Large Scale Applications of Machine Learning using Remote Sensing for Building Agriculture Solutions
March 5, 12, 19, 2024

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: Would Keras and the data loader still be required if this model were implemented with PyTorch instead of Tensorflow?
Answer 1: The data loader is separate from Keras and there isn't a reason you'd need to use the same package to implement the data loader AND the model. Just make sure your data loader functionality is similar to what we showed in part-2, but you can certainly use PyTorch for data loading and model training.

Question 2: Is anything similar to Tensorboard available for PyTorch?
Answer 2: Not that I'm aware of - Tensorboard sort of stands alone in that area.
https://pytorch.org/docs/stable/tensorboard.html

Question 3: Will it work the same with later versions like long-term support (LTS) 14.x?
Answer 3: Maybe… you'd have to try it. We just make sure to clearly state this to avoid any reasons why things might now run properly for folks (changing Databricks runtimes can sometimes be a problem).

Question 4: I know this was covered before, but is there a way to not lose the cluster created once it is "terminated" for not being used after 60 minutes?
Answer 4: Not with Databricks community - it terminates the instance entirely. The best thing to do is save everything to persistent storage (DBFS) for recovery when restarting the cluster.

Question 5: The homework in this course requires more coding than most courses. Is there an office hour for help? Or is there a way to ask for help if we get stuck on some code?

Answer 5: If you have any issues just fire us off an email. We tried to mostly make the homework based on the standard output of running the given scripts (not having to modify things much at all… just understand what's roughly going on). Both Erik and my emails are on slide 28 in the PDF presentation for today's slides: https://appliedsciences.nasa.gov/sites/default/files/2024-03/Part_3_FINAL.pdf.

Please note: due to the large number of participants, we are not always able to help troubleshoot individual issues.

Question 6: I missed the Part 2 live session. Is it still possible to get the certificate?
Answer 6: If you missed one live session but attend the rest and complete the homework by the due date, you will be eligible to receive a certificate.

Comment 7: (From Xavier) Here is some feedback on my experience if useful to some of you: I had issues downloading some of the datasets from Google Drive, especially DataSet1, as I got error messages. Found the way around with Google Drive top right menu option "Download all" instead of separately. It created just one file which then downloaded correctly. Same for the Python files.

Question 8: What criteria would make you say the model is "overfitting"?
Answer 8: Primarily by monitoring a separate validation dataset during training. We talked a bit about that in Part 2 as well (and showed an example training and validation curve).

Question 9: On some plots you showed, the loss function had a narrow minimum. Is it possible to go back to the specific set of model parameters which achieved this minimum, and maybe assess whether this was an actual narrow minimum vs. some kind of computation uncertainty?
Answer 9: Yes, we have the code setup to save the best model so you can return to it after training. It's quite trivial to do it (just a flag in the training function).

Question 10: Since 12 Sentinel-2 bands are used in the model, I am wondering if it is possible to analyze which bands are not significant for the prediction.
Answer 10: This analysis could be done by doing a “leave-one-out” validation by training the model multiple times removing different bands each time. Then each model

performance could be compared to see which bands have the largest impact on model performance.

Question 11: Are there ways to create accurate early stage predictions? For example, using Sentinel-1 or other sources?
Answer 11: Other data sources could certainly be added to try to improve prediction performance using the same data pipeline we provided here.

Question 12: Could you improve the model including phenology analysis?
Answer 12: Do you mean by including things like the plant growth stage? I would say probably yes, but that's hard information to obtain so we assume here we don't have it.

Question 13: Is the model trained per image or looks on the time series? Are the spatial errors happening only on single image training or did I misunderstand?
Answer 13: Time series - the predictions are shown on the most recent image in the time series.

Question 14: Does the model handle the season of the year in order to get adapted to the life cycle of the different cultivated plants?
Answer 14: Yes, implicitly by the fact that we are using the time series, it should learn the season/crop stage information and use it for prediction.

Question 15: Just out of curiosity, have you tried the traditional machine learning approach (e.g., random forest)? How does deep learning improve the prediction accuracy?
Answer 15: We didn't, but you could easily substitute that for the model here and try it! That's what the cropland data layer uses for prediction.

Question 16: How many hidden layers are in the network structure?
Answer 16: We used 2 1dCNN layers with 1 dense layer with 50 hidden units. It's not very complicated at all. You all can tune it to probably get better results (we didn't do much tuning at all).

Question 17: What extra inputs could we use to make our crop classification more accurate? Also, what other advanced machine learning models could we try to get better results?

Answer 17: Perhaps more images from other sources such as Sentinel-1 and Landsat would be the easiest next step. If you know more things that are relevant (like growth stage) you could use them, but that information is hard to obtain.

Question 18: What would be the minimum coefficient or value in a confusion matrix to accept a classification of a time series with multispectral satellite images?
Answer 18: This depends on what your criteria are for success.

Question 19: In case of cloudy days, what is the date interval you use to get the better training pixel values from satellite images?
Answer 19: We use 5 days, but you could use whatever frequency you think you'll have decent resolution at. The cropland data layer only has an image every 2 weeks.

Question 20: My question is related to generalizing this pipeline for locations outside the US. My consideration would be to use CDL data from locations in the US that have similar growing conditions as, say, a country like Nigeria. It would be interesting to see how well this pipeline generalizes.
Answer 20: We have made suggestions in Part 1 and 2. Use the CDL first and train your model on it if you know the crops for your area.

Question 21: Can I run all the code locally? Or is there a specific dependency on Databricks? I ran part and for the parquet file locally.
Answer 21: Yes, with some changes to the code you can run this locally. In Part 1, there are more dependencies to use Databricks.

Question 22: Do you think recurrent neural networks (RNNs) would perform better than 1D convolutional neural networks (CNNs) for this task?
Answer 22: No I don't - I've never seen RNNs get very good performance relative to CNNs.

Question 23: For temporal sequence handling, people prefer recurrent neural networks (RNN), long short-term memory (LSTM), and attention-based models, right? Can you elaborate more on this?
Answer 23: Please refer to question 23.

Question 24: Do you believe that treating this task as pixel-based rather than object-based causes us to miss out on some spatial information?

Answer 24: Yes. Adding spatial information would be useful.

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: What should we do when our target for training and testing does not exist? Is there any algorithm that can do the task? Also, if we have a series of data for a latitude longitude, how can we use it as target data knowing that the inputs data are GeoTIFF data?
Answer 1: If you have an idea of what types of crops are in your area, you can still sample the CDL and train your model.

Question 2: How do you create the "/FileStore/" directory in Databricks?
Answer 2: If the FileStore button on the top left of the Workspaces tab isn't there, you can enable the FileStore in Databricks Community by following these instructions: https://docs.databricks.com/en/administration-guide/workspace-settings/dbfs-browser.html.

Question 3: Can you please provide more information about the CDL?
Answer 3: Sure. We provided some links in the first part of this series (in the PDF)... but the best source is the FAQs:
https://www.nass.usda.gov/Research_and_Science/Cropland/sarsfaqs2.php.

Question 4: Building such a model requires Remote Sensing Context, Agronomy, DL Theory, Tensorflow Usage, and MLops. Most of the time thematic scientists are needed to give hardware/software specifications to DevOps/MLops to initiate model building. Some pointers on how to give this spec would be appreciated.
Answer 4: We hesitate to give much advice here. If you do Parts 1 and 2 of this training, the model will be somewhat forgiving.

Question 5: Could you comment on how a trained model may perform when there is varying presence of clouds in images that are used for prediction? How best should we handle the presence of clouds in images?

Answer 5: We ignore heavy cloud cover (very opaque ones). Since we use a time series we generally believe the model can do a good job of interpolating where there is missing data due to clouds, or where light cirrus clouds have somewhat warped the colors.

Question 6: Is the training stratified so that it is equally exposed to all types?
Answer 6: The train/val/test split was done by year for the same region to avoid data leakage, which resulted in a similar label distribution for each set. More details of this process can be found in Part 2 of this training.

Question 7: How can they monitor crops temporarily? Do you train models with spectral signatures?
Answer 7: We use a time series of the imagery, where the prediction date is considered the latest image available (and the previous images are part of the "context" of the input, similar to LLMs or any other time series data).

Question 8: Can I use this machine learning module for agricultural drought monitoring and risk analysis of an area?
Answer 8: Yes we would expect that you can use the same process for drought monitoring. Just change out the labels.

Question 9: How does the model know it is accurately predicting no crop growing? Is that included in the CDL training data? I thought there was only one label associated with the CDL data.
Answer 9: The “No Crop Growing” label is our custom label that we defined in Part 2 that defines when the CDL says there is a crop growing in that location, but the last two images in the time-series don’t show any vegetation as detected by the SCL. Our measure of accuracy then is measured by comparing accuracy on our test set using our custom labeler function.

Question 10: If we have labels in vector format (.shp), can we customize the code? And if so, how?
Answer 10: If you look at the first script in Part 1 (acquiring CDL data), format your labels in the same final format we have there. Then you can run the rest of the scripts without change.

Question 11: Do you consider that the choice of an optimizer different from Adam's can modify the prediction results of the case study?
Answer 11: Yes, different optimizers can have an impact on the prediction results. The choice of optimizer is something to tune, but Adam is a popular one.

Question 12: To fine-tune a model, is there a rule of thumb on how much labeled data is needed against the model used?
Answer 12: Unfortunately not....you kind of have to gather a little and try it out (try to avoid training too many epochs on your fine-tuning dataset, and see how it performs by reserving a testing set).

Question 13: What extra inputs could we use to make our crop classification more accurate? Also, what other advanced machine learning models could we try to get better results?
Answer 13: Perhaps more images from other sources such as Sentinel-1 and Landsat would be the easiest next step. If you know more things that are relevant (like growth stage) you could use them, but that information is hard to obtain.

Question 14: On some plots you showed, the loss function had a narrow minimum. Is it possible to go back to the specific set of model parameters which achieved this minimum, and maybe assess whether this was an actual narrow minimum vs. some kind of computation uncertainty?
Answer 14: Yes, we have the code setup to save the best model so you can return to it after training. It's quite trivial to do it (just a flag in the training function).

Question 15: Since 12 Sentinel-2 bands are used in the model, I am wondering if it is possible to analyze which bands are not significant for the prediction.
Answer 15: This analysis could be done by doing a “leave-one-out” validation by training the model multiple times removing different bands each time. Then each model performance could be compared to see which bands have the largest impact on model performance.

Question 16: How accurate will our prediction be for a country like Ethiopia, where one farmland has mixed crops? Is there a different method we could use?
Answer 16: If you have mixed crops, we are predicting on a pixel level. Your limitation is your resolution.

Question 17: Can machine learning algorithms be trained directly on satellite imagery data to generate maps without relying on SIG (Système d'information géographique, or GIS)?
Answer 17: Yes.

Question 18: What does the .zip data that we imported into the Databricks database refer to? They are not directly created in the code, in case we want to adjust the code with our labels?
Answer 18: It's the data that would be created if you ran the entire data acquisition script to completion. We ran it all and gave it to you since it would take a long time in Databricks community.

Question 19: Could you share your thoughts on using a Hyperspectral sensor for this analysis? I understand that the current status is less frequent, so even building the time series data would be challenging though.
Answer 19: Depends on the source of your data. This may be outside the scope of this training.