
October 29-31, 2024
ALCF Hands-on
HPC Workshop

SCIKIT-LEARN WITH DASK AND INTEL EXTENSION

Outline

- Current state of GPU utilization in Scikit-learn
- How to enable GPU support in Scikit-learn

  Intel Extension for Scikit-learn
- Nvidia RAPIDS

SCIKIT-LEARN WITH DASK AND INTEL EXTENSION

Scikit-learn

- ✓ Designed to provide simple and efficient tools for data mining and data analysis
- ✓ Built on top of NumPy, SciPy, and Matplotlib
- ✓ Emphasizes ease of use, performance, and interoperability with other libraries

SCIKIT-LEARN WITH DASK AND INTEL EXTENSION

Current state of GPU Utilization in Scikit-learn

Scikit-learn does not natively support running on GPUs

- **Algorithmic Constraints:** Many of the algorithms in scikit-learn are designed and optimized for CPU-based computation. Adapting these algorithms to leverage GPU would require significant changes and may not always lead to performance boost.
- **Software Dependencies:** Introducing GPU support would require additional software dependencies and hardware-specific configurations, complicating the installation and maintenance process for users and developers.
- **Design Constraints:** Scikit-learn focuses on providing a unified API for basic machine learning tasks. Adding GPU support would require a redesign of the package.

SCIKIT-LEARN WITH DASK AND INTEL EXTENSION

Recent Developments and Partial GPU Support

There have been some efforts towards enabling partial GPU support in scikit-learn

Array API Support: Scikit-learn has introduced partial GPU support via the Array API, enabling certain estimators to run on GPUs if the input data is provided as a PyTorch or CuPy array.

Intel Extension for Scikit-learn: Intel has developed an extension for scikit-learn that accelerates computations on Intel CPUs and GPUs. This extension patches scikit-learn estimators, improving performance without changing the existing API.

SCIKIT-LEARN WITH DASK AND INTEL EXTENSION

Enabling GPU Support in Scikit-learn

Using Intel Extension for Scikit-learn

✓ Provides better performance without relying on a different library, so you don't need to change your code

✓ Support for Intel's oneAPI concepts, your code can easily run on different devices like CPU and GPU

✓ Enable from command line:

```
python -m sklearnex my_application.py
```

✓ Inside script:

```
from sklearnex import patch_sklearn
patch_sklearn()
```

SCIKIT-LEARN WITH DASK AND INTEL EXTENSION

Enabling GPU Support in Scikit-learn

Using Intel® Extension for Scikit-learn

Algorithms supported on GPU include:

Clustering

- DBSCAN
- K-Means

Classification

- Random Forest Classifier
- Logistic Regression
- KNN
- SVC

Regression

- Random Forest Regressor
- Linear Regression
- KNN

Dimensionality Reduction

- PCA

See list of supported algorithms on CPU: https://intel.github.io/scikit-learn-intelex/latest/algorithms.html

SCIKIT-LEARN WITH DASK AND INTEL EXTENSION

Enabling GPU Support in Scikit-learn

Intel Developer Cloud Access

1. Redeem coupon

https://console.cloud.intel.com/docs/guides/get_started.html - cloud-credits-and-coupons

Coupon code: TLSL-ND74-KZ2S

2. Connect and Launch Jupyter Lab

3. Open Terminal in Jupyter Lab and git clone the training repo

git clone https://github.com/argonne-lcf/ALCF_Hands_on_HPC_Workshop/

4. Access the Notebook

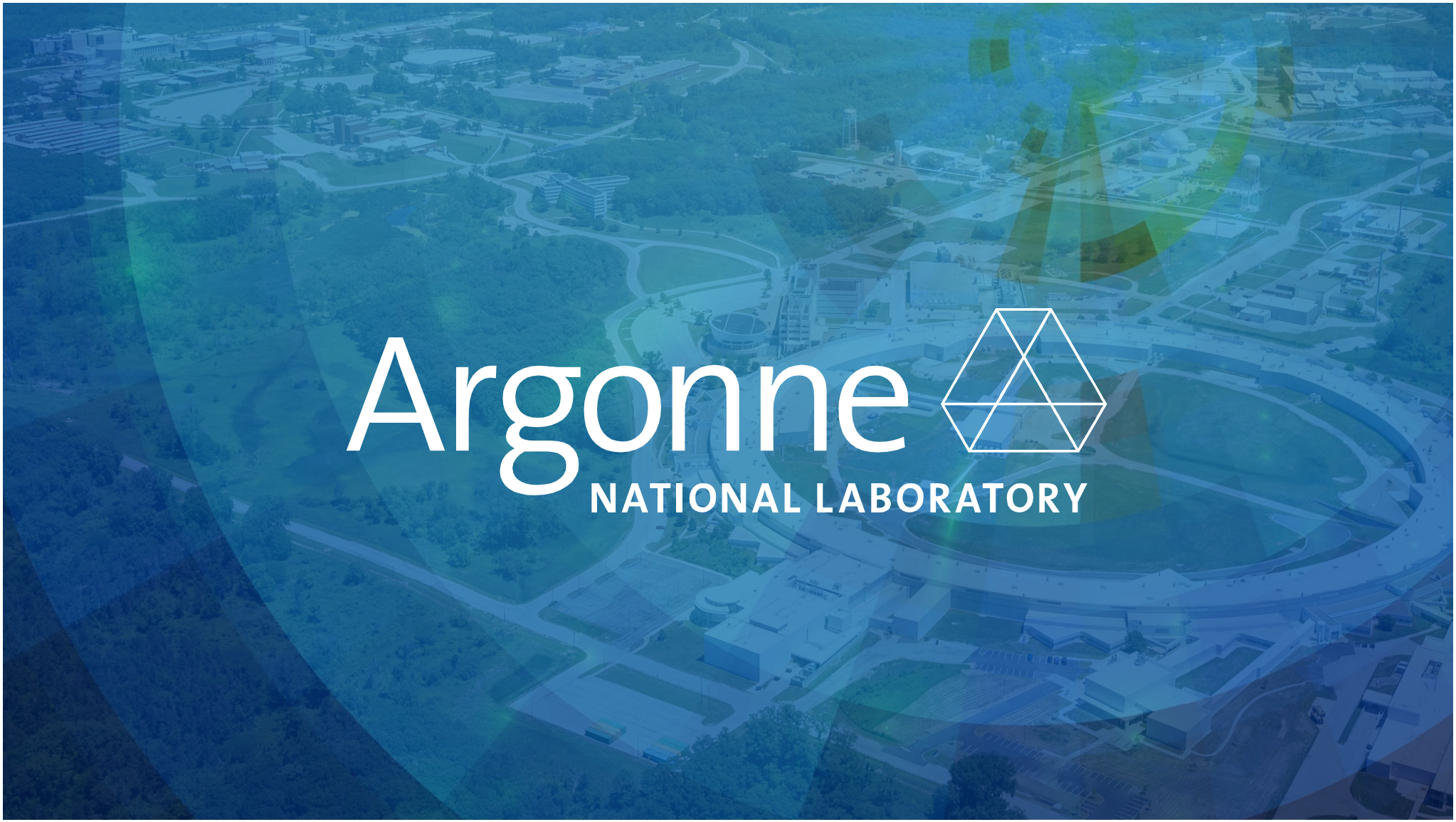
Scikit-learn/Scikit-learn_Intel_ext.ipynb

SCIKIT-LEARN WITH DASK AND INTEL EXTENSION

RAPIDS

- An open-source data analytics and machine learning acceleration platform that leverages GPUs to accelerate computations.
- Based on Python, has pandas-like and Scikit-learn-like interfaces
- Scalable with Dask integration
- Rapids APIs - cuDF, cuML , dask-cuDF


Argonne
NATIONAL LABORATORY
Argonne Leadership
Computing Facility
U.S. DEPARTMENT OF
ENERGY


Argonne
NATIONAL LABORATORY