Accelerating Deep Learning Adoption in Biomedicine With the CANDLE Framework

Wozniak, J.M.,[*] Brettin, T.,[†] Stevens, R.[†]

[*] Data Science and Learning, Argonne National Laboratory, Lemont, IL, USA

[†] Computing, Environment, and Life Sciences, Argonne National Laboratory, Lemont, IL, USA

The Cancer Deep Learning Environment (CANDLE) is an open framework for rapid development, prototyping, and scaling deep learning applications on high-performance computing (HPC) systems. CANDLE was initially developed to support a focused set of three pilot applications jointly developed by cancer researchers and deep learning / HPC experts, but is now generalizable to a wide range of use cases. It is designed to ease or automate several aspects of the deep learning applications development process. CANDLE runs on systems from individual laptops to OLCF Summit, the most powerful supercomputer in the world, and enables researchers to scale application workflows to the largest possible scale.

Key components of the CANDLE architecture are shown in Figure 1. A handful of top-level workflows have been developed by the CANDLE team, but these can easily be modified or extended. These include:

Hyperparameter Search: This is the process of refining the architecture of the underlying neural network in the deep learning application. Given the basic design of a neural network, there are many flexible parameters that can be tuned to produce a particular result, such as accuracy or performance. These are typically optimized using generic optimization routines. Such searches

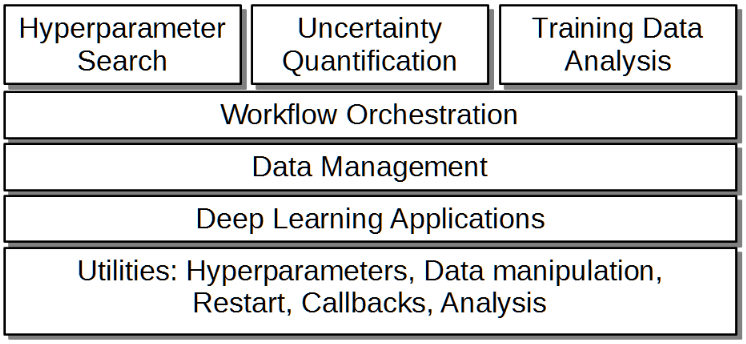
Hyperparameter Search
Uncertainty Quantification
Training Data Analysis
Workflow Orchestration
Data Management
Deep Learning Applications
Utilities: Hyperparameters, Data manipulation, Restart, Callbacks, Analysis

Fig. 1. CANDLE: Key components. Top-level workloads drive large, parallel computing runs, supported by lower-layer functionality and reusable abstractions.

produce a great many (thousands) of parallel trial training cycles, which run for minutes to hours.

Uncertainty Quantification (UQ): This is the analysis of the sensitivity of the deep learning method to small changes in the training data or other quantities in the neural network. UQ provides estimates on how robust the overall method is in the presence of data errors and other variability. UQ studies also produce a great many (thousands) trial training and inferencing runs , which are then analysed using statistical techniques.

Training Data Analysis: This is the analysis of the given training data sets for their applicability to target application problems, generalizability across problems (transfer learning), or unexpected negative impacts. Training data analysis is performed by training neural networks on various subsets of the whole training data corpus, then performing cross-comparisons of the neural networks produced, or applying transfer learning to so far unused parts of the training data corpus or other problems entirely. Since training data can be broken up and reordered in a great many ways, this methodology also produces a great many (thousands) of training cycles.

Supporting infrastructure: Under the top-level workflows, several supporting components are provided. The Workflow Orchestration component is based on the Swift/T workflow system [1], [2], a previously MPI-based dataflow programming language and runtime. The Data Management component includes capabilities for data input and output, data caching, and metadata storage in databases [3]. The Deep Learning Applications are somewhat independent of CANDLE as they are developed by focused applications teams, but they are built around common abstractions and APIs designed for integration into CANDLE workflows. These include options for handling hyperparameters formatted in a common way. As most of our applications are based on Keras [4] / Tensorflow, CANDLE provides wrappers and simplifications for managing neural network weights data, input and output tools, features to manage restartability and callbacks (a Keras feature to execute Python code dynamically during neural network training), and so on. CANDLE also includes an analysis library for evaluating UQ.

Summary: CANDLE is a comprehensive framework for managing many aspects of deep learning application development in biomedicine and other areas.

References

[1] J. M. Wozniak, T. G. Armstrong, M. Wilde, D. S. Katz, E. Lusk, and I. T. Foster, “Swift/T: Scalable data flow programming for distributed-memory task-parallel applications,” in *Proc. CCGrid*, 2013.

[2] T. G. Armstrong, J. M. Wozniak, M. Wilde, and I. T. Foster, “Compiler techniques for massively scalable implicit task parallelism,” in *Proc. SC*, 2014.

[3] J. M. Wozniak, P. E. Davis, T. Shu, J. Ozik, N. Collier, M. Parashar, I. Foster, T. Brettin, and R. Stevens, “Scaling deep learning for cancer with advanced workflow storage integration,” in *Proc. Machine Learning in High Performance Computing Environments (MLHPC) at SC*, 2018.

[4] F. Chollet *et al.*, “Keras,” https://github.com/fchollet/keras, 2015.