Triton Inference Server in Azure ML

Agenda

- Quick Overview of Inferences and Challenges
- Overview of NVidia Triton Inference Server
- REALTIME Usecase - Microsoft Translate
- **Demo: Azure ML with Triton - Densnet**

Ayyanar J

ABOUT ME

I started my career as humble Hardware and Networking engineer in 2005 in HCL Infosystem.

Over the last 17 years of my IT Career, I have worked as Wintel, Linux Middleware Engineer, Infrastructure Architect, Cloud Solution Architect, Bigdata Manager, DevOps Lead, and DataPlatform Lead Engineer. I worked in Europe, Canada, and the US for a brief period of time.

I have 50+ Technical Certification in AWS, Azure, IBM Cloud, CKA/CKAD, TOGAF Level 2 Certified

I always love to learn - Unlearn - Relearn, Motivated to Share knowledge with peers, community and learn from them.

Follow me on https://www.linkedin.com/in/jayyanar/

https://cloudnloud.com **info@cloudnloud.com**

CloudnLoud **Community**

MACHINE LEARNING LIFECYCLE

Training

Inference

Problem Statement -

How we address via ML/AI

Value out of ML

1

Gather Required data

Explore the data

Set the Baseline Accuracy ~

2

No

Identify the Model to be used

Train the Model using training data

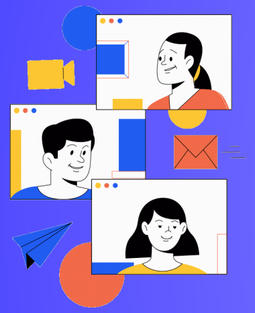
Evaluate the Model with test data

3

Accuracy / Prediction Achieved as Expected

Yes

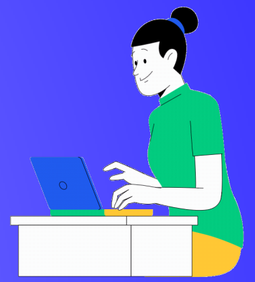
4

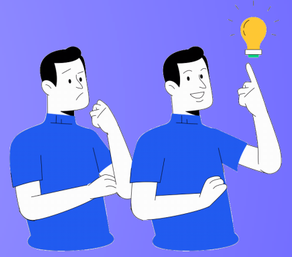


Deploy the Model as Batch Processing.

Deploy the Model as Realtime Inference

5





For Realtime Inference -

Monitor the Accuracy / Throughput - Collect

The Realtime data collected

6

CORE MLOPS CAPABILITIES

Model Serving (Inference)

Model Registry

Online Experimentation

Machine learning Metadata

and

Artifacts Repository

MACHINE LEARNING INFERENCE

Inference - Using Trained models to predict outcomes from new observations in efficient deployment

- Online inference in near real time for high-frequency singleton requests (or mini batches of requests), using interfaces like REST or gRPC.
- Streaming inference in near real time, such as through an event-processing pipeline.
- Offline batch inference for bulk data scoring, usually integrated with extract, transform, load (ETL) processes.
- Embedded inference as part of embedded systems or edge devices.

WHY WE NEED TO TALK ABOUT INFERENCE NOW

- In the next 5 years, large models like GPT-3/4, BERT, and T5, BioBert, BioGPT capable of producing impressive results on a variety of tasks in Natural Language Processing (NLP), including text classification, language translation, and question answering will bring value to business and a lot of innovation.

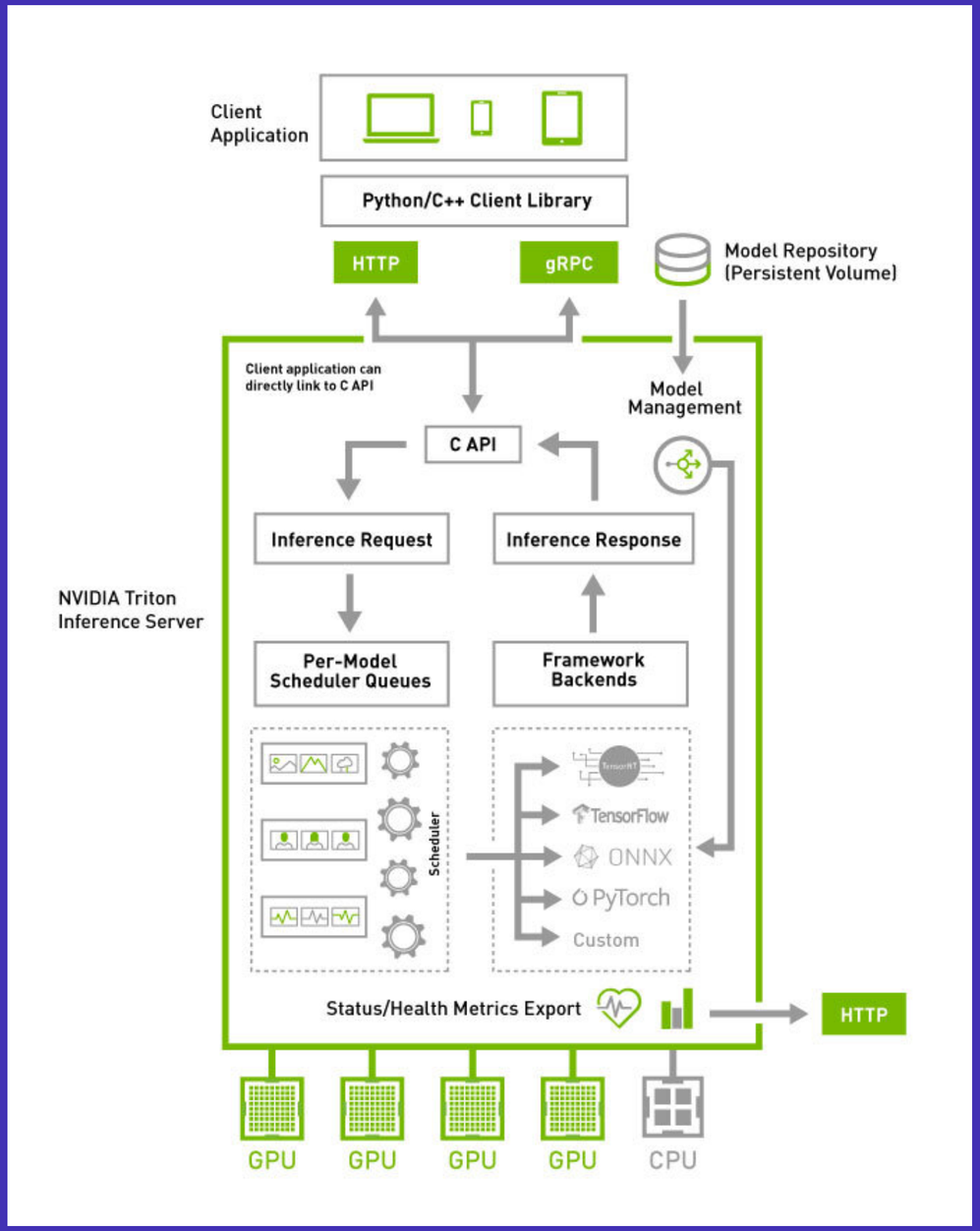
- Also Large computer image models are deep neural networks trained to perform tasks related to image processing, such as image classification, object detection, segmentation, and generation. Some examples include ResNet, Inception, and GANs. These models typically have a large number of parameters, trained on massive datasets of images, and can produce state-of-the-art results on a wide range of image-related tasks.

- Now we have a huge array of foundation models. Business or Organization need to host these large model inferences to scale cost effectively using their own data with FineTune options.

INFERENCES CHALLENGES FOR MLOPS ENGINEER

- My Journey - Local, VM+LB, Azure Container Instance, Azure Kubernetees Service
- Multiple Model Frameworks and Serving Options - Pytorch, Tensorflow, XGBoost, ONNX, TensorRT.
- Deployment Options -Single Model, Multi Model, Multi Containers, Serial Inference, OnPrem, Selecting Cloud vendor.
- Migrating Legacy Models
- Performance - Latency, Throughput, Monitoring the Metrics.
- Diverse Compute Option - CPU, GPU, TPU

ARCHITECTURE OF NVIDIA TRITON SERVER



https://docs.nvidia.com/deeplearning/triton-inference-server/user-guide/docs/_images/arch.jpg

NVIDIA TRITON SERVER - BENEFITS

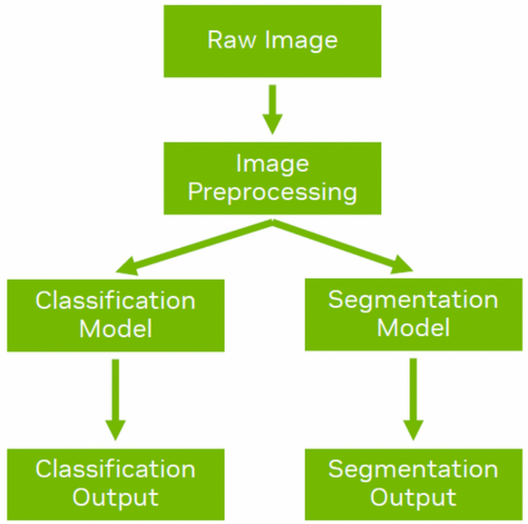
- **Support for Multiple frameworks: Triton can be used to deploy models from all major frameworks. Triton supports TensorFlow GraphDef, TensorFlow SavedModel, ONNX, PyTorch TorchScript, TensorRT, RAPIDS FIL for tree based models, and OpenVINO model formats.**

- **MultiModel pipelines: Triton model ensemble represents a pipeline of one or more models or pre/post processing logic and the connection of input and output tensors between them. A single inference request to an ensemble will trigger the execution of the entire pipeline.**

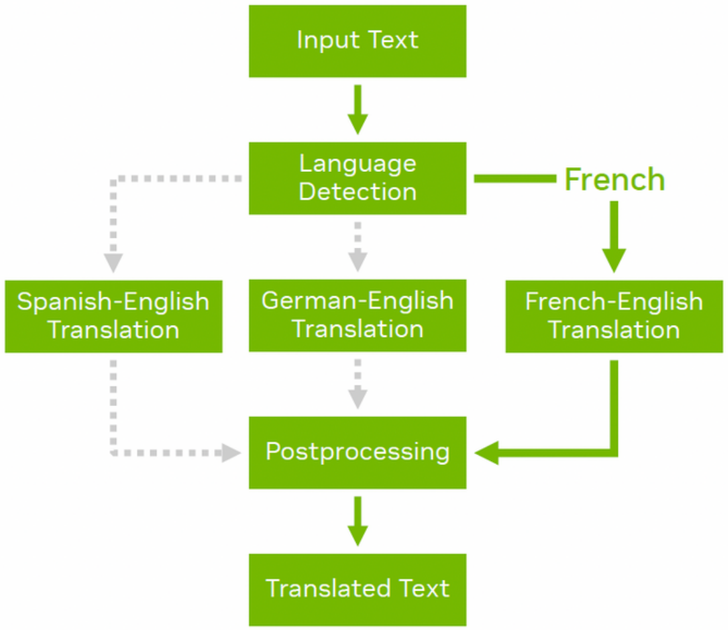
- **Concurrent model execution: Multiple models can run simultaneously on the same GPU or on multiple GPUs for different model management needs.**

- **Dynamic batching: For models that support batching, Triton has multiple built-in scheduling and batching algorithms that combine individual inference requests together to improve inference throughput. These scheduling and batching decisions are transparent to the client requesting inference.**

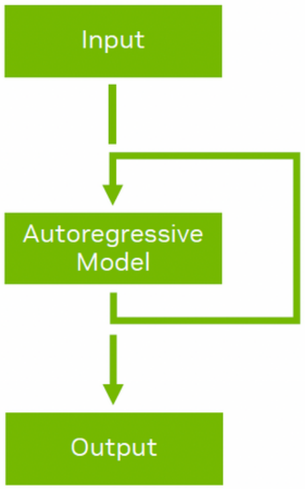
NVIDIA TRITON SERVER - MODEL PIPELINES



Model Ensemble



Conditional Execution



Looping execution

✓**Models from any framework**

✓**GPU shared memory for optimal performance**

✓**Run on GPU or CPU**

https://docs.nvidia.com/deeplearning/triton-inference-server/user-guide/docs/_images/arch.jpg

REALTIME USECASE - MICROSOFT TRANSLATE

- **Microsoft is using NVIDIA GPUs and Triton Inference Server to deploy and scale the Z-code models efficiently for high-performance inference.**

- **Microsoft has developed a new family of AI models called Z-code to improve the efficiency and quality of its Azure AI services, including Translator.Z-code uses transfer learning to improve quality for machine translation and language understanding tasks by taking advantage of shared linguistic elements across multiple languages.**

- **Z-code employs a sparse “Mixture of Experts” approach that is more efficient to run because it only needs to engage a portion of the model to complete a task, allowing massive scale in the number of model parameters while keeping the amount of compute constant.**

- **The Z-code Mixture of Experts models consistently improve translation quality and can even extend capabilities to underrepresented languages that have less available training data.The new Z-code-based translation model is now available, by invitation initially, to customers using document translation in Translator, and it has shown improvement in translation quality over current production models according to industry metrics.**

https://news.microsoft.com/source/features/ai/new-z-code-mixture-of-experts-models-improve-quality-efficiency-in-translator-and-azure-ai/

DEMO

https://github.com/triton-inference-server/server

NVIDIA GPU Cloud (NGC) Container Registry

https://github.com/NVIDIA/DeepLearningExamples

Reference Article

Azure Machine Learning

Single Model - Deployment - https://learn.microsoft.com/en-us/azure/machine-learning/how-to-deploy-with-triton?view=azureml-api-2&tabs=azure-studio%2Cendpoint

MultiModel - Deployment - https://github.com/jayyanar/azure-triton-mme/blob/main/triton-mme.ipynb