Министерство образования Республики Беларусь
Учреждение образования
Белорусский государственный университет
информатики и радиоэлектроники

УДК 004.93:004.382.4

Ковбаса
Галина Александровна

Программный модуль распознавания образов в реальном времени для микрокомпьютеров

АВТОРЕФЕРАТ
на соискание степени магистра технических наук
по специальности 1-40 80 01 «Компьютерная инженерия»

_____

Научный руководитель
Азаров Илья Сергеевич
доктор технических наук, доцент

_____

Минск 2022

ВВЕДЕНИЕ

Проблема распознавания образов приобрела в настоящее время большой масштаб в связи с развитием технологий распознавания образов для авторизации мобильных устройств, автомобильных автопилотов, персональных ассистентов и иных целей. Нейронные сети представляют собой перспективную вычислительную технологию, дающую новые возможности в задачах распознавания образов, детектирования лиц, классификации изображений и иных сферах.

Ввиду невысокой производительности нейронных сетей в целях распознавания образов на встраиваемых решениях, предлагается облегченная модель на основе YOLOv5 для работы в видеопотоке реального времени на маломощных периферийных устройствах, таких как микрокомпьютеры, мобильные телефоны. Разрабатываемый программный модуль должен иметь высокую производительность как на устройствах с производительным GPU, так и без. Целесообразно иметь возможность устанавливать и запускать модуль на различных устройствах как для тестирования, так и дальнейшего применения в готовых системах.

Таким образом, концепция разработки программного модуля распознавания образов велась с учетом следующих принципов:

1.Кроссплатформенность модуля.

2.Применение современных подходов для детектирования объектов.

3.Высокая скорость распознавания на встраиваемых системах без мощного GPU.

Результаты данного научного исследования позволят в дальнейшем применять разработанный технологический подход в системах контроля доступа и мониторинга помещений, для персональных роботов-ассистентов для личного пользования и применения в сферах образования и услуг.

Популярность нейросетевых решений для распознавания образов обеспечивает высокую актуальности работы. Предлагаемая в диссертационной работе комбинация архитектуры сети и последующего способа оптимизации ранее не была представлена, из чего состоит новизна данного подхода

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Цель и задачи исследования

Цель работы – разработка программного модуля распознавания образов для микрокомпьютеров.

Задачи исследования – исследование и анализ существующих подходов компьютерного зрения на основе классических методов и сверточных нейронных сетей, таких как R-CNN, SSD и YOLO, как наиболее перспективных методов распознавания образов в растровом изображении; разработка и внедрение архитектурных и алгоритмических усовершенствований для оптимизации времени исполнения и точности распознавания; реализация программного модуля для решения задачи распознавания образов видеопотока в режиме реального времени для микрокомпьютеров.

Объектом исследования является проблема выполнения существующих методов распознавания образов в режиме реального времени на маломощных и встраиваемых платформах, таких как микрокомпьютеры. В частности, исследуется влияние предлагаемых архитектурных решений нейронных сетей на точность распознавания образов.

Новизна научной работы заключается в решении проблем распознавания образов в режиме реального времени для маломощных и встроенных систем.

Предметом исследования являются архитектурные и алгоритмические решения для сжатия и увеличения производительности программного обеспечения на основе сверточных нейронных сетей.

Основной гипотезой диссертационной работы является возможность оптимизации нейронной сети для использования в качестве детектора и классификатора образов в режиме реального времени для систем на базе микрокомпьютеров.

В качестве критерия наибольшей эффективности алгоритма будут применяться как стандартные методы оценки распознавания изображения, так и оценка производительности программного модуля в рамках микрокомпьютера. В качестве аппаратной базы для исследования была использована система на базе микрокомпьютера Raspberry Pi 4 Model B.

Личный вклад соискателя

Результаты, приведенные в диссертации, получены соискателем лично. Вклад научного руководителя И. С. Азарова, заключается в формулировке целей и задач исследования.

Апробация результатов диссертации

Основные положения диссертационной работы докладывались и обсуждались на 58-ой научно-технической конференции аспирантов, магистрантов и студентов БГУИР; III Международной научно-практической конференции «Компьютерные технологии и анализ данных» (CTDA’2022).

Опубликованность результатов диссертации

По теме диссертации опубликовано 2 печатные работы в сборниках трудов и материалов конференций различного уровня.

Структура и объем диссертации

Диссертация состоит из введения, общей характеристики работы, пяти глав, заключения, списка использованных источников, списка публикаций автора и приложений.

Общий объем работы составляет 69 страниц, из которых основного текста 56 страниц, 31 рисунок на 25 страницах, 4 таблицы на 4 страницах, список использованных источников из 54 наименований на 5 страницах и 4 приложения на 4 страницах.

КРАТКОЕ СОДЕРЖАНИЕ РАБОТЫ

В первой главе представлен анализ предметной области, обозначены основные существующие проблемы в рамках исследования и направления их решения.

Проведен обзор передовых технологий распознавания образов и различных подходов по оптимизации нейронных сетей. С 2012 года начали стремительно появляться новые варианты сверточных нейронных сетей [3]. Современные СНС, применяемые для распознавания образов, такие как Fast R-CNN, Faster R-CNN, RetinaNet, SSD, YOLO, имеют высокую точность распознавания образов. Применение того или иного варианта нейронной сети обосновано различными факторами, такими как целевое устройство, на котором будет в дальнейшем запускаться обученная модель, набор данных, на которых необходимо обучить сеть, время, необходимое на обучение сети. В данном случае сохранение баланса между точностью распознавания и скоростью работы будет иметь решающее значение для выбора сети в качестве основы разрабатываемого программного модуля Для наибольшей достоверности при сравнении результатов выполнения нейронных сетей в данной работе целесообразно сравнивать их на одинаковом наборе данных, таком как Microsoft COCO [26].

Рассматриваемые методы оптимизации будут включать в себя как архитектурные решения, так и различные алгоритмы по уменьшению размера сети и ее параметров.

Один из наиболее эффективных и популярных методов оптимизации архитектуры нейросетей — замена любых сетевых блоков аналогичными, более быстрыми вариантами. СНС состоит из множества слоев: входного и выходного слоев, а также скрытых слоев, которые в свою очередь могут содержать слои активации, слои свертки, слои пулинга и другие.

Для дальнейшей оптимизации нейронной сети будут выполняться методы обрезки сети и квантования. Обрезка — это процесс удаления параметров из существующей нейронной сети, который может включать удаление отдельных параметров или групп параметров, таких как нейроны. Эта процедура направлена на сохранение точности сети при одновременном повышении ее эффективности.

Применение данных способов сжатия нейронной сети потенциально позволит уменьшить вычислительную сложность сети, уменьшив количество параметров и операций, ускорив работу сети и не потеряв при этом высокую точность распознавания.

Для программного обеспечивания, работающего в реальном времени необходимо иметь наилучшее соотношение скорости работы к точности распо-

знавания. Таким образам, в качестве базы для разрабатываемого программного модуля была выбрана нейронная сеть YOLOv5, так как:

– Показывает одно из лучших соотношений скорости работы к точности распознавания на датасете Microsoft COCO 2017.

– Обладает простой архитектурой, которую легко модифицировать.

– Существует ряд высоко оценённых работ по различным вариациям YOLO, такие как tiny-YOLO, YOLOx, YOLOR и другие [4, 8, 30].

– Авторы YOLOv5 обеспечивают постоянную поддержу репозитория с исходным кодом работы [22, 23]. Часто выходят обновления с новыми возможностями для обучения или валидации сети.

Во второй главе описано проектирование требований и методик тестирования для разрабатываемого программного модуля, а также приводятся методы оценки показателей сети.

Были определены средства разработки, настройки окружения и параметры системы, на которой будет производиться разработка программного модуля, обучение нейронной сети, тестирования и валидация полученной модели.

В качестве аппаратной базы для исследования использована система на базе микрокомпьютера Raspberry Pi 4 Model B 4ГБ. Модуль распознавания образов проектируется с учетом возможности встраивания в программную систему проекта «Многозадачный робот. Программная часть» и будет обладать соответствующим интерфейсом, структура которого представлена на рисунке 1.

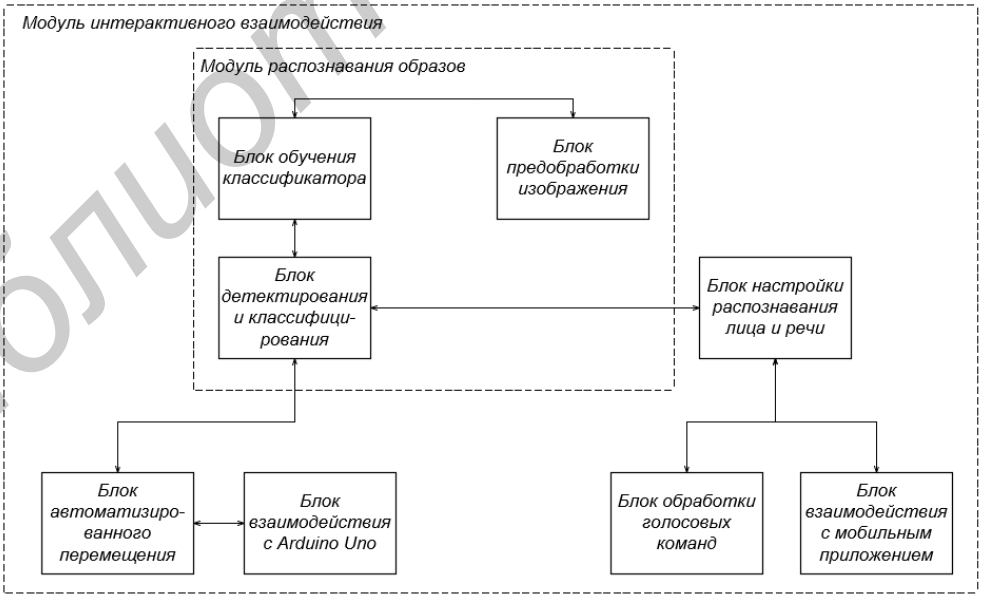

Рисунок 1 – Интеграция структуры программного модуля в аппаратно-программную платформу «Мультизадачный робот»

Данный модуль будет включать в себя такие функциональные возможности как обучение модели, сжатие модели, тестирование на подготовленном да-

тасете, распознавание образов в реальном времени и на отдельных изображениях, функциональные возможности для валидации модели.

Предварительное обучение нейронной сети будет осуществляться на устройстве с GPU NVIDIA GeForce 3060TI 8Gb.

Тестирование обученных моделей и дальнейшая оптимизация производились на устройстве со следующими характеристиками:

Процессор Intel i7-8550U;

Размер оперативной памяти – 16ГБ;

Видеокарта NVIDIA GeForce MX150 1ГБ;

Размер встроенной памяти – 1ТБ;

ОС – Windows 10 x64 версии 21H1.

После выбора конкретной топологии необходимо выбрать гиперпараметры обучения нейронной сети. Стандартные параметры, предлагаемые для сети YOLOv5s, будут взяты за основу для дальнейшего подбора, который будет осуществляться по результатам обучения модели в главе 3.

Были проанализированы различные методы оптимизации процесса обучения сети. Среди которых наиболее важным параметром является регуляризация по норме L1. Для обучения модели был выбран оптимизатор Adam, преимущества которого также были описаны в данной главе.

Были рассмотрены различные методы валидации модели для оценки качества обучения. Будут оцениваться такие метрики модели, как mAP, Inference, количество параметров, размер полученного файла весов, а также графики распределения коэффициентов масштабирования.

В следующей главе описаны эксперименты для изучения влияния предлагаемых архитектурных изменений на качество распознавания сети и время распознавания изображений на тестовой выборке.

По результатам анализа наиболее перспективных решений по распознаванию образов на основе нейронных сетей было решено разрабатывать программное решение на основе нейронной сети YOLOv5 [4]. Представленная в диссертации нейронная сеть будет использовать ShuffleNet V2 в качестве backbone сети для экстракции признаков.

Анализируя методы оптимизации вычислений, примененные в данной сети, следует отметить следующие особенности. Например, в ShuffleNet V2 поэлементное сложение заменено конкатенацией, что аналогично изменению PANet в YOLOv5. Также в ShuffleNet V2 представлена сетевая архитектура быстрого доступа, аналогичная применяемой в сети DenseNet [8]. Кроме того, такие операции как активация ReLU или свертки по глубине, существуют только в одной ветви базового структурного блока, а поэлементные операции конкатенации, перетасовки и разделения каналов объединяются в одну поэлементную операцию.

В экспериментах на влияние архитектурных и алгоритмических модификаций нейронной сети гиперпараметры были выбраны следующие:

– Эпохи. Начальное значение количества эпох – 300 эпох. В большинстве случаев было продолжено обучение до 400-450 эпох.

– Размер изображения. Модель обучается на датасете MS COCO 2017 с исходным разрешением --img 640.

– Размер пакета. В данном случае на устройстве NVIDIA GeForce 3060TI 8Gb максимально возможный размер пакета был равен 8.

– Начальная скорость обучения была выбрана равной 0,01

– Затухание импульса и веса установлены равными 0,9 и 0,0005 соответственно.

– Типы аугментации данных — Mosaic, Crop, Brightness, Exposure, Noise.

Результаты тестирования моделей устройстве с процессором Intel i7-8550U приведены в таблице 1.

Таблица 1 – Результаты тестирования моделей

Модель	Размер изображения	mAP@0.5	mAP@0.5:0.95	Inference	Параметры	Размер файла
YOLOv5s	640*640	56.0	37.2	131ms	7.23M	14M
YOLOv5-Shuffle	640*640	51.6	35.7	120ms	5.39M	10.9M

В результате замены backbone сети в YOLOv5 с CSPDarknet на ShuffleNet V2 количество параметров модели сети было уменьшено на 22 процента, при том, что метрика mAP@0.5:0.95 уменьшилась менее чем на 5 процентов. Время выполнения модели на устройства с Intel i7-8550U было уменьшено с 131мс до 109мс. Таким образом, можно утверждать, что при сохранении высокого уровня точности распознавания сети был достигнут существенный прирост в скорости работы модели на CPU.

В четвертой главе предложена реализация подходов для оптимизирования работы программного модуля для режима реального времени, приводятся результаты экспериментальных исследований.

Структурированная обрезка после обучения сети является наиболее популярным вариантов сжатия предварительно обученной модели. В соответствии с данной стратегией обрезки абсолютные значения каждого веса каждого слоя собираются в список. Там значения веса сортируются в порядке возрастания и выбирается порог, ниже которого будет x% весов, где x — процент прореживания. Затем каждый вес в рассматриваемых слоях обнуляется, если его

абсолютное значение меньше порога.

Однако такой вариант сжатия сети предполагает удалить все веса и связи нейронов за один проход, в соответствии с чем мы не можем корректно оценить, произвели ли мы максимально возможное сжатие сети или превысили порог разреженности, при котором будет частично утеряна точность распознавания.

Поэтому наилучшим вариантом в данном случае будет применение стратегии итеративного обрезки нейронной сети, последовательность операций которого отображена на рисунке 2.

Рисунок 2 – Последовательность операций

В соответствии с работой «Стратегия сжатия нейронных сетей семейства YOLOv5 для встраиваемых решений» данное решение позволит «снизить требования к оборудованию для выполнения детектирования в режиме реального времени» [2-А.].

В данной работе обрезка будет подразделяться на обрезку каналов и обрезку слоев. Данные процедуры анализируют коэффициент γ, поэтому чем выше разреженность весов модели, тем выше будут показатели сжатия сети после обрезки.

Таблица 2 – Результаты обрезки каналов и слоев

Модель	Размер изображения	mAP@0.5	mAP@0.5:0.95	Параметры	Размер файла
YOLOv5s	640*640	56.0	37.2	7.23M	14M
YOLOv5-Shuffle	640*640	51.6	35.7	5.39M	10.9M
YOLOv5-Shuffle-pruned&tuned	640*640	50.1	32.5	3.58M	6.4M

В результате данных процедур размер модели был успешно сжат до 45% от исходной величины, точность осталась практически неизменной, что отображено в таблице 2.

Выполнение квантования оптимизированной нейронной сети позволяет еще больше уменьшить количество параметров и тем самым уменьшить время на обработку кадра. В данной работе квантование нейронной сети будет производиться при помощи методов фреймворка PyTorch.

Таблица 3 – Результаты тестирования моделей на MS COCO 2017

Модель	mAP@ 0.5	mAP@ 0.5:0.95	Параметры	Размер (M)	Inference	GFlops
YOLOv5l	67.3	49.0	46.5M	91.4	320ms	109.1
YOLOv5s	56.8	37.4	7.2M	14.5	131ms	16.5
YOLOv4-tiny	42.0	22.0	5.9M	23.1	125ms	6.9
YOLOv5-Shuffle-P&T	50.1	32.5	3.6M	6.4	65ms	3.4
YOLOv5-Shuffle-q16	49.1	32.0	3.6M	6.4	54ms	2.6
YOLOv5-Shuffle-q8	47.4	27.6	3.6M	3.2	37ms	1.7

Экспериментальные результаты, представленные в таблице 3, свидетельствуют о том, что точность распознавания сети после обрезки каналов остается на уровне 94,2% от базового уровня в том же наборе данных. При этом количество параметров уменьшено на 63,9%, а скорость обнаружения составляет 123,45 кадров в секунду на GPU и 30 кадров в секунду на CPU.

В пятой главе описаны испытания конечного программного модуля на целевом устройстве, а также особенности его реализации. По результатам работы был получен автономный программный модуль, состоящий из трех блоков (блок детектирования и классифицирования, блок обучения классификатора и блок предобработки изображений) и включающий в себя следующие методы:

– методы обучения;

– методы оптимизации нейронной сети в целях снижения вычислительной сложности модели;

– реализация модели нейронной сети;

– методы тестирования полученной модели;

– методы детектирования объектов на входном изображении;

– набор процедур подготовки изображений тестовой и обучающей выборки.

Дальнейшее тестирование программного модуля производилось на целевом устройстве Raspberry Pi 4 Model B 4ГБ. Для тестирования также использовался датасет MS COCO 2017, чтобы сравнить показатели из экспериментов с результатами на целевом устройстве.

Результаты тестирования различных моделей нейронных сетей представлены в таблице 4.

Таблица 4 – Результаты тестирования моделей на MS COCO 2017 на целевом устройстве

Модель	mAP@0.5	mAP@0.5:0.95	Параметры	Размер (M)	Inference
YOLOv5l	64.3	48.0	46.5M	91.4	460ms
YOLOv5s	51.8	36.4	7.2M	14.5	154ms
YOLOv4-tiny	46.0	21.0	5.9M	23.1	130ms
YOLOv5-Shuffle-P&T	49.2	30.5	3.6M	6.4	79ms
YOLOv5-Shuffle-q16	49.1	31.9	3.6M	6.4	72ms
YOLOv5-Shuffle-q8	47.5	28.1	3.6M	3.2	40ms

Результаты работы модуля продемонстрированы ниже на рисунках 3, 4, 5.

Рисунок 3 – Результат работы модуля

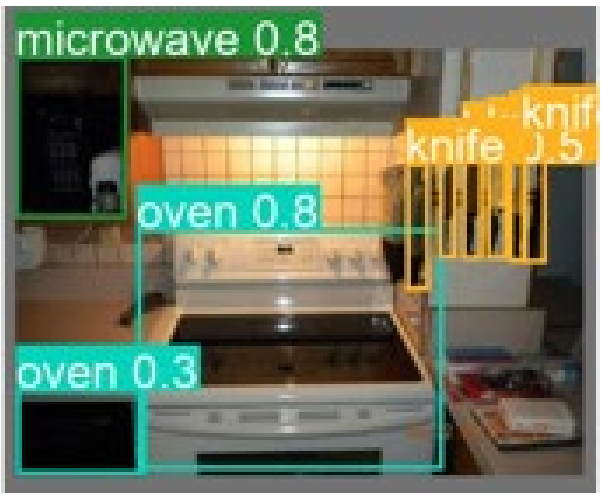

Рисунок 4 – Результат работы модуля

Рисунок 5 – Результат работы модуля

В будущих исследованиях будет продолжаться изучение эффективности алгоритмов квантования и итеративной обрезки для уменьшения их влияния на точность обнаружения.

По результатам проведенного тестирования на целевом устройстве можно сделать вывод, что предложенные методы оптимизации нейронной сети позволяют существенно увеличить скорость обработки изображений на CPU. Наиболее эффективной для применения на микрокомпьютере оказалась модель YOLOv5-Shuffle-q8, квантованная в целочисленный Int8.

ЗАКЛЮЧЕНИЕ

Основные научные результаты диссертации

Нейронные сети – это одна из самых перспективных областей в настоящее время, поскольку в будущем они будут применяться практически повсеместно, в разных областях науки и техники, так как они способны значительно облегчить труд, а иногда и обезопасить человека.

В ходе выполнения магистерской диссертации были получены следующий результаты:

– Изучены и проанализированы современные методы получения признаковых представлений изображений с помощью обучения без учителя.

– На основе проведённого исследования была предложена и реализована новая архитектура сети на основе современной нейронной сети YOLOv5.

– Проведено экспериментальное исследование предложенной архитектуры сети и сравнение ее результатов работы с уже существующими вариантами в изучаемой области.

– Предложенный метод итеративного сжатия нейронной сети алгоритмически реализован.

– Проведено экспериментальное исследование алгоритма, сравнение показателей сжатой нейронной сети на разных этапах выполнения алгоритма с существующими вариантами нейронных сетей семейства YOLO.

Были глубоко изучены сверточные нейронные сети и их основные компоненты. Вся информация об источниках, использованных в работе, приведена в соответствующих главах, результаты анализа предметной области включены в текст работы.

Были рассмотрены различные архитектурные решения и их влияние на производительность и точность распознавания нейронной сети. Выяснилось, что выбранный подход приводит не только к уменьшению размера файла весов, но и к уменьшению количества параметров сети и, как следствие, увеличению скорости работы. Была выбрана наиболее подходящая стратегия для прореживания весов сверточной нейронной сети, алгоритм которой был разработан и применен на практике в дальнейших главах.

Программное обеспечение было спроектировано и разработано для обучения нейронной сети, последующего сжатия и квантования, а также тестирования и валидации параметров полученной модели. Перечисленные функциональные возможности программного модуля были визуально отлажены и протестированы в текущей работе.

Был поставлен эксперимент по изучению зависимости точности распознавания модели от использованной топологии экстрактора признаков. Результаты были обработаны, проанализированы и представлены в тексте диссертации. По результатам первого эксперимента можно сделать вывод, что использование ShuffleNet V2 дает высокие показатели точности распознавания образов и увеличение скорости работы модели, превосходящее показатели исходной сети YOLOv5.

Был проведен эксперимент по изучению влияния разреженности сверточных и Batch Norm слоев на время обработки изображения и точность распознавания объектов в рамках усовершенствованной модели YOLOv5, полученной в результате применения алгоритма итеративной обрезки каналов и слоев. По результатам этого эксперимента было определено наилучшее пороговое значение для разреженного обучения данной сети, разработан алгоритм итеративной обрезки каналов и слоев, проведено сравнение результатов работы сжатой сети с оригинальными моделями. На основе проведенного эксперимента можно сделать вывод, что за счет использования данной стратегии сжатия сети можно добиться более чем двукратного повышения производительности модуля.

По результатам исследования, новая архитектура сети после применения предлагаемых методов оптимизации дает лучшие показатели, чем архитектуры сетей в подобных работах по оптимизации, использующих MobileNetv3 и EfficientNet [55].

Рекомендации по практическому использованию результатов

1. Архитектура YOLO с backbone сетью ShuffleNet V2 показывает высокие результаты в проведенных экспериментах. Данная модель сети может быть использована как самостоятельно в рамках программного модуля для распознавания образов в режиме реального времени, так и для продолжения развития моделей семейства YOLOv5.
2. Результаты показывают, что предложенный в работе метод сжатия нейронной сети семейства YOLOv5 значительно уменьшает количество параметров сети и, как следствие, вычислительную сложность модели, сохраняя при этом баланс между Inference time и mAP@0,5:0,95. Наилучшие результаты при этом показывают модели YOLOv5-Shuffle-q16 на Intel i7-8550U и YOLOv5-Shuffle-q8 на Raspberry Pi 4 Model B 4ГБ. Данный метод сжатия сети также будет работать на любой модели архитектуры YOLO, начиная с 3 версии.
3. Возможность интеграции данного программного модуля в комплексное аппаратно-программное решение повышает ценность данной работы.

Дальнейшее развитие исследования

Полученные экспериментальные данные позволяют сделать вывод о необходимости дальнейших исследований в области экстракторов признаков нейронных сетей для распознавания образов для применения в системах реального времени, а также возможности использования комбинированных подходов в нейронных сетях типа One-Stage Detector.

Запланированы дальнейшие эксперименты по изучению влияния степени разреженности нейронных сетей семейства YOLOv5 на точность распознавания образов.

СПИСОК ОПУБЛИКОВАННЫХ РАБОТ

1- А. Ковбаса, Г.А. Стратегия сжатия нейронных сетей семейства YOLOv5 для встраиваемых решений / Г.А. Ковбаса // Компьютерные технологии и анализ данных (CTDA’2022) : материалы III Междунар. науч.-практ. конф., Минск, 21–22 апр. 2022 г. / Белорус. гос. ун-т ; редкол.: В. В. Скакун (отв. ред.) [и др.]. – Минск : РИВШ, 2022. – С. 49-52.

2- А. Ковбаса, Г.А. Программный модуль для распознавания образов на основе нейронных сетей семейства yolov5 для встраиваемых решений/ Г.А. Ковбаса // Компьютерные системыи сети: материалы 58-ой научной конференции аспирантов, магистрантов и студентов. – Минск: БГУИР, 2022. – с. 24-26.