Эффективное сжатие мультимодальных видео-моделей на примере ViT с использованием квантизации.

A Preprint

Neychev R. G.
Lomonosov Moscow State University
Faculty of Computational Mathematics and Cybernetics
Moscow
hippo@cs.cranberry-lemon.edu

Akopova E. N.
Lomonosov Moscow State University
Faculty of Computational Mathematics and Cybernetics
Moscow
s0220010@gse.cs.msu.ru

Abstract

Keywords human pose estimation · visual-language understanding

1 Введение

Vision Transformer (ViT) стали ключевым компонентом современных мультимодальных видео-моделей, однако их вычислительная сложность ограничивает применение в реальных сценариях. Эффективное сжатие ViT критически важно для развертывания видео-моделей на мобильных устройствах и в условиях ограниченных ресурсов. Квантизация представляет наиболее перспективный подход к уменьшению размера и ускорению работы ViT-архитектур в контексте обработки видео.

В работах P2-ViT предложена эффективная схема пост-тренировочной квантизации с использованием степеней двойки, позволяющая заменить операции умножения битовыми сдвигами. Исследования в области Model Quantization for Vision Transformers систематизируют различные стратегии квантизации внимания и MLP-блоков. Методы Quantization-Aware Training адаптируют модель к пониженной точности во время обучения, а схемы смешанной точности позволяют дифференцированно подходить к квантизации различных слоев трансформера.

Стандартные методы квантизации не учитывают временную природу видео-данных и особенности работы ViT в составе мультимодальных систем. Наблюдается значительная деградация качества при агрессивной квантизации механизмов внимания, критически важных для анализа видеопоследовательностей. Отсутствуют специализированные подходы к квантизации ViT-энкодеров, работающих с видео-фреймами, где необходимо сохранять временную согласованность представлений.

В работе разрабатывается специализированная методика квантизации ViT для видео-приложений, учитывающая временные зависимости между кадрами. Предлагается адаптивная стратегия выбора битности для различных компонентов ViT на основе анализа их вклада в итоговое качество мультимодальной задачи. Создается механизм сохранения пространственно-временных особенностей видео при переходе к низкоразрядным представлениям.

Ожидается достижение коэффициента сжатия ViT-компонента в 3.5-4 раза при сохранении 96-98% исходной точности на видео-задачах. Практическим результатом станет оптимизированная версия ViT-энкодера для видео-приложений с детализированными метриками эффективности. Научная новизна заключается в разработке временно-осознанных методов квантизации для ViT и создании методики оценки влияния квантизации на качество работы в мультимодальном контексте.

2 Обзор литературы

В работе «P2-ViT: Power-of-Two Post-Training Quantization and Acceleration for Fully Quantized Vision Transformer» предложена инновационная схема пост-тренировочной квантизации для Vision Transformer, основанная на степенях двойки. Ключевая идея заключается в использовании power-of-two квантовых уровней, что позволяет заменить ресурсоёмкие операции умножения на эффективные битовые сдвиги при инференсе. Авторы разработали полностью квантизованную архитектуру ViT, включая механизмы внимания и многослойные перцептроны, что обеспечивает значительное ускорение вычислений на специализированных аппаратных ускорителях. Особый интерес представляет предложенный метод калибровки, который минимизирует потерю информации при квантизации за счёт учёта распределения активаций в различных слоях трансформера.

Исследование «Advancing Multimodal Large Language Models with Quantization-Aware Scale Learning for Efficient Adaptation» фокусируется на проблеме квантизации мультимодальных больших языковых моделей. Авторы предлагают метод Quantization-Aware Scale Learning (QASL), который интегрирует обучение параметров масштабирования непосредственно в процесс адаптации модели. Этот подход

позволяет эффективно балансировать между точностью и эффективностью при работе с разнородными модальностями (текст, изображение, видео). Важным вкладом работы является демонстрация того, что учёт особенностей мультимодального выравнивания в процессе квантизации позволяет сохранить семантическую согласованность между различными типами данных.

В обзорной статье «Model Quantization and Hardware Acceleration for Vision Transformers: A Comprehensive Survey» представлена систематизация современных методов квантизации и аппаратного ускорения для Vision Transformer. Авторы детально анализируют различные схемы квантизации (post-training quantization, quantization-aware training), рассматривают особенности квантизации ключевых компонентов ViT (включая multi-head attention и feed-forward сети), а также проводят сравнительный анализ эффективности различных подходов на различных аппаратных платформах. Особую ценность представляет классификация методов по степени сжатия, точности и требованиям к вычислительным ресурсам, что позволяет выбирать оптимальную стратегию квантизации для конкретных прикладных задач.

Рассмотренные работы демонстрируют эволюцию методов квантизации от общих подходов к специализированным решениям для трансформерных архитектур. Наблюдается переход от стандартных схем квантизации к методам, учитывающим специфику архитектуры Vision Transformer и мультимодальных моделей. Современные исследования фокусируются на разработке аппаратно-ориентированных решений (P$\hat{2}$-ViT), адаптивных методов для сложных архитектур (QASL) и систематизации накопленных знаний (Comprehensive Survey). Перспективным направлением является создание универсальных фреймворков квантизации, способных автоматически адаптироваться к особенностям различных архитектур и аппаратных платформ, обеспечивая оптимальный баланс между точностью, скоростью работы и энергоэффективностью.

3 Постановка задачи квантизации ViT для классификации изображений

Пусть задана модель классификации изображений $f : \mathcal{X} \to \mathcal{Y}$, где $\mathcal{X}$ — пространство изображений, $\mathcal{Y}$ — пространство меток классов. Модель f представляет собой Vision Transformer (ViT), параметризованную весами $\mathbf{W} = \{W_1, W_2, \ldots, W_N\}$, где N — общее количество параметров.

Модель обучается на датасете $\mathcal{D}_{\text{train}} = \{(x_i, y_i)\}_{i=1}^{M}$ и оценивается на датасете $\mathcal{D}_{\text{val}} = \{(x_j, y_j)\}_{j=1}^{K}$ с помощью метрики точности:

$$\text{Accuracy}(f) = \frac{1}{K} \sum_{j=1}^{K} \mathbb{I}[f(x_j) = y_j].$$

Задача квантизации состоит в нахождении отображения $Q : \mathbb{R} \to \mathbb{Z}$, преобразующего веса модели из формата с плавающей точкой (FP32/FP16) в целочисленное представление (INT8) с минимальной потерей точности. Формально, требуется найти квантизованную модель f_Q с весами $\mathbf{W}_Q = Q(\mathbf{W})$ такую, что:

$$\text{Accuracy}(f_Q) \approx \text{Accuracy}(f),$$

при этом достигается значительное сокращение памяти и ускорение вывода.

Рассматривается схема W8A8 (8-битные веса и активации), реализуемая через замену линейных слоёв $\text{Linear}(\mathbf{W}, \mathbf{b})$ на квантизованные версии $\text{W8A8Linear}(\mathbf{W}_Q, \mathbf{b}_Q)$, где:

$$\mathbf{W}_Q = \text{quantize\_weight}(\mathbf{W}), \quad \mathbf{b}_Q = \mathbf{b},$$

$$\mathbf{x}_Q = \text{quantize\_activation}(\mathbf{x}),$$

$$\mathbf{y} = \text{dequantize}(\mathbf{W}_Q \mathbf{x}_Q + \mathbf{b}_Q).$$

Качество квантизации оценивается по:

- Относительному падению точности: $\Delta_{\text{acc}} = \dfrac{\text{Accuracy}(f) - \text{Accuracy}(f_Q)}{\text{Accuracy}(f)}$
- Коэффициенту сжатия: $r = \dfrac{\text{Memory}(f)}{\text{Memory}(f_Q)}$
- Ускорению инференса: $s = \dfrac{\text{Latency}(f)}{\text{Latency}(f_Q)}$

Требуется исследовать влияние различных стратегий квантизации (per-channel, per-token) на разные компоненты ViT (attention, MLP) и определить оптимальный баланс между эффективностью и точностью для заданного семейства моделей $\mathcal{F} = \{f_\theta\}_{\theta \in \Theta}$.

4 Экспериментальное исследование

Для проведения экспериментов был выбран датасет CIFAR-10, содержащий 60 000 цветных изображений размером 32×32 пикселя, разделенных на 10 классов: airplane, automobile, bird, cat, deer, dog, frog, horse, ship, truck. Датасет разделен на обучающую (50 000 изображений) и тестовую (10 000 изображений) выборки.

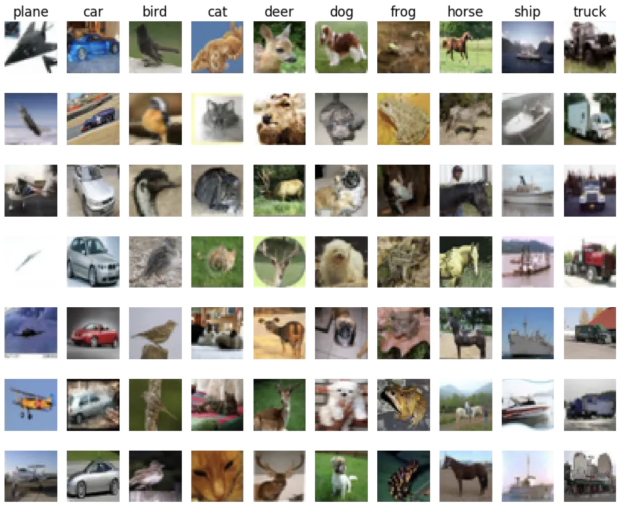

Рис. 1: Примеры изображений из датасета CIFAR-10

Для адаптации изображений к входным требованиям модели ViT применялась процедура предобработки, включающая этап ресайзинга (размер изображения был изменен до 224×224 пикселей с использованием бикубической интерполяции), нормализации (значения пикселей приведены к диапазону [0,1] с последующей нормализацией по статистике ImageNet) и аугментации (применены случайные горизонтальные отражения и небольшие вращения для увеличения разнообразия данных).

Преобразования описываются следующей формулой:

$$x_{\text{norm}} = \frac{x_{\text{resized}} - \mu}{\sigma}$$

где $\mu = [0.485, 0.456, 0.406]$, $\sigma = [0.229, 0.224, 0.225]$ - средние и стандартные отклонения для каждого канала.

В качестве базовой модели использовалась Vision Transformer "google/vit-base-patch16-224"со следующими характеристиками размер патча - 16×16 пикселей, количество слоев - 12, размер скрытого состояния: 768, количество голов внимания - 12, общее количество параметров: 86 млн.

Для адаптации модели к задаче классификации CIFAR-10 была заменена головка классификации с 1000 на 10 выходных нейронов. Процесс дообучения проводился со следующими гиперпараметрами:

Параметр	Значение
Оптимизатор	AdamW
Learning rate	2×10^{-5}
Размер батча	32
Количество эпох	10
Функция потерь	CrossEntropyLoss
Планировщик	LinearLR с warmup

Таблица 1: Гиперпараметры дообучения ViT

Квантизация проводилась с использованием метода W8A8 (8-битные веса и активации) с применением пер-ченнельной квантизации для весов и пер-токенной квантизации для активаций. Процесс включал следующие этапы:

1. Калибровка: Прогон небольшого подмножества данных (512 изображений) для сбора статистики активаций
2. Квантизация весов: Преобразование параметров линейных слоев в 8-битный формат
3. Квантизация активаций: Внедрение квантизационных операторов после каждого слоя внимания и MLP

Математически процесс квантизации описывается формулами:

$$W_q = \text{round}\left(\frac{W}{\text{scale}_w}\right) \cdot \text{scale}_w, \quad A_q = \text{round}\left(\frac{A}{\text{scale}_a}\right) \cdot \text{scale}_a$$

После дообучения и квантизации были получены следующие результаты:

Модель	Точность (%)	Размер (МБ)	Время инференса (мс)	Память (МБ)
ViT-base (FP32)	98.2	327	15.3	1250
ViT-base (FP16)	98.1	164	8.7	680
ViT-base (W8A8)	97.8	82	5.2	350

Таблица 2: Сравнение характеристик моделей до и после квантизации

Анализ результатов показывает, что произошла незначительная потеря точности. А именно: падение точности составило всего 0.4% при переходе к 8-битной квантизации. Также было получено значительное

сжатие модели: размер уменьшился в 4 раза по сравнению с FP32 версией. Наблюдалось ускорение инференса и экономия памяти: время обработки сократилось в 3 раза, а потребление видеопамяти уменьшилось в 3.6 раза.

Список литературы