Д. К. Зитцер
Тюменский государственный университет, г. Тюмень
УДК 004.932

КРИТЕРИИ ЭФФЕКТИВНОЙ ОРГАНИЗАЦИИ РАБОТЫ ВИДЕОАНАЛИТИКИ

Аннотация. Статья посвящена формированию критериев эффективной организации работы видеоаналитики. Будут рассмотрены несколько вариантов организации вместе с рассуждениями о достоинствах и недостатках каждого. Статья будет полезна всем, кто интересуется вопросами применения видеоаналитики в различных областях деятельности.

Ключевые слова: видеоаналитика, производительность видеоаналитики, эффективная организация работы видеоаналитики, сопутствующие вычисления.

Введение. Видеоаналитика (ВА) — это технология, использующая методы компьютерного зрения для автоматизированного получения данных на основании анализа изображений или последовательностей изображений (видеопотоков) [1]. В последние годы бурному развитию ВА способствуют несколько факторов. Во-первых, системы ВА могут превосходно использовать широко развитую инфраструктуру систем безопасности видеонаблюдения [2]. Во-вторых, в большинстве задач, связанных с визуальным восприятием, алгоритмы компьютерного зрения уже превзошли человека в качестве работы [3]. Именно в связи с широким распространением систем ВА, вопросы, связанные с эффективной в вычислительном плане организацией работы таких систем, постепенно выходят на первый план и бросают новые вызовы инженерному и научному сообществу. Цель данной статьи — сформулировать критерии эффективной организации работы систем ВА.

Устройство видеоаналитики. Типовое устройство конвейера ВА представлено на рис. 1. Работа ВА начинается с камеры, которая является источником данных. Камера генерирует видеопоток, который сжимается с помощью кодеков, чаще всего реализующих стандарт H264, и пакуется в медиаконтейнер. Получая видеопоток в сжатом виде, декодер занимается извлечением из него кадров, то есть процессом декодирования. Далее кадры при помощи препроцессора готовятся для подачи в нейронную сеть. Например, может изменятся разрешение кадра и порядок его цветовых каналов. Подготовленный кадр пропускается через нейронную сеть (НС). НС — это обязательный и самый важный компонент любой современной видеоаналитики. Получившиеся в ходе вычислений результаты обрабатываются постпроцессором. Например, на этом шаге может применяться алгоритм подавления немаксимумов [4]. Далее отрабатывает эвристика, задача которой сформировать понятное для оператора системы ВА событие. И в конце происходит уведомление оператора системы с помощью заранее настроенного канала коммуникации.

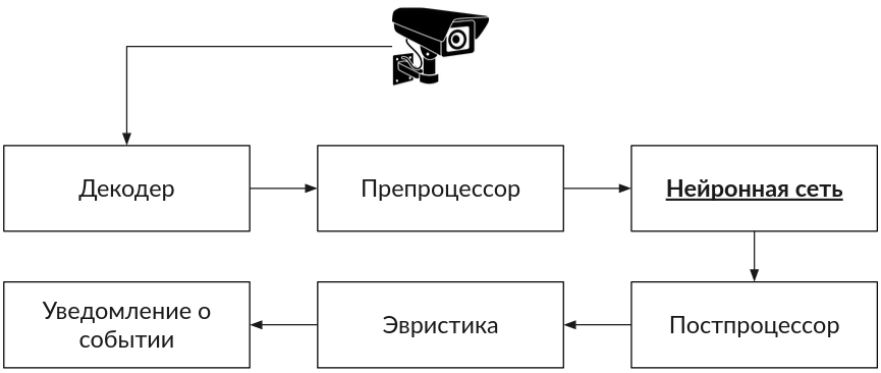

Рис. 1. Типовое устройство конвейера ВА

Производительность видеоаналитики. Производительность ВА замеряется при помощи двух метрик. Задержка измеряется в миллисекундах и показывает время обработки одного изображения. Пропускная способность измеряется количеством обрабатываемых кадров в секунду.

Существует четыре вложенных друг в друга уровня ВА, из которых складывается общая производительность системы [5]. Уровни представлены на рис. 2. На самом нижнем уровне находятся слои НС. На этом уровне, например, можно пытаться написать более вычислительно эффективные реализации операций перемножения матриц или свертки. Уровнем выше находится уровень архитектуры нейронной сети. На нем определяется структура и организация НС, то есть то, как НС будет обрабатывать входные данные. Например, в задаче детекции объектов существуют разные архитектурные подходы [6, 7], обладающие разным балансом в скорости и качестве работы. Следующий уровень — это уровень сопутствующих вычислений. Это те вычисления, без которых не может функционировать НС и создаваться события ВА. Именно к этому уровню в данной статье будут формироваться критерии эффективности в вычислительном плане. Самый верхний уровень — это уровень программного продукта. Он уже включает в себя весь ландшафт системы видеонаблюдения, в рамках которого видеоаналитика работает.

Рис. 2. Уровни ВА, из которых складывается общая производительность системы ВА

В ряде недавних публикаций зарубежных авторов роль сопутствующих вычислений была переосмыслена. В работе [5] на примере видеоаналитики по распознаванию лиц было установлено, что инфраструктура современных центров обработки данных, в которой функционирует ВА, может вносить до 30% общей задержки работы системы ВА. Авторы статьи [8] при помощи ряда экспериментов установили, что в суммарной задержке системы видеоаналитики доля сопутствующих вычислений и шагов по перемещению данных может составлять до 56% для среднеформатных изображений и около 80% для крупноформатных.

Варианты организации работы сопутствующих вычислений. К уровню сопутствующих вычислений относятся четыре компонента конвейера ВА: препроцессор, нейронная сеть, постпроцессор, эвристика. Стоит отметить, что эти компоненты неоднородны в используемых для вычислений аппаратных ресурсах. Все компоненты, за исключением НС, вычисляются

с помощью центрального процессора, а вот НС чаще всего производит вычисления на графическом процессоре. Такая неоднородность оказывает влияние на эффективную организацию работы сопутствующих вычислений. Для примера рассмотрим некоторые базовые варианты.

Начнем с самого простого варианта. Все этапы сопутствующих вычислений запускаются последовательно в рамках одного процесса операционной системы. К достоинствам такого варианта стоит отнести то, что он очень прост в реализации. Также полностью отсутствуют накладные расходы на передачу данных между компонентами. Однако этот вариант обладает существенным недостатком, связанным с простоем графического процессора на время работы препроцессора, постпроцессора и эвристики. Видеоаналитика может быть «облачной», а это значит, что за графический процессор платится довольно существенная арендная плата, соответственно, недопустимо, чтобы графический процессор находился в простое.

Во втором варианте этап, связанный с вычислениями нейронной сети, выносится в отдельный веб-сервис. Пересылка подготовленных кадров происходит по сети. К достоинствам такого варианта стоит отнести то, что он все еще очень прост в реализации. Однако, в силу того, что данные передаются по сети, то, соответственно, возрастает сетевой трафик. В некоторых компаниях, предоставляющих услуги на базе облачных технологий, подобный трафик тарифицируется. Второй недостаток проявляется в том, что для передачи кадра в веб-сервис его нужно сериализовать, что создает накладные вычислительные расходы.

В третьем варианте работа вычислительно сложных компонентов, таких как НС и препроцессор, выносятся в отдельные процессы операционной системы. В этом варианте для обмена кадрами используются механизмы операционной системы для межпроцессного взаимодействия. Например, использование механизма разделяемой памяти позволит избежать дополнительных накладных расходов на сериализацию кадра при его передаче на обработку. К достоинствам варианта стоит отнести то, что он лишен всех недостатков ранее рассмотренных вариантов. Однако, реализация подобного рода схемы потребует от инженера-программиста довольно высокого уровня квалификации.

Выводы. Эффективная организация сопутствующих вычислений ВА — это получение максимально возможной пропускной способности ВА при высоком проценте утилизации аппаратных ресурсов сервера и отсутствии неоправданных накладных расходов. Для эффективной организации работы сопутствующих вычислений ВА необходимо, чтобы работа вычислительно сложных компонентов разносилась по отдельным процессам операционной системы, при этом очень важно под каждую аппаратную платформу подбирать наиболее быстрый механизм межпроцессного взаимодействия, учитывая немалый размер кадра.

Заключение. В данной статье объектом исследования выступала ВА, а предметом — эффективная организация ее работы. Были рассмотрены три базовых варианта организации работы вместе с рассуждениями о достоинствах и недостатках каждого. В конце были сформулированы критерии эффективной организации работы ВА. В дальнейшем статья может быть полезна при выборе или разработке программного обеспечения для организации работы ВА. Также стоит провести ряд экспериментов по количественной оценке пропускной способности ВА для описанных в статье вариантов организации.

СПИСОК ЛИТЕРАТУРЫ

1. ГОСТ Р 59385-2021. Информационные технологии. Искусственный интеллект. Ситуационная видеоаналитика. Термины и определения.
2. Ameen M., Stone R. Advancements In Crowd-Monitoring System: A Comprehensive Analysis of Systematic Approaches and Automation Algorithms: State-of-The-Art // arXiv preprint arXiv:2308.03907. — 2023.
3. Wang Y. et al. Computation-efficient deep learning for computer vision: A survey // Cybernetics and Intelligence. — 2024.
4. Non Maximum Suppression: Theory and Implementation in PyTorch. URL: https://learnopencv.com/non-maximum-suppression-theory-and-implementation-in-pytorch (дата обращения: 12.05.2024).
5. Richins D. et al. AI tax: The hidden cost of AI data center applications //ACM Transactions on Computer Systems (TOCS). — 2021. — Т. 37, № 1-4. — С. 1-32.
6. Liu W. et al. Ssd: Single shot multibox detector // Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14. — Springer International Publishing, 2016. — С. 21-37.
7. Redmon J. et al. You only look once: Unified, real-time object detection // Proceedings of the IEEE conference on computer vision and pattern recognition. — 2016. — С. 779-788.
8. AbouElhamayed A. F. et al. Beyond Inference: Performance Analysis of DNN Server Overheads for Computer Vision //arXiv preprint arXiv:2403.12981. — 2024.