Google Discussions

Exam Professional Machine Learning Engineer All Questions

View all questions & answers for the Professional Machine Learning Engineer exam

Go to Exam

EXAM PROFESSIONAL MACHINE LEARNING ENGINEER TOPIC 1 QUESTION 74 DISCUSSIO..

Actual exam question from Google's Professional Machine Learning Engineer

Question #: 74

Topic #: 1

[All Professional Machine Learning Engineer Questions]

You have deployed a model on Vertex AI for real-time inference. During an online prediction request, you get an “Out of Memory” error. What should you do?

A. Use batch prediction mode instead of online mode.

B. Send the request again with a smaller batch of instances.

C. Use base64 to encode your data before using it for prediction.

D. Apply for a quota increase for the number of prediction requests.

Show Suggested Answer

by Sivaram06 at *Dec. 11, 2022, 10:05 a.m.*

Comments

Type your comment...

Submit

⊟ **hiromi** Highly Voted 1 year, 10 months ago

Selected Answer: B

B is the answer
429 - Out of Memory
https://cloud.google.com/ai-platform/training/docs/troubleshooting

upvoted 22 times

tavva_prudhvi 1 year, 7 months ago

Upvote this comment, its the right answer!

upvoted 4 times

PhilipKoku Most Recent 5 months ago

Selected Answer: B

B) Use smaller set of tokens

upvoted 1 times

pmle_nintendo 8 months, 1 week ago

Selected Answer: B

By reducing the batch size of instances sent for prediction, you decrease the memory footprint of each request, potentially alleviating the out-of-memory issue. However, be mindful that excessively reducing the batch size might impact the efficiency of your prediction process.

upvoted 1 times

M25 1 year, 6 months ago

Selected Answer: B

Went with B

upvoted 1 times

tavva_prudhvi 1 year, 7 months ago

B. Send the request again with a smaller batch of instances.

If you are getting an "Out of Memory" error during an online prediction request, it suggests that the amount of data you are sending in each request is too large and is exceeding the available memory. To resolve this issue, you can try sending the request again with a smaller batch of instances. This reduces the amount of data being sent in each request and helps avoid the out-of-memory error. If the problem persists, you can also try increasing the machine type or the number of instances to provide more resources for the prediction service.

upvoted 2 times

BenMS 1 year, 8 months ago

Selected Answer: C

This question is about prediction not training - and specifically it's about _online_ prediction (aka realtime serving).

All the answers are about batch workloads apart from C.

upvoted 1 times

BenMS 1 year, 8 months ago

Okay, option D is also about online serving, but the error message indicates a problem for individual predictions, which will not be fixed by increasing the number of predictions per second.

upvoted 1 times

Antmal 1 year, 7 months ago

@BenMS this feels like a trick question.... makes on to zone to the word batch. https://cloud.google.com/ai-platform/training/docs/troubleshooting .... states then when an error occurs with an online prediction request, you usually get an HTTP status code back from the service. These are some commonly encountered codes and their meaning in the context of online prediction:

429 - Out of Memory
The processing node ran out of memory while running your model. There is no way to increase the memory allocated to prediction nodes at this time. You can try these things to get your model to run:

Reduce your model size by:
1. Using less precise variables.
2. Quantizing your continuous data.
3. Reducing the size of other input features (using smaller vocab sizes, for example).
4. Send the request again with a smaller batch of instances.

upvoted 1 times

koakande 1 year, 10 months ago

Selected Answer: B

https://cloud.google.com/ai-platform/training/docs/troubleshooting

upvoted 2 times

ares81 1 year, 11 months ago

The correct answer is B.

upvoted 1 times

LearnSodas 1 year, 11 months ago

Selected Answer: B

answer B as reported here: https://cloud.google.com/ai-platform/training/docs/troubleshooting

upvoted 1 times

Sivaram06 1 year, 11 months ago

Selected Answer: B

https://cloud.google.com/ai-platform/training/docs/troubleshooting#http_status_codes

upvoted 1 times

ExamTopics doesn't offer Real Microsoft Exam Questions. ExamTopics doesn't offer Real Amazon Exam Questions. ExamTopics Materials do not contain actual questions and answers from Cisco's Certification Exams.

CFA Institute does not endorse, promote or warrant the accuracy or quality of ExamTopics. CFA® and Chartered Financial Analyst® are registered trademarks owned by CFA Institute.