AgroScout Cultivates New Levels of Efficiently with ClearML Experiment Management Tools

With massive collections of complex imagery to analyze, tag, and feed into models, AgroScout has embedded ClearML into its infrastructure at critical stages in model development workflows

Forsight: Client Overview

AgroScout has created an automated, AI-driven scouting platform for the early detection of pests and disease in vast agricultural areas. With precise, data-driven insight (literally impossible to attain through periodic, manual sampling), a farmer can proactively prevent substantial loss of crops and corresponding revenues as he drives increased productivity across his acreage, All the while reducing his use of pesticides because he can more precisely target problematic areas.

The Challenge

AgroScout uses drone and camera footage to gather crop data to identify the subtle indicators of pest and disease. If this were all simply the same Big Data challenge common to companies focused on AI-based imagery analysis (medicine, construction, traffic, etc.), they would be able to handle it by simply building models and throwing plenty of computing resources at them. But working with imagery analysis for agriculture poses a few unique challenges:

The team needed to focus on their technology, not infrastructure managing it in the background.

- Annotation isn't trivial; even experts can't find all plant defects
- Analysis requires engaging a team of specialized, global experts to annotate; managing their work and standardizing their terminology needs to be precise
- High data modality. Each field – and even each field's dataset – has unique and multi-input features to be learned and merged (e.g., changing lighting or flight angles)
- Scale - Just a single image can yield thousands of annotations, and annotation QA is difficult; there are often tens of thousands of these images.

In short, AgroScout's technology involved some technical heavy lifting, and their workflows needed to be optimized to take manual work off their hands. Naturally, they wanted a scalable, self-contained solution that would take little effort to integrate and maintain. The team needed to focus on their technology, not infrastructure managing it in the background.

The Solution

AgroScout's research led them to ClearML as a way to track research, control their cloud resources, and manage data throughout their development cycles.

ClearML helped in six ways:

1. **Processing hi-res images**
   As processing massive high-resolution images isn't feasible, AgroScout's approach is to cut them into sub-images before they are fed into the model for training. Naturally, this dramatically increases the number of images to store and track for analysis. ClearML's dataset management feature helped them to create versions for each data set, including one version that had been preprocessed and was ready for training. Facilitating this workflow was already practically sufficient to justify their integration of ClearML, but it was just the beginning.

2. **Managing divergent annotations from multiple agronomists**
   Every ML model begins with human input to teach the system what to look for in its initial analysis of datasets it is given to work on. When it comes to agriculture, skilled annotators are hard to find; qualified individuals and companies were recruited from around the globe. AgroScout quickly discovered that no matter how directly and clearly they requested standardized annotations, they were receiving data with various label names and terminologies, even from multiple employees within the same vendor. Using dataset statistics in ClearML's UI, they identified mismatches, created unification strategies to map synonymous terms, and finally created alias rules for labels with the same meaning.

Just a single image can yield thousands of annotations. Even for a human, it is hard to detect all the diseases and pests, and to decide how exactly to classify what's been detected.

3. **Annotating Incredibly complex, multi-faceted datasets**
   AgroScout cannot succeed with a single model to cover all agricultural scenarios. Their clients are located around the globe, and naturally, every field has its own unique crop type, weather, age, soil, and other variables. These factors make the analysis of each particular location different than the analysis of another. All these vectors create a management problem when trying to keep metadata readily available without crawling through multiples folders for every analysis. ClearML's Hyperdatasets allowed AgroScout to organize their data versions and associate them to model training tasks for post-training analysis.

   The software itself must handle immense quantities of metadata … ClearML was designed for precisely this type of scalability challenge.

4. **Annotation Mass**
   A single image can sometimes require literally thousands of annotations to record details like ground lift (bumps that indicate that a plant is about to emerge), first emergence, adult plant, and gaps in the field. Not only is this an intense process for the annotators, but the software itself must handle immense quantities of metadata, with full editing features along the way. This type of data management can present a challenge to many data management tools, but ClearML was designed for precisely this type of scalability challenge.

5. **Managing workloads on GPU instances**
   Training and testing the type of models used for agricultural analysis – even when optimized – require vast GPU resources. AgroScout built data ingestion pipelines triggered by AWS Lambda functions. Integrating with ClearML Orchestrate, which manages workloads on EC2 instances for training use; executed tasks are then monitored and managed using ClearML's Orchestrate UI. This was a classic example of a heavy, time-consuming, manual task that would have diverted them from core development.

6. **Choosing the winning models**

   This particular type of model comparison is exceptionally fluid and dynamic because of the "fuzzy" biological data representing an organic landscape's continuously changing and unpredictable nature. It's rarely a case of "this model succeeded, and this one failed." Conducting the comparisons – deciding which experiments to continue running and which to end – is at the core of ClearML's suite of tools, and proved powerful enough to conduct active, ongoing comparisons to streamline the process of refining the models.

The Results

While overall productivity and accelerated time to market are hard to measure precisely, AgroScout's team feels the improvement across their development lifecycle:

- They increase their data volume over 100x without growing the data team

- Increased experiment volume 50x with the same team size

- Shortened the time to production (time from experimentation start to model in production) by over 50%.

ClearML helps Data Scientists, ML Engineers, and DevOps improve their ML workflows with solutions to manage their experiments and data. Their open-source MLOps platform automates and simplifies the development, management, and processes for thousands of global data science teams and brands including NVIDIA, NetApp, Samsung, Hyundai, Bosch, Microsoft, Intel, IBM, Philips, and more.

Contact us to learn how we can help you: info@clear.ml