Evaluation of IA³ and QLoRA Fine-Tuning Strategies on Llama 3.2 for Sentiment Analysis

Team Members: Ahmad Agah, Ness Blackbird, Marchelle Le, Simrah Saleem

This study evaluates two PEFT strategies, IA^3 and QLoRA, for adapting Llama-3.2 models to three-way sentiment analysis. Llama 3.2 allowed for ease of deployment on small devices while lowering computational cost but its compact size makes it challenging to balance between accuracy and memory efficiency. IA^3fine-tuned models on a massive dataset of 944,423 samples for 3-way sentiment classification and achieved significant performance convergence, correcting the base model's "positivity bias" by adjusting internal activation gates. Next, QloRa was analyzed on a UMSAB English subset of 1,840 samples and a TweetEval subset of 45,615 samples. After attempting to increase precision, it continued to underfit neural sentiment and suffer quantization degradation, suggesting that 4-bit quantization may harm the semantic resolution of sub-3B parameter models. In conclusion, lightweight models, such as vector-based scaling like IA³, offer robust adaptation strategies compared to quantized low-rank adaptation.

Methodology for IA³

Data & Configuration: A Dataset of 944,423 labeled examples was aggregated from TweetEval, Sentiment140, SST, IMDB, and Yelp. We implemented IA³ with PEFT library to target linear layers, allowing the model to adjust attention mechanisms and feedforwarding outputs with minimal parameter overhead; q_proj, k_proj, o_proj, gate_proj, up_proj, and down_proj.

Training Hyperparameters: The model was trained for 3 epochs with a batch size of 98 and gradient accumulation steps of 2, which stabilizes the updates over large batches, and a learning rate of 5e-3. We used the adamw_torch optimizer with a max sequence length of 160 tokens, sufficient to capture both tweets and the core sentiment of longer reviews.

Quantitative Performance: IA³ efficiently achieved validation accuracy of 78%, completing in ~3 hours on a single GPU using bfloat16. A separate validation run on binary classification yielded higher accuracy, isolating the "neutral" class as the primary challenge in the 3-way task.

Qualitative Shift: pre- vs post-training confusion matrices' behavioral correction.

- **Zero-shot Baseline**: The pre-trained model exhibited extreme "positive bias", classifying nearly all inputs as positive. It lacked the specific decision boundaries for neutral content.
- **Post-IA³ Correction**: After fine-tuning, the confusion matrix was diagonalized, aligning predictions with true labels. By adjusting the "gates" (vectors) across all projection layers, IA³ successfully inhibited the generic positive drift and amplified the specific structural definitions of "neutral" and "negative" present in the 944k training examples.

Methodology for QLoRA

QloRA was implemented to reduce memory usage by quantizing the base model while training additive low-rank adapters. It loads Llama-3.2 in 4-bit NF4 precision with bfloat16, stabilized by prepare_model_for_kbit_training to preserve full precision and enabling gradient checkpointing. Training was run on T4 GPU after repeated failures on the A100 for large datasets and Ti 4070-Super (16GB RAM). UMSAB dataset failed with a batch size of 1, produced unstable ~3.0 loss. In contrast, increasing the batch size to 4 significantly stabilized learning and dropped below ~1.0 loss, allowing QloRA to behave as expected. Next, the larger dataset of 45,615 samples tested with causal learning improved from 63% to 71% accuracy. While the sequential model reached 74% accuracy after 2 epochs, the small dataset reached 72%. From adjusting different parameters and the number of epochs, it consistently struggled with neutral sentiment, where many tweets are ambiguous, political, and rhetorical. Confusion matrices showed strong learning of positive and negative sentiment. For example, emotional tweets were grouped correctly, while informational tweets were mislabeled as negative. In conclusion, QloRA converged efficiently but combining quantization and small model size hindered its performance compared to IA^3 on nuanced sentiment boundaries.

Analysis of Performance Degradation. We attribute this performance drop to three factors:

- **Learning Rate vs. Dataset Size**: Combining a lower learning rate with fewer epochs resulted in underfitting. QLoRA did not escape the local minimum of the pre-trained "positivity bias". IA^3's aggressive learning rate allowed for rapid, decisive updates to the "gates", whereas QLoRA's conservative updates were insufficient for this volume of data in the allotted time.
- **Quantization Sensitivity:** QLoRA with 4-bit quantization can be destructive for lightweight models. Compressing the already compact weights into a 4-bit precision likely eroded the semantic resolution required to distinguish subtle sentiment nuances (like sarcasm in tweets), which the full precision vectors of IA^3 preserved.
- **Complexity Mismatch**: IA^3 updates a very small set of parameters, making it easier to optimize with a high learning rate without instability. QLoRA must optimize dense low-rank matrices on top of quantization, making it harder to optimize complex structures with 4-bit base weights. QLoRA would struggle to overcome pretrained positivity bias and convergence speed compared to the lightweight IA^3 approach.

Ethical Considerations when adapting LLMs for sentiment analysis:

The base Llama 3.2 model suffered from a significant "positivity bias" due to its generic pre-training. In high-stakes applications, a model that defaults to "positive" or "neutral" classification risks of suppressing critical concerns. IA^3 corrects this by creating clearer decision boundaries showing to be a safer option for accurate sentiment detection. Next, since the datasets came from social media and review platforms, skewing towards specific demographics and specific communication styles. As a result, IA^3 and QloRa underperformed on neutral sentiment, linguistic groups or cultural dialects. Fairness checks must be included to ensure equitable performance. Lastly, IA^3 updated less than 0.1% of the parameters and required only a few hours on a single GPU. QLoRA uses 4-bit quantization. Both showed a decrease in energy use and carbon footprint, aligning with Green AI, reducing the energy cost of adapting AI to new tasks.

Conclusion

This project demonstrated the difference between IA^3 and QLoRa's behoviors when adapting with Llama 3.2 to sentiment analysis, though we could not compare them directly as they were trained on different datasets, hardware setups, and parameters. IA^3 fine tuned was fast, efficient, extremely light weight, and benefited from larger batch sizes. On the other hand, QLoRa took more effort to stabilize as it is sensitive to batch size, quantization settings and GPU environment. But once we stablized it by improving configuration and increasing batch size, we ultimately achieved a higher accuracy score on the 3 way classification when trained with a larger dataset. A main takeaway from this experiments is that QLoRA does not train reliably with a batch size of 1 and demands more careful hyperparameter tuning than IA^3. Both methods showed small LLMs can be effectively adapted through PEFT and their performance depends largely on data scale, precision of model and implementation choices. Future work in evaluating IA^3 vs. QLoRA under matched conditions are needed to enable a fair comparison.