted values

Future Work

1. Custom Model to predict execution time for target application
2. Predicting execution time for inference data with missing IO features
3. Establishing an execution time prediction specific cost-models that consider the trade-off between under-over predictions based on execution environments.

References

1. https://www.osc.edu/; https://www.olcf.ornl.gov/olcf-resources/compute-systems/titan/
2. https://github.com/keichi/gray-scott; https://github.com/usadellab/Trimmomatic
3. https://github.com/CODARcode/cheetah
4. Park, N., Mohammadi, M., Gorde, K., Jajodia, S., Park, H., & Kim, Y. (2018). Data synthesis based on generative adversarial networks. *arXiv preprint arXiv:1806.03384*.
5. Marcos Amarís, Raphael Y. de Camargo, Mohamed Dyab, Alfredo Goldman, and Denis Trystram. 2016. A comparison of GPU execution time prediction using machine learning and analytical modeling. In 2016 IEEE 15th International Symposium on Network Computing and Applications (NCA). https://doi.org/10.1109/NCA.2016.7778637

THE OHIO STATE UNIVERSITY

Contact: vallabhajosyula.2; chanana.5 ; duckett.17; Carsten.12; chen.8028; ramnath.6}@buckeyemail.osu.edu; wolfmd@ornl.gov