Exercise 1 - Analyze compression technique combinations for optimal efficiency

Estimated Time: 15 minutes.

The scenario

You are an AI Optimization Specialist at a company renowned for deploying efficient deep learning models. A new intern on the team has been experimenting with compressing a Computer Vision model for a crucial edge device deployment. They have shared some of their approaches and are seeking your expertise to understand why some attempts failed and how to strategize better.

Your challenge

Part 1: Insights on compression techniques interactions

Your intern has started a table summarizing potential issues when combining compression techniques

Your task: Complete the table by filling missing entries in the “Compability” and "Key Interaction Consideration / Potential Issue” sections.

💡*Hint:* *Think about what information each technique needs to work effectively.*

1st Technique Applied	**2nd Technique Applied**	**Compatibility** *[poor, good, moderate]*	**Key Interaction Consideration / Potential Issue**
1. INT4 Post-Training Quantization	Magnitude Pruning	Poor	Extreme quantization destroys weight importance signals needed for pruning decisions
2. Magnitude Pruning	Post-Training Quantization	Good	Pruning creates narrower weight distributions, making quantization more effective
3. Knowledge Distillation	Pruning (on the student model)	Moderate	Student model is already smaller; aggressive pruning may remove essential knowledge
4. Quantization-Aware Training	Graph Optimizations	Good	QAT models are already optimized for quantized ops; graph optimizations enhance performance
5. Post-training Pruning (on the teacher model)	Knowledge Distillation	Poor	Pruning removes knowledge that the teacher needs to transfer; distillation should come first

Part 2: Analyzing a flawed pipeline

The intern shares the following pipeline they used for a ResNet-50 model. Despite achieving 80% size reduction, accuracy dropped by 23%, making it unusable.

Failed pipeline: ResNet-50 for edge vision task

1. *Initial Step:* Applied **Post-Training INT4 Quantization** to the pre-trained ResNet-50.
2. *Second Step:* Performed **Magnitude Pruning** (unstructured, aiming for 50% sparsity).
3. *Third Step:* Attempted **Knowledge Distillation** to a smaller, custom CNN architecture.
4. *Final Step:* Applied **Graph Optimizations** (e.g., layer fusion).

A. Review pipeline design practices *(2 sentences)***:** Why is it not a good idea typically to combine so many compression methods in the same pipeline?

Compression pipelines typically use two (or three, at most) techniques, as each additional stage introduces error and complexity, often with diminishing returns. Beyond a point, the accumulated degradation in model quality and increased implementation difficulty outweigh the benefits.

B. Identify the primary flaw *(1-2 sentences)***:** What is the most critical misstep in this pipeline's design concerning technique interaction?

Starting with aggressive INT4 quantization destroyed weight precision, making subsequent techniques unable to identify important vs unimportant parameters.

C. Explain the negative cascade *(2-3 sentences)***:** How did the initial steps negatively impact the subsequent compression techniques?

INT4 quantization eliminates subtle weight differences needed by pruning to identify redundant connections. This degraded model becomes a poor teacher for distillation, compounding quality loss at each stage.

D. Propose a more robust pipeline order *(List the 4 techniques in a better order)***:**
1. Knowledge Distillation
2. Magnitude Pruning
3. Post-Training Quantization (INT8)
4. Graph Optimizations

Briefly justify why your first technique goes first *(1 sentence)*:

Distillation should go first to transfer knowledge while the model has full precision and capacity.

Bonus: Could any technique be omitted from your improved pipeline? Why? *(1-2 sentences)*:

Pruning could be omitted by directly creating a smaller student model during distillation. In fact, additional pruning after distillation could harm the carefully learned representations.

Part 3: Quick check: Core principles

1. Which statements about compression technique ordering are TRUE?

(Select all that apply)

a. Always apply the most aggressive compression first for maximum size reduction
b. Training-time techniques (like QAT) generally preserve more information than post-training methods
c. Graph optimizations should always be the first step
d. Techniques that preserve model information often work better early in pipelines
e. The order rarely matters as long as all techniques are applied

Explanation: Training-time techniques like QAT and gradual pruning can adapt during training to minimize information loss, making them more effective than post-training alternatives. Similarly, techniques that preserve information (like distillation or careful pruning) should typically come early in the pipeline, as later techniques need this preserved information to make good compression decisions.

2. When compressing Large Language Models (LLMs) vs CNNs, what is an accurate consideration for compression pipeline design?

a. Both architectures use identical compression techniques with the same effectiveness
b. LLMs can use attention head pruning while CNNs typically use channel/filter pruning
c. Knowledge distillation only works effectively for CNN architectures
d. The order of compression techniques matters only for LLMs, not CNNs

Explanation: Different architectures have different structural elements that can be compressed. LLMs have attention heads that can be pruned, while CNNs have channels and filters. This architectural difference influences which compression techniques are most effective for each model type.

Self-reflection checklist - Your key takeaways

Before finishing, ensure you understand:

- ☐ Why compression technique order significantly impacts results
- ☐ How aggressive early compression undermines later techniques
- ☐ Why preserving information early in pipelines is often crucial
- ☐ How architecture differences (CNN vs LLM) affect compression strategies