Вопросы создания аппаратно-программного комплекса распознавания видеоизображений с технологией искусственного интеллекта на базе отечественной платформы

Пузийчук С.И., Суменков Н.А.

В статье рассматриваются общие подходы процесса создания аппаратно-программного комплекса для распознавания видеоизображений с использование искусственного интеллекта. Рассматриваются особенности отечественной технологической платформы ODANT для решения задач, связанных с обработкой изображений и распознавания образов.

Создание аппаратно-программного комплекса распознавания объектов предусматривает работы с большими объемами данными и требует привлечение целого ряда новых сквозных технологий. Среди них - объектно-семантическое представление гетерогенных данных, обеспечивающее связность единого цифрового пространства и повышающего его безопасность за счет сокращения «поверхности атаки» и технологии искусственного интеллекта для анализа изображений, основанные на глубоком обучении нейронных сетей. Не претендуя на освещение всех существующих сквозных технологий, разрабатываемых и используемых в развивающейся экономике данных, авторы в данной статье ставят своей целью описать практические шаги по применению указанных технологий и привести частные примеры и возможности их использования.

Основными этапами работ создания аппаратно-программного комплекса распознавания видеоизображений являются [10]:

1) постановка задачи распознавания объектов в части номенклатуры распознаваемых объектов, условий работы системы распознавания, основными параметрами и характеристиками ее основных элементов.

2) выбор архитектуры нейронной сети исходя из оптимального на момент формирования сети сочетание ее сложности, быстродействия и эффективности работы непосредственно с заданными классами объектов.

3) выбор или создание датасета для обучения и тестирования нейронной сети. При создании датасета выполняется поиск исходных данных в виде изображений (фотографий, последовательности действий), разметка полученных изображений, а также их размещение по каталогам в соответствии со структурой датасета.

4) обучения нейронной сети с выбранной архитектурой с использованием созданного или выбранного датасета (его части —

обучающего набора). Результатом обучения будет массив весов нейронной сети, который загружается в выбранную архитектуру сети.

5) проверка качества обучения нейронной сети, её тестирование. При этом используется вторая часть датасета — тестовый или проверочный набор. Процесс тестирования целесообразно проводить совместно с этапом обучения путем их чередования в рамках каждой эпохи обучения, а также выполнения проверки на обнаружение факта переобучения.

6) перенос модели обученной нейронной сети на целевую архитектуру аппаратных средств.

Постановка задачи распознавания объектов.

В состав системы технического зрения для распознавания нескольких типов объектов, находящихся в поле зрения, как правило, входят одна или несколько камер, видеосигнал с которых оцифровывается и передается в цифровое устройство обработки на базе персонального компьютера или другой подходящей архитектуры. В готовом изделии довольно часто используется процессор для встроенных применений на основе архитектуры ARM (Advanced RISC Machines), который работает под управлением операционной системы (ОС) Linux.

При этом в процессе разработки программного обеспечения (ПО) для такого процессора обычно используется персональный компьютер (ПК). Для успешного переноса ПО требуется максимальная совместимость ПО для архитектур ARM и ПК. Поэтому средства разработки на ПК чаще всего также используют ОС Linux.

Любая система технического зрения имеет несколько основных параметров, таких как разрешающая способность, минимальная освещенность на объекте или чувствительность камеры, отношение сигнал/шум, скорость смены кадров и некоторые другие. Эти параметры выбираются исходя из особенностей работы системы, количества и типа обнаруживаемых объектов.

Соответственно, выбирается тип камеры, камерного модуля или микросхемы КМОП-фотоприемника изображения. Далее исходя из предполагаемой сложности архитектуры нейронной сети, требований к быстродействию и к энергопотреблению, выбирается вычислительное устройство на базе ПК, микропроцессора ARM или других аппаратных средств.

Выбор архитектуры нейронной сети.

Одной из эффективных технологий распознавания объектов в видеопотоке (на изображении) является использование сверточных нейронных сетей (далее – CNN) и рекуррентных нейронных сетей (далее – RNN).

CNN используются для извлечения признаков из изображения, а RNN - для анализа последовательности изображений.

Классическая структура сверточной нейросети VGG16 представлена на рисунке 1 [5].

Сверточная сеть представляет собой комбинацию трех типов слоев:

- слои, которые выполняют функцию свертки над двумерными массивами данных (сверточные слои), используются для извлечения признаков изображений;

- слои, выполняющие функцию уменьшения формата данных (слой субдискретизации или пулинг слои), используются для уменьшения размерности и извлечения ключевых признаков;

- полносвязные слои, завершающие процесс обработки данных, используются для классификации и определения нарушений.

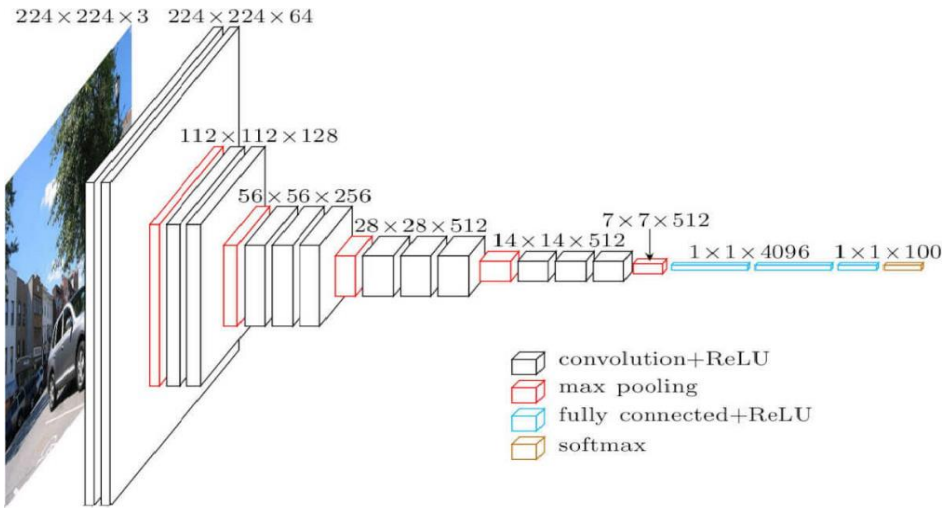

Рисунок 1 – Структура сверточной нейросети - VGG16
для выделения признаков изображений

Структура CNN принципиально многослойная. Работа CNN обычно интерпретируется как переход от конкретных особенностей изображения к более абстрактным деталям, и далее к ещё более абстрактным деталям вплоть до выделения понятий высокого уровня. При этом сеть самонастраивается и вырабатывает необходимую иерархию абстрактных признаков (последовательности карт признаков), фильтруя маловажные детали и выделяя существенное [7].

Сеть VGG-16 имеет 16 слоев и способная работать с изображениями достаточно большого формата 224x224 пикселя [5]. В своей стандартной

топологии эта сеть способна работать с датасетом изображений ImageNet, содержащем более 15 млн. изображений, разбитых на 22000 категорий [1].

RNN отличаются от многослойных сетей тем, что могут использовать свою внутреннюю память для обработки последовательностей произвольной длины. Благодаря направленной последовательности связей между элементами рекуррентных сетей они применимы в таких задачах, где нечто целостное разбито на сегменты, например, распознавание рукописного текста или распознавание речи.

В процессе сегментации выполняется разделение видео на отдельные кадры или группы кадров. На основе определенных критериев, таких как движение объектов, изменение освещенности или изменение фокуса камеры производится автоматическая сегментация видео. Таким образом RNN позволяют моделировать блоки памяти для сохранения данных и моделирования краткосрочных зависимостей.

Выбор (обучение) датасета и обучение нейронной сети.

Для работы с нейронными сетями требуется их обучение под конкретную задачу. В частности, для решения задачи распознавания объектов на изображении требуется обучение сети по специально подготовленному набору данных, который содержит изображения всех классов распознаваемых объектов, сгруппированных в соответствующие разделы. Такой тип данных носит название датасет (набор данных, Data set) [7, 8, 9].

Существует большое количество уже собранных и подготовленны датасетов для решения различных задач с использованием нейронных сетей (не только для задач распознавания объектов) [7]. Более того, существуют уже заранее обученные под решение конкретной задачи нейронные сети, которые можно взять в готовом виде [1]. Но перечень таких сетей и датасетов не очень большой, и в общем случае перед разработчиком может стоять задача выбора конфигурации нейронной сети под конкретную задачу и создание соответствующей базы данных (датасета) для ее обучения [2, 3].

Формирование датасета является наиболее трудоемкой частью процесса разработки.

Процесс обучения нейронных сетей представляет собой сложный процесс обработки данных, который включает в себя последовательное предъявление данных на вход нейронной сети и сравнение выходных данных с их истинным значением, после чего вносится коррекция весовых коэффициентов нейронов в сторону уменьшения ошибки выходных данных. Этот процесс производится многократно с использованием данных из датасета. Проверка производительности обученной модели проводится на тестовых данных. После оценки результатов проводится дополнительной обучение или тюнинг модели.

Один цикл обучения с использованием всего датасета носит название эпоха. Как правило, для качественного обучения сети требуется много эпох.

Процесс обучения нейронных сетей, имеющих много скрытых слоев, часто носит название глубокого обучения.

Контроль качества обучения нейронной сети.

В процессе обучения и контроля за качеством обучения широко используются методы градиентного спуска и обратного распространения ошибки [10]. Метод градиентного спуска (рисунок 2) позволяет найти локальный минимум функции (которая является функцией ошибки) от нескольких переменных (весов нейронов) путем последовательного малого изменения этих переменных.

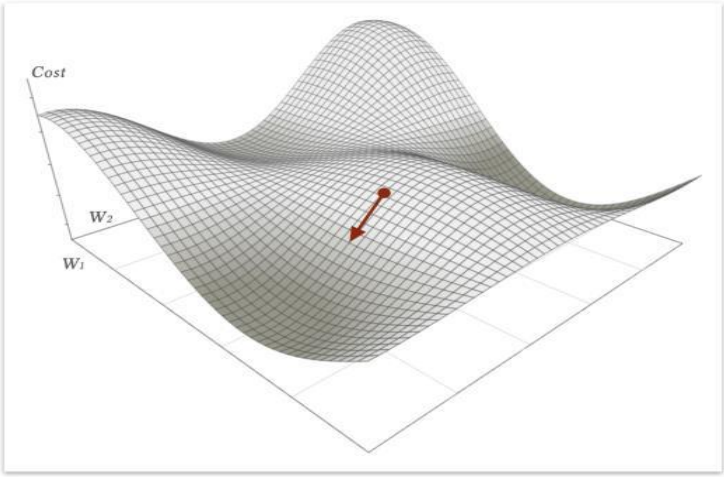

Рисунок 2 – Иллюстрация метода градиентного спуска

Метод обратного распространения ошибки позволяет выполнить процедуру коррекции весов по направлению от выхода ко входам нейронов путем использования частных производных.

Качество обучения можно контролировать путем анализа ошибок сети на ее выходе при сравнении с истинными значениями датасета (рисунок 3).

В процессе обучения, как правило, эти ошибки уменьшаются по экспоненте, но может наступить момент, когда ошибка сети начинает расти. С этого момента фиксируется так называемый эффект переобучения сети, когда процесс обучения целесообразно остановить.

Для анализа качества обучения используется специальный раздел данных датасета, который носит название тестовый набор. Это часть данных, аналогичная тренировочному набору, но не использованная в процессе обучения сети. То есть эта часть данных сети не знакома, и тестирование качества обучения при предъявлении новых данных является более корректным.

Процесс обучения сети занимает значительное время и как правило разбивается на этапы по обучению сети в целом и ее отдельных частей.

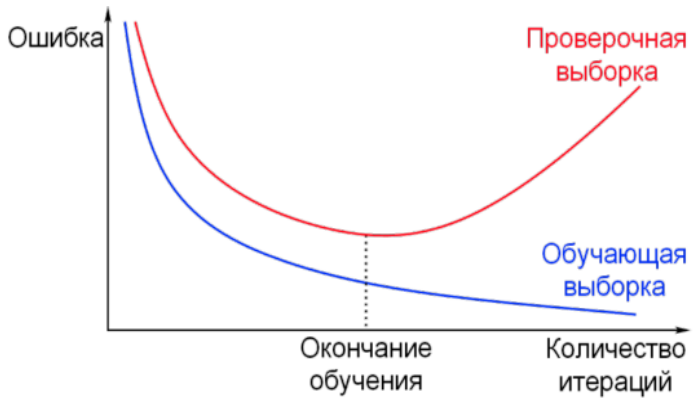

Рисунок 3 – Диаграмма контроля качества обучения нейросети

Структурно нейронные сети постоянно усложняются, появляются новые архитектуры сетей со все большим количеством слоев. Построение сети становится все более сложной задачей, требующей введения большого количества параметров, а главное, знания о том, как эти параметры должны быть настроены. Для упрощения такой работы используются специальные фреймворки, которые резко упрощающие работу по созданию, настройки и обучению нейронных сетей.

Вопросы переноса модели нейронной сети на платформу ODANT.

Отечественной платформой, представляющей новый подход к объектному хранению, передаче и обработке данных в сети, позволяющей создавать семантические сети из распределенных информационных сетей различного масштаба и сложности, пригодных для машинной обработки является технологическая платформа ODANT.

В состав технологической платформы ODANT входят [4]:

- средства хранения в виде объектно-файловой системы управления базами данных;
- средства обработки информации – framework программная платформа, упрощающая разработку программного продукта;
- средства разработки конечных решений.

Структурная схема платформы ODANT представлена на рисунке 1.

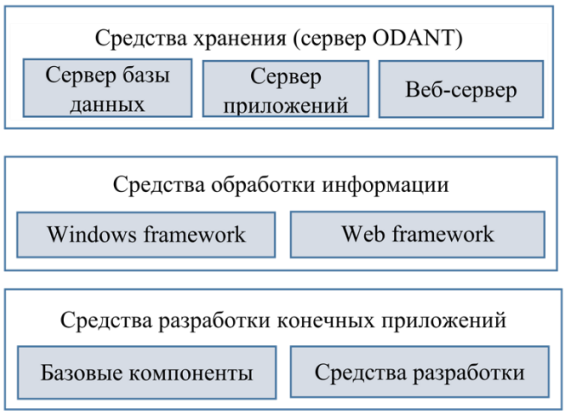

Рисунок 1 – Структурная схема платформы ODANT

Отличительными особенностями платформы ODANT являются [6]:

1) Бесшовная совместимость систем – интероперабельность. Глобальная совместимость информации – в рамках всех подсистем и баз данных позволяет значительно снизить сложность построения больших распределенных информационных систем. Платформа характеризуется применением объектно-ориентированной базы данных и хранением данных в открытом формате XML (eXtensible Markup Language), ODBM (Object DataBase Markup Language), что позволяет эффективно описывать информационные модели любой сложности в виде объектных классов. На основе метаданных ODBML появляется возможность создания цифровых двойников из физического мира. Кроме этого протокол ODANT поддерживает не только описание моделей, но и их поведение и визуализацию [4]. Характеризуется: единым форматом данных, объединением кластеров систем в глобальные сети, единым протоколом взаимодействия, беспроблемной интеграцией сетей, предоставляет возможность автодиагностики систем.

2) Объектно-ориентированный подход (далее – ООП), предусматривает антропогенный подход к построению моделей; древовидную архитектуру моделей данных; гибкую модификацию работающих систем; многократное использование созданных моделей; обработку методов, свойств и событий интегрированных в данные предоставляет возможность отображения информации о сложных взаимосвязях объектов. Объектно-ориентированная модель данных позволяет идентифицировать отдельную запись базы данных и определять функции их обработки. Однако к недостаткам объектно-ориентированного подхода следует отнести высокую понятийную сложность, неудобство обработки данных и низкую скорость выполнения запросов [7].

3) Моделирование вместо программирования. Преимущественное использования визуальных интерфейсов при разработке решений позволяет перейти от программирования к конструированию (моделированию). Таким образом, функции программистов по разработке решений начинают выполнять специалисты предметной области, конструируя систему в визуальном интерфейсе. При этом, за счет глобальной совместимости данных и максимального использования готовых модулей специалист может собирать решение из готовых блоков – ранее разработанных моделей данных. Применение такого подхода позволяет кардинально повысить скорость создания и модификации решений.

Технология ODANT позволяет создавать цифровые платформы, с возможностью построения распределенных информационных сетей, при этом программы, сервисы и системы бесшовно совмещаются между собой и имеют свойство объединяться в единую кибер-сеть. Технология объектной работы с данными повышает производительность в процессе взаимодействия между людьми и устройствами в едином информационном пространстве за счет создания информационных систем из готовых компонентов, как из

конструктора [4], рисунок 2. Сборку информационной системы из готовых модулей выполняет бизнес-аналитик без привлечения программистов.

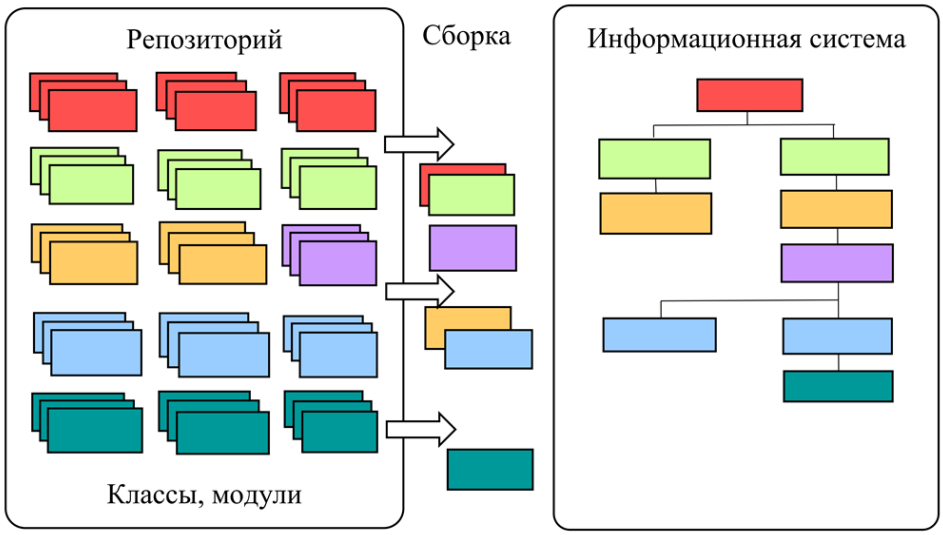

Рисунок 2 – Создание информационной системы из готовых компонентов

4) Информационно-компонентная модель позволяет многократно сократить участие программистов при разработке систем за счет использования визуальной среды разработки и моделирования информационных сущностей; разработка систем без программирования, предоставляет возможность неограниченного наращивания функционала, многократно-повторного использования наработанных участниками сети информационных моделей и сущностей; обеспечивается снижение требований к вычислительным ресурсам и простота создания программного обеспечения.

5) Сетецентрическая архитектура платформы ODANT ориентирована на распределенную работу. Установленные на отдельных серверах объектно-ориентированные базы данных по факту являются одной распределенной базой данных с единой системой безопасности, адресации, взаимодействия компонентов. Каждый элемент сети является сервером для элементов ниже, обеспечивается сохранение работоспособности при обрыве связи и высокая согласованность действий. Это позволяет кардинально снизить время на обеспечение взаимодействия различных территориальных подсистем. Сетецентрическая модель организации сети не требует крупных дата-центров устойчива к атакам, временным разрывам связи и выходам из строя или отключениям части узлов.

6) Неограниченная масштабируемость систем предусматривает: адаптационное партиционирование данных; поэтапный процесс шардинга; балансировку на стороне клиента; естественную система шардинга; репликацию данных на фолловера; изменение масштаба данных.

7) Возможность оптимизации разработанной модели за счет применения:

- архитектурной оптимизации (уменьшение сложности модели, использование легковестных структур, применение методов компрессии);

- оптимизации процесса обработки видео (параллельная обработка, использование аппаратного ускорения, предварительная обработка);

- обучение с учителем (аугментация данных для увеличения разнообразия тренировочного набора данных и улучшения обобщающей способности модели, обучение на частях видео для более эффективного обучения);

- оптимизация гиперпараметров (поиск по сетке, случайный поиск, нахождение оптимальных значений гиперпараметров модели);

- оценка производительности и повторное обучение модели на расширенном наборе данных с новыми методами обучения для улучшения точности.

Преимуществами технологической платформы ODANT является: полный технологический стек разработки; 100% российская разработка; единое информационной пространство; распределенное хранилище данных; платформа индустриального интернета; технология цифровых двойников [6].

Вывод: Таким образом, технологическая платформа ODANT может быть использована, как сквозная технология для создания аппаратно-программного комплекса, выполняющего обработку изображений и распознавание образов с применением технологий искусственного интеллекта.

Список информационных источников

1. Классификация изображений ImageNet - URL: https://gist.github. com/yrevar/942d3a0ac09ec9e5eb3a (дата обращения 29.05.2024).
2. Перечень датасетов, встроенных в Keras - URL: https://keras.io/api/datasets/ (дата обращения 29.05.2024).
3. Перечень предобученных сетей в Keras - URL: https://keras.io/api/applications/, (дата обращения 29.05.2024).
4. Перепелкин Р.А., Чумаков Г.В., Витухин В.В., Марочкин М.В., Трофимов И.М. ODANT – Интернет нового поколения // Проектирование будущего. Проблемы цифровой реальности: труды 2-й Международной конференции (7-8 февраля 2019 г., Москва). — М.: ИПМ им. М.В.Келдыша, 2019. — С. 112-117. — URL: https://keldysh.ru/future/2019/11.pdf doi:10.20948/future-2019-11.

5. Примеры архитектур сверточных сетей VGG-16 и VGG-19 - URL: https://proproprogs.ru/neural_network/primery-arhitektur-svertochnyh-setey-vgg16-i-vgg19, (дата обращения 29.05.2024).
6. Технологическая платформа ODANT [Электронный ресурс]. URL: https://legacy.odant.ru/ODANT-PLATFORM.pdf.
7. Франсуа Шолле, Эрик Нильсон, Стэн Бэйлесчи, Шэкуинг Цэй JavaScript для глубокого обучения: TensorFlow.js. – СПб.: Питер, 2021. – 576 с.: ил. – Серия «Библиотека программиста»). ISBN 978-5-4461-1697-3.
8. Шарден, Б. Крупномасштабное машинное обучение вместе с Python: учебное пособие / Б. Шарден, Л. Массарон, А. Боскетти; перевод с английского А. В. Логунова. — Москва : ДМК Пресс, 2018. — 358 с. —ISBN 978-5-97060-506-6. — Текст: электронный // Лань : электронно-библиотечная система. — URL: https://e.lanbook.com/book/105836 (дата обращения: 29.05.2024)
9. Ян Лекун. Как учится машина. Революция в области нейронных сетей и глубокого обучения. (Библиотека Сбера: Искусственный интеллект). — М.: Альпина нон-фикшн, 2021. — ISBN 978-5-907394-29-2.
10. Ярышев С.Н., Рыжова В.А., Технологии глубокого обучения и нейронных сетей в задачах видеоанализа – СПб: Университет ИТМО, 2022. – 82 с.

Ключевые слова
Аппаратно-программный комплекс, датасет, информационно-компонентная модель, искусственный интеллект, нейронная сеть, платформа ODANT, рекуррентная нейронная сеть, сверточная нейронная сеть.

Пузийчук Сергей Иванович, начальник ЦСТС НИИСТ ФКУ НПО «СТиС» МВД России, spuziichuk@mvd.ru
Суменков Николай Александрович, д.т.н., ведущий научный сотрудник ЦСТС НИИСТ ФКУ НПО «СТиС» МВД России, NSumenkov@mvd.ru

Puziychuk S.I., Sumenkov N.A.

Issues of creating a hardware and software complex for video image recognition with artificial intelligence technology based on a domestic platform

Keywords

Hardware and software complex, dataset, information component model, artificial intelligence, neural network, ODANT platform, recurrent neural network, convolutional neural network.

Abstract

The article discusses general approaches to the process of creating a hardware and software complex for recognizing video images using artificial intelligence. The

features of the ODANT technology platform for solving problems related to image processing and image recognition are considered.