EXAM PROFESSIONAL MACHINE LEARNING ENGINEER TOPIC 1 QUESTION 101 DISCUSSI...

Actual exam question from Google's Professional Machine Learning Engineer
Question #: 101
Topic #: 1
[All Professional Machine Learning Engineer Questions]

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.

Show Suggested Answer

by ares81 at *Dec. 16, 2022, 11:34 a.m.*

Comments

Type your comment...

Submit

mymy9418 Highly Voted 1 year, 10 months ago

i think should be D
https://cloud.google.com/tpu/docs/bfloat16

upvoted 7 times

fitri001 Most Recent 6 months, 1 week ago

Selected Answer: D

Configuring bfloat16 instead of float32 (D): This offers a good balance between speed, memory usage, and minimal code changes. Bfloat16 uses 16 bits per float value compared to 32 bits for float32.

pen_spark
expand_more This can significantly reduce memory usage while maintaining similar accuracy in many machine learning models, especially for image recognition tasks.expand_more It's a quick change with minimal impact on the code and potentially large gains in training speed.

upvoted 2 times

pinimichele01 6 months, 3 weeks ago

Selected Answer: D

"the Google hardware team chose bfloat16 for Cloud TPUs to improve hardware efficiency while maintaining the ability to train deep learning models accurately, all with minimal switching costs from float32"

upvoted 1 times

pico 1 year, 1 month ago

Selected Answer: B

while reducing the global batch size (Option B) and configuring your model to use bfloat16 (Option D) are both valid options, reducing the global batch size is typically a safer and more straightforward choice to quickly iterate and make minimal changes to your code while still achieving reasonable model performance.

upvoted 1 times

pico 1 year, 1 month ago

Why not D:
Numerical Precision: bfloat16 has a lower numerical precision compared to float32
Compatibility: Not all machine learning frameworks and libraries support bfloat16 natively.
Hyperparameter Tuning: When switching to bfloat16, you may need to adjust hyperparameters, such as learning rates and gradient clipping thresholds, to accommodate the lower numerical precision
Model Architecture: Some model architectures and layers may be more sensitive to reduced precision than others.

upvoted 1 times

tavva_prudhvi 12 months ago

TPUs are optimized for operations with bfloat16 data types. By switching from float32 to bfloat16, you can benefit from the TPU's hardware acceleration capabilities, leading to faster computation and reduced memory usage without significant changes to your code.

While bfloat16 offers a lower precision compared to float32, it maintains a similar dynamic range. This means that the reduction in numerical precision is unlikely to have a substantial impact on the accuracy of your model, especially in the context of image classification tasks like bone fracture risk assessment in X-rays.

While reducing the batch size can decrease memory usage, it can also affect the model's convergence and accuracy. Additionally, TPUs are highly efficient with large batch sizes, so reducing the batch size might not fully leverage the TPU's capabilities.

upvoted 2 times

Voyager2 1 year, 5 months ago

Selected Answer: D

I think it should be D since they are using a TPU.https://cloud.google.com/tpu/docs/bfloat16

upvoted 1 times

M25 1 year, 6 months ago

Selected Answer: D

Went with D

upvoted 1 times

tavva_prudhvi 1 year, 7 months ago

Selected Answer: D

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

upvoted 1 times

TNT87 1 year, 8 months ago

Selected Answer: D

Selected Answer: D

Answer D

upvoted 2 times

ailiba 1 year, 8 months ago

"the Google hardware team chose bfloat16 for Cloud TPUs to improve hardware efficiency while maintaining the ability to train deep learning models accurately, all with minimal switching costs from float32" so since its already trained on TPU, D maybe has no effect?

upvoted 2 times

John_Pongthorn 1 year, 9 months ago

Selected Answer: D

I go with D exactly, primarily. the rest don't make any sense at all

upvoted 2 times

ares81 1 year, 10 months ago

Selected Answer: D

It should be D.

upvoted 1 times

hiromi 1 year, 10 months ago

Selected Answer: D

D

Agree with mymy9418

upvoted 2 times

mil_spyro 1 year, 10 months ago

Selected Answer: D

Agree with D

upvoted 1 times

ares81 1 year, 10 months ago

Selected Answer: B

It should be B.

upvoted 1 times