Квантизация Aria

Andrew

1 Введение

В работе представлен сравнительный анализ инструментов Quanto и BitsAndBytes для квантизации архитектуры Mixture-of-Experts (MoE) с компонентом Vision Transformer (ViT). Особое внимание уделяется модификации модели Aria, методам пост-тренировочной квантизации (PTQ) и их применению к различным частям больших языковых моделей (LLM).

2 Модификация модели Aria

Квантованная версия модели Aria пока отсутствует, и научных работ на эту тему также нет. Однако разработчики предоставили модифицированную версию модели на основе PyTorch, которая облегчает процесс квантизации.

Модифицированная версия модели Aria представляет собой форк оригинальной модели `rhymes-ai/Aria`. Основное изменение заключается в замене группированных операций GEMM (General Matrix Multiply) на последовательные многоуровневые персептроны (MLP). В этой конфигурации каждый эксперт реализован как слой `torch.nn.Linear`, выполняемый последовательно.

Такая перестройка архитектуры имеет два преимущества:

- **Упрощенная квантизация**: Последовательные слои `nn.Linear` лучше поддерживаются текущими открытыми библиотеками для квантизации, такими как BitsAndBytes и Quanto. Эти библиотеки оптимизированы для работы с линейными слоями, что позволяет применять эффективные методы квантизации без значительных изменений в структуре модели.
- **Снижение сложности**: Удаление группированных GEMM операций упрощает вычислительный граф модели, что может способствовать уменьшению задержек и повышению производительности при выполнении модели на GPU.

3 Методы квантизации

Модель объемом 50 ГБ можно эффективно квантовать с помощью инструментов BitsAndBytes и Quanto. Для этого потребуется GPU типа A100, около 60-65 ГБ оперативной памяти и не более 1.5 часов на загрузку, квантизацию и сохранение модели.

BitsAndBytes и Quanto относятся к пост-тренировочной квантизации (PTQ). Существуют также методы Quantization-aware Training (QAT), которые учитывают квантизацию во время обучения модели для улучшения её устойчивости к снижению битности. Для QAT одной GPU не хватит, так как QAT предполагает обучение.

4 Квантизация различных частей LLM

Модель Aria состоит из компонента `google/siglip-so400m-patch14-384`, дообученного вместе с LLM, имеющей 28 слоев. В каждом слое по 64 эксперта и два общих эксперта. Весовые коэффициенты данного компонента имеют размер 800M параметров, поэтому их можно оставить в первозданном виде.

Сосредоточимся на квантизации LLM. Различные части могут по-разному реагировать на снижение битности.

4.1 Весовые коэффициенты FFN экспертов

Весовые коэффициенты feed-forward network (FFN) экспертов являются основными кандидатами для квантизации. Они устойчивы к низкоразрядной квантизации благодаря меньшему количеству выбросов по сравнению с плотными слоями FFN. Согласно исследованию, квантизация этих весов с помощью PTQ до 4 бит может повысить производительность модели и ускорить её в 1.24 раза.

> "Expert FFN layers are more robust to low-bit quantization due to fewer outliers compared to dense FFN layers. PTQ to 4 bits can make the model 1.24 times faster while maintaining or even improving performance."

4.2 Весовые коэффициенты слоев внимания

Квантизация весов слоев внимания возможна, но требует более высокой битности для сохранения производительности. Пост-тренировочная квантизация с 3 битами допустима, однако 2-битная квантизация негативно влияет на результаты.

> "Quantizing attention layer weights is feasible with 3-bit PTQ, but 2-bit quantization significantly degrades performance."

4.3 Весовые коэффициенты разделённых и общих экспертов

В архитектурах MoE с разделёнными и общими экспертами рекомендуется приоритизировать квантизацию разделённых весов экспертов. Общие эксперты активируются для каждого входного токена, поэтому им требуется более высокая точность.

> "Non-shared expert weights should be prioritized for quantization over shared experts, as shared experts are activated for every input token and require higher precision."

4.4 Первые блоки MoE

Первые блоки MoE в модели критически важны для производительности, поэтому рекомендуется использовать более высокую битность для их квантизации. Исследования показывают, что выделение большего количества битов первым блокам приводит к лучшим результатам по сравнению с последними блоками.

> "Allocating higher bit widths to the first few MoE blocks yields better performance compared to the last blocks, as the initial layers are more critical for overall model performance."

4.5 Плотные слои FFN

Квантизация плотных слоев FFN наиболее негативно сказывается на производительности модели. Пост-тренировочная квантизация с 3 битами может значительно снизить точность, а 2-битная квантизация делает модель практически неэффективной.

> "Quantizing dense FFN layers adversely affects model performance. PTQ with 3 bits leads to a significant accuracy drop, and 2-bit quantization renders the model nearly ineffective."

5 Частота использования экспертов как эвристика

Частота использования экспертов рекомендуется использовать как эвристику для направления процесса квантизации. Менее часто используемые эксперты могут быть квантованы до меньшей битности без значительного влияния на производительность модели. Например, квантование весов топ-16 наиболее часто используемых экспертов на каждый блок MoE до 4 бит с использованием PTQ, а остальных весов до 2 бит обеспечивает приемлемую производительность. Разница в производительности при различных битностях активаций является незначительной.

6 Потенциальные решения

Для преодоления перечисленных проблем предлагаются различные техники и направления исследований:

- **Пост-тренировочная квантизация**: Специализированные методы PTQ, учитывающие разреженность и структуру экспертов в MoE моделях, помогают снизить память и повысить скорость инференса. Такие методы включают распределение битов на основе частоты использования экспертов и приоритизацию квантизации слоев внимания.
- **Квантизация с учётом обучения**: Методы Quantization-aware Training (QAT) могут улучшить производительность квантизированных MoE моделей, особенно при низкой битности. Например, модель MoE с 2-битными весами экспертов, обученная с использованием QAT, может достичь лучшей производительности, чем плотная модель, обученная на тех же данных.
- **Оптимизация времени выполнения**: Оптимизированные реализации времени выполнения для квантизированных MoE моделей необходимы для достижения значительных ускорений. Например, оптимизация времени выполнения для 4-битной квантизированной MoE модели может привести к ускорению на 1.24 раза по сравнению с моделью FP16.

7 Дополнительные нюансы квантизации и компонентно-специфические соображения

Различные части MoE моделей требуют особого подхода к квантизации. Ниже представлены основные принципы и рекомендации:

- **Слои внимания более эффективно квантизируются**: Квантизация слоев внимания до более высоких битностей сохраняет производительность лучше, чем распределение этих битов на FFN слои внутри блоков MoE.
- **Первые блоки MoE более важны**: Выделение большего количества битов первым блокам MoE улучшает общую производительность модели.
- **Разделённые эксперты требуют более высокой точности**: Общие эксперты, активируемые для каждого входного токена, требуют более высокой битности для сохранения точности.
- **Линейные слои с большими выбросами**: Линейные слои с большим диапазоном весов, включая выбросы, сложнее квантизировать эффективно. Рекомендуется использовать оценщики выбросов для выделения таких слоев и выделения им большего количества битов.
- **Квантизация весов и активаций**: Выводы из экспериментов по квантизации весов применимы и к сценариям, где квантизируются как веса, так и активации.
- **Квантизация роутера**: Маршрутизаторы играют критическую роль в распределении задач между экспертами, поэтому их работа должна быть максимально точной. По этой причине лучше избегать квантизации маршрутизаторов, сохраняя их в исходном, более точном виде.
- **Квантизация слоя эмбеддингов**: Эмбеддинги напрямую влияют на представление данных, передаваемых модели. Слои эмбеддингов рекомендуется тоже не квантизовать чтобы сохранить максимальную точность и качество обработки данных.

8 Заключение

Квантизация архитектуры Mixture-of-Experts (MoE) с компонентом Vision Transformer (ViT) возможна, но нужно качественно распланировать конфигурацию для неё. Важно учитывать частоту использования экспертов и особенности отдельных слоёв модели, чтобы не снизить черезмерно возможности модели