

- Expert Verified, Online, Free.

≡ MENU

G 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 131 DISCUSSI...

Actual exam question from Google's Professional Machine Learning Engineer

Question #: 131

Topic #: 1

[All Professional Machine Learning Engineer Questions]

You are an ML engineer at a mobile gaming company. A data scientist on your team recently trained a TensorFlow model, and you are responsible for deploying this model into a mobile application. You discover that the inference latency of the current model doesn't meet production requirements. You need to reduce the inference time by 50%, and you are willing to accept a small decrease in model accuracy in order to reach the latency requirement. Without training a new model, which model optimization technique for reducing latency should you try first?

- A. Weight pruning
- B. Dynamic range quantization
- C. Model distillation
- D. Dimensionality reduction

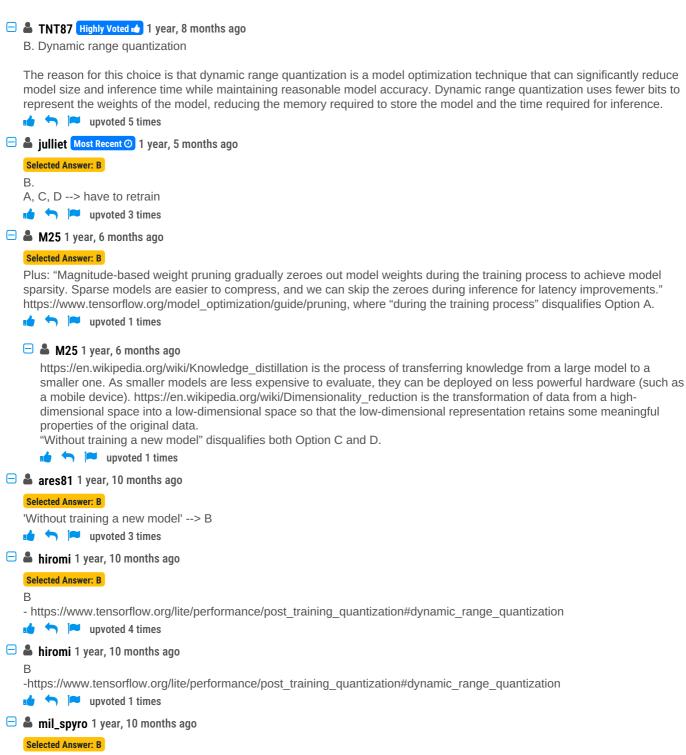
Show Suggested Answer

by Amil_spyro at Dec. 13, 2022, 7:53 p.m.

Comments

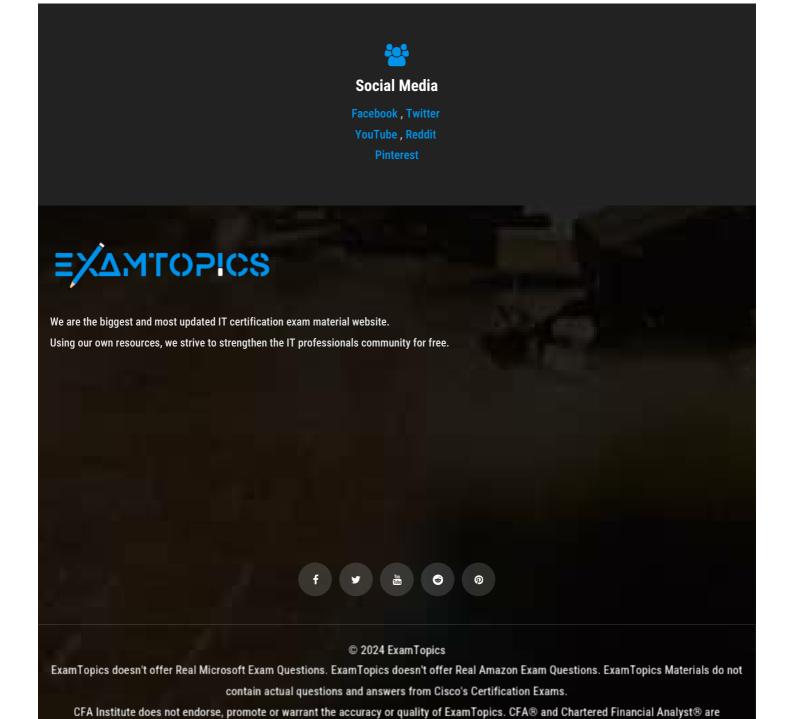
Type your comment...

Submit



The requirement is "Without training a new model" hence dynamic range quantization. https://www.tensorflow.org/lite/performance/post training quant

upvoted 3 times



registered trademarks owned by CFA Institute.