Congratulations! You passed!

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

Go to next item

Quantization and Pruning

Total points 7

1. Today, due to the developments in machine learning research and mobile and edge devices, there exists a wide range of alternatives to deploy a machine learning solution locally. **1 / 1 point**

- (•) Yes
- ( ) No

Correct
That's right! With the advent of ML research and improved mobile devices' capabilities, there are more opportunities to deploy on-device machine learning solutions. Additionally, these devices have evolved over a period to be built upon cheap hardware, thereby allowing its mass production.

2. Which of the following are reasons for Improving mobile & IoT business with ML? (Select all that apply) **1 / 1 point**

- [x] Automating operational efficiency.

Correct
That's right! Mobile and IoT deployments 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.

- [ ] Eliminate risk.
- [x] Strengthened security.

Correct
That's right! With the current increase in breaches and confidential data theft, companies want to strengthen their security. Employing AI in mobile and IoT security can help detect third parties to intrude, shield your private information, and respond to incidents automatically.

- [x] 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 consumer experiences and users.

3. ML Kit brings Google's machine learning expertise to mobile developers. With this tool, you can _____ . (Select all that apply) **1 / 1 point**

- [x] use to access 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.

- [ ] use it to train your model on-device.
- [x] use a pre-trained model.

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.

- [x] use it to customize your models

Correct
That's right! With ML, you can apply different customizations on your device ML features, such as facial detection, bar-code scanning, and object detection, among others.

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].
- (•) [-127, 127], equal to 0
- ( ) [-128, 127], equal to 0
- ( ) [-128, 127], in range [-128, 127].

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 impact is there on the behavior of the model? (Select all that apply) **1 / 1 point**

- [ ] You can increase precision as a result of the optimization process.
- [x] You can have 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.

- [x] You can decrease the 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.

- [x] You can have change layer weights and activations networks

  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. One such family of optimizations known as pruning aims to remove neural network connections, increasing the number of parameters involved in the computation. **1 / 1 point**

- ( ) Yes
- (•) No

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

7. Which ones of the following describe the benefits of applying sparsity with a pruning routine? (Select all that apply) **1 / 1 point**

- [ ] Method perform well at a large scale
- [x] 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.

- [x] 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.

- [x] 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.