Efficient machine learning with scalable aerial data pipelines

Learn how leading aerial imagery intelligence companies transform 1500 terabytes of data into vital insights with scalable plug-and-play data pipelines.

Overview

Data scientists spend more than half of their time cleaning up the mix of structured and unstructured data and preparing to input it into the machine learning / AI models rather than actual big data analysis. A few quick steps have been shown to help cut down both machine learning costs and time spend on the project.

Needs of a typical ML company utilizing aerial/drone imagery

- Continuously collect & pre-process data
- Need to develop, test, scale, and deploy our best-performing models to the cloud
- Train performant models rapidly due to external factors (e.g. time-to-insight limited by seasons in agricultural tech)

Technical aspects of the implementation

- Imagery is high-resolution (e.g. 10cm/pixel)
- Data is multi-spectral and multi-sensor (e.g. RGB and infrared imagery; thermal, topography, soil composition, weather, and management data)
- Data is temporal (e.g. need to understand how field health evolves and to capture how management decisions impact yield)
Problem

- Building a solid data foundation to confidently obtain high-quality models
- Inefficiencies in data management and pipelining result in unnecessary costs due to the sheer size of the data
- Need to experiment with different data architectures on the fly and iterate on experiments
- Identify the most appropriate data for training our models and optimizing whether to train locally or in the cloud
- Balancing the cost of compute versus network (computing while one is downloading data so that no time or resources are wasted)

Solution

- Utilizing a simple API for creating, storing and collaborating on AI datasets of any size
- Created datasets represent the aerial imagery datasets in an AI-native format (see the illustration)
- Streaming-friendly format (e.g. Hub by Activeloop) allows training models at scale, without having to download the entire dataset
- Rapidly transforming & experimenting on the fly

Results

- Model accuracy increases via active learning, on average +12% from the baseline model
- Inference speed typically increases 3x from weeks to 1-2 hours
- On average, compute costs drop by at least 50%
- -30% decrease in infrastructure cost due to optimization yielded by the dataset format and efficient pipelines