Introduction to MLOps and LLM Pipelines

November 3rd, 2023
About our Speakers

Eduardo Alvarez – Senior AI Solutions Engineer, Intel
Eduardo Alvarez is a Senior AI Solutions Engineer at Intel, specializing in architecting AI/ML solutions, MLOps, and deep learning. With a background in the energy tech startup space, he managed a team focused on delivering SaaS applications for subsurface AI in hydrocarbon and renewable energy exploration and production. Now at Intel, Eduardo collaborates across technical teams, designing impactful solutions highlighting the Intel software and hardware stack’s influence on Deep Learning and GenAI workloads. He is the author of Intel’s MLOPs Professional Developer course, where he brings his expertise in the production deployments of AI tools to a broad audience of student and enterprise developers.
Agenda

- Introducing Intel® Certified Developer Program – MLOps Professional Certification
- Role of MLOps in Production AI/ML Solutions
- Considerations for Performance and Optimization
- Future of Operational AI, Sustainability, and Ethical AI
- How to get certified
Introducing Intel® Certified Developer – MLOps Professional

✓ Develop a marketable skillset focused on incorporating compute awareness into the AI solution design process

✓ Increase your hireability through the creation of a project that showcases competency in designing and implementing performant AI solutions

MLOps Professional Training Package
• 26 Video Lessons
• 8 Hands on Labs using the Intel® Developer Cloud
• 1 capstone project
• Office Hours
• Practice Certification Exam
Why Focus on Operational AI?

Operational AI Engineering Evolution Model = Data Model Operations

Traditional ML Era
The early 2010s

Heavy experimentation with forests, trees, linear regression, etc.

Neural Net Mania
2015 - 2019

Stronger focus on data to train data-hungry neural networks (FFNN, LSTMs, RNNs, CNNs, etc.) with significant focus on model development and optimization.

GenAI Revolution
2020 - Current

Open-source foundational models reduce focus on pre-training and architecture design. Data becomes a major focus in training LLMs and VT and for fine-tuning and RAG applications.

Democratized AI
2025 -

New high-quality data sources are scarce. Transformers technology has peaked, and the upper bound becomes operational optimizations driven by HW/SW.
This is how we teach Operational AI

With careful consideration for compute and software optimizations across the ML Lifecycle and the foundational knowledge required to architect and implement solutions in production.
Understanding MLOps

How It Enhances Application Development and Adds Value
What is MLOps?

MLOps, short for Machine Learning Operations, is a practice that focuses on the integration of machine learning models into operational processes to ensure the reliability, scalability, and maintainability of AI systems. It combines DevOps, software engineering, and machine learning principles to streamline and automate the end-to-end machine learning lifecycle.
**MLOps Architecture**

- **Application Development**
  - Data Analysis
  - DS Code Development
  - CI/CD Stage: Build, test, package, and deploy pipelines

- **Core MLOps**
  - Feature Store
  - Data Engineering
  - Model Engineering
  - Automated Pipelines
  - Pipeline Trigger
  - ML Metadata Store
  - Performance Monitoring
  - Model Registry
  - Staging Environment (CD stage)
  - Inference Service
Initial Approach to Model Management

- Novices often save models as simple files without tracking their evolution.
- Basic tracking might include using spreadsheets for model names, parameters, and metrics.
- As AI initiatives expand, model management becomes more complex with frequent iterations and larger datasets.
- MLOps presents advanced practices to transition from basic to scalable and efficient AI system management.

<table>
<thead>
<tr>
<th>Name</th>
<th>Accuracy</th>
<th>Data Version</th>
<th>Path</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model A</td>
<td>86%</td>
<td>1.0</td>
<td>C://home/my_models/modelA.h5</td>
</tr>
<tr>
<td>Model B</td>
<td>67%</td>
<td>1.4</td>
<td>C://home/my_models/modelB.h5</td>
</tr>
<tr>
<td>Model C</td>
<td>78%</td>
<td>3.0</td>
<td>C://home/my_models/modelC.h5</td>
</tr>
</tbody>
</table>
Model Metadata

- Model Training
- Model Validation

Code Registry
- Data Versioning
- Container Registry
- Code Object Store

Model Object Storage

ML Metadata Store

Model A
- Hyperparameters
- Training and eval data versions
- Performance metrics
- Infrastructure, optimizations, package versions.

Model B
- Hyperparameters
- Training and eval data versions
- Performance metrics
- Infrastructure, optimizations, package versions.
Storing Models

- Model A: Model A Artifacts, Checkpoints, etc.
- Model B: Model B Artifacts, Checkpoints, etc.

ML Model Development:
- Data Versioning
- Container Registry
- Code Registry
- Code Object Store
- Model Object Storage
- ML Metadata Store
- Model Training
- Model Validation
ML Prediction Service
The two primary inference modes are batch prediction and real-time/online prediction. Batch prediction processes multiple inputs together, which is beneficial for large datasets or cost-saving. Real-time prediction focuses on low-latency responses to individual data points.
Fancy Orchards: Real-time Inference

- Fancy Orchards" uses real-time inference for production line QA/QC processes.
- Camera sensors continuously capture apple images, which the AI system instantly analyzes for defects.
- Swift real-time decisions enable rapid detection and removal of defective apples.
"Fancy Orchards" uses batch inference to enhance apple cultivation based on historical weather patterns. The model processes chunks of new weather data for water and fertilizer recommendations. This batch processing optimizes resources by processing data in larger, more efficient sets.

**Fancy Orchards: Batch Inference**

- **Batch Data Inference Request**: 72hrs worth of weather and soil sensor data
- **Batch Processing Endpoint**: Batch Data Inference Request -> Batch Processing
- **Irrigation and Fertilization Optimization Tool**: Post-Proc, Inference, Pre-Proc
- **Optimized schedule of optimized irrigation and fertilization**
Real-time Monitoring for AI in Production

Monitoring AI workloads provides continuous insights into performance, resource utilization, and system health. Utilizing monitoring tools, metrics such as GPU utilization, memory consumption, and latency can be tracked, with alerts set up for a proactive response.

User Reviews a Product

Database

Recommendation Engine API

Movie Recommendations Updated in User Interface

Real-time Monitoring Dashboard

Warnings

Runtime data is logged, processed, and displayed on the telemetry dashboard.
Hands-on #1 - Setup

1. Navigate to cloud.intel.com
2. In the Training and Workshops Section, select “Launch Jupyter Lab”
4. Navigate 01_model_development_basics in the workshops folder and open model_development_basics.ipynb
5. Start workshop!
Common Bottlenecks in AI Workloads and Opportunities for Optimization

Exploring Optimizations for AI Workloads
Performance Bottlenecks

- 92% Thread Utilization
- 85% Thread Utilization
- 96% Thread Utilization
- 99% Thread Utilization
- 82% Thread Utilization
- 99% Thread Utilization

97% Memory Utilization

Inputs/Outputs are Bottlenecked
Software-Hardware Co-Design

Advanced Matrix Extensions, available in Intel 4th Generation Xeon processors, have dedicated instructions to improve memory management and matrix operations for deep learning workloads.

Upstreamed optimizations from Intel into PyTorch, enabling Advanced Matrix Extensions optimizations during model training with auto mixed precision.
<table>
<thead>
<tr>
<th>Engineer Data</th>
<th>Create Machine Learning &amp; Deep Learning Models</th>
<th>Deploy</th>
</tr>
</thead>
</table>
| Container Repository  
oneContainer | oneAPI powered AI Reference Kits | MLOps  
Cnvr.g.io | Developer Sandbox  
Intel® Developer Cloud | Annotation/Training/Optimization  
Intel® GETI |
| Connect AI to Big Data | BigDL (previously "Analytics Zoo") | |
| Data Analytics Scale | Optimized Frameworks and Middleware | Optimize Models |
| MODIN  
SciPy | TensorFlow  
PyTorch  
mxnet  
PaddlePaddle | Automate Model Tuning  
AutoML |
| pandas  
NumPy | LightGBM  
XGBoost  
CatBoost | Automate Low-Precision Optimization  
SigOpt |
| w/ Intel Optimizations | | Intel Neural Compressor |
| SYCLomatic | oneDAL | oneDNN  
oneCCL  
oneMKL | SynapseAI™ |

Note: not all components are necessarily compatible with all other components in other layers.
Example: Intel® Optimization for PyTorch

**ECOSYSTEM**
- torchvision
- TorchServe
- Hugging Face
- PyTorch Lightning
- ...

**FRAMEWORKS**
- PyTorch
- Intel® Extension for PyTorch*

**LIBRARIES**
- oneDNN
- oneCCL

*Other names and brands may be claimed as the property of others*
Future of Operational AI
Considerations for LLMs
Understanding Hallucinations in Generative AI

User

"Can I trust this tool?"

LLM Hallucination

"The moon is made of green cheese"

"Is the moon made of green cheese?"
Role of RLHF in Addressing Hallucinations

- Human Reviewer
- Trained LLM
- Sample Output
- Reward
- State
- Action
- Supervised Learner
- Reinforcement Learning Reward Model
- Reinforcement Learning Agent
- Weight Update
- Reward Prediction
Prompt: Create a new brownie recipe.

Brownie making typically involves mixing a batter of ingredients like chocolate, sugar, butter, and flour, and then baking it in an oven until it has a fudgy or cake-like texture. The key to a great brownie often lies in the quality of chocolate used and the baking time, which can vary depending on whether you prefer gooey or cakey brownies.

Retrieved Context

Prompt Template

This is the {user input}

Here is some information from our database about this topic {context}

Add context to the prompt template

Retrieved Context

Pass templated prompt to inference service

Add user input to prompt template

Pass prompt to vector database

Vector Database

Embedding Model

Vector Embedding
Hands-on #2 - Setup

1. Navigate to cloud.intel.com
2. In the Training and Workshops Section, select “Launch Jupyter Lab”
4. Navigate `02_llm_pipelines` in the workshops folder and open `llm_pipelines.ipynb`
5. Start workshop!
Impact of MLOps on Sustainability & Ethical AI

Role of MLOps in Responsible AI
How does MLOps Impact Sustainability of a Solution?

Using compute-aware principles, AI/ML developers can help in the right-sizing of computational resources to reduce the idle time of servers. Balance hardware and software from edge to cloud—utilizing a heterogenous infrastructure with a combination of AI computing chipsets that meet specific application needs can ensure that compute is optimized, and energy used efficiently.

Ensuring models are performant in production so that users get the right answer the first time, reducing the computational load of multiple requests to servers.

Optimized models with built-in Intel® AI Engines like Intel® Advanced Matrix Extensions (Intel® AMX) or model compression techniques improve energy efficiency of workloads.
How does MLOps Impact Ethical AI?

Clear benefits

- Standardizes model development for ethical compliance.
- Enables traceability through version control.
- Automated checks for bias and ethical behavior.

Profound Impacts

- Enables deep, scientific inquiry into AI ethics.
- Facilitates ethical ‘peer review’ through collaboration.
- Allows real-time ethical evaluation metrics.
- Promotes a culture of continuous ethical reflection.
- Makes ethics an integral part of AI, not just a checklist.
Scan to Register for the Course