Аппаратно-ориентированная адаптация YOLOS-tiny к BF16-аккумулированию

Андреев Д. О.

Магистрант

МИЭМ НИУ «Высшая школа экономики», Москва, Россия

E-mail: doandreev@edu.hse.ru

Научный руководитель *— Нефедов С.И.*

Современные режимы вычислений пониженной точности обычно используют формат BF16 на входах матричных операций и более широкий аккумулятор, как правило FP32, для накопления промежуточных сумм [1, 2]. Это позволяет почти не терять качество модели при заметном ускорении инференса. Однако в энергоограниченных ускорителях точность аккумулятора может быть ограничена самим форматом BF16, и тогда ошибки округления накапливаются внутри длинных сумм. Для трансформеров такой режим особенно опасен, поскольку искажает вычисление механизма внимания и последующих линейных преобразований.

В работе исследуется влияние BF16-аккумулирования на трансформерный детектор YOLOS-tiny [4] на валидационной выборке COCO 2017 [5]. Под BF16-аккумулированием понимается режим, в котором промежуточная сумма скалярного произведения после каждого шага округляется к BF16: $s_{i+1} = \mathrm{round}_{\mathrm{bf16}}(s_i + \mathrm{round}_{\mathrm{bf16}}(a_i b_i))$. Для моделирования такого инференса реализованы специализированные вычислительные ядра для линейных слоёв и батчевых матричных умножений в блоках внимания. Для адаптации модели предложена схема дообучения: в прямом проходе используется целевая BF16-арифметика, а в обратном — STE-подобная аппроксимация градиента [3] через стандартные высокоточные матричные операции.

Эксперименты на 5000 изображениях показывают, что переход от FP32 к стандартному BF16-режиму с аппаратным широким накоплением уменьшает AP только с 0.2692 до 0.2652. Однако при принудительном BF16-аккумулировании во всех целевых компонентах AP снижается до 0.2404, то есть на 2.88 п.п. относительно FP32. Компонентная абляция показывает, что наибольший вклад в деградацию дают блоки внимания и линейные проекции/MLP: в изоляции они уменьшают AP на 1.62 и 1.37 п.п. соответственно, тогда как вклад первого слоя разбиения на патчи слабее. Эффект неоднороден по размерам объектов: APS падает с 0.075 до 0.065, APM — с 0.285 до 0.258, APL — с 0.450 до 0.410, что указывает на повышенную чув-

ствительность малых объектов к ошибкам округления.

Короткое дообучение в целевом режиме BF16-accum частично компенсирует потери качества. При немедленном включении всех низкоточных компонентов AP возрастает до 0.2490, что соответствует восстановлению 0.86 п.п. или около 30% потерянного качества; постепенная стратегия даёт 0.2462. При этом в контрольном многосидовом BF16-режиме без BF16-accum стандартное отклонение составляет порядка $2 \cdot 10^{-4}$ AP, поэтому наблюдаемая деградация на два порядка превышает стохастическую вариативность. Полученные результаты показывают, что точность аккумулятора является критическим параметром при аппаратно-ориентированном переносе трансформерных детекторов на энергоэффективные ускорители: выигрыш в ресурсах промежуточных сумм требует специальной адаптации модели уже на этапе обучения.

Литература

1. Wang S., Kanwar P. BFloat16: The secret to high performance on Cloud TPUs. 2019.
2. Micikevicius P., Narang S., Alben J. et al. Mixed Precision Training // ICLR. 2018.
3. Bengio Y., Leonard N., Courville A. Estimating or Propagating Gradients Through Stochastic Neurons. 2013.
4. Fang Y., Zhang S., Wang C. et al. You Only Look at One Sequence: Rethinking Transformer in Vision through Object Detection // NeurIPS. 2021.
5. Lin T.-Y., Maire M., Belongie S. et al. Microsoft COCO: Common Objects in Context // ECCV. 2014.