ECE661 Quiz 5

Name:________________________ **UniqueID:**________________ **Score:**

This quiz is closed-book. By signing your name above, you agree to follow Duke Community Standard. For True/False and multiple-choice questions, no justification is needed.

1. (1pt) (T/F) When implementing quantization-aware training with straight-through estimator (STE), gradients are quantized during backpropagation, ensuring that updates are consistent with the quantized weights.

False. Gradients are typically not quantized, but the forward pass uses the quantized weights.

2. (3pt) (Short Answer) Describe the potential regularization effects of weight quantization in neural networks. Can it perform a role similar to dropout? Provide a brief explanation supporting your answer.

No. Quantization and dropout serve different purposes (1pt); while quantization reduces model size and computational complexity (1pt), dropout is a technique for preventing overfitting by randomly zeroing network outputs during training (1pt).

3. (3pt) (Short Answer) Explain the concept of mixed-precision quantization in neural networks. What are the primary benefits of it?

This involves using different quantization precisions for different layers or parts of the network (1pt). It can provide a balance between efficiency (lower memory and computation) (1pt) and maintaining high accuracy, as critical parts of the network can be kept at higher precision (1pt).

4. (3pt) (Short Answer) Discuss the impact of inducing sparsity in the structure of deep neural networks. What are the primary benefits of sparsity in terms of architecture and performance?

Sparsity in a neural network refers to the presence of many zero-valued or near-zero weights (1pt). This reduces the model's size and can improve inference efficiency by skipping unnecessary calculations (1pt). However, achieving optimal sparsity without sacrificing performance requires careful tuning (1pt).