Questions & Answers Session A

Please type your questions in the Question Box. We will try our best to get to all your questions. If we don't, feel free to email Sean McCartney (sean.mccartney@nasa.gov), John Just (JustJohnP@JohnDeere.com) or Erik Sorensen (SorensenErik@JohnDeere.com).

Question 1: Have you thought about raster vision & Pytorch? I feel like the industry overall is shifting towards Pytorch. Or do you see otherwise?
Answer 1: Pytorch is great -- and agreed you could definitely shift to that. We had originally used tensorflow due to its support of parquet files being a bit better, but I think even Pytorch has good support for that now.

Question 2: Where can I see the past class recording?
Answer 2: All training recordings will be posted to the training webpage within 48 hours of the live training session. Scroll down to the Part you are interested in: https://appliedsciences.nasa.gov/get-involved/training/english/arset-large-scale-applications-machine-learning-using-remote-sensing
Also:
https://youtu.be/wwhb14hDhEQ?si=_ZYcOT1oCLO3nHLY

Question 3: Will this tutorial include an explanation of interleave? I've got a TensorFlow data pipeline that reads NetCDF files directly that uses flat_map, but it never felt like it was really doing anything in parallel. Maybe tf.py_function and tf.Data.interleave don't play well together?
Answer 3: Interleave and flap_map work similarly, but interleave supports parallel processing and allows some additional capabilities for optimizing how it splits the processing using the block_length and cycle_length parameters. We actually use interleave when not doing our processing on Databricks, but we had trouble getting interleave to work on Databricks Community Edition. Interleave does work well but there are some tricks to it -- i.e. you have to make sure to close the files, and be careful that you don't have too many [large] ones open at once.

Question 4: How can this approach for crop classification can be applied outside the US (since it's using the cropland data layer [CDL] layer as the training/validation dataset)?
Answer 4: The Cropland Data Layer only makes predictions for the Contiguous United States. However, if there are other crop classification models that provide crop predictions in other locations around the world, those can easily be swapped in place of the CDL in this pipeline. E.g., if you have boundaries and you have ground truth already for those areas, then you would just want to format your ground truth into a table similar to the format we use from the script in Part-1. Then from there use that table in part-2 for obtaining the data. Just be careful of the CRS (coordinate reference system) of your boundaries when querying the STAC API for sentinel-2 data. If no alternative crop classification model exists for your AOI, an alternative approach could be looking at areas in the US that are similar in both agricultural practices and climate, training a model for that area, and applying it to your area of interest. If you pre-train your model on the CDL in such a way and the crop types are mostly covered, then it is likely that the amount of ground truth you need can be significantly reduced (i.e. you may just need a little to “fine tune” your model), or eliminated entirely if performance looks acceptable.

Question 5: Does the validation depend on which 80% you select (e.g., in which order you decide to present the data to be trained on)?
Answer 5: Yes, in our case we split by year to avoid data-leakage since there may be information shared within the year (e.g. similar weather/climate conditions). In our case our data split is more like 60/40 train/validation.

From [participant] to everyone:
Yes, the order of data presentation can impact model performance, and k-fold cross-validation is a strategy to address this. By randomly dividing the data into 20% test chunks and rotating them, we ensure a more robust evaluation of the model's generalization across various data subsets, reducing the risk of bias introduced by a specific data order.

thanks [participant]. It's absolutely true that you want to be careful about how you present the data during training. I've empirically found that the more random/mixing the better (e.g. during training try to mix up the crop types, areas, and years as much as possible).

Question 6: What are the main considerations when deciding which development environment to use, i.e. Databricks vs. Colab vs. local Jupyter notebooks?
Answer 6: Databricks community has limited storage compared to your local Jupyter notebook, but if you want to use PySpark it's MUCH easier to use DataBricks. Your local Jupyter notebook also won't have a time limit and will [likely] have more computational power. Colab doesn't have any kind of persistent storage, but you can link to it publicly so as long as you don't need persistent storage and you want your code publicly available it may be fine.

Question 7: When we are using satellite imagery, do we need to perform radiometric calibration? Can you discuss the radiometric calibration for satellite data?
Answer 7: We are utilizing the Sentinel-2 Level-2A imagery which is radiometrically calibrated which applies atmospheric corrections and gives surface reflectance values.

Question 8: How do you consider the variation of weather conditions across years? For example, the data for trains is from a wet year, the validation data is from a dry year.
Answer 8: This is exactly why we use multiple years for training and a separate year for testing to make sure it can handle such seasonal variations. Ideally you'd also use multiple areas for the same reason.

Question 9: Will the)'s affect the mean? Did you calculate this before?
Answer 9: Yes, the mean and standard deviation are calculated before the normalization of the features is applied.

Question 10: Can you pad in median values instead of zeros for those missing dates?
Answer 10: Yes, just be sure the median padded values are normalized to the same scale as the rest of the normalized data.

Question 11: Why is the train/val/test split against the timeline – if we want to predict the future, isn't 2019 the train, 2020 is the val and 2021 is the test?
Answer 11: Yes this is correct -- when we train we are still loading in shorter time series from the growing season and teaching our model to predict the CDL on a limited

set of data from that year (so we're teaching it to predict CDL crop class in "real-time" so to speak...as long as there is vegetation present).

Question 12: Do the 0s in the training affect the average in the normalization, or this is after?
Answer 12: After so they don't affect normalization -- we pre-compute the mean and standard deviations prior to padding.

Question 13: Regarding one hot encoding: suppose we have 25 independent variables, each with 6 categories. Does this mean that after one-hot encoding, we would end up with a dataset with 150 columns?
Answer 13: No - one-hot encoding is just for the targets, so let's say our input was 25 bands (multiple satellites…sentinel-2 only has 12)...and our time length had 20 bins/time slots, then our input size would be (batch_size, time_length, channels) = (batch_size, 20, 25) and our output target size is still (batch_size, 6).

Question 14: How can I use this for a smallholder agricultural dataset with almost no ground truth?
Answer 14: This is a common question, when you do not have ground truth, you will still need to know some things about the crop. The best thing to do is use data as broadly sampled as possible and train the model on that.

Question 15: The scene classification may fall sometimes. For each bucket (5-day window), incorrect scene classification may introduce the outliers, is this going to impact the accuracy of the training process?
Answer 15: Yes there absolutely can be errors in the CDL. Of course that can impact the model and for this case we state in part-1 that we assume those errors will mostly transfer to our model. Of course that's true for any "ground truth" that you use -- errors in ground truth may transfer to your model if they are learnable errors. If the errors are outliers that don't fit the data patterns well then your model may effectively ignore them during training.

Question 16: How can I use this for a smaller agricultural dataset with almost no ground truth?
Answer 16: You will want to at least know the distribution of the data in the area you wish to apply this. For your case, figure out generally what kinds of crops and in what

proportions they occur for that area, and then find them from the CDL and train your model focused on those crops.

Question 17: What is the rationale to use 5-day as the bucket window? Is this based on the revisit time for Sentinel-2? Is crop phenology considered?
Answer 17: Primarily availability of the Sentinel-2 data. The crop can change very rapidly during growth and senescence phases and having data available even more frequently may be useful during those time frames if possible depending on your use case.

Question 18: Suppose we know the cropping pattern in that area. Then, how can we use CDL for this area of interest, let's say, for India?
Answer 18: See question 4.

Question 19: Do the results in values of standard deviation and mean vary depending on computer capability, or just the parallelization?
Answer 19: No it shouldn't be related to compute capability or parallelization. You can compute running means and standard deviations across a dataset too large for memory.

Question 20: I am wondering why the procedure does not use any indices (e.g. NDVI) in the training/prediction. Traditionally, those indices were frequently used to improve the classification accuracy.
Answer 20: Those are derived from the raw bands and so they are essentially perfectly correlated with our other inputs and don't add any new information. Our model can learn "NDVI" or other indices implicitly because we are using nonlinear models (neural networks).

Question 21: Can we use GPUs on databricks platform?
Answer 21: Yes but only in the paid version.

Question 22: We've found SCL has trouble with thin, wispy clouds. Do you have any other products you use to mask out clouds?
Answer 22: No -- but for thin clouds there's still "information" that is present (we can usually see some things on the ground) so the model can still learn from those at times.

Question 23: What kind of general task is expected for the homework? In particular, will it be about applying the methods presented, or will it require updating / modifying the python codes?
Answer 23: The homeworks will largely involve running the code provided and making minor changes to it to answer some questions.

Question 24: With the introduction of short corn, if that hybrid becomes popular, how could that affect the model?
Answer 24: We assume you mean the shorter growing season. It really depends on how the crop is different from the regular season corn. If it has some similarities to regular corn and is different enough from other crops, it should perform well.

Question 25: Can you explain the need of adding the zeros?
Answer 25: Adding the zeroes is a necessary step when converting the irregularly-spaced time-series to a bucketed regularly-spaced time-series of satellite imagery. E.g. if there are no images that fall in the specified bucket, we will add a 0 value to “pad” that bucket with the mean of each band in the imagery dataset.

Question 26: Can you explain some concept of multi-date of sowing and harvest?
Answer 26: If you are referring to areas with more than one crop, we do have one of those examples. We try to avoid using that in the sample data. If you have trained on those individual crops, you will be more likely to predict on the correct crop.

Question 27: Would wispy clouds introduce some bias in the mode by changing reflectance values?
Answer 27: It has a likely chance, but may not be a big issue if the model is trained to account for that.

Question 28: I did understand the reasoning for the 5-day and 25-day time frames. But can you (re-?)explain the 1028 figure?
Answer 28: The 1028 is the batch size, which specifies how many rows or time-series the model will process during each “batch”. This can be changed to optimize the training of the model.

Questions & Answers Session B

Please type your questions in the Question Box. We will try our best to get to all your questions. If we don't, feel free to email Sean McCartney (sean.mccartney@nasa.gov), John Just (JustJohnP@JohnDeere.com) or Erik Sorensen (SorensenErik@JohnDeere.com).

Question 1: How do we determine the best data set for training and testing the machine learning model? (The partition of the data into training and testing is usually done randomly using a random state).
Answer 1: You don't want to do this randomly in a naive sense. You should select your dataset (training & testing) such that it matches your goals. So if your goal is to focus on an area with predominantly maize and soy, you wouldn't want to train/test on an area with the predominant distribution of land cover being something other than those. So first you should select your dataset so that the distribution of targets matches your goals. Then you should make sure to get enough data across seasons to account for variation in things like temperature and precipitation that can be quite different between years, in order to make your model more robust.

Question 2: On slide 22, regarding the Overfitting chart, he mentioned that at a certain point, you can early-stop the model in order to avoid overfitting. How could the model be stopped when it is already running - would this not crash the process? Do you need to include the stopping condition in the parameters of the model so it stops by itself?
Answer 2: Tensorflow has callbacks that add additional functionality during the training process. One of these callbacks adds early stopping to the training process, which will track if the validation loss does not improve after a specified number of training steps and will stop training, saving the best performing model.

Question 3: In the last slide, how is the value of NDVI between 0 to 100? The value of NDVI lies between -1 to 1. I think I missed something, so I just want to clear this up.

Answer 3: In the code we scaled the NDVI values to the range of -100 to 100 by multiplying the values by 100. Usually NDVI is clipped below zero, and in this case if it's showing 0-100 that would be the same intention as 0-1.

Question 4: I don't have access to the DBF database on databricks. Could you check the access?
Answer 4: Look in the Catalog → then you can see the option to toggle to “DBFS” (it defaults to “Database Tables”. If it is not there, follow these instructions to enable it: https://docs.databricks.com/en/administration-guide/workspace-settings/dbfs-browser.html.

Question 5: Does this data work if we want to focus on flood damage assessment on crop monitoring?
Answer 5: Yes, but your labels would need to change to focus on water instead of crop type. The first script of part-1 of this series, where we get CDL data, would be where you set up your areas that have flood damage and target labels for that. Then the rest of the data acquisition process and modeling all apply.

Question 6: Do we need an AWS account to run these notebooks in Databricks?
Answer 6: No. Databricks Community uses Azure (paid DataBricks subscriptions can use AWS or Azure which will require an AWS or Azure account).

Question 7: I want to focus on Flood Monitoring and Crop Damage Assessment Using Remote Sensing (Sentinel-2 MSI- Sentinel-1 SAR Data, and High-Resolution Imagery with Planet Smallsat Constellations). Are crop types important?
Answer 7: Yes, it is important. Please see the answer to the question above.

Question 8: Out of curiosity: why do you use the term 'bucket' instead of the more commonly used 'bin' or temporal window composite?
Answer 8: No particular reason we use the word bucket vs bin, the meaning here is the same.

Question 9: Can we use the same code like in the example with our sample data?
Answer 9: All of this code should work with your own custom data, as long as you match the format that we put the targets in for the CDL. I.e., run the first script from

part-1 to acquire CDL data and look at that format, and do something similar for your targets, then the rest of this should similarly work (including acquiring associated Sentinel-2 data).

Question 10: Did you try to normalize with all data versus only training data?
Answer 10: Calculate the normalization factors (mean, std) from the training data, and apply those same ones to the validation/test data.

Question 11: Will CDL monthly data be released soon? Is it feasible?
Answer 11: I'm not aware of CDL being released monthly (or if there is a plan for that). They've said it's costly to compute across the entire US so they've limited the resolution and rate that they calculate it. It's also a sensitive thing for markets so I think they want to avoid estimating it and then having markets rely on it.

Question 12: Are the same means and standard deviations used for both the training and test set, or is it recomputed given a new raster?
Answer 12: Calculate the normalization factors (mean, std) from the training data and apply those same ones to the validation/test data. They should not be recomputed given a new raster.

Question 13: What sensor do you recommend using (radar or optical) to monitor/predict land use and vegetation in the central zone of Mexico? And what other sources of information do you recommend combining?
Answer 13: Start with optical data (one source, easy to understand) and get your pipeline and modeling working for the area you're interested in (Mexico central zone) first, then you can make it more complicated by adding more inputs/sources. E.g. move on to using Harmonized Landsat/Sentinel dataset afterwards for higher temporal resolution (although you may lose some spatial resolution moving from 10m to 30m pixel size).

Question 14: Is Planet data with 3 meter resolution better for crop type, or sentinel 1?
Answer 14: It depends on your use case and limitations. Cost considerations are also important to consider.

Question 15: About the approach on training one year, testing another, and validating another year: this has the complication of drought years which can complicate the classification. Is there an approach that considers training, validation and test with mixed years considering same region sampling?
Answer 15: It is important to have some domain knowledge. If you can use 3-5 years in your training dataset, you are generally going to be covering a variety of season differences and your training data will be more robust.

Question 16: Does it make sense to create an environment for future machine learning coding? Or do you just install the libraries onto your main system?
Answer 16: In Databricks we just put all the pip installs in the notebook and you have to reinstall each time. Those don't persist because the compute/cluster terminates and deletes all storage upon shutdown for community.

Question 17: Have you made comparisons between Dask and PySpark for your use case? Why did you decide to go with PySpark?
Answer 17: Both Dask and PySpark utilize distributed computing to parallelize computation optimizing efficiency. We haven’t done performance testing comparing the two, but we utilize PySpark as it takes no additional setup in Databricks and scales extremely well for this use-case.