← Back Quantization and Pruning
Practice Quiz • 20 min

Congratulations! You passed!

Grade received 100% **To pass** 80% or higher

Go to next item

Quantization and Pruning

Total points 7

1. True Or False: Today, due to developments in machine learning research, and performance improvements for mobile and edge devices, there exists a wide range of options to deploy a machine learning solution locally. **1 / 1 point**

○ False

● True

Correct
That's right! With mobile devices becoming increasingly more powerful and at the same time cheaper, these devices are now able to deploy machine learning solutions at the edge.

2. Which of the following benefits does machine learning provide to mobile & IoT businesses that use it? (Select all that apply) **1 / 1 point**

☑ Automating operational efficiency.

Correct
That's right! Mobile and IoT deployments can streamline your business and help you make accurate predictions. Also, the automation of some processes can decrease the time of information analysis, and therefore, can be crucial to improve operational efficiency.

□ Eliminating risk.

☑ Strengthening security.

Correct
That's right! With ever increasing number of breaches and confidential data theft, companies want to strengthen their security. Employing ML in mobile and IoT security can help detect intrusions, protect your data, and respond to incidents automatically.

☑ Improving user experience with data.

Correct
That's right! Businesses with a mobile or IoT strategy know how technology can capture and transform data to offer greater access to consumer information and therefore devise better means to enhance their user experiences.

3. ML Kit brings Google's machine learning expertise to mobile developers. Which of the following are features of ML Kit? (Select all that apply) **1 / 1 point**

☑ Access to cloud-based web services

Correct
That's right! With ML, you can upload your models through the Firebase console and let the service take care of hosting and serving them to your app users.

□ On-device model training

☑ Model customization

Correct
That’s right! With ML, you can customize your device ML features, such as facial detection, bar-code scanning, and object detection, among others.

☑ Pre-trained model compatibility

Correct
That’s right! With ML, you can use a pre-trained TensorFlow Lite set of vetted models, provided they meet a set of criteria.

4. In per-tensor quantization, weights are represented by int8 two’s complement values in the range ____________ with zero-point ____________

1 / 1 point

○ [-127, 127], in range [-128, 127].

○ [-128, 127], in range [-128, 127].

● [-127, 127], equal to 0

○ [-128, 127], equal to 0

Correct
That's right! In per-tensor weights, there are two complement values in the range [-127, 127], with zero-point equal to 0 in approximates.

5. Quantization squeezes a small range of floating-point values into a fixed number. What are the impacts of quantization on the behavior of the model?

1 / 1 point

☑ Decreased interpretability of the ML model

Correct
That's right! In the case of ML interpretability, there are some effects imposed on the ML model after quantization. This means it’s hard to evaluate whether transforming a layer was going in the right. Therefore, the interpretability of the model may decrease.

□ Increased precision as a result of the optimization process

☑ Changes in transformations and operation

Correct
That's right! You could have transformations like adding, modifying, removing operations, coalescing different operations, and so on. In some cases, transformations may need extra data.

☑ Layer weights changes and network activations

Correct
One of the significant impacts is the change of static parameters such as layer weights, and others could be dynamic such as activations within networks.

6. True Or False: One family of optimizations, known as pruning, aims to remove neural network connections, increasing the number of parameters involved in the computation.

1 / 1 point

○ True

● False

Correct
That's right! The pruning optimization aims to eliminate neural network connections, but instead of increasing the number of parameters, you have to

reduce them. With pruning, you can lower the overall parameter count in the network and reduce their storage and computational cost.

7. Which of the following describe the benefits of applying sparsity with a pruning routine? (Select all that apply)

1 / 1 point

☑ Can be used in tandem with quantization to get additional benefits

✓ **Correct**
That's right! In some experiments, weight pruning is compatible with quantification, resulting in compounding benefits. Therefore, it is possible to further compress the pruned model by applying post-training quantization.

☑ Better storage and/or transmission

✓ **Correct**
That's right! An immediate benefit that you can get out of pruning is disk compression of sparse tensors. Thus, you can reduce the model's size for its storage and transmission by applying simple file compression to the pruned checkpoint.

□ Method perform well at a large scale

☑ Gain speedups in CPU and some ML accelerators

✓ **Correct**
That's right! You can even gain speeds in the CPU and ML throttles that fully exploit integer precision efficiencies in some cases.