

УДК 332.14

РОЛЬ ОЦЕНКИ НАКЛАДНЫХ РАСХОДОВ ПАМЯТИ И CPU ПРИ ИНТЕГРАЦИИ НЕЙРОСЕТЕВЫХ МОДЕЛЕЙ СЕМЕЙСТВА YOLO

Д. А. ВЕТРОВ

Научный руководитель П. С. ХАРЛАМОВ

Смоленский филиал Российской академии народного хозяйства
и государственной службы при Президенте Российской Федерации
Смоленск, Россия

Архитектуры семейства YOLO в настоящее время являются стандартом в области детекции объектов в режиме реального времени. Скорость обработки кадров в наиболее современных моделях может достигать 0,001 с. Однако существует проблема внедрения таких систем в промышленные объекты и охранные системы: возрастает нагрузка и ухудшается производительность. Основная причина подобных расходов чаще всего находится в накладных расходах центрального и графического процессора (рис. 1). Наиболее часто такие расходы возникают при препроцессинге, изменении разрешений и преобразовании цветовых пространств, а также при передаче подготовленных тензоров из системной памяти в видеопамять перед инференсом и обратно.



Рис. 1. Схема потока данных и накладных расходов нейросетевых моделей семейства YOLO

Постобработка и препроцессинг при отсутствии оптимизации модели и настройки ее под конкретные аппаратные ресурсы могут занимать до 60 %...70 % от общего времени обработки кадра. Неправильная и некорректная работа с памятью может вызвать сбои при обработке параллельно получаемых видеоданных, связанные с утечкой памяти. Особые проблемы представленные упущения могут вызвать на встраиваемых системах и умных камерах в условиях жёсткой ограниченности ресурсов. Оценка накладных расходов памяти и CPU позволяет обоснованно выбрать технологический стек (TensorRT, OpenVINO, C++, например, вместо Python), вынести этап устранения дублирующих кадров прямо на GPU и спроектировать энергоэффективные архитектуры видеоаналитики, что поможет значительно улучшить качество работ систем на базе YOLO, позволяя довести показатели распознавания до заявленных разработчиками.