Containerized or Serverless Learning

Comparing the Prediction Latency of Machine Learning Model Deployments on Containerized and Serverless Environments

Diya Ramasamy | Iris Ma | Krithika Balasubramanyam
Raj Shreyas Penukonda | Yuyu Lai

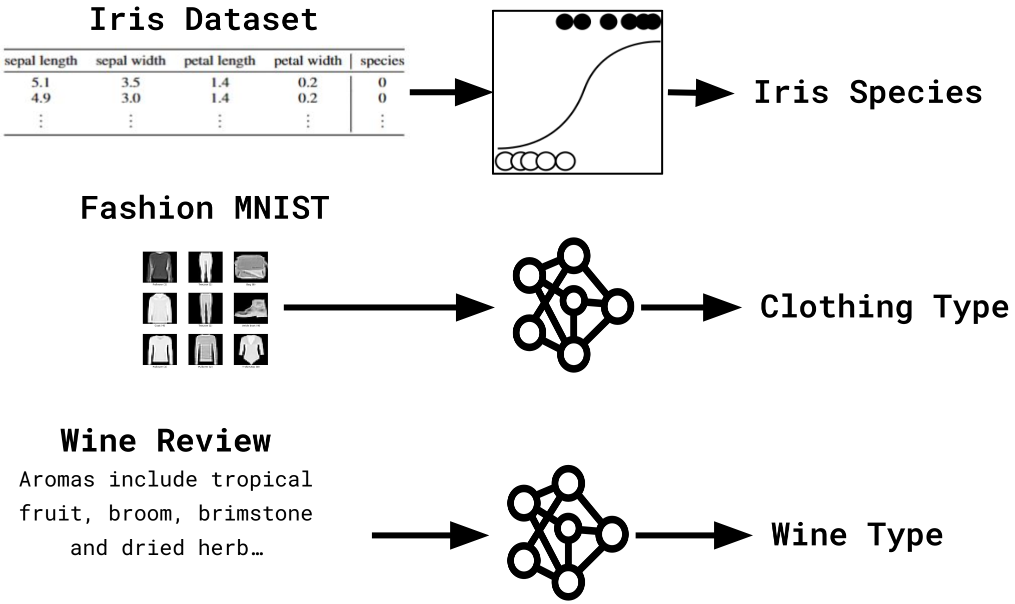
System Purpose and Functionality

Deployed three different machine learning models on containerized and serverless environments as APIs

Logistic Regression (Scikit-Learn)
- ➔ Input: Tabular Data
- ➔ Output: Iris Species

Computer Vision Model (TensorFlow)
- ➔ Input: Grayscale Image
- ➔ Output: Clothing Type

NLP Model (TensorFlow)
- ➔ Input: Wine Review
- ➔ Output: Type of Wine



Iris Dataset

sepal length	sepal width	petal length	petal width	species
5.1	3.5	1.4	0.2	0
4.9	3.0	1.4	0.2	0
⋮	⋮	⋮	⋮	⋮

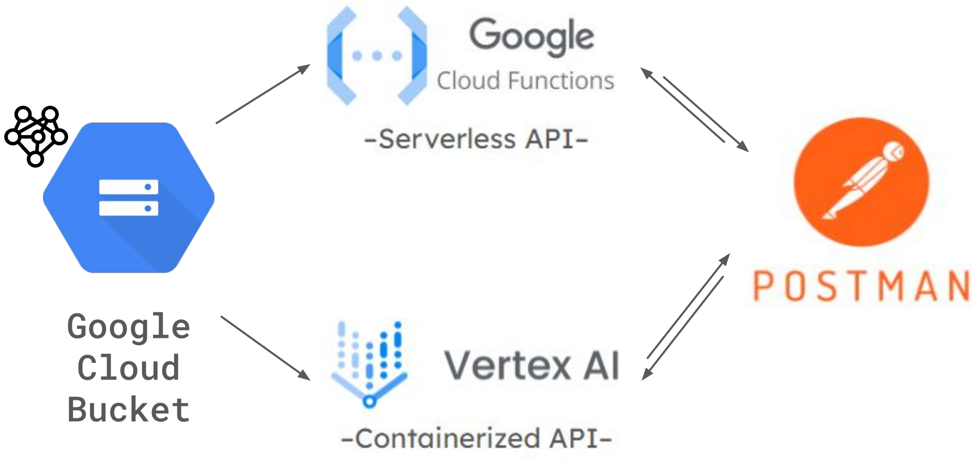
Fashion MNIST

Wine Review

Aromas include tropical fruit, broom, brimstone and dried herb…

Design

Pipeline

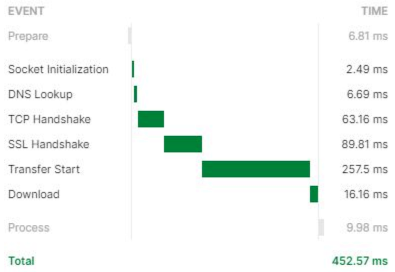


Google Cloud Bucket

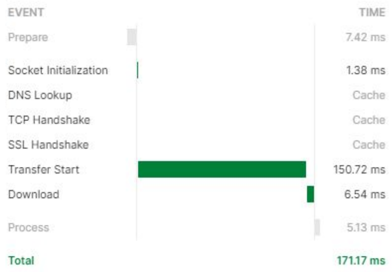
-Serverless API-

-Containerized API-

Metrics



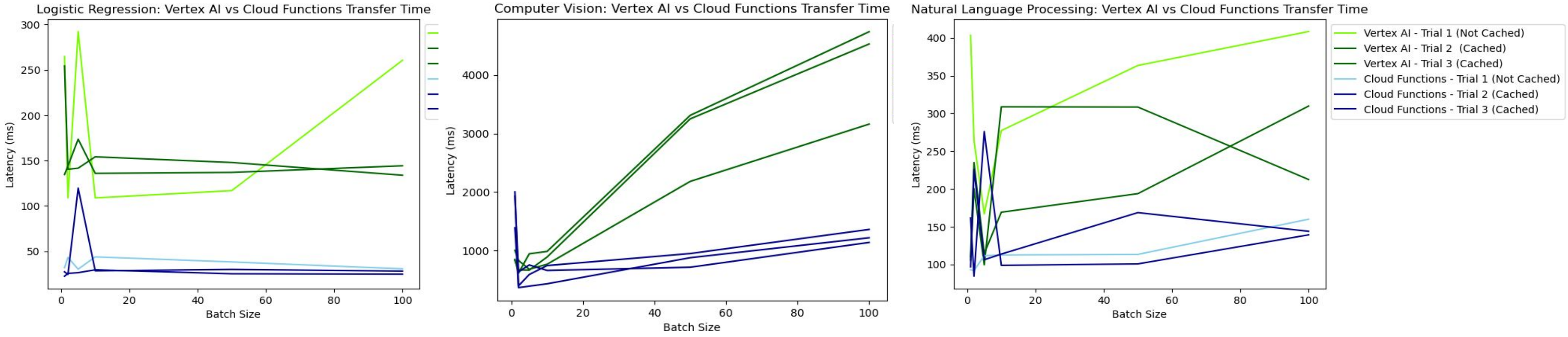
EVENT	TIME
Prepare	6.81 ms
Socket Initialization	2.49 ms
DNS Lookup	6.69 ms
TCP Handshake	63.16 ms
SSL Handshake	89.81 ms
Transfer Start	257.5 ms
Download	16.16 ms
Process	9.98 ms
Total	452.57 ms



EVENT	TIME
Prepare	7.42 ms
Socket Initialization	1.38 ms
DNS Lookup	Cache
TCP Handshake	Cache
SSL Handshake	Cache
Transfer Start	150.72 ms
Download	6.54 ms
Process	5.13 ms
Total	171.17 ms

Batch Sizes	Transfer Time (#1)	Transfer Time (#2)	Transfer Time (#3)
1 2 5 10 50 100	…	…	…

Results and Challenges



Logistic Regression: Vertex AI vs Cloud Functions Transfer Time

Computer Vision: Vertex AI vs Cloud Functions Transfer Time

Natural Language Processing: Vertex AI vs Cloud Functions Transfer Time

GCF outperforms Vertex AI; dynamic resource allocation allows for consistently low latency

Challenge	Resolution
Out of memory issues for large models	Trained another model from scratch with Tensorflow
Framework version conflicts	Trained models with the exact version available on GCP
Authentication issues when calling endpoints	Regenerated tokens every hour