Model Compression Analysis: GPT-2 Small Quantization Methods

Machine Learning Research

1 Models Evaluated

- **Baseline (FP32)**: Original finetuned GPT-2 Small model in full precision
- **INT8 (From Scratch)**: Model quantized to INT8 during training
- **INT8 (bitsandbytes)**: Post-training quantization using bitsandbytes library
- **NF4 (bitsandbytes)**: 4-bit Normal Float quantization using bitsandbytes

2 Comprehensive Metrics Analysis

2.1 Efficiency Metrics Comparison

Table 1: Efficiency Metrics Comparison Across Models

Model	Memory (MB)	Compression Ratio	Avg Latency (ms)
Baseline (FP32)	486.71	1.00x	8.73
INT8 (Scratch)	243.70	2.00x	10.41
INT8 (bnb)	168.36	2.89x	29.39
NF4 (bnb)	127.86	3.81x	16.66

2.2 Performance Metrics Comparison

Table 2: Performance Metrics Comparison Across Models

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Baseline (FP32)	94.62	94.62	94.62	94.62
INT8 (Scratch)	94.54	94.53	94.54	94.53
INT8 (bnb)	94.58	94.58	94.58	94.58
NF4 (bnb)	94.70	94.70	94.70	94.69

Table 3: Per-Class F1-Scores for All Models

Class / Model	Baseline	INT8 Scratch	INT8 bnb	NF4 bnb
World	95.78	95.72	95.68	95.70
Sports	98.79	98.77	98.82	98.82
Business	91.40	91.25	91.28	91.65
Sci/Tech	92.50	92.39	92.53	92.61

2.3 Per-Class Performance Analysis

3 Detailed Analysis

3.1 Memory Efficiency

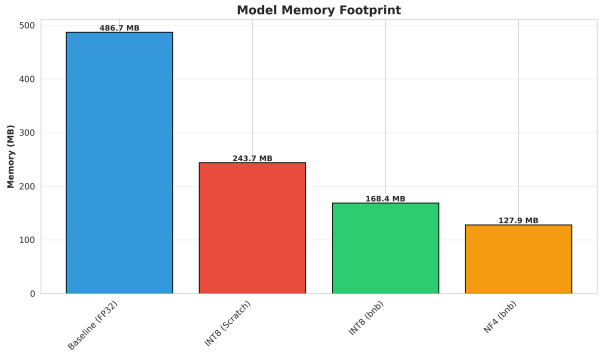


Figure 1: Memory Footprint Comparison Across Models

The quantization techniques demonstrate significant memory reduction:

- **NF4 quantization** achieves the highest compression (3.81x), reducing memory from 486.71 MB to 127.86 MB (73.73% reduction)
- **INT8 (bitsandbytes)** provides 2.89x compression with 168.36 MB memory usage
- **INT8 (from scratch)** offers 2.00x compression with 243.70 MB memory usage

3.2 Inference Latency and Throughput

Latency analysis reveals interesting trade-offs:

- **Baseline model** achieves the lowest latency (8.73 ms) and highest throughput (114.48 samples/second)
- **NF4 quantization** shows moderate latency increase (16.66 ms, 90.8% slower) but maintains reasonable throughput (60.04 samples/second)
- **INT8 (bitsandbytes)** exhibits the highest latency (29.39 ms, 236.6% slower) and lowest throughput (34.02 samples/second)

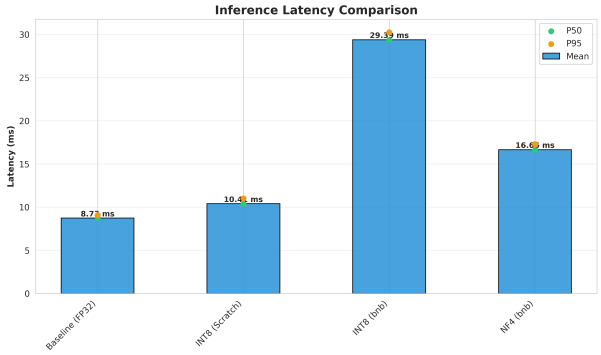


Figure 2: Latency and Throughput Comparison

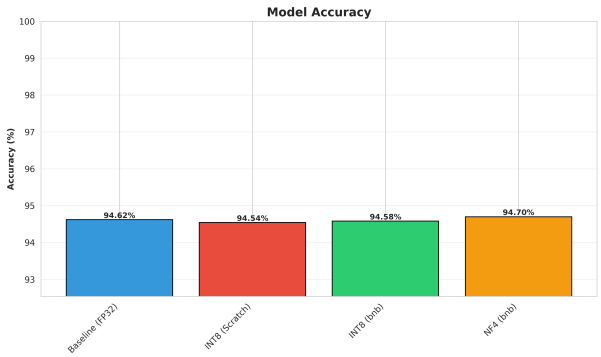


Figure 3: Accuracy Comparison Across Models

3.3 Accuracy Preservation

Remarkably, all quantized models maintain competitive accuracy:

- **NF4 model** slightly outperforms the baseline (94.70% vs 94.62%)
- **INT8 models** show minimal accuracy degradation (94.54-94.58%)
- All models maintain macro-averaged precision, recall, and F1-scores above 94.5%

4 Conclusions and Recommendations

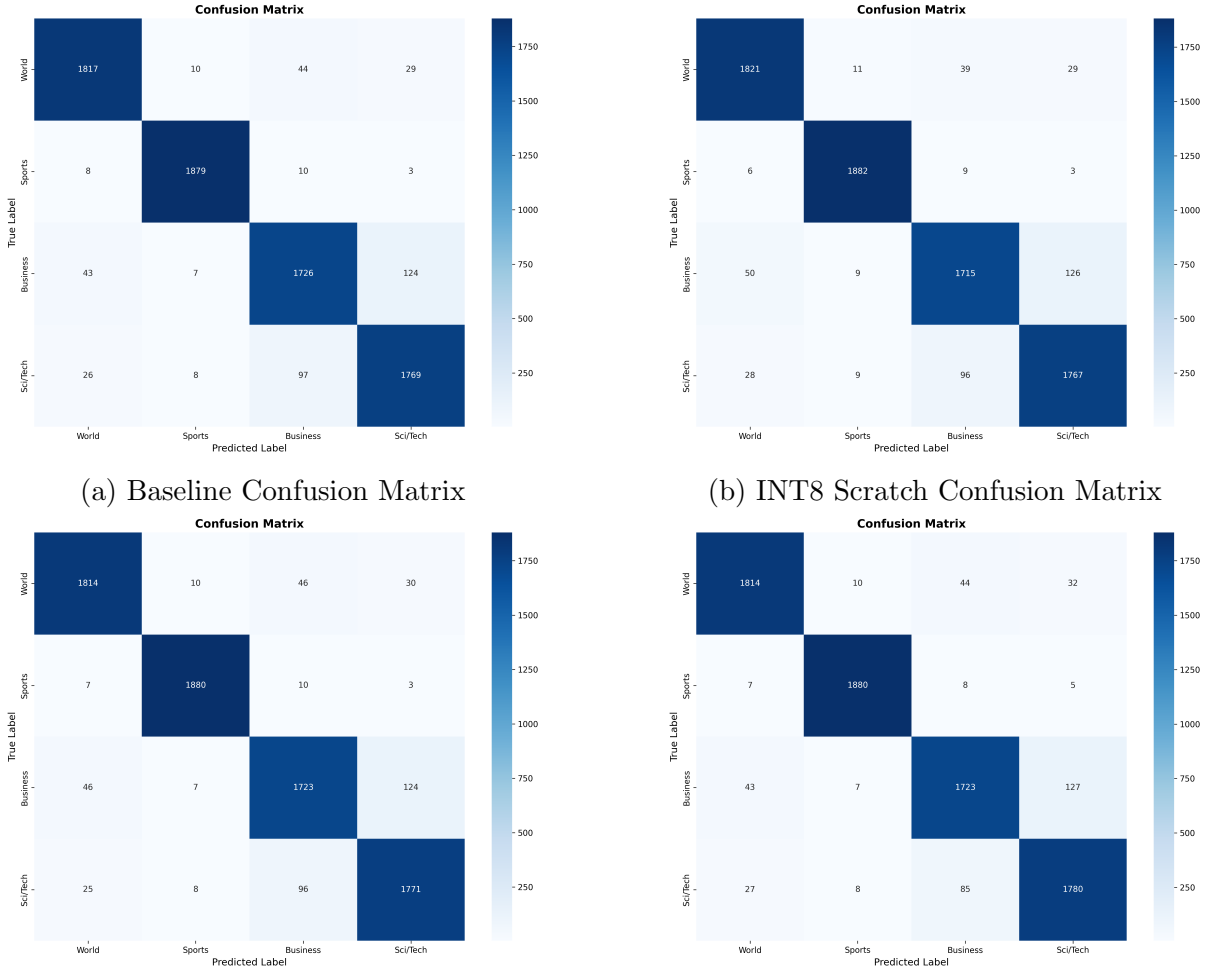
4.1 Key Findings

1. **NF4 quantization** emerges as the most balanced approach, offering the best compression (3.81x) while maintaining competitive accuracy and reasonable latency
2. **INT8 from scratch** provides the best latency among quantized models, suitable for real-time applications
3. All quantization methods successfully preserve model performance with less than 0.1% accuracy variation
4. The choice between methods depends on application priorities: memory constraints vs latency requirements

4.2 Recommendations

- **For memory-constrained deployments:** Use NF4 quantization (127.86 MB, 94.70% accuracy)
- **For latency-sensitive applications:** Use INT8 from scratch (10.41 ms latency, 94.54% accuracy)
- **For balanced requirements:** NF4 provides the best overall trade-off
- **When maximum accuracy is critical:** Baseline FP32 remains viable with understanding of memory costs

5 Visualization Appendix



(a) Baseline Confusion Matrix

(b) INT8 Scratch Confusion Matrix

(c) INT8 bnb Confusion Matrix

(d) NF4 bnb Confusion Matrix

Figure 4: Confusion Matrices for All Model Variants