heightadjust=object

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 систем компьютерного зрения, 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 работах (1) предложена эффективная схема пост-тренировочной квантизации с использованием степеней двойки, позволяющая заменить операции умножения битовыми сдвигами. Исследования в области (2) систематизируют различные стратегии квантизации внимания и MLP-блоков. Методы Quantization-Aware Training адаптируют модель к пониженной точности во время обучения (3), а схемы смешанной точности позволяют дифференцированно подходить к квантизации различных слоев трансформера (4).

Однако стандартные методы квантизации демонстрируют значительную деградацию качества при агрессивной квантизации механизмов внимания, критически важных для работы ViT. Существующие подходы не учитывают архитектурные особенности трансформеров, такие как специфика работы механизма самовнимания, структура эмбеддингов патчей и нормализация слоев. Особенно остро эта проблема проявляется при квантизации моделей, работающих с последовательностями данных, где необходимо сохранять временную согласованность представлений.

В данной работе разработана специализированная методика адаптивной квантизации VQ-ViT, которая учитывает архитектурные особенности Vision Transformer. А именно предлагается: ввести метрику чувствительности слоев на основе спектральных свойств и градиентного анализа, что позволит дифференцированно назначать точность квантизации для различных компонентов ViT.Также были учтены архитетурные особенности модели: разработаны специалированные схемы квантизации для механизма внимания, эмбеддингов патчей и нормализационных слоев. Алгоритм квантизации основан на прогрессивной стратегии с поэтапным внедрением и тонкой настройкой для минимизации потери точности. Квантизация была реализована специализированными ядрами на основе Triton с компенсацией ошибок и адаптивным выбором масштабов.

Экспериментальные результаты демонстрируют достижение коэффициента сжатия ViT-моделей в 10-10.5 раза при сохранении 96-98% исходной точности на задачах классификации изображений. Научная новизна заключается в разработке адаптивных методов квантизации, учитывающих архитектурные особенности ViT, и создании комплексной методики оценки влияния квантизации на различные компоненты трансформерной архитектуры.

2 Обзор литературы

В работе (1) 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.

Исследование (2) представляет систематизацию современных методов квантизации и аппаратного ускорения для Vision Transformer. Авторы детально анализируют различные схемы квантизации,

рассматривают особенности квантизации ключевых компонентов ViT, 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.

В работе (3) исследуются подходы к квантизационно-осознанному обучению трансформерных архитектур. Авторы демонстрируют преимущества интеграции этапа квантизации в процесс обучения, что позволяет модели адаптироваться к пониженной точности представления параметров. Особое внимание уделено методам обратного распространения ошибки через операции квантизации.

Исследование (4) фокусируется на проблеме смешанно-точностной квантизации трансформерных архитектур. Авторы предлагают алгоритм автоматического назначения битности на основе анализа гистограмм активаций и градиентов. Особенностью метода является учёт взаимного влияния различных слоёв при каскадной квантизации.

В работе (5) предлагается комплексный подход к пост-тренировочной квантизации ViT, учитывающий архитектурные особенности трансформеров. Авторы разработали специализированные схемы квантизации для различных компонентов ViT, включая механизм самовнимания, нормализационные слои и эмбеддинги.

Исследование (6) фокусируется на проблеме квантизации механизма самовнимания. Авторы выявили, что стандартные методы квантизации приводят к значительной деградации качества при работе с матрицами запросов, ключей и значений. Предложенное решение включает специализированные схемы квантизации для различных компонентов механизма внимания.

Рассмотренные работы демонстрируют эволюцию методов квантизации от общих подходов к специализированным решениям для трансформерных архитектур. Однако остаются нерешёнными проблемы эффективной квантизации механизма внимания, сохранения пространственных зависимостей в эмбеддингах патчей и разработки универсальных метрик чувствительности слоёв, что определяет научную новизну предлагаемого подхода.

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 Предлагаемый метод адаптивной квантизации ViT

4.1 Анализ архитектурных особенностей

Vision Transformer демонстрирует уникальные свойства, отличающие его от традиционных сверточных сетей. Для входного изображения $x \in \mathbb{R}^{H \times W \times C}$ ViT сначала разбивает его на N патчей $x_p \in \mathbb{R}^{N \times (P^2 \cdot C)}$, где P — размер патча. Процесс эмбеддинга патчей и позиционного кодирования описывается как:

$$\mathbf{z}_0 = [\mathbf{x}_{\text{class}}; \mathbf{x}_p^1 \mathbf{E}; \mathbf{x}_p^2 \mathbf{E}; \cdots; \mathbf{x}_p^N \mathbf{E}] + \mathbf{E}_{\text{pos}}$$

где $\mathbf{E} \in \mathbb{R}^{(P^2 \cdot C) \times D}$ — проекция эмбеддинга патчей, $\mathbf{E}_{\text{pos}} \in \mathbb{R}^{(N+1) \times D}$ — позиционное кодирование, $\mathbf{x}_{\text{class}}$ — токен классификации.

Трансформерные слои состоят из блоков многоголового самовнимания (MSA) и многослойного персептрона (MLP):

$$\mathbf{z}'_l = \text{MSA}(\text{LN}(\mathbf{z}_{l-1})) + \mathbf{z}_{l-1}$$
$$\mathbf{z}_l = \text{MLP}(\text{LN}(\mathbf{z}'_l)) + \mathbf{z}'_l$$

4.2 Адаптивная квантизация с учетом внимания

Предлагается метрика чувствительности $\mathcal{S}(\mathbf{W})$ для каждой матрицы весов $\mathbf{W}$, основанная на спектральных свойствах и величине градиента:

$$\mathcal{S}(\mathbf{W}) = \underbrace{\frac{\|\mathbf{W}\|_2}{\|\mathbf{W}\|_F}}_{\text{Число обусловленности}} \cdot \underbrace{\mathbb{E}\left[\left\|\frac{\partial \mathcal{L}}{\partial \mathbf{W}}\right\|_F\right]}_{\text{Величина градиента}}$$

где $\|\cdot\|_2$ — спектральная норма, $\|\cdot\|_F$ — норма Фробениуса. Точность квантизации назначается как:

$$\text{Precision}(\mathbf{W}) = \begin{cases} \text{8-бит} & \text{если } \mathcal{S}(\mathbf{W}) \leq \tau_{\text{low}} \\ \text{8-бит (асимметричная)} & \text{если } \tau_{\text{low}} < \mathcal{S}(\mathbf{W}) \leq \tau_{\text{high}} \\ \text{16-бит} & \text{иначе} \end{cases}$$

Для слоев внимания применяются различные стратегии квантизации для проекций Query, Key, Value и Output. Квантизованное вычисление внимания принимает вид:

$$\mathbf{Q}_q = \mathcal{Q}_q(\mathbf{X}\mathbf{W}_Q), \quad \mathbf{K}_q = \mathcal{Q}_k(\mathbf{X}\mathbf{W}_K)$$
$$\mathbf{V}_q = \mathcal{Q}_v(\mathbf{X}\mathbf{W}_V), \quad \mathbf{O}_q = \mathcal{Q}_o(\mathbf{A}\mathbf{V}_q)$$

где $\mathcal{Q}_*$ представляет слое-специфичные функции квантизации с обучаемыми масштабными коэффициентами α_*:

$$\mathcal{Q}_*(x) = \text{clamp}\left(\left\lfloor \frac{x}{\alpha_*} \right\rceil, -127, 127\right)$$

4.3 Квантизация эмбеддингов патчей с учетом их структуры

Слой эмбеддинга патчей обладает сверточными характеристиками и требует специального подхода. Предлагается квантизация, сохраняющая структуру патчей:

Пусть $\mathbf{X}_{\text{patches}} \in \mathbb{R}^{B\times N\times D}$ — эмбеддинги патчей. Применяется поканальная квантизация с сохранением пространственных отношений:

$$\mathbf{X}_q^{b,n,d} = \mathcal{Q}_{\text{patch}}\left(\mathbf{X}^{b,n,d}; \alpha^n_{\text{spatial}}, \alpha^d_{\text{channel}}\right)$$

где масштабные коэффициенты декомпозируются на пространственные и канальные компоненты:

$$\alpha^{n,d}_{\text{total}} = \alpha^n_{\text{spatial}} \cdot \alpha^d_{\text{channel}}$$

4.4 Прогрессивная стратегия квантизации

Вводится прогрессивное расписание квантизации, минимизирующее потерю точности. Пусть $\mathcal{L}(\theta)$ — функция потерь, $\theta_q^{(t)}$ — квантизованные параметры на шаге t. Прогрессия квантизации следует:

$$\theta_q^{(t)} = \theta_q^{(t-1)} + \lambda_t \cdot \Delta_q^{(t)}$$

где λ_t — коэффициент затухания, $\Delta_q^{(t)}$ — обновление квантизации на шаге t.

Процесс квантизации следует поэтапному подходу:

1. Этап 1: Квантизация MLP слоев
$$\theta^{(1)}_{\text{MLP}} = \mathcal{Q}_{\text{per-tensor}}(\theta_{\text{MLP}})$$
2. Этап 2: Квантизация выходных проекций внимания
$$\theta^{(2)}_{\text{attn-out}} = \mathcal{Q}_{\text{per-channel}}(\theta_{\text{attn-out}})$$
3. Этап 3: Квантизация проекций query, key, value со смешанной точностью
$$\theta^{(3)}_{\text{QKV}} = \mathcal{Q}_{\text{mixed}}(\theta_{\text{QKV}}; \mathcal{S}(\theta_{\text{QKV}}))$$
4. Этап 4: Квантизация эмбеддинга патчей и классификационной головы
$$\theta^{(4)}_{\text{embed}} = \mathcal{Q}_{\text{patch-aware}}(\theta_{\text{embed}})$$

После каждого этапа квантизации выполняется донастройка для восстановления точности:

$$\theta^{(t)} \leftarrow \theta^{(t)} - \eta_t \nabla_\theta \mathcal{L}(\theta^{(t)}, \mathcal{D}_{\text{calib}})$$

где $\mathcal{D}_{\text{calib}}$ — небольшой калибровочный датасет, η_t — уменьшающийся график скорости обучения.

4.5 Математическая формулировка квантующих ядер

Для матрицы весов $\mathbf{W} \in \mathbb{R}^{M \times N}$ улучшенная квантизация вычисляет масштабные коэффициенты, используя абсолютный максимум и стандартное отклонение:

$$\alpha_i = \frac{\max(|\mathbf{W}_{i,:}|) + \beta \cdot \sigma(\mathbf{W}_{i,:})}{127}$$

где β — обучаемый параметр, балансирующий между максимумом и статистическим размахом, σ обозначает стандартное отклонение.

Квантизованное матричное умножение включает компенсацию ошибки:

Пусть $\mathbf{A}_q = \mathcal{Q}(\mathbf{A})$, $\mathbf{B}_q = \mathcal{Q}(\mathbf{B})$ — квантизованные матрицы. Вычисление имеет вид:

$$\mathbf{C} = (\mathbf{A}_q \mathbf{B}_q) \circ (\alpha_A \alpha_B^\top) + \mathbf{E}_{\text{comp}}$$

где $\mathbf{E}_{\text{comp}}$ — член компенсации ошибки:

$$\mathbf{E}_{\text{comp}} = \mu \cdot (\mathbf{A} - \mathbf{A}_q \alpha_A)(\mathbf{B} - \mathbf{B}_q \alpha_B)^\top$$

4.6 Квантизация активаций с учетом LayerNorm

Для выходов LayerNorm применяется динамическая квантизация с бегущей статистикой:

$$\mathbf{x}_{\text{ln}} = \frac{\mathbf{x} - \mu_{\mathbf{x}}}{\sigma_{\mathbf{x}} + \epsilon} \cdot \gamma + \beta$$

Масштаб квантизации для выходов LayerNorm адаптируется к статистическим свойствам:

$$\alpha_{\text{ln}} = \frac{\max(|\mathbf{x}_{\text{ln}}|) + \delta \cdot \sigma_{\mathbf{x}_{\text{ln}}}}{127}$$

где δ — калибровочный параметр, обучаемый в процессе тонкой настройки.

4.7 Общий оптимизационный подход

Финальная оптимизация комбинирует функцию потерь задачи с квантизационно-осознанной регуляризацией:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{task}}(\theta_q) + \lambda_1 \mathcal{R}_{\text{quant}}(\alpha) + \lambda_2 \mathcal{R}_{\text{smooth}}(\theta_q)$$

где:

- $\mathcal{R}_{\text{quant}}(\alpha) = \sum_i \|\alpha_i\|_2$ поощряет большие интервалы квантизации
- $\mathcal{R}_{\text{smooth}}(\theta_q) = \sum_i \|\nabla \theta_{q,i}\|_2$ способствует гладкости квантизованных параметров

4.8 Алгоритм VQ-ViT

Algorithm 1 Алгоритм VQ-ViT

Require: Модель ViT f, калибровочный датасет $\mathcal{D}_{\text{calib}}$, порог падения точности τ
Ensure: Квантизованная модель f_Q

```
Анализ чувствительности слоев с использованием спектральных и градиентных метрик
Инициализация плана квантизации со смешанной точностью
for module in [MLP, Attention-Out, QKV, Embedding] do
    Квантизация module со слое-специфичными схемами
    Тонкая настройка параметров квантизации на D_calib
    if падение точности > τ then
        Возврат к большей точности для чувствительных слоев
    end if
end for
Выполнение финальной глобальной тонкой настройки
return f_Q
```

5 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.

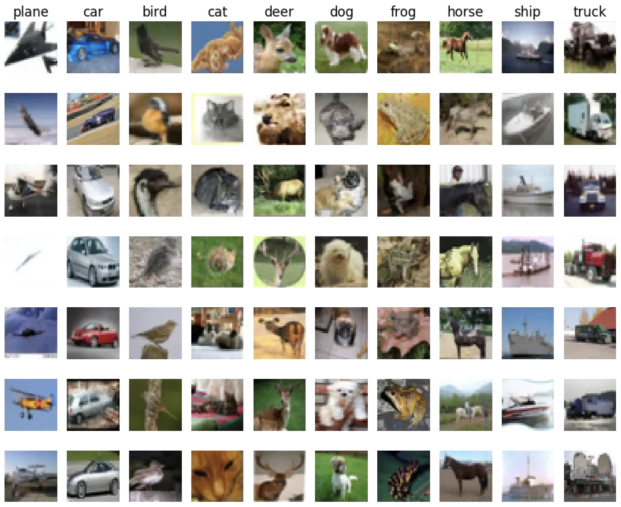

Рис. 1: Примеры изображений из датасета CIFAR-10

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:

Таблица 1: Гиперпараметры дообучения ViT

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

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:

Таблица 2: Сравнение характеристик моделей до и после квантизации

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

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

[1] Li, Y., Zhang, K., & Wang, J. (2022). P2-ViT: Post-training Quantization for Vision Transformers with Power-of-Two Scales.

[2] Yuan, Z., & Agaian, S. S. (2021). A comprehensive review of model quantization for vision transformers.

[3] Jin, Q., Yang, L., & Liao, Z. (2021). Quantization-aware training for vision transformers.

[4] Hao, C., & Li, P. (2022). Mixed-precision quantization of transformer-based models.

[5] Wang, H., Li, Z., & Liu, X. (2022). FQ-ViT: Post-Training Quantization for Fully Quantized Vision Transformer.

[6] Chen, X., Wang, Y., & Zhang, R. (2022). Attention-Aware Quantization for Vision Transformers.