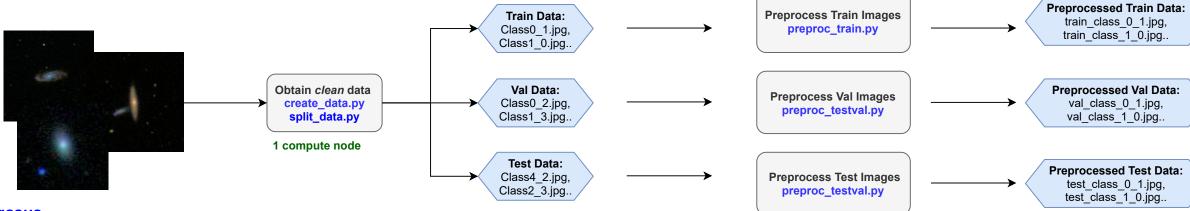| Data Acquisition	Clean Data	Data Preprocessing
Data is downloaded and unzipped in cwd	Reduce to 28790 images. 8434, 8069,578, 3903, 7806	Separate jobs for train data, and test/val data

Pegasus

Download **galaxy-zoo-the-galaxy-challenge.zip**
File size: 1.9 GB

The images are assigned to classes based on their properties. Subset of the images is used as a dataset going forward. Total 28,793 images each between 8 kB to 16 kB. Total: 386.0 MB

Pegasus

Downloads the data from an online repository (1.9 GB)

Obtain *clean* data
create_data.py
split_data.py

1 compute node

Train Data:
Class0_1.jpg,
Class1_0.jpg..

Val Data:
Class0_2.jpg,
Class1_3.jpg..

Test Data:
Class4_2.jpg,
Class2_3.jpg..

Preprocess Train Images
preproc_train.py

Preprocess Val Images
preproc_testval.py

Preprocess Test Images
preproc_testval.py

N compute nodes in parallel

Preprocessed Train Data:
train_class_0_1.jpg,
train_class_1_0.jpg..

Preprocessed Val Data:
val_class_0_1.jpg,
val_class_1_0.jpg..

Preprocessed Test Data:
test_class_0_1.jpg,
test_class_1_0.jpg..

The data preprocessing jobs can be parallelized based on number of available nodes. Total Input: 386.0 MB

Pegasus

launches multiple data preprocessing jobs in parallel.

Hyperparameter Optimization	Train Model	Test Model
Tune model to find the best setting (Optimizer, Activation)	Use the best optimization parameters to train model	7 classification metrics accuracy, precision, recall, F1, confusion matrix, ROC, AUC

Decaf

Pegasus
Launches one big HPO job inside which Decaf runs communication between nodes (?)

Decaf
Decaf transfers in-memory the test images from the train model job

Each worker needs access to the train and validation set of images.

Preprocessed Train Data:
train_class_0_1.jpg,
train_class_1_0.jpg..

Preprocessed Val Data:
val_class_0_1.jpg,
val_class_1_0.jpg..

Manager or DB
maintains all the results of the trials

WORKER 1
hpo_vgg16.py

WORKER 2
hpo_vgg16.py

WORKER 3
hpo_vgg16.py

WORKER 4
hpo_vgg16.py

WORKER 5
hpo_vgg16.py

WORKER 6
hpo_vgg16.py

N compute nodes in parallel with communication via DB or Decaf

Decaf

Decaf manages communication between the worker nodes.

best_vgg16_hpo_params.txt

Size: 108 bytes

Preprocessed Test Data:
test_class_0_1.jpg,
test_class_1_0.jpg..

Train Model
train_model_vgg16.py
(utilizes checkpointing)

1 compute node

Preprocessed Train Data:
train_class_0_1.jpg,
train_class_1_0.jpg..

Preprocessed Val Data:
val_class_0_1.jpg,
val_class_1_0.jpg..

Decaf
Decaf transfers in-memory all the train and validation images from one of the HPO jobs

vgg16_weights.pkl

Size: at least 336.1 MB depends on model and number of layers fixed

Decaf
Decaf transfers in-memory the weights of the trained model for evaluation. The weights are the output of the train model job.

Preprocessed Test Data:
test_class_0_1.jpg,
test_class_1_0.jpg..

Size: ~60 MB

Test Model
vgg16_test.py

1 compute node

Plot Figures Results

Size: around 1MB