El Oued University
Faculty of Exact Science

Academic Year: 2025/2026
Solution: ML for IoT Exam

Exam Solution Key

Questions

1. **Knowledge Distillation Objective:** The primary goal is to transfer the performance and "knowledge" of a large, complex model (Teacher) into a smaller, faster model (Student). This allows high-accuracy performance on resource-constrained IoT devices.
2. **Memory: Training vs. Inference:** Training requires significantly more memory because it must store **Gradients** for backpropagation, **Optimizer States** (like momentum or variance in Adam), and **Full Layer Activations** from the forward pass to compute derivatives.
3. **Autonomous Vehicle Constraint:** The primary constraint is **Latency (Real-time response)**. Safety requires the system to detect and react to obstacles in milliseconds; high throughput is useless if the decision arrives after a collision.
4. **NAS vs. KD Trade-offs: NAS** (Neural Architecture Search) explores the structural design space to find the most efficient topology, which is computationally expensive during the search phase. **KD** (Knowledge Distillation) keeps the architecture fixed and focuses on the training process to maximize the smaller model's accuracy.
5. **Accuracy in Imbalanced Data:** Accuracy is misleading here. If 99.9% of transactions are legitimate, a model that labels *everything* as "Legitimate" achieves 99.9% accuracy but fails to detect a single fraudulent transaction, which is the system's actual purpose.

Exercice Solution

Part 1: Architecture & Memory Analysis

Q1: Total Parameters (P_{total})

- **Layer 1 (Conv2D):** $(3 \times 3 \times 1 + 1) \times 16 = 160$
- **Layer 3 (Conv2D):** $(3 \times 3 \times 16 + 1) \times 32 = 4,640$
- **Layer 6 (Dense):** $(28,800 + 1) \times 3 = 86,403$
- **Total:** $160 + 4,640 + 86,403 = \mathbf{91,203}$ parameters

Q2: Memory for Parameters (M_{params}) Using FP32 (4 bytes per parameter):

$$M_{params} = \frac{91,203 \times 4}{1024} = \mathbf{356.26} \text{ KB}$$

Q3: Peak Activation Memory (M_{hidden}) Identifying the largest output feature map:

- Layer 1: $126 \times 126 \times 16 \times 4$ bytes ≈ 992.25 KB
- Layer 3: $61 \times 61 \times 32 \times 4$ bytes ≈ 465.12 KB

Peak Memory (Layer 1): $\mathbf{992.25}$ KB

Q4: Total Inference Memory (M_{total})

$$M_{total} = 356.26 + 992.25 = \mathbf{1,348.51} \text{ KB}$$

Q5: Deployment feasibility: No. $1,348.51$ KB $>$ 520 KB SRAM.

Part 2: Structural Pruning & Optimization

Q1: Independent Scenarios

- **Only Structural Pruning (FP32):**
  - New $P_{total} = (10 \times 8) + (3 \times 3 \times 8 + 1) \times 8 + (7,200 + 1) \times 3 = 80 + 584 + 21,603 = 22,267$.
  - $M_{params} = (22,267 \times 4)/1024 = 86.98$ KB.
  - New Peak Activation (L1): $(126 \times 126 \times 8 \times 4)/1024 = 496.12$ KB.
  - **Total:** $86.98 + 496.12 = \mathbf{583.1}$ KB.
- **Only Int8 Quantization:** Total Memory / 4 = $1,348.51/4 = \mathbf{337.13}$ KB.

Q2: Pruned + Quantized (Int8)

$$M_{total} = \frac{583.1 \text{ KB}}{4} = \mathbf{145.77} \text{ KB}$$

Does it fit? Yes, 145.77 KB $<$ 520 KB.

Part 3: Computational Cost & Latency

Q1: Total Operations (Millions)

- **Base Model:** $(4.57\text{M} + 34.26\text{M} + 0.17\text{M}) = \mathbf{39}$ MFLOPs
- **Pruned Model:** $(2.28\text{M} + 4.27\text{M} + 0.04\text{M}) = \mathbf{6.59}$ MFLOPs

Q2: Latency Results (Seconds)

Model	**FP32 (10 MFLOPS)**	**Int8 (80 MOPS)**
Base (39M)	3.9 s	0.487 s
Pruned (6.59M)	0.659 s	0.082 s

Q3: Client 1 (¡ 0.2s): Best config is **Pruned + Int8**. Accuracy: $93\% - 5\% - 3\% = \mathbf{85}\%$.
Q4: Client 2 (¡ 0.5s): Best config is **Base + Int8**. Accuracy: $93\% - 3\% = \mathbf{90}\%$.