Your grade: 100%

Your latest: **100%** • Your highest: **100%** • To pass you need at least 75%. We keep your highest score.

Next item

1. Which of the following reasons makes selective fine-tuning less effective for transformer architectures? 1 / 1 point

- ○ Because it allows adding layers to a pre-trained transformer model between the attention blocks.
- ◉ Because of the higher number of parameters in transformer architectures.
- ○ Because it involves updating the neural parameters, layers, and neurons.
- ○ Because it allows for task-specific customization in transformers.

Correct
Selective fine-tuning updates only a subset of layers or parameters, which works for other networks. It is less effective for transformer architectures due to their higher number of parameters and the need for more extensive updates. This limitation has led to the development of alternative methods.

2. Which of the following is the key aspect of the low-rank adaptation (LoRA) for enhancing the efficiency of fine-tuning large language models (LLMs)? 1 / 1 point

- ○ LoRA parallelizes the model across the GPUs to increase the training speed.
- ○ LoRA introduces a new architecture for replacing transformers in LLMs.
- ○ During the training process, LoRA eliminates large datasets.
- ◉ LoRA uses row-rank decomposition for trainable parameters for fine-tuning LLMs using weight matrices.

Correct
LoRA reduces the number of trainable parameters decomposing weight matrices and makes fine-tuning process efficient.

3. Which of the following statements is correct with respect to quantization to quantized low-rank adaptation (QLoRA)? 1 / 1 point

- ◉ Quantization to QLoRA helps the model be fine-tuned using less memory; however maintains the same performance.
- ○ Quantization to QLoRA reduces the model's accuracy by increasing inference speed
- ○ Quantization to QLoRA streamlines the model architecture, making it simpler to interpret.
- ○ Quantization to QLoRA helps the model handle more complex tasks and increases the calculation's precision.

Correct
Using less memory helps to handle large-scale models on limited hardware resources and maintains the model's performance with full precision.

4. In the context of training LoRA with PyTorch, identify the correct statement for the initialization of low-rank matrices AAA and BBB. 1 / 1 point

- ○ Initialize the matrices AAA and BBB with zeros to prevent initial influence on the model's predictions.
- ◉ Initialize the matrices AAA and BBB using any distribution method by ensuring stable training and proper scaling of the values.
- ○ Initialize the AAA and BBB matrices as identity matrices, ensuring they start as neutral elements in matrix multiplication.
- ○ Initialize the matrices AAA and BBB using random values and leveraging standard normal distribution.

Correct
Proper scaling and stable training are important to initialize AAA and BBB matrices. This helps maintain balance gradients and prevent challenges, such as exploding and vanishing gradients.