OCTO
an IBM Company

EDGE MACHINE LEARNING OPERATIONS:

The Key to Operationalizing AI & ML

> *Machine learning methods are usually based on the assumption that the data generation mechanism does not change over time. Yet real-world applications of machine learning, including image recognition, natural language processing, speech recognition, robot control, and bioinformatics, often violate this common assumption. Dealing with non-stationarity is one of modern machine learning's greatest challenges."*
>
> ***MACHINE LEARNING IN NON-STATIONARY ENVIRONMENTS, MIT PRESS.***

ML Models Decay in Production Environments

Machine Learning (ML) models are inherently brittle and they all decay at various rates and for different reasons when deployed. Accordingly, all ML models require operations and maintenance (O&M) that takes infrastructure, people, and tools to keep production-grade ML models performing as well as possible in operational environments. Machine Learning Operations (MLOps) is the capability that underpins successful AI. Most MLOps platforms available today focus on AutoML and are well-suited for model development and training, packaging, test/validation, and deployment. However, few MLOps platforms are designed to work with dynamic DOD and IC sensor data, which also requires the ability to monitor model performance in austere environments and trigger semi-automated or automated retraining when model decay is detected.

One example of problematic dynamic sensor data is computer vision (CV) for object detection or aided target recognition. With computer vision models, as with all ML models, new patterns continuously emerge in the real world—or battlefield—over time. The model lacks approximate function mappings for these newly emerged patterns, because they were not present in its training data, and the net result is degraded predictive ability. **Figure 1** illustrates the deterioration in a CV model designed for object detection. The blue line that starts horizontally represents model accuracy. Initially accuracy is consistently high. However, as the environment in which the model operates experiences change, the model's accuracy decreases, also known as model "drift" or "decay." In this case, the drifts are demarcated by the figure's 17 green vertical lines.

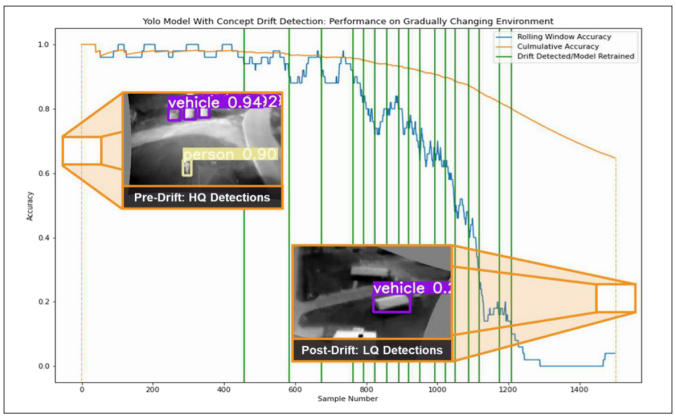

Figure 1: Computer Vision Model Concept Drift

Model drift over time is inevitable, underscoring the importance of MLOps processes and tools designed to support constantly evolving operational environments confronted by the DOD and IC.

They tell the all too common story of gradual initial decay, followed by a precipitous plunge in the model's predictive competence as the real-time environment increasingly differs from the model's training data.

Notably, these 17 drifts also represent 17 opportune moments to take remedial action, such as leveraging MLOps' continuous integration/continuous deployment (CI/CD) power by triggering a rebuild of the ML pipeline and refitting the model. If the DOD were to simply purchase ML models without the ability to retrain them on the fly, the models would degrade over time and the operational user would be put at risk. Additionally, the cost of continually purchasing new models developed for very similar operational environments is wasteful and cost prohibitive. Therefore, Octo created the Hatteras MLOps platform to automate model retraining, minimize model cost, and maximize operational capability.

Octo Maintains Production Grade ML Model Performance

Hatteras

To maximize the economic and operational value of ML, the DOD and IC must think of ML models not as widgets trained and deployed in a "set-it-and-forget-it" way, but as a dynamic system that must be continuously maintained.

To combat model decay in operational environments, Octo's oLabs team developed the Hatteras MLOps platform. Hatteras is a low-cost but essential MLOps solution that enables AI at scale, even in austere environments. Hatteras' containerized architecture enables it to be installed on low Size Weight and Power (SWaP) devices, such as the DOD's Tactical Cloud Package, ground stations in a box, and or robust enterprise cloud resources. It empowers end users by preventing model decay and automatically retraining ML models through a user interface that requires no DevOps experience or advanced ML expertise. Hatteras' processes and benefits are illustrated in **Figure 2.**

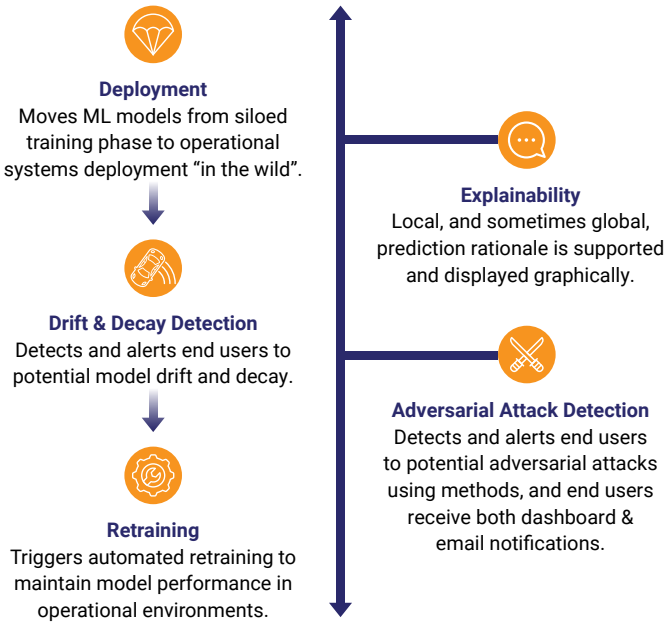

Figure 2: Hatteras Benefits

Hatteras offers several benefits to users over traditional MLOps tools, which are specifically relevant to DOD and IC end-users.

> ***If the DOD were to simply purchase ML models without the ability to retrain them on the fly, the models would degrade over time and the operational user would be put at risk."***

The Hatteras solution also:

- Embraces open-source frameworks and integrates open-source tools with permissive licenses. This both prevents vendor lock-in and provides a continuity path when personnel turnover.
- Empowers users to deploy & remediate models through an intuitive user interface; no advanced training is necessary.
- Enables users to iteratively develop ML models, deploy custom and existing models, and monitor/continuously retrain models—all through one tool.
- Is designed to maintain operational AI and ML model performance.

Our Hatteras MLOps platform delivers production-grade AI and ML through an open-source tool stack and helps automate the data labeling process (we call it Robotic Real-Time Battlefield Labeling). These are just a few of Hatteras' intuitive, open-source features for training, deploying, monitoring, and actively retraining ML models.

Octo's Investment in Self-Funded R&D

OLABS

Octo recently made a multimillion-dollar investment to solidify and enhance its commitment to providing innovative approaches to artificial intelligence (AI) and machine learning (ML) operationalization; our 14,000 sq. ft. emerging technology research and development (R&D) facility, oLabs™, offers our federal mission partners capabilities not previously available in the National Capital Region.

oLabs provides 20 petaflops of AI compute, several petabytes of flash storage, tactical communications devices, end-user devices, and a close quarter battle (CQB) facility to enable innovative yet operationally relevant data pipeline prototyping and experimentation with emerging data science and machine learning operations (MLOps) tools and data-centric technologies. Leveraging oLabs' infrastructure, standards, and the oLabs AI operationalization process, our AI—kmw 12 and ML engineers, data scientists, architects, and software developers partner with federal clients and domain and industry experts to ensure AI and ML models and their associated data pipelines are properly prepared, tested, and ready for operational deployment.