Homework: Large Scale Applications of Machine Learning using Remote Sensing for Building Agriculture Solutions

Total points 27/28 ?

This homework includes questions from all **3** parts of the training series. You might want to record your answers on a sheet of paper or word document before submitting them here. You will not be able to save your answers and return to complete this form at a later time.

To receive a certificate of completion, you must have attended all **3** parts and have completed this homework by **April 1, 2024.** We are aware of registration technical difficulties for Session A and will account for this with certificates. Once you submit the homework, you will receive an email with a copy of your responses.

A free Databricks Community Edition account is needed to complete this homework assignment. Please follow these instructions provided on the training webpage before attempting the homework.

Once you click submit, you can click "View Score" to see how you did.

Email *

takataka.ju@gmail.com

Name (First Last) *

Jumpei Takami

Part 1: Data Preparation of Imagery for Large-Scale ML Modeling

For Part 1 homework questions, participants will need to run all of the files in Part 1 from the data folder, including the Part-1_CDL-Acquisition.py, Part-2_Sentinel-2_Acquisition.py, Part-3_Final_Step_Processing.py, and Final_CDL_S2_Data_Quality_Checks.py files. Instructions for setting up the Databricks Community Edition environment to run these files can be found in the Databricks Setup Instructions. The entirety of this homework should take approximately 30 minutes to 1 hour to complete.

✓ **CDL-related questions (Part-1_CDL-Acquisition.py)** *1/1

Which crop type is the most represented (by % area) across all years in the training dataset? (Hint: output of cmd 32):

○ Woody Wetlands

○ Peanuts

○ Corn

● Soybeans ✓

○ Dbl Crop Winter Wheat/Soybeans

○ Rice

○ Cotton

✓ Given that each pixel (or row) of the dataset is a 30m x 30m (900m^2, or 0.09 hectares) area, what is the total size in hectares of corn planted in 2019 in the training data? (Hint: output of cmd 32): *1/1

○ 1032

● 708 ✓

○ 311

○ 129

✓ For the [dense] test dataset, what are the 5 largest land covers (see cmd 16 output)? *1/1

● Cotton, Rice, Soybeans, Corn, Fallow ✓

○ Woods, Cotton, Rice, Soybeans, Corn

○ Rice, Winter Wheat, Soybeans, Fallow, Cotton

○ Developed Area, Soybeans, Cotton, Rice, Fallow

✓ What URL do we use to retrieve CDL data from an API in the code? * 1/1

○ https://www.nass.usda.gov/Research_and_Science/Cropland/docs/CDL_codes_names_colors.xlsx

○ https://CDL.gov

○ https://www.nass.usda.gov/Research_and_Science/Cropland/sarsfaqs2.php

● https://nassgeodata.gmu.edu/axis2/services/CDLService/GetCDLFile ✓

✗ **Sentinel-2 related questions (Part-2_Sentinel-2_Acquisition.py & Part-3_Final_Step_Processing.py - NOTE: these notebooks will take a long time to run all the way through. These questions do not require the entire notebooks to be run in their entirety. Investigating the code should be sufficient.):** *0/1

Some areas of the US receive more coverage than others (higher geometric revisit frequency) due to adjacent orbit path overlap. Given there are two Sentinel-2 satellites with nominal revisit frequencies of 5 days, what is the ***nominal*** maximum possible visits across both Sentinel-2 satellites for any area of the US in a **10-day period** based on figure 2 on the Sentinel 2 Revisit and Coverage page?

○ 10

○ 6

○ 4

● 2 ✗

Correct answer

● 4

Feedback

*In the US the nominal revisit time per satellite is 2 per 5 days, so in a 10 day period the nominal maximum possible visits across both satellites is 2*2=4.*

✓ What is the source of our Sentinel-2 data used in the demo? * 1/1

● AWS Open Data Registry ✓

○ Google Earth Engine

○ SentinelHub

○ Planetary Computer

✓ What library and function do we use to sample the Sentinel-2 geotiffs? * 1/1

● rasterio sample ✓

○ shapely sample

○ pyproj sample

○ geopandas sample

✓ How many classes does the Sentinel-2 scene classification layer have? * 1/1

○ 6

○ 8

○ 10

● 12 ✓

✓ **Quality Check Related Questions (Final_CDL_S2_Data_Quality_Checks.py):** *1/1

What is the primary DataBricks function we use to interactively view tables and subsequently create plots?

○ Show

○ Visualize

● Display ✓

○ Plot

✓ What composite band index do we use to view the data as a time series and ensure our data is processed correctly? *1/1

○ SWIR

● NDVI ✓

○ NDMI

○ NDWI

✓ How do we read a parquet table in as a PySpark dataframe into DataBricks? *1/1

● spark.read.parquet ✓

○ pandas.read_parquet

○ Dask.dataframe.read_parquet

Part 2: Data Loaders for Training ML Models on Irregularly-Spaced Time-Series of Imagery

For Part 2 homework questions, participants will need to run all of the files in Part 2 from the data folder, including the Part2_tensorflow_dataloader.py file, with the associated data provided in the s2_final.zip and s2_dense_test_final.zip files. After unzipping this file in the code, the data is extracted and stored as train_val_data.parquet. Instructions for setting up the Databricks Community Edition environment to run these files is shown above.

✓ How many Rice pixels are contained in the train_val_data.parquet dataset for the year 2020? Hint: df.groupby('CDL', 'year').count().display() *1/1

○ 9315

● 10882 ✓

○ 8534

○ 25179

✓ Which labels have only 1 pixel in the train_val_data.parquet dataset? Hint: df.groupby('CDL').count().orderBy('count', ascending=False).display() **(select all that apply)** *1/1

☐ Potatoes

☑ Dry Beans ✓

☑ Alfalfa ✓

☑ Cantaloupes ✓

☑ Dbl Crop WinWht/Cotton ✓

☐ Aquaculture

✓ How many bbox partitions are included in the train_val_data.parquet? Hint: you can view how the parquet file is partitioned by doing %sh ls /tmp/train_val_data.parquet/. *1/1

○ 2

○ 5

● 7 ✓

○ 50

○ 100

✓ How many pixels-timeseries or locations (lat/lon combinations) are in the train_val_data.parquet for the year 2019? Hint: df.groupby(‘year’).count().display()) *1/1

○ 80238

○ 63

○ 102942

● 80617 ✓

✓ What are the average number of images taken for the year 2020 in the train_val_data.parquet? Hint: df.groupby(‘year’).mean().display() *1/1

● 141 ✓

○ 125

○ 77

○ 365

✓ What is the shape of a single training batch if the DAYS_IN_SERIES is changed from 120 to 100? *1/1

● (1028, 21, 12) ✓

○ (1028, 28, 12)

○ (1028, 18, 12)

○ (1028, 20, 12)

✓ What is the shape of a single training batch if the DAYS_PER_BUCKET is changed from 5 to 10? *1/1

○ (1028, 50, 12)

○ (1028, 5, 12)

○ (1028, 12, 12)

● (1028, 13, 12) ✓

✓ What is the shape of a single training batch if the BATCH_SIZE is changed from 1028 to 512? *1/1

○ (512, 20, 12)

○ (1028, 20, 12)

● (512, 25, 12) ✓

○ (1028, 25, 12)

✓ How many “No Crop Growing” Labels are in the training dataset? Hint: use the np.histogram on the tf.argmax(all_labels) from the label histogram to find this. *1/1

● 28877 ✓

○ 4408

○ 14401

○ 970

✓ How many “Cultivated” labels are in the training dataset? Hint: use the np.histogram on the tf.argmax(all_labels) from the label histogram to find this *1/1

○ 28877

○ 4408

○ 14401

● 970 ✓

Part 3: Training & Testing ML Models for Irregularly-Spaced Time Series of Imagery

For Part 3 homework questions, participants will need to run all of the files in Part 3 from the data folder, including the Part3_model_training_and_evaluation.py file, with the associated data provided in the s2_final.zip and s2_dense_test_final.zip files. After unzipping this file in the code, the data is extracted and stored as train_val_data.parquet. Additionally some questions will be asked about the model’s results stored in the provided model_120days_results.parquet file. The model that generated these results is stored in the provided model_120days.keras file. Instructions for setting up the Databricks Community Edition environment to run these files is shown above.

✓ Dropout helps prevent overfitting (is a regularization parameter). Hint: see tensorflow docs. *1/1

● True ✓

○ False

✓ With a kernel_size of 5 and DAYS_PER_BUCKET set to 5 how large is the window for the Conv1D in days? *1/1

○ 10 days

● 25 days ✓

○ 20 days

○ 1 day

✓ What was the accuracy of the model on 2019-5-30? Hint: look at the code from the results time-series (cmd 48). *1/1

● 67.29% ✓

○ 69.59%

○ 82.38%

○ 75.67%

✓ Which optimizers are available in the tf.keras.optimizers module? Hint: keras docs **(select all that apply)** *1/1

☑ Adam ✓

☑ SGD ✓

☑ Adagrad ✓

☑ AdamW ✓

☑ Lion ✓

✓ How many times did the model misclassify Rice as Cotton in the model_120days_results.parquet file? (Hint: to see raw prediction counts look at the "normalize" parameter of the sklearn.metrics.confusion_matrix page). *1/1

○ 22

○ 11000

● 50 ✓

○ 160

✓ How many times did the model correctly classify the “Cultivated” class in the model_120days_results.parquet file? (Hint: to see raw prediction counts look at the “normalize” parameter of the sklearn.metrics.confusion_matrix page). *1/1

○ 0

○ 50

● 1 ✓

○ 910

✓ What is the **micro** f1 score of the model throughout the entire year. Hint: see sklearn.metrics.f1_score and the “average” parameter. *1/1

● 75% ✓

○ 81%

○ 62%

○ 89%