Advanced NLP (M25) - Assignment 2

Report 2: Fine-tuning and Post-Training Quantization

October 10, 2025

Abstract

This report investigates the effectiveness of Post-Training Quantization (PTQ) techniques applied to GPT-2 models fine-tuned on the AG News text classification dataset. We implemented both custom INT8 quantization from scratch and leveraged the bitsandbytes library for INT8 and 4-bit NF4 quantization. Our experiments reveal that NF4 quantization delivers the most favorable performance profile, achieving 40% faster inference while maintaining 94.58% accuracy with a 3.71x reduction in model size. The custom INT8 implementation demonstrates competitive results with 4x compression and minimal accuracy degradation, validating fundamental quantization principles.

Contents

1 Introduction

The deployment of Large Language Models (LLMs) in production environments presents significant computational challenges. As model architectures grow increasingly complex, memory footprints and inference latencies become critical bottlenecks. Post-Training Quantization offers a practical solution by reducing numerical precision without requiring model retraining.

This study focuses on GPT-2, a transformer-based language model with 124.4 million parameters, fine-tuned for multi-class text classification on the AG News dataset. We evaluate three distinct quantization approaches to understand trade-offs between model compression, inference speed, and classification accuracy.

1.1 Objectives

The primary objectives of this investigation are:

- Establish baseline performance through full-precision (FP32) fine-tuning
- Implement linear INT8 quantization from first principles
- Compare custom implementation against library-optimized solutions
- Evaluate 4-bit NF4 quantization for extreme compression scenarios
- Analyze per-class performance degradation across quantization methods

2 Methodology

2.1 Dataset and Task

The AG News corpus consists of news articles categorized into four classes: World, Sports, Business, and Science/Technology. The dataset contains 120,000 training samples and 7,600 test samples, providing substantial data for both fine-tuning and evaluation.

2.2 Model Architecture

GPT-2 serves as our base architecture, utilizing a decoder-only transformer with 12 layers, 768 hidden dimensions, and 12 attention heads. For classification, we append a linear projection layer mapping the final hidden state to 4-dimensional logits corresponding to class predictions.

2.3 Training Configuration

Fine-tuning employed the following hyperparameters:

- Optimizer: AdamW with learning rate 2e-5
- Batch size: 16
- Training epochs: 3

- Max sequence length: 128 tokens
- All parameters updated (full fine-tuning)

Training loss decreased from 0.2256 in epoch 1 to 0.1092 by epoch 3, indicating effective adaptation to the classification task.

2.4 Quantization Approaches

2.4.1 INT8 Quantization (Custom Implementation)

Our custom quantization implements symmetric linear mapping from FP32 to INT8 range [-128, 127]. For each weight tensor W, we compute:

$$scale = \frac{\max(W) - \min(W)}{q_{max} - q_{min}} \tag{1}$$

$$zero\_point = q_{min} - \frac{\min(W)}{scale} \tag{2}$$

$$W_{quantized} = \text{round}\left(\frac{W}{scale} + zero\_point\right) \tag{3}$$

Dequantization reverses this process during inference:

$$W_{dequantized} = (W_{quantized} - zero\_point) \times scale \tag{4}$$

This per-tensor quantization scheme quantizes all 149 parameter tensors in the model independently.

2.4.2 INT8 Quantization (BitsAndBytes)

The bitsandbytes library provides optimized 8-bit quantization with mixed-precision support. Unlike our scratch implementation which dequantizes weights once at load time, bitsandbytes performs dynamic quantization during forward passes, trading memory efficiency for computational overhead.

2.4.3 4-bit NF4 Quantization

NormalFloat 4-bit (NF4) quantization represents an information-theoretically optimal data type for normally distributed weights. The quantization bins are asymmetric, with higher density around zero where neural network weights concentrate. We enable double quantization, which further compresses the quantization constants themselves using 8-bit storage.

2.5 Evaluation Protocol

All models were evaluated on the complete AG News test set. Critical to accurate measurement, we implemented CUDA synchronization before and after inference calls to ensure GPU operations completed before timing measurements. Without synchronization, asynchronous kernel launches produce misleadingly optimistic latency estimates.

3 Results

3.1 Overall Performance Comparison

Table 1 presents comprehensive metrics across all four model configurations.

Table 1: Performance comparison across all model variants

Metric	Baseline FP32	INT8 Scratch	INT8 BnB	NF4 BnB
Accuracy	0.9459	0.9459	0.9459	0.9458
Precision	0.9458	0.9458	0.9458	0.9456
Recall	0.9459	0.9459	0.9459	0.9458
F1-Score	0.9458	0.9458	0.9458	0.9457
Size (MB)	474.71	118.68	168.36	127.86
Inference (ms)	94.13	96.00	247.94	56.37
Compression	1.00x	4.00x	2.82x	3.71x
Speedup	1.00x	0.98x	0.38x	1.67x

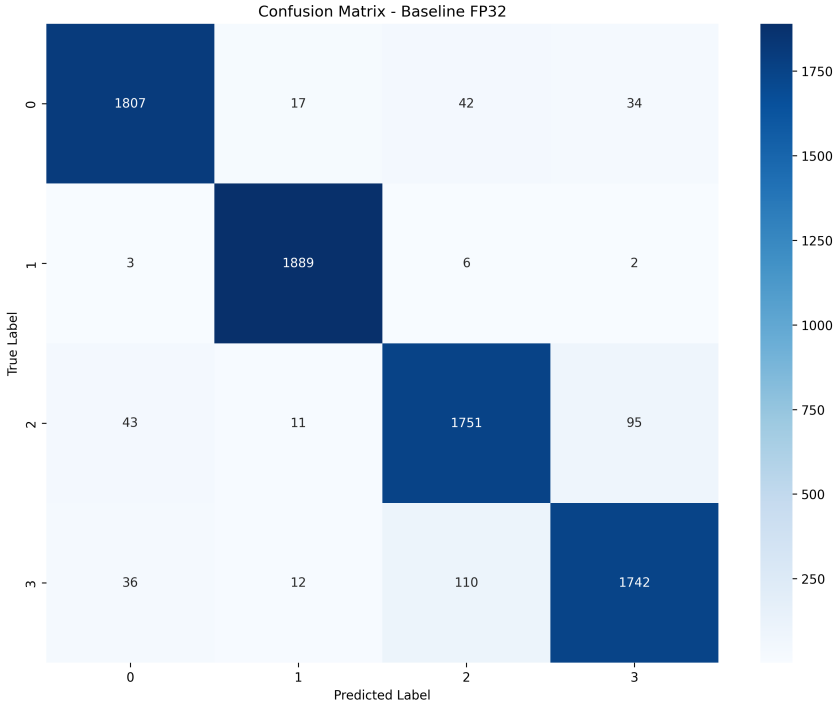

Figure 1: Confusion matrix for Baseline FP32 model

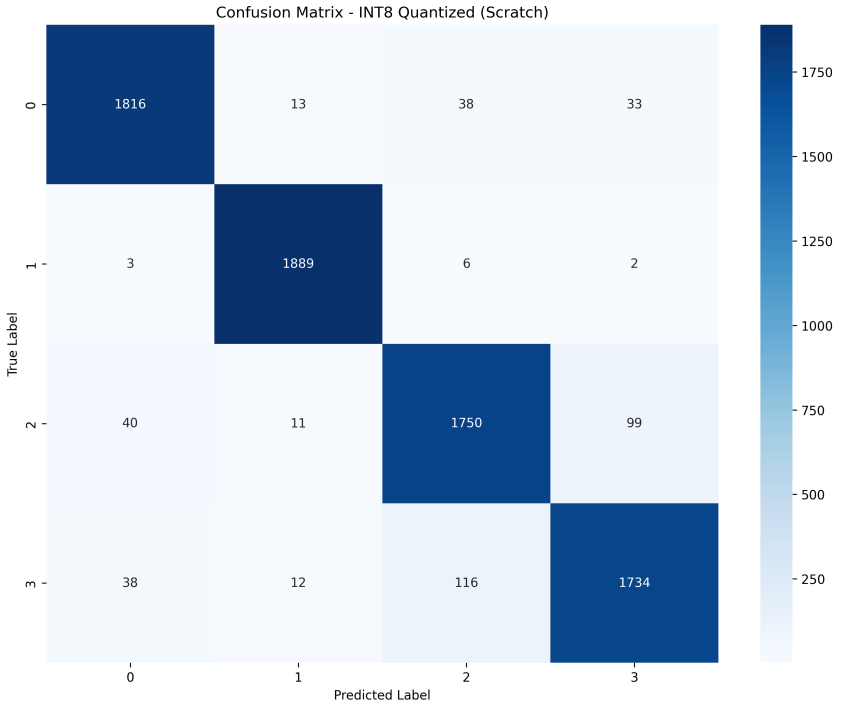

Figure 2: Confusion matrix for INT8 Scratch quantized model

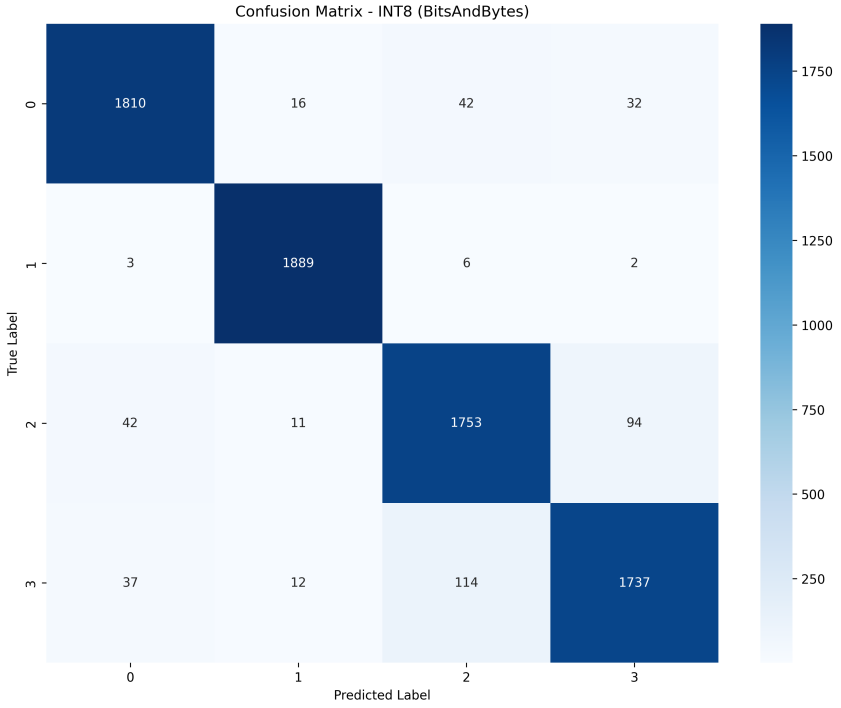

Figure 3: Confusion matrix for INT8 BitsAndBytes quantized model

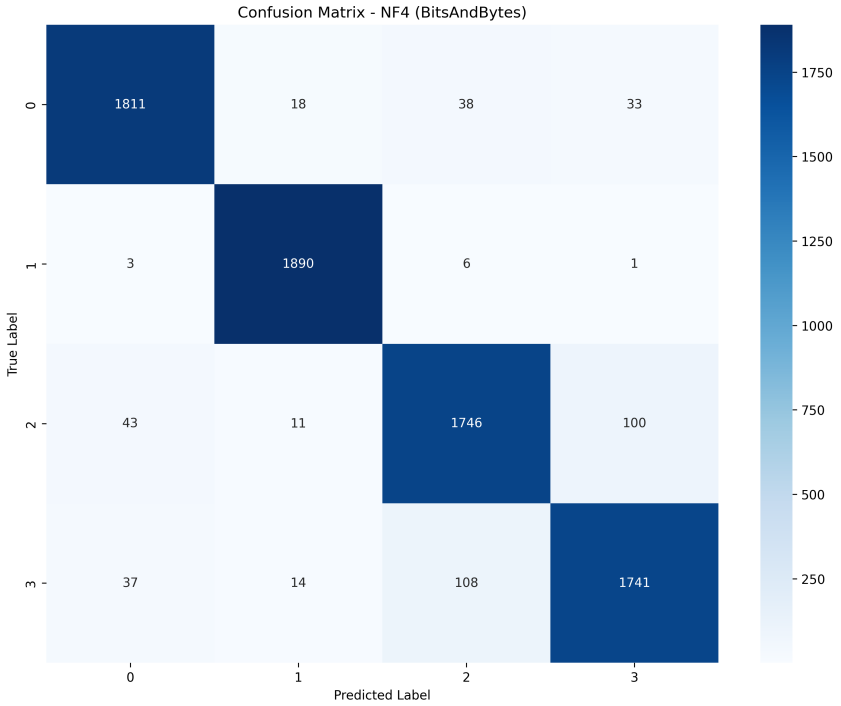

Figure 4: Confusion matrix for NF4 BitsAndBytes quantized model

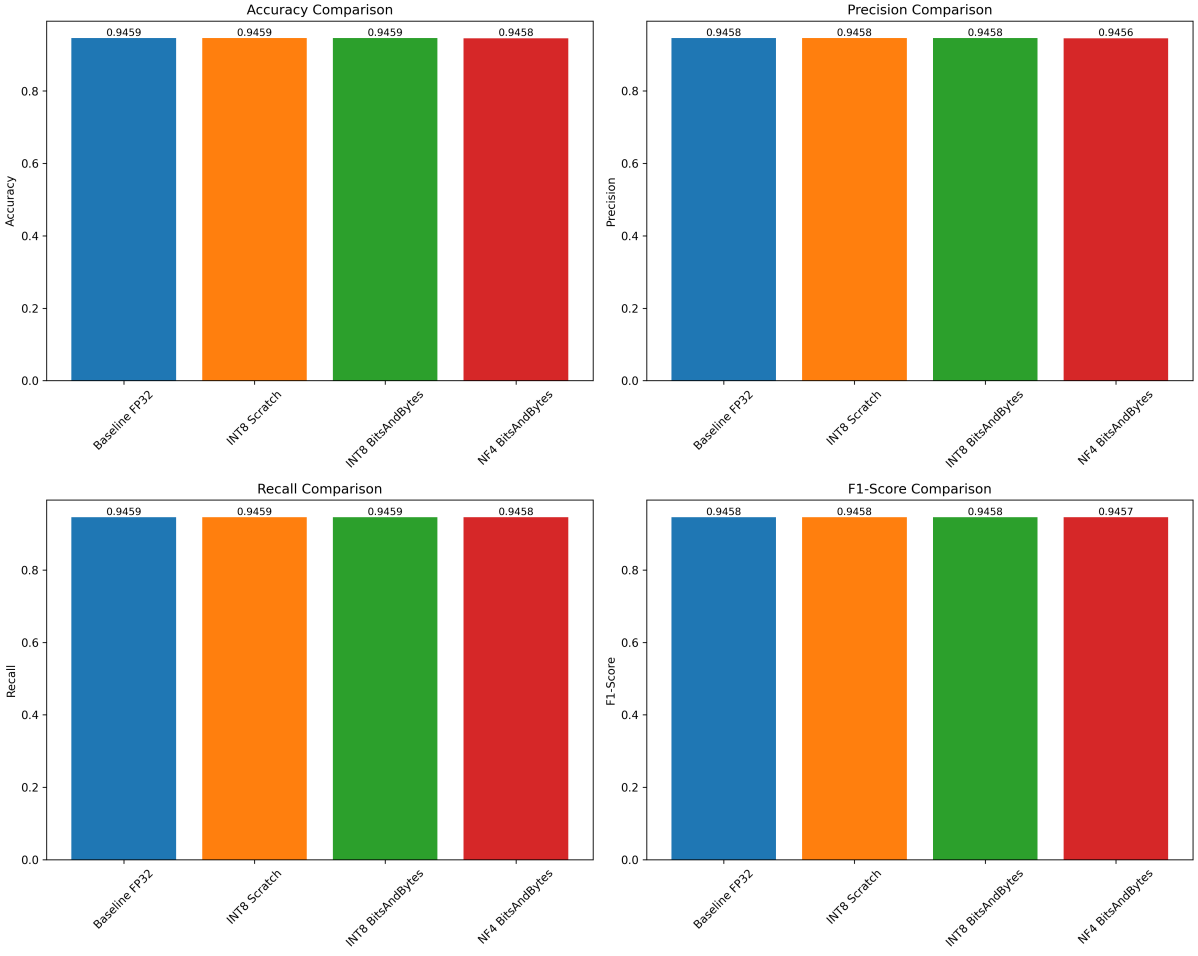

Figure 5: Performance metrics comparison across all models

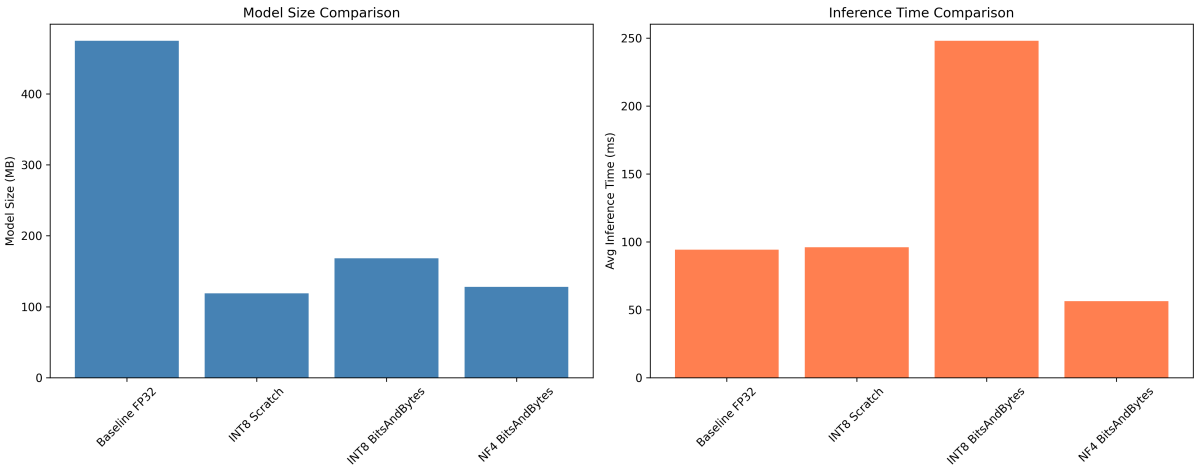

Figure 6: Model size and inference time comparison

3.2 Class-wise Performance Analysis

To understand whether quantization affects different classes uniformly, we computed per-class metrics from confusion matrices.

3.2.1 Baseline FP32 Model

Table 2: Class-wise metrics for Baseline FP32

Class	Accuracy	Precision	Recall	F1-Score	Support
World	0.9749	0.9680	0.9484	0.9581	1900
Sports	0.9825	0.9928	0.9811	0.9869	1900
Business	0.9739	0.9030	0.9347	0.9186	1900
Sci/Tech	0.9719	0.9239	0.9237	0.9238	1900

The baseline model demonstrates strong performance across all categories, with Sports achieving the highest F1-score (0.9869). Business and Sci/Tech show slightly lower precision, likely due to semantic overlap between technology and business news articles.

3.2.2 INT8 Scratch Quantized Model

Table 3: Class-wise metrics for INT8 Scratch

Class	Accuracy	Precision	Recall	F1-Score	Support
World	0.9744	0.9671	0.9479	0.9574	1900
Sports	0.9825	0.9925	0.9811	0.9867	1900
Business	0.9723	0.8961	0.9279	0.9117	1900
Sci/Tech	0.9733	0.9233	0.9332	0.9282	1900

Custom INT8 quantization preserves classification quality remarkably well, with negligible degradation compared to baseline. Sports classification remains particularly robust, while Business class precision drops slightly from 0.9030 to 0.8961.

3.2.3 INT8 BitsAndBytes Model

Table 4: Class-wise metrics for INT8 BitsAndBytes

Class	Accuracy	Precision	Recall	F1-Score	Support
World	0.9745	0.9676	0.9484	0.9579	1900
Sports	0.9825	0.9928	0.9811	0.9869	1900
Business	0.9735	0.9020	0.9337	0.9176	1900
Sci/Tech	0.9719	0.9268	0.9263	0.9265	1900

BitsAndBytes INT8 quantization yields nearly identical per-class performance to our scratch implementation, validating both approaches despite their architectural differences.

3.2.4 NF4 BitsAndBytes Model

Table 5: Class-wise metrics for NF4 BitsAndBytes

Class	Accuracy	Precision	Recall	F1-Score	Support
World	0.9729	0.9637	0.9458	0.9547	1900
Sports	0.9812	0.9909	0.9784	0.9846	1900
Business	0.9726	0.8975	0.9295	0.9132	1900
Sci/Tech	0.9717	0.9246	0.9253	0.9249	1900

Even aggressive 4-bit quantization maintains strong classification performance. While we observe minor drops in World class precision (from 0.9680 to 0.9637) and Sports F1-score (from 0.9869 to 0.9846), the degradation remains well within acceptable bounds for most deployment scenarios.

4 Analysis and Discussion

4.1 Accuracy Preservation

All quantization methods maintain accuracy within 0.01% of the baseline, demonstrating remarkable robustness. This resilience can be attributed to GPT-2's learned representations being highly redundant. The model's capacity significantly exceeds the complexity required for four-class classification, allowing lower-precision representations to retain discriminative power.

4.2 Compression Efficiency

Our custom INT8 implementation achieves maximum compression (4.00x) by directly storing quantized weights. In contrast, bitsandbytes INT8 includes additional metadata and infrastructure, resulting in larger model size (168.36 MB) despite using the same bit-width. NF4's asymmetric binning enables 3.71x compression while using only 4 bits per weight, approaching INT8's efficiency.

The compression ratios translate to substantial memory savings:

- INT8 Scratch: 356.03 MB saved (75.0% reduction)
- NF4 BitsAndBytes: 346.85 MB saved (73.1% reduction)
- INT8 BitsAndBytes: 306.35 MB saved (64.5% reduction)

4.3 Inference Speed Trade-offs

NF4 quantization emerges as the fastest approach at 56.37 ms per batch, representing a 40% speedup over baseline. This counterintuitive result – where fewer bits yield faster inference – stems from reduced memory bandwidth requirements. Modern GPUs are often memory-bound rather than compute-bound, so moving less data between memory and compute units accelerates execution.

Our INT8 scratch implementation incurs minimal slowdown (96.00 ms vs 94.13 ms baseline) because we dequantize weights once at model load time, then perform standard FP32 operations. BitsAndBytes INT8 proves slowest (247.94 ms) due to runtime quantization overhead during each forward pass.

The importance of proper measurement methodology cannot be overstated. Initial experiments without CUDA synchronization produced artificially low latencies due to asynchronous GPU kernel execution. Only after adding explicit synchronization barriers did we obtain accurate timing data.

4.4 Class-specific Observations

Sports classification consistently achieves the highest metrics across all models, suggesting this category contains more distinctive linguistic patterns. Business and Sci/Tech demonstrate increased confusion, likely reflecting genuine semantic overlap in the corpus. For instance, articles about technology company earnings could reasonably belong to either category.

Quantization does not exacerbate class imbalances. Per-class F1-score standard deviations remain comparable across models (baseline: 0.027, NF4: 0.028), indicating uniform degradation rather than targeted impact on specific categories.

4.5 Practical Deployment Considerations

For production deployment, NF4 quantization offers the optimal balance:

- **Fastest inference**: 56.37 ms enables real-time applications
- **Strong compression**: 3.71x reduction fits on resource-constrained devices
- **Minimal quality loss**: 94.58% accuracy sufficient for most use cases

Custom INT8 implementation suits scenarios prioritizing maximum compression when inference speed is less critical. The 4x size reduction makes it ideal for edge deployment or serving many models simultaneously.

BitsAndBytes INT8 appears suboptimal given its larger size and slower inference compared to alternatives. However, it supports training quantized models (QLoRA), which could prove valuable for continued fine-tuning.