УДК 004.85

ИССЛЕДОВАНИЕ МЕТОДОВ УСКОРЕНИЯ ИНФЕРЕНСА МОДЕЛЕЙ НЕЙРОННЫХ СЕТЕЙ ДЛЯ ДЕТЕКЦИИ ОБЪЕКТОВ НА ИЗОБРАЖЕНИИ

Иванов Д.А. (АО Концерн "Вега")

Научный консультант – кандидат технических наук, доцент Никонов А.Н.
(АО Концерн "Вега")

Введение. В последние годы количество данных, получаемых со спутников дистанционного зондирования Земли, стремительно увеличивается [1]. Ежедневно спутники фиксируют сотни терабайт данных, что требует мощных вычислительных ресурсов для их обработки. В традиционных фреймворках, таких как PyTorch, инференс моделей детекции объектов может занимать значительное время, что делает обработку таких данных ресурсозатратной задачей. Для обеспечения высокой скорости обработки данных необходимо применение методов ускорения нейросетевых моделей.

Одним из наиболее перспективных подходов к ускорению является использование специализированных фреймворков, таких как TensorRT [2]. Их использование позволяет значительно ускорять процесс инференса за счёт оптимизаций вычислительного графа, квантизации весовых коэффициентов и других методов компрессии. В данной работе рассматриваются возможности применения этих методов для ускорения детекции объектов на спутниковых снимках.

Основная часть. В рамках исследования были проведены эксперименты по ускорению инференса модели YOLOv9, обученной на спутниковых снимках оптического диапазона, с использованием различных методов оптимизации:

1. Исследование ускорения за счет следующих оптимизаций:
   1) Объединение слоёв свёртки, batch normalization и ReLU для сокращения количества вычислительных операций.
   2) Автонастройка вычислительных ядер в зависимости от модели, структуры входных данных и характеристик используемого GPU.
   3) Встраивание алгоритма, подавления дублирующихся предсказаний (Non-Maximum Suppression).
2. Исследование влияния понижения точности вычислений (переход с режима с плавающей запятой FP32 на пониженную точность FP16, на режим с уменьшенной мантиссой BF16 и на режим с целочисленными вычислениями INT8) на качество и скорость инференса [3].

Результаты исследования показали, что использование графовых и аппаратных оптимизаций на примере TensorRT позволяет увеличить скорость детекции при неизменном качестве. Точность распознавания в режиме FP16 позволяет ускорить инференс почти в 11 раз (с 31.74 мс в PyTorch до 2.83 мс в TensorRT) при минимальной потере точности (mAP50_95 = 0.8914 против 0.8926). Режим «Best» (автоматически подобранный режим TensorRT), включающий INT8, также показал высокие результаты в скорости, но снижение точности делает его менее универсальным в практическом применении. FP16 обеспечивает баланс между скоростью и точностью, что делает его предпочтительным выбором для задач детекции объектов, где важно сохранить качество предсказаний.

Выводы. В ходе работы разработан экспериментальный стенд для исследования методов ускорения инференса моделей детекции объектов на примере фреймворка TensorRT, проведено исследование различных методов ускорения инференса. Результаты показывают, что наилучший компромисс между точностью и скоростью обеспечивает использование TensorRT с квантизацией FP16. Данный подход позволяет значительно снизить затраты на вычислительные ресурсы при обработке больших объёмов спутниковых данных.

Список использованных источников:

1. [Электронный ресурс]. – Режим доступа: https://www.roscosmos.ru/24707/ (дата обращения: 10.02.2025).
2. Zhou Y., Yang K. Exploring TensorRT to Improve Real-Time Inference for Deep Learning // 2022 IEEE 24th International Conference on High Performance Computing & Communications; 8th International Conference on Data Science & Systems; 20th International Conference on Smart City; 8th International Conference on Dependability in Sensor, Cloud & Big Data Systems & Application (HPCC/DSS/SmartCity/DependSys). – Hainan, China, 2022. – С. 2011–2018.
3. Wang M., Sun H., Shi J., Liu X., Zhang B., Cao X. Q-YOLO: Efficient Inference for Real-time Object Detection // arXiv preprint arXiv:2307.04816. – 2023.