Vendor: Google

Exam Code:PROFESSIONAL-MACHINE-LEARNING-ENGINEER

Exam Name: Professional Machine Learning Engineer

Version: Demo

QUESTION 1

You work for a gaming company that has millions of customers around the world. All games offer a chat feature that allows players to communicate with each other in real time. Messages can be typed in more than 20 languages and are translated in real time using the Cloud Translation API. You have been asked to build an ML system to moderate the chat in real time while assuring that the performance is uniform across the various languages and without changing the serving infrastructure.

You trained your first model using an in-house word2vec model for embedding the chat messages translated by the Cloud Translation API. However, the model has significant differences in performance across the different languages. How should you improve it?

A. Add a regularization term such as the Min-Diff algorithm to the loss function.

B. Train a classifier using the chat messages in their original language.

C. Replace the in-house word2vec with GPT-3 or T5.

D. Remove moderation for languages for which the false positive rate is too high.

Correct Answer: B

Since the performance of the model varies significantly across different languages, it suggests that the translation process might have introduced some noise in the chat messages, making it difficult for the model to generalize across languages. One way to address this issue is to train a classifier using the chat messages in their original language.

QUESTION 2

You recently built the first version of an image segmentation model for a self-driving car. After deploying the model, you observe a decrease in the area under the curve (AUC) metric. When analyzing the video recordings, you also discover that the model fails in highly congested traffic but works as expected when there is less traffic. What is the most likely reason for this result?

A. The model is overfitting in areas with less traffic and underfitting in areas with more traffic.

B. AUC is not the correct metric to evaluate this classification model.

C. Too much data representing congested areas was used for model training.

D. Gradients become small and vanish while backpropagating from the output to input nodes.

Correct Answer: A

The most likely reason for this result is the model is overfitting in areas with less traffic and underfitting in areas with more traffic. Probably because the model was trained on a dataset that did not have enough examples of congested traffic. As a result, the model is not able to generalise well. When the model is validated on congested traffic, it makes mistakes because it has not seen this type of data before.

QUESTION 3

You are building a real-time prediction engine that streams files which may contain Personally Identifiable Information

(PII) to Google Cloud. You want to use the Cloud Data Loss Prevention (DLP) API to scan the files. How should you ensure that the PII is not accessible by unauthorized individuals?

A. Stream all files to Google Cloud, and then write the data to BigQuery. Periodically conduct a bulk scan of the table using the DLP API.

B. Stream all files to Google Cloud, and write batches of the data to BigQuery. While the data is being written to BigQuery, conduct a bulk scan of the data using the DLP API.

C. Create two buckets of data: Sensitive and Non-sensitive. Write all data to the Non-sensitive bucket. Periodically conduct a bulk scan of that bucket using the DLP API, and move the sensitive data to the Sensitive bucket.

D. Create three buckets of data: Quarantine, Sensitive, and Non-sensitive. Write all data to the Quarantine bucket. Periodically conduct a bulk scan of that bucket using the DLP API, and move the data to either the Sensitive or Non-Sensitive bucket.

Correct Answer: D

https://cloud.google.com/architecture/automating-classification-of-data-uploaded-to-cloud-storage#building_the_quarantine_and_classification_pipeline

---

QUESTION 4

You deployed an ML model into production a year ago. Every month, you collect all raw requests that were sent to your model prediction service during the previous month. You send a subset of these requests to a human labeling service to evaluate your model\\'s performance. After a year, you notice that your model\\'s performance sometimes degrades significantly after a month, while other times it takes several months to notice any decrease in performance. The labeling service is costly, but you also need to avoid large performance degradations. You want to determine how often you should retrain your model to maintain a high level of performance while minimizing cost. What should you do?

A. Train an anomaly detection model on the training dataset, and run all incoming requests through this model. If an anomaly is detected, send the most recent serving data to the labeling service.

B. Identify temporal patterns in your model\\'s performance over the previous year. Based on these patterns, create a schedule for sending serving data to the labeling service for the next year.

C. Compare the cost of the labeling service with the lost revenue due to model performance degradation over the past year. If the lost revenue is greater than the cost of the labeling service, increase the frequency of model retraining; otherwise, decrease the model retraining frequency.

D. Run training-serving skew detection batch jobs every few days to compare the aggregate statistics of the features in the training dataset with recent serving data. If skew is detected, send the most recent serving data to the labeling service.

Correct Answer: D

https://cloud.google.com/blog/topics/developers-practitioners/monitor-models-training-serving-skew-vertex-aiew-vertex-aiandved=2ahUKEwiRg_aoj9n8AhWb7TgGHcGCDREQFnoECAwQAQandusg=AOvVaw197NneIJM0ra7fLq2zsOin

---

QUESTION 5

You work with a data engineering team that has developed a pipeline to clean your dataset and save it in a Cloud Storage bucket. You have created an ML model and want to use the data to refresh your model as soon as new data is available. As part of your CI/CD workflow, you want to automatically run a Kubeflow Pipelines training job on Google Kubernetes Engine (GKE). How should you architect this workflow?

A. Configure your pipeline with Dataflow, which saves the files in Cloud Storage. After the file is saved, start the training job on a GKE cluster.

B. Use App Engine to create a lightweight python client that continuously polls Cloud Storage for new files. As soon as a file arrives, initiate the training job.

C. Configure a Cloud Storage trigger to send a message to a Pub/Sub topic when a new file is available in a storage bucket. Use a Pub/Sub-triggered Cloud Function to start the training job on a GKE cluster.

D. Use Cloud Scheduler to schedule jobs at a regular interval. For the first step of the job, check the timestamp of objects in your Cloud Storage bucket. If there are no new files since the last run, abort the job.

Correct Answer: C

https://cloud.google.com/architecture/architecture-for-mlops-using-tfx-kubeflow-pipelines-and-cloud-build#triggering-and-scheduling-kubeflow-pipelines

---

QUESTION 6

You are training an object detection machine learning model on a dataset that consists of three million X-ray images, each roughly 2 GB in size. You are using Vertex AI Training to run a custom training application on a Compute Engine instance with 32-cores, 128 GB of RAM, and 1 NVIDIA P100 GPU. You notice that model training is taking a very long time. You want to decrease training time without sacrificing model performance. What should you do?

A. Increase the instance memory to 512 GB and increase the batch size.

B. Replace the NVIDIA P100 GPU with a v3-32 TPU in the training job.

C. Enable early stopping in your Vertex AI Training job.

D. Use the tf.distribute.Strategy API and run a distributed training job.

Correct Answer: D

QUESTION 7

You are training a deep learning model for semantic image segmentation with reduced training time. While using a Deep Learning VM Image, you receive the following error: The resource \\'projects/deeplearning-platforn/ zones/europe-west4c/acceleratorTypes/nvidia-tesla-k80\\' was not found. What should you do?

A. Ensure that you have GPU quota in the selected region.

B. Ensure that the required GPU is available in the selected region.

C. Ensure that you have preemptible GPU quota in the selected region.

D. Ensure that the selected GPU has enough GPU memory for the workload.

Correct Answer: B

https://cloud.google.com/deep-learning-vm/docs/troubleshooting#resource_not_found
https://cloud.google.com/compute/docs/gpus/gpu-regions-zones

QUESTION 8

You are developing an ML model intended to classify whether X-ray images indicate bone fracture risk. You have trained a ResNet architecture on Vertex AI using a TPU as an accelerator, however you are unsatisfied with the training time and memory usage. You want to quickly iterate your training code but make minimal changes to the code. You also want to minimize impact on the model\\'s accuracy. What should you do?

A. Reduce the number of layers in the model architecture.

B. Reduce the global batch size from 1024 to 256.

C. Reduce the dimensions of the images used in the model.

D. Configure your model to use bfloat16 instead of float32.

Correct Answer: D

https://cloud.google.com/tpu/docs/bfloat16

QUESTION 9

You are working on a Neural Network-based project. The dataset provided to you has columns with different ranges. While preparing the data for model training, you discover that gradient optimization is having difficulty moving weights to a good solution. What should you do?

A. Use feature construction to combine the strongest features.

B. Use the representation transformation (normalization) technique.

C. Improve the data cleaning step by removing features with missing values.

D. Change the partitioning step to reduce the dimension of the test set and have a larger training set.

Correct Answer: B

https://developers.google.com/machine-learning/data-prep/transform/transform-numeric

---

QUESTION 10

You are the Director of Data Science at a large company, and your Data Science team has recently begun using the Kubeflow Pipelines SDK to orchestrate their training pipelines. Your team is struggling to integrate their custom Python code into the Kubeflow Pipelines SDK. How should you instruct them to proceed in order to quickly integrate their code with the Kubeflow Pipelines SDK?

A. Use the func_to_container_op function to create custom components from the Python code.

B. Use the predefined components available in the Kubeflow Pipelines SDK to access Dataproc, and run the custom code there.

C. Package the custom Python code into Docker containers, and use the load_component_from_file function to import the containers into the pipeline.

D. Deploy the custom Python code to Cloud Functions, and use Kubeflow Pipelines to trigger the Cloud Function.

Correct Answer: A

https://kubeflow-pipelines.readthedocs.io/en/stable/source/kfp.components.html?highlight=func_to_container_op%20#kfp.components.func_to_container_op

---

QUESTION 11

You work for an online travel agency that also sells advertising placements on its website to other companies. You have been asked to predict the most relevant web banner that a user should see next. Security is important to your company. The model latency requirements are 300ms@p99, the inventory is thousands of web banners, and your exploratory analysis has shown that navigation context is a good predictor. You want to Implement the simplest solution. How should you configure the prediction pipeline?

A. Embed the client on the website, and then deploy the model on AI Platform Prediction.

B. Embed the client on the website, deploy the gateway on App Engine, and then deploy the model on AI Platform Prediction.

C. Embed the client on the website, deploy the gateway on App Engine, deploy the database on Cloud Bigtable for writing and for reading the user\\'s navigation context, and then deploy the model on AI Platform Prediction.

D. Embed the client on the website, deploy the gateway on App Engine, deploy the database on Memorystore for writing and for reading the user\\'s navigation context, and then deploy the model on Google Kubernetes Engine.

Correct Answer: C

https://medium.com/google-cloud/secure-cloud-run-cloud-functions-and-app-engine-with-api-key-73c57bededd1

QUESTION 12

Your team is building a convolutional neural network (CNN)-based architecture from scratch. The preliminary experiments running on your on-premises CPU-only infrastructure were encouraging, but have slow convergence. You have been asked to speed up model training to reduce time-to-market. You want to experiment with virtual machines (VMs) on Google Cloud to leverage more powerful hardware. Your code does not include any manual device placement and has not been wrapped in Estimator model-level abstraction. Which environment should you train your model on?

A. AVM on Compute Engine and 1 TPU with all dependencies installed manually.

B. AVM on Compute Engine and 8 GPUs with all dependencies installed manually.

C. A Deep Learning VM with an n1-standard-2 machine and 1 GPU with all libraries pre-installed.

D. A Deep Learning VM with more powerful CPU e2-highcpu-16 machines with all libraries pre-installed.

Correct Answer: C

https://cloud.google.com/deep-learning-vm/docs/cli#creating_an_instance_with_one_or_more_gpus
https://cloud.google.com/deep-learning-vm/docs/introduction#pre-installed_packages