УДК 004.93

# ПРОГРАММНЫЙ МОДУЛЬ РАСПОЗНАВАНИЯ ОБРАЗОВ ДЛЯ МИКРОКОМПЬЮТЕРОВ НА ОСНОВЕ НЕЙРОННЫХ СЕТЕЙ СЕМЕЙСТВА YOLOV5

*Ковбаса Г.А.[1], магистрант гр.055741*

*Белорусский государственный университет информатики и радиоэлектроники[1]*
*г. Минск, Республика Беларусь*

*Азаров И.С. – докт. техн. наук*

**Аннотация.** Проблема распознавания образов приобрела в настоящее время большой масштаб в связи с развитием автомобильных автопилотов, персональных ассистентов, технологий виртуальной и дополненной реальности. Однако существующие технологии не обладают достаточной эффективностью для быстрого обучения моделей и детектирования образов множества объектов в условиях получения выборки в реальном времени. В данной работе представлен компактный программный модуль на базе комбинации сети YOLOv5 и ShuffleNet V2, размер и количество параметров которого было оптимизированно для последующей работы на встраиваемых системах. Программный интерфейс модуля спроектирован с учетом возможности портирования на различные платформы.

**Ключевые слова.** Распознавание образов, YOLOv5, оптимизация нейронной сети, трансферное обучение.

Существующие технологии детектирования объектов на основе нейронных сетей не обладают достаточной эффективностью на микрокомпьютерах в условиях получения выборки в реальном времени в связи с высокой вычислительной сложностью сети и большим количеством параметров.

Целью данного исследования является решение проблемы распознавания образов в режиме реального времени для маломощных и встроенных систем, а также применение подхода трансферного обучения для быстрого переобучения нейронной сети.

Концепция разработки программного модуля распознавания образов велась с учетом следующих принципов:

1. Кроссплатформенность модуля.
2. Применение современных технических подходов для детектирования объектов.
3. Высокая скорость детектирования на встраиваемых системах без мощного GPU.

Целевой платформой для применения программного модуля распознавания образов является аппаратно-программная платформа на базе микрокомпьютера Raspberry Pi Model 4B «Мультизадачный робот» [1]. Структурная схема интеграции модуля представлена на рисунке 1. Интерфейс для взаимодействия с модулем также предоставляет возможность автономной работы и тестирования вне системы «Мультизадачный робот».

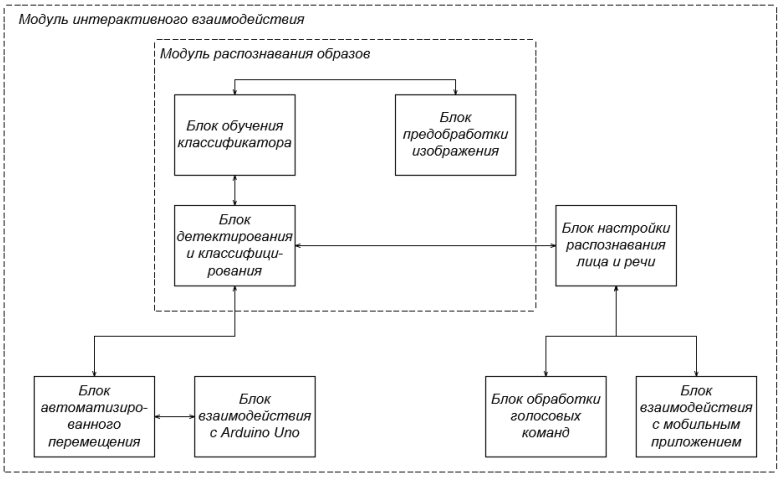


Рисунок 1 – Интеграция структуры программного модуля в аппаратно-программную платформу «Мультизадачный робот»

Программный модуль распознавания образов, как отображено на рисунке 1, состоит из трех блоков. Блок детектирования и классифицирования включает в себя реализацию модели нейронной сети, методы тестирования полученной модели и детектирования объектов входного изображения или видеопотока. Блок обучения классификатора предоставляет методы обучения и

оптимизации нейронной сети. Блок предобработки изображений выделен для процедур подготовки тестовой и обучающей выборки.

1.Архитектура нейронной сети

В основе механизма распознавания образов данного программного модуля лежит нейронная сеть YOLOv5 [2]. В отличии от иных передовых подходов на основе сверточных нейронных сетей, таких как Fast R-CNN и SSD, YOLO предоставляет наиболее эффективную сеть по соотношению скорости работы и точности детектирования, что подтверждается испытаниями модели на датасете MS COCO 2017 [3].

Для обеспечения высокой производительности детектирования на микрокомпьютерах предлагается облегченный вариант сети YOLOv5, backbone часть которой будет заменена на архитектуру сети ShuffleNet V2 [4]. Оригинальная архитектура нейронной сети YOLOv5 представлена на рисунке 2.

Из рисунка 2 можно увидеть, что оригинальная модель использует CSPDarknet в качестве backbone сети. Модели CSP, среди которых и CSPDarknet, основаны на DenseNet [5]. Сеть DenseNet содержит в себе компактно соединенные блоки (dense), соединяющие между собой слои нейронной сети, что позволяет уменьшить проблему исчезающего градиента, поддержать распространение функций и уменьшить количество сетевых параметров [6].

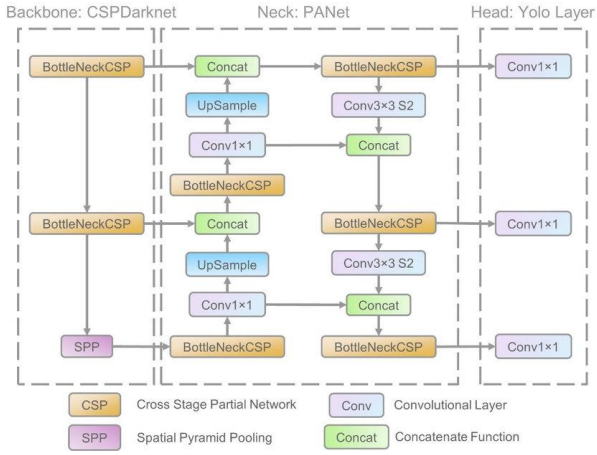


Рисунок 2 – Архитектура оригинального YOLOv5 [7]

ShuffleNet V2 — эффективная CNN для мобильных устройств и встраиваемых систем. Она заимствует сетевую архитектуру быстрого доступа, аналогичную DenseNet. Но в отличии от DenseNet, в ShuffleNet V2 поэлементное сложение заменено конкатенацией, аналогично изменению PANet в YOLOv5 [5]. Использование ShuffleNetV2 в представленной работе позволяет уменьшить среднее время детектирования образов на изображении, при этом точность детектирования изменяется не более чем на 5 процентов по результатам тестирования, что отображено в таблице 1.

2.Сжатие нейронной сети

Для уменьшения количества параметров сети и снижения требований к вычислительной мощности оборудования в данном исследовании используется метод итеративной обрезки каналов и слоев, последовательность которого продемонстрирована на рисунке 3.


Рисунок 3 – Последовательность операций

В соответствии с рисунком 3 процесс сжатия нейронной сети состоит из следующих шагов:

1. Разреженное обучение. Разреженность слоев сети для обрезки каналов обеспечивается применением L1 регуляризации к коэффициентам масштабирования [8]. Степень разреженности сети после обучения напрямую влияет на результат обрезки каналов и слоев.
2. Обрезка каналов.
3. Обрезка слоев.
4. Тонкая подстройка весовых коэффициентов.
5. Квантование.

В процессе обрезки каналов оценивается коэффициент γ заданного слоя и производится удаление весов, которые меньше порогового значения. При обрезке слоев для каждого слоя сортируется среднее значение γ, для сокращения выбирается слой с наименьшей величиной данного коэффициента.

Итерации обучения и обрезки каналов и слоев производятся до достижения максимально возможного сжатия модели при сохранении исходной точности. В результате данных процедур размер модели был успешно сжат до 45% от исходной величины, точность осталась практически неизменной, что отображено в таблице 1.

После сжатия модели было применено статическое квантование при помощи средств фреймворка PyTorch [9]. Для последующей оценки результатов производилось квантование в Float16 и Int8, полученные данные приведены в таблице 1.

## 3.Результаты тестирования на MS COCO

В результате данной работы были получены обученные модели, способные детектировать до 80 классов, представленных в датасете MS COCO 2017 [10]. Для распознавания дополнительных категорий объектов возможно применение трансферного обучения на предварительно обученной модели.

Как показано в таблице 1, значение mAP_0.5 сети YOLOv5-Shuffle-P&T после процедур сжатия уменьшилось менее чем на 5% по сравнению с исходной сетью YOLOv5-Shuffle. В свою очередь точность распознавания квантованной модели YOLOv5-Shuffle-q16 остается на уровне 87,1% от уровня YOLOv5s на датасете MS COCO 2017. При этом количество параметров было уменьшено на 78%. Время детектирования стало на 56,8% меньше, чем у YOLOv4-Tiny, что соответствует требованиям для выполнения детектирования в реальном времени.

Таблица 1 – Результаты тестирования моделей на MS COCO 2017 на устройстве с Intel i7-8550U

| Model | Params | File size | mAP@0.5 | mAP@0.5:0.95 | Inference time | GFlops |
|---|---|---|---|---|---|---|
| YOLOv5l | 46.5M | 91.4 | 67.3 | 49.0 | 320ms | 109.1 |
| YOLOv5s | 7.2M | 14.5 | 56.8 | 37.4 | 131ms | 16.5 |
| YOLOv4-tiny | 5.9M | 23.1 | 42.0 | 22.0 | 125ms | 6.9 |
| YOLOv5-Shuffle | 5.39M | 10.9M | 51.6 | 35.7 | 84ms | 5.9 |
| YOLOv5-Shuffle-P&T | 3.6M | 6.4 | 50.1 | 32.5 | 65ms | 3.4 |
| YOLOv5-Shuffle-q16 | - | 6.4 | 49.1 | 32.0 | 54ms | 2.6 |
| YOLOv5-Shuffle-q8 | - | 3.2 | 47.4 | 27.6 | 37ms | 1.7 |

В будущих исследованиях эффективность алгоритмов квантования и итеративной обрезки каналов и слоев будет продолжать изучаться для уменьшения влияния на точность обнаружения. Результаты данной работы позволят в дальнейшем применять разработанный технологический подход в системах мониторинга помещений, для персональных роботов-ассистентах для личного пользования, в сферах образования и слуг.

***Список использованных источников:***

*1. Мультизадачный робот с функцией слежения за объектом / Ковбаса Г. А. [и др.] // Компьютерные системы и сети : сборник тезисов докладов 56-й научной конференции аспирантов, магистрантов и студентов, Минск, апрель-май 2020 года / Белорусский государственный университет информатики и радиоэлектроники. - Минск : БГУИР, 2020. - С. 20-21.*

*2. YOLOv5. [Электронный ресурс]. – Режим доступа: https://github.com/ultralytics/yolov5. – Дата доступа: 28.12.2021.*

*3. A review: Comparison of performance metrics of pretrained models for object detection using the TensorFlow framework / Sánchez Hernández, Sergio & Romero, H & Morales, A.// IOP Conference Series: Materials Science and Engineering. 844. 012024, June 2020. DOI: 10.1088/1757-899X/844/1/012024.*

*4. ShuffleNet V2: Practical Guidelines for Efficient CNN Architecture Design / Ningning Ma, Xiangyu Zhang, Hai-Tao Zheng and Jian Sun, // arXiv:1807.11164v1 [cs.CV] 30 Jul 2018. DOI: 10.48550/ARXIV.1807.11164*

5. YOLOv5 New Version - Improvements And Evaluation. [Электронный ресурс]. – Режим доступа: https://blog.roboflow.com/yolov5-improvements-and-evaluation/. – Дата доступа: 08.02.2022.
6. Review: DenseNet — Dense Convolutional Network (Image Classification). [Электронный ресурс]. – Режим доступа: https://towardsdatascience.com/review-densenet-image-classification-b6631a8ef803. – Дата доступа: 02.01.2022.
7. A Forest Fire Detection System Based on Ensemble Learning /. Xu, Renjie & Lin, Haifeng & Lu, Kangjie & Cao, Lin & Liu, Yunfei .// Forests. 12. 217, February 2021. DOI:10.3390/f12020217.
8. Регуляризация L1 и L2: определение и применение в разных сферах. [Электронный ресурс]. – Режим доступа: https://codernet.ru/articles/drugoe/regulyarizacziya_l1_i_l2_opredelenie_i_primenenie_v_raznyix_sferax/. – Дата доступа: 06.01.2022.
9. PyTorch. Quantization. [Электронный ресурс]. – Режим доступа: https://pytorch.org/docs/stable/quantization.html. – Дата доступа: 08.02.2022.
10. COCO – Common Objects in Context. [Электронный ресурс]. – Режим доступа: https://cocodataset.org/. – Дата доступа: 15.02.2022.

UDC 004.93

# SOFTWARE MODULE FOR PATTERN RECOGNITION FOR MICROCOMPUTERS BASED ON NEURAL NETWORKS OF YOLOV5 FAMILY

*Kovbasa G.A.*[1]

*Belarusian State University of Informatics and Radioelectronics*[1]*, Minsk, Republic of Belarus*

*Azarov E.S. – Doctor of Technical Sciences*

**Annotation.** The problem of pattern recognition has now acquired a large scale in connection with the development of car autopilots, personal assistants, virtual and augmented reality technologies. However, existing technologies are not efficient enough for fast training of models and detection of images of many objects in the conditions of obtaining a sample in real time. This paper presents a compact software module based on a combination of the YOLOv5 network and ShuffleNet V2, the size and number of parameters of which have been optimized for subsequent work on embedded systems. The program interface of the module is designed with the possibility of porting to different platforms.

**Keywords.** Pattern recognition, YOLOv5, neural network optimization, transfer learning.