Google Discussions

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 mil_spyro at *Dec. 13, 2022, 7:53 p.m.*

Comments

Type your comment...

Submit

TNT87 Highly Voted 1 year, 8 months ago

B. Dynamic range quantization

The reason for this choice is that dynamic range quantization is a model optimization technique that can significantly reduce model size and inference time while maintaining reasonable model accuracy. Dynamic range quantization uses fewer bits to represent the weights of the model, reducing the memory required to store the model and the time required for inference.

upvoted 5 times

julliet Most Recent 1 year, 5 months ago

Selected Answer: B

B.
A, C, D --> have to retrain

upvoted 3 times

M25 1 year, 6 months ago

Selected Answer: B

Plus: “Magnitude-based weight pruning gradually zeroes out model weights during the training process to achieve model sparsity. Sparse models are easier to compress, and we can skip the zeroes during inference for latency improvements.” https://www.tensorflow.org/model_optimization/guide/pruning, where “during the training process” disqualifies Option A.

upvoted 1 times

M25 1 year, 6 months ago

https://en.wikipedia.org/wiki/Knowledge_distillation is the process of transferring knowledge from a large model to a smaller one. As smaller models are less expensive to evaluate, they can be deployed on less powerful hardware (such as a mobile device). https://en.wikipedia.org/wiki/Dimensionality_reduction is the transformation of data from a high-dimensional space into a low-dimensional space so that the low-dimensional representation retains some meaningful properties of the original data.
“Without training a new model” disqualifies both Option C and D.

upvoted 1 times

ares81 1 year, 10 months ago

Selected Answer: B

'Without training a new model' --> B

upvoted 3 times

hiromi 1 year, 10 months ago

Selected Answer: B

B
- https://www.tensorflow.org/lite/performance/post_training_quantization#dynamic_range_quantization

upvoted 4 times

hiromi 1 year, 10 months ago

B
-https://www.tensorflow.org/lite/performance/post_training_quantization#dynamic_range_quantization

upvoted 1 times

mil_spyro 1 year, 10 months ago

Selected Answer: B

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

upvoted 3 times