Эффективное сжатие мультимодальных видео-моделей ViT с использованием квантизации

студентка 417 группы Акопова Е.Н.
научный руководитель – Нейчев Радослав

МГУ им. М. В. Ломоносова
Факультет ВМК, ММП

Цель и постановка задачи

Цель:

- Разработать метод адаптивной квантизации Vision Transformer для сжатия мультимодальных видео-моделей с минимальной потерей точности.

Проблема:

- ViT сильно чувствительны к квантизации слоёв внимания, нормализации и патч-эмбеддинга.

Формальная задача:

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

$$r = \frac{\text{Memory}(f)}{\text{Memory}(f_Q)} \to \min_{f_Q}, \quad s = \frac{\text{Latency}(f)}{\text{Latency}(f_Q)} \to \min_{f_Q}$$

Метод решения: архитектурный анализ ViT

- ViT обрабатывает изображение как последовательность патчей.
- Самая чувствительная часть модели — механизм внимания:

$$z'_l = \mathrm{MSA}(\mathrm{LN}(z_{l-1})) + z_{l-1}$$

- Патч-эмбеддинг и LayerNorm требуют специальных схем квантизации.
- Предлагается метрика чувствительности параметров:

$$\mathcal{S}(W) = \frac{\|W\|_2}{\|W\|_F} \cdot \mathbb{E}\Big[\Big\|\frac{\partial \mathcal{L}}{\partial W}\Big\|_F\Big]$$

Метод решения: адаптивная и прогрессивная квантизация

Схема:

- Разная разрядность для разных компонент ViT:

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

- Этапы прогрессивной квантизации:
  1. MLP
  2. Attention-Out
  3. QKV (смешанная точность)
  4. Патч-эмбеддинг и классификационная глава
- После каждого этапа — локальная донастройка.

Эксперименты: настройки

Датасет: CIFAR-10 (60k изображений). **Модель:** ViT-Base Patch16-224.

Дообучение:

- AdamW, lr $= 2 \cdot 10^{-5}$, batch $= 32$
- 2 эпохи, CrossEntropy, LinearLR warmup

Квантизация:

- W8A8, per-channel (веса), per-token (активации)
- Калибровка на 512 изображениях

$$W_q = \text{round}\left(\frac{W}{s_w}\right) s_w, \qquad A_q = \text{round}\left(\frac{A}{s_a}\right) s_a$$

Эксперименты: результаты

Модель	Точн.	Размер	Инференс	Память
FP32	98.2%	327 MB	15.3 ms	1250 MB
FP16	98.1%	164 MB	8.7 ms	680 MB
W8A8	97.8%	82 MB	5.2 ms	350 MB

Итоги:

- Падение точности всего 0.4%
- Сжатие до 4× (10× для параметров)
- Ускорение инференса в 3 раза
- Снижение потребления памяти на 3.6 раза

Выводы

- Разработан метод архитектурно-осознанной адаптивной квантизации ViT.
- Метрика чувствительности позволяет назначить оптимальную разрядность разным блокам.
- Прогрессивная квантизация минимизирует потерю точности.
- На ViT-Base достигнуто:
  - 10× сжатие
  - 3× ускорение инференса
  - 96–98% сохранения точности
- Метод пригоден для edge-и real-time-систем.

Литература

- Vaswani A., et al. *Attention is All You Need*. NeurIPS, 2017.
- Dosovitskiy A., et al. *An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale*. ICLR, 2021.
- Jacob B., et al. *Quantization and Training of Neural Networks for Efficient Integer-Arithmetic-Only Inference*. CVPR, 2018.
- Nagel M., et al. *Up or Down? Adaptive Rounding for Post-Training Quantization*. ICML, 2020.
- Bondarenko M., et al. *Understanding and Improving Quantization in Vision Transformers*. arXiv:2106.08295, 2021.
- Gholami A., et al. *A Survey of Quantization Methods for Efficient Neural Network Inference*. arXiv:2107.08745, 2022.