Contribution ID: **3** Type: **Full presentation: long (25 mins.)**

Workflow Platform for Machine Learning and Non-machine Learning Applications

Monday, 2 November 2020 13:15 (30 minutes)

ML research usually starts with a prototype on a single machine with CPU only. As a project grows, it would have to experience two major transitions: from laptop to data centre, and from data centre to clouds. Major rework is often required for each transition. Researchers often have little expertise in both core facilities and clouds. Many projects experience unnecessary growing pain or even fail to reach production-quality products.

Kubeflow has largely unified the computing in three different environments. We still need federated data access to unify the storage in different environments. This is particularly important for data-intensive ML applications, such as image classification as well as non-ML applications, such as genomic sequence analysis. On the one hand, the algorithms require large amounts of data in TB to PB range. On the other hand, many implementations assume that all data is stored on a local disk. We then use Onedata, a FUSE-based utility to present data to workflows as files stored in a local POSIX-like file system. When we shift the computing between the three different environments, we do not copy data into respective environments. We do not change implementation for different data storage, either.

We have run a workflow of image classification notebook in three environments. The source images are cardiomyocyte tissues from Image Data Repository (IDR). The results clearly demonstrate how the training and verification become significantly faster from a single machine to local data center with CPU only, and to cloud with GPU. Without making any changes to the Python script in the notebook, we are able to make use of more resources in the core facility, and GPU in Google cloud to accelerate the training and validation significantly. With much improved throughput, we were able to experiment with various image augmentation, many different image classification models, and hyper parameters quickly in the core facility and in the clouds with different GPU models.

We have also run a hand-crafted non-ML pipeline for classic variant calling on the same platform. We express the directed acyclic graph (DAG) as Python functions and function calls with DSL. The functions are backed by our custom containers created with any programming languages. The Kubeflow DSL compiler turns them into the highly detailed YAML files for Argo on Kubernetes. The time and effort to create a pipeline from scratch is greatly reduced.

Kubeflow, as a brand new ML workflow platform, is little known to the Computational Biology and Bioinformatics communities. We have successfully enhanced it with the integration with FUSE-based utilities. This gives us the flexibility to leverage more computing resources, faster network on internet backbones and the latest GPU models without changing our implementation. It allows us to use commercial cloud resources in a cost-efficient manner, where GPUs may be charged by the second, instead of having them reserved for the whole duration of a batch job. It also allows us to combine ML workflows and classic non-ML workflows on a single unified platform.

Primary authors: Dr YUAN, David (European Bioinformatics Institute); Dr WILDISH, Tony (European Bioinformatics Institute)

Presenter: Dr YUAN, David (European Bioinformatics Institute)

Session Classification: Workflow Management solutions