CS7.501: Advanced Natural Language Processing
Assignment Report

Chetan Mahipal

Course Instructor - Dr. Manish Shrivastava

IIIT Hyderabad - November 16, 2024

1 Quantization from Scratch

This report analyzes the impact of quantization on the GPT-2 model's memory usage, inference speed (latency), and performance (measured by perplexity). The evaluation focuses on three configurations: the original model, 8-bit whole quantized, and 8-bit selectively quantized (decoder block layers).

1.1 Memory Footprint

Quantization significantly reduces the model's memory usage, as shown in Table 1.

Model Type	**Memory Footprint (MB)**	**Reduction (%)**
Original Model	486.70	–
8-bit Whole Quantized	118.68	75.6
8-bit Selective Quantized	271.91	44.1

Table 1: *Memory Footprint Comparison.*

1.2 Latency

Inference latency improves significantly with quantization. Table 2 summarizes the results.

Model Type	**Latency (s)**	**Improvement (%)**
Original Model	0.3263	–
8-bit Whole Quantized	0.1884	42.2
8-bit Selective Quantized	0.1856	43.1

Table 2: *Latency Comparison.*

1.3 Perplexity

While quantization slightly increases perplexity, selective quantization demonstrates a better trade-off. Table 3 shows the results.

Model Type	Perplexity	Change (%)
Original Model	49.62	–
8-bit Whole Quantized	51.61	+4.0
8-bit Selective Quantized	51.08	+2.9

Table 3: *Perplexity Comparison.*

1.4 Key Insights

- **Memory Savings:** Whole quantization achieves the greatest memory reduction, while selective quantization balances memory efficiency and precision.
- **Speed:** Both quantized models significantly reduce latency, with selective quantization performing as efficiently as whole quantization.
- **Performance Trade-off:** Perplexity degradation is minimal, with selective quantization showing better performance than whole quantization.

1.5 Conclusion

Selective quantization of critical layers (decoder block) strikes a balance between memory efficiency, speed, and model performance. This makes it a viable approach for resource-constrained deployment scenarios.

2 BitsAndBytes Quantization

This report presents an analysis of the effects of different quantization techniques on the GPT-2 model's performance. Specifically, we compare Linear Quantization (8-bit and 4-bit) with Nonlinear 4-bit (NF4) Quantization, focusing on memory usage, latency, and perplexity.

2.1 Quantization Results Summary

The results from the experiments are summarized below in terms of memory footprint, latency, and perplexity:

Model Type	Memory Footprint (MB)	Perplexity	Latency (s)
Original GPT (FP32)	486.70	49.62	0.3661
8-bit Quantized	168.35	49.84	0.6903
4-bit Quantized	127.85	58.07	0.2403
4-bit NF Quantized	127.85	53.21	0.2769

Table 4: *Comparison of Memory Footprint, Perplexity, and Latency.*

2.2 Analysis of NF4 Quantization vs Linear Quantization

2.2.1 1. Concept of NF4 Quantization

NF4 (Nonlinear 4-bit Quantization) applies a nonlinear scaling to the weights of the model, prioritizing more precision for small values. This nonlinear approach differs from linear quantization, where all values are scaled equally within the target bit-width. The primary goal of NF4 quantization is to minimize accuracy loss, particularly for weight values close to zero, which can have a

disproportionate impact in language models.

2.2.2 2. Impact on Perplexity

The comparison of perplexity across the models is as follows:

- **8-bit Quantized Model:** There is a slight increase in perplexity from 49.62 to 49.84, suggesting that the 8-bit quantization retains much of the original model's performance with a significant reduction in memory usage.
- **4-bit Quantized Model:** The perplexity increases significantly (from 49.62 to 58.07), indicating that the reduced precision leads to a substantial loss in performance. The 4-bit quantization offers greater memory savings but at the cost of model accuracy.
- **4-bit NF Quantized Model:** The NF4 quantization helps preserve more accuracy compared to the standard 4-bit model, with perplexity rising to 53.21, which is an improvement over the linear 4-bit model (58.07).

2.2.3 3. Impact on Latency

The latency results are as follows:

- **8-bit Quantized Model:** The latency increases from 0.3661s to 0.6903s, suggesting that quantization, while reducing memory usage, introduces a higher computational cost.
- **4-bit Quantized Model:** This model shows a reduction in latency to 0.2403s, significantly faster than both the original and 8-bit quantized models.
- **4-bit NF Quantized Model:** The NF4 quantized model shows a slight improvement in latency compared to the linear 4-bit model, with a latency of 0.2769s.

2.3 Key Insights

- **Memory Savings**: The 8-bit and 4-bit quantization approaches provide significant reductions in memory usage, with the 8-bit model achieving a 65.4% reduction and the 4-bit models achieving a 73.8% reduction.
- **Latency Improvements**: The 4-bit quantized models provide the fastest inference times, with the standard 4-bit model achieving the lowest latency.
- **Perplexity Trade-off**: While 4-bit quantization leads to a notable increase in perplexity, the NF4 quantization reduces this degradation, making it a better option in terms of preserving model accuracy.

2.4 Conclusion

Quantization methods such as 8-bit and 4-bit reduce the memory footprint significantly, but they also lead to performance degradation, particularly in perplexity. NF4 Quantization, with its nonlinear scaling approach, provides a better trade-off between memory savings and model accuracy compared to standard 4-bit quantization. However, there is still a noticeable increase in perplexity, which is a common trade-off when quantizing models to lower bit-widths.