Quantization of Logistic Regression Model

Link:
https://colab.research.google.com/drive/1Z3D2Q5OzEBnJlodP8otJFRzmqfk2e2dx?authuser=1

1. Objective

The objective of this lab is to provide hands-on experience in applying quantization techniques to optimize machine learning models. The focus is on:

- **Understanding Dynamic Quantization**: Exploring how 8-bit dynamic quantization can reduce model size and improve computational efficiency in PyTorch.
- **Optimizing Logistic Regression**: Applying quantization to a logistic regression model and evaluating its performance.
- **Performance Comparison**: Measuring and comparing the following aspects of the model before and after quantization:
  - **Model Accuracy**: Understanding the impact of quantization on predictive performance.
  - **Model Size**: Observing reductions in memory usage after scaling model weights.
  - **Inference Time**: Comparing inference speed to assess potential improvements in real-time applications.
- **Exploring Trade-offs**: Identifying the balance between model efficiency and precision when deploying in resource-constrained environments.

This approach demonstrates how quantization can make machine learning models more suitable for deployment on edge devices and in low-compute settings.

2. Data Preparation

• **Dataset**: MNIST Digits Dataset
• **Number of samples**: 1797
• **Number of classes**: 10 (Digits 0-9)
• **Data Split**: 80% training set, 20% test set.

3. Model Performance Comparison

Metric	Original Model	Quantized Model
Accuracy	97.22	67.22
Model Size	5.98 KB	0.62 KB
Inference Time	0.0012 seconds	0.0013 seconds

3.1 Accuracy

• **Original Model Accuracy**: 97.22 %
• **Quantized Model Accuracy**: 67.22%
The quantized model attained 67% accuracy, lower than the original model however still within acceptable parameters.

3.2 Model Size

• **Original Model Size**: 59,832 bytes
• **Quantized Model Size**: 6232 bytes
Here, Quantization decreased the model size, rendering the quantized model considerably more efficient regarding memory utilization.

3.3 Inference Time

• **Original Model Inference Time**: 0.0012 seconds
• **Quantized Model Inference Time**: 0.013 seconds
Here, Quantization enhanced inference time. This enhancement is essential for real-time applications requiring rapid inference.

4. Conclusion

In this lab, we applied 8-bit dynamic quantization to a logistic regression model and evaluated its impact on model size, inference time, and accuracy. The results demonstrate the significant trade-offs between model efficiency and performance after quantization.

- **Model Size**: The quantized model achieved a notable reduction in size, shrinking from **5.98 KB** to **0.62 KB**—an almost **90% decrease**. This is a clear advantage for deployment in memory-constrained environments.

- **Inference Time**: While quantization typically aims to improve inference speed, in this case, the quantized model's inference time slightly increased from **0.001176 seconds** to **0.001360 seconds**. The small overhead might be attributed to the dynamic quantization process and the model's structure.
- **Model Accuracy**: The most significant trade-off was in accuracy. The original model maintained a strong accuracy of **97.22%**, whereas the quantized model’s accuracy dropped to **67.22%**, indicating a substantial loss in predictive performance. This loss is likely due to the reduced precision from 32-bit to 8-bit weights, which can affect the model’s ability to learn fine-grained distinctions.