Supporting a Community of Cancer Models with the CANDLE Checkpoint Module

Rajeev Jain
Mathematics & Computer Science
Argonne National Laboratory
Lemont, IL, USA
jain@mcs.anl.gov

Justin M. Wozniak
Data Science & Learning
Argonne National Laboratory
Lemont, IL, USA
woz@anl.gov

Jamaludin Mohd-Yusof
Computer, Computational & Statistical Sciences
Los Alamos National Laboratory
Los Alamos, NM, USA
jamal@lanl.gov

George Zaki
Bioinformatics and Computational Science
Frederick National Laboratory for Cancer Research
Frederick, MD, USA
george.zaki@nih.gov

Sunita Menon
Bioinformatics and Computational Science
Frederick National Laboratory for Cancer Research
Frederick, MD, USA
sunita.menon@nih.gov

Motivation: The development and use of high-performance machine learning (ML) models for cancer is accelerated by streamlining the ability of researchers to share information, including the ability to cross-validate models across data sets. As models become larger and more complex, the ability to leverage the compute capability of exascale machines to optimize training and hyperparameters will be more beneficial to researchers without access to such resources. As illustrated in Figure 1, we propose herein checkpoint conventions and an associated library to ease the generation, validation, distribution, and reuse of ML models for cancer science.

For example, the National Cancer Institute (NCI) has deployed the Predictive Oncology Model and Data Clearinghouse (MoDaC) [1] repository that contains both data and models, but requires additional metadata beyond that stored in standard checkpointing. The Innovative Methodologies and New Data for Predictive Oncology Model Evaluation (IMPROVE) collaborative NCI-DOE project [2] among other goals, works to validate, understand and improve the latest state-of-art drug-response models. By integrating the required MoDaC metadata (and providing a template for other repositories) into model checkpointing, we enhance the ability of cancer researchers who develop neural networks to make their models more widely available.

The CANDLE Checkpoint Module: The Cancer Deep Learning Environment (CANDLE) [3] is a collection of cancer mini-applications called “Benchmarks” and a workflow framework around the Benchmarks called “Supervisor.” The **`candle_lib`** package is a pip-installable library designed to standardize and streamline machine learning code development and deployment. Originally developed as part of the CANDLE Benchmarks suite, it is now a independently-installable library that provides various utilities including integration with the CANDLE Supervisor layer to automate running complex workflows on exascale machines. At the core of CANDLE software is the notion of CANDLE compliance, a simple API that can be added to user models to support standardized CANDLE *hyperparameters*, which control both network architecture *and* system-level functionality including checkpoints.

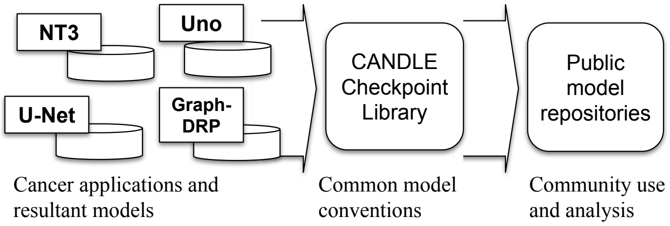
NT3
Uno
U-Net
Graph-DRP
CANDLE Checkpoint Library
Public model repositories
Cancer applications and resultant models
Common model conventions
Community use and analysis

Fig. 1. Conceptual flow of community models into common checkpoint format for analysis and ingest into public repositories, including series (pl.) of checkpoints for capturing model behavior over epochs.

The CANDLE Checkpoint module inside `candle_lib` automates several checkpoint functions useful to both stand-alone Benchmarks and Supervisor workflows. It provides callback interfaces for the popular deep learning frameworks, along with standardized methods for controlling the frequency and number of saved checkpoints, and automatically generates metadata to satisfy the requirements of the various repositories. Additionally, it avoids modifying checkpoints in place, and uses a hard-linking scheme to ensure data consistency if a run crashes during a checkpoint operation.

The associated MoDaC utilities allow the user to interface with the repository directly.

Summary: By including the `candle_lib` package, the IMPROVE project enables checkpointing in a straightforward standard way. It supports multiple ML frameworks and new frameworks can be added as required while providing a consistent interface. This approach enables easy comparison, validation, hyper-parameter optimization and other studies across all the community ML models. We believe that this is critical to advance and standardize rapidly growing research area of cancer drug discovery.

References

[1] "Predictive oncology model and data clearinghouse." [Online]. Available: https://modac.cancer.gov

[2] "Innovative methodologies and new data for predictive oncology model evaluation (improve)." [Online]. Available: https://jdacs4c-improve.github.io/

[3] J. M. Wozniak, R. Jain, P. Balaprakash, J. Ozik, N. Collier, J. Bauer, F. Xia, T. Brettin, R. Stevens, J. Mohd-Yusof, C. G. Cardona, B. V. Essen, and M. Baughman, "CANDLE/Supervisor: A workflow framework for machine learning applied to cancer research," *BMC Bioinformatics*, vol. 19, no. 18, p. 491, 2018. [Online]. Available: https://doi.org/10.1186/s12859-018-2508-4

I. Acknowledgments

This work was supported by the U.S. Department of Energy, Office of Science, Advanced Scientific Computing Research, under contract number DE-AC02-06CH11357.

The following text will be removed in the final submission: