Deforestation Hackathon - Schneider Electric

Omar Aburayyan

Architecture and project setup

- An AlexNet CNN was trained on the dataset from scratch using Pytorch
  - AlexNet is one of the most significant cnn architectures and it's good to use it as benchmark model
  - I used Pytorch because of its very flexible when it comes to debugging and tweaking models to your liking
  - Since the data is imbalanced, I used the out of the box pytorch DataSampler to handle data imbalance

- I used DVC (Data Version Control) to manage the ML pipeline
  - dvc is a great tool for managing machine learning pipelines;
  - It allows us to split the pipeline into different stages, easily track, run and do experiments

What would I have added if I had more time?

- Pay more attention to the imbalance in the dataset
  - Instead of using `ImbalancedDatasetSampler`, I would have explored different image augmentation techniques and applied them within the pytorch dataloader. This way we would know exactly how the data is over/undersampled
  - Grassland data is only available after 2012
    - There might be some deterministic behaviour that could be predicted using other attributes than the image (latitude, longitude, year)
    - One idea is to extend the problem solution to a multimodal classification model where the image model is combined together with another model based on other features than the image
- Error analysis to understand which pictures are being misclassified from the validation set
- Test other architectures and compare their performance with AlexNet
- Optimise for different hyper parameters such as the batch size and learning rate

Thank You for the great hackathon 🌲🌳🌴