Cancer Research

CANDLE: Exascale Deep Learning–Enabled Precision Medicine for Cancer

The US Department of Energy (DOE) has entered into a partnership with the National Cancer Institute (NCI) of the National Institutes of Health (NIH). This partnership has identified three key science challenges that the combined resources of DOE and NCI can accelerate. The first challenge (called the "drug response problem") is to develop predictive models for drug response that can be used to optimize preclinical drug screening and drive precision medicine–based treatments for cancer patients. The second challenge (called the "RAS pathway problem") is to understand the molecular basis of key protein interactions in the RAS/RAF pathway that is present in 30% of cancers. The third challenge (called the "treatment strategy problem") is to automate the analysis and extraction of information from millions of cancer patient records to determine optimal cancer treatment strategies across a range of patient lifestyles, environmental exposures, cancer types, and health care systems. While these challenges are at different scales and have specific scientific teams collaborating on the data acquisition, data analysis, model formulation, and scientific runs of simulations, they also share several common threads. The CANDLE project focuses on the machine learning aspect of the challenges and in particular builds on a single scalable deep neural network (DNN) code called CANDLE (CANcer Distributed Learning Environment).

The CANDLE challenge problem is to solve large-scale machine learning problems for three cancer-related pilot applications: the drug response problem, the RAS pathway problem, and the treatment strategy problem. For the drug response problem, unsupervised machine learning methods are used to capture the complex, nonlinear relationships between the properties of drugs and the properties of tumors to predict response to treatment, with the goal of developing a model that can provide treatment recommendations for a given tumor. For the RAS pathway problem, multiscale MD runs are guided through a large-scale state-space search using unsupervised learning to determine the scope and scale of the next series of simulations based on the history of previous simulations. For the treatment strategy problem, semi-supervised machine learning is used to automatically read and encode millions of clinical reports into a form that can be computed upon. Each problem requires a different approach to the embedded learning problem, but all approaches are supported with the same scalable deep learning code in CANDLE.

The CANDLE software suite broadly consists of two distinct, interoperating levels: the DNN codes and the Supervisor portion, which handles work distribution across a distributed network. At the DNN level, the CANDLE utility library provides a series of utility functions that streamline the process of writing CANDLE-compliant code. This enables the essential functionality for network hyperparameters to be set either from a default-model file or from the command line. This in turn enables experiments to be designed that sweep across a range of network hyperparameters in an efficient manner. The Supervisor framework provides a set of modules to enable various hyperparameter optimization (HPO) schemes and to automatically distribute the workload across available computing resources. Together, these capabilities allow users to efficiently perform HPO on the large compute resources available across the DOE complex, as well as on any local compute resources.

The challenge for exascale manifests in the need to train large numbers of models. A need inherent to each pilot application requires production of high-resolution models that cover the space of specific predictions (i.e., individualized in the precision medicine sense). Take, for example, training a model that is specific to a certain drug and individual cancer. Starting with 1,000 different cancer cell lines and 1,000 different drugs, a leave-one-out strategy to create a high-resolution model for each drug by cancers requires approximately 1 million models. These models are similar enough that we can use a transfer learning strategy, where weights are shared during training in a way that avoids information leakage, which significantly reduces the time needed to train a large set of models.

PI: Rick Stevens, Argonne National Laboratory

Collaborators: Argonne National Laboratory, Oak Ridge National Laboratory, Lawrence Livermore National Laboratory, Los Alamos National Laboratory, Frederick National Laboratory for Cancer Research, National Cancer Institute

Progress to date

- Demonstrated an improved DNN that adds drug target descriptions to the input and improves cell line properties set. This work extends CANDLE's ability to integrate multimodal molecular and drug feature types across multiple data sources in a deep learning framework for drug response.
- Delivered an initial implementation of the CANDLE library to streamline the writing of CANDLE-compliant codes, as well as provide time-saving functionality to improve programmer productivity.
- Delivered a capability to simultaneously extract four critical pieces of information from unstructured text documents, namely, identifying (1) the cancer primary subsite (six classes, three breast cancer subsites, and three lung cancer subsites); (2) histologic grade (three classes); (3) behavioral type (two classes); and (4) laterality (two classes).
- Demonstrated a new multitask deep learning model for document classification with a hierarchical convolutional attention network (MT-HCAN).
- Demonstrated the design of optimal hyperparameters for the MT-HCAN using the CANDLE's mlr-MBO and HyperSpace workflows on Summit (Oak Ridge Leadership Computing Facility).

CANDLE, a partnership between DOE and NCI, is developing highly efficient DNNs optimized for the unique architectures provided by next-generation exascale platforms to address three significant science challenge problems in cancer research.