ГИБРИДНЫЕ МОДЕЛИ СВЕРТОЧНЫХ НЕЙРОННЫХ СЕТЕЙ YOLO ДЛЯ ОБНАРУЖЕНИЯ И КЛАССИФИКАЦИИ ЛЕТАЮЩИХ ОБЪЕКТОВ НА ИЗОБРАЖЕНИЯХ

Демлер И.С.[1]
Научный руководитель: Марков Н.Г.[2]
[1] *НИ ТПУ, ИШИТР, студент гр. 8ПМ32, e-mail: isd8@tpu.ru*
[2] *НИ ТПУ, д.т.н., профессор ОИТ ИШИТР, e-mail: markovng@tpu.ru*

Аннотация

Для решения задачи объектного детектирования летающих объектов на изображениях предложены две гибридные модели на основе стандартной сверточной нейронной сети YOLOv8s с использованием блоков трансформеров и механизма внимания. Показано, что эти модели по точности детектирования объектов по метрикам $AP_{0,5}$, $mAP_{0,5}$ и $mAP_{0,5\text{-}0,95}$ превосходят стандартную модель, а их производительность близка к ее производительности.

Ключевые слова: детектирование летающих объектов, модель сверточной нейронной сети YOLOv8, визуальный трансформер, механизм внимания

Введение

Объектное детектирование (обнаружение, локализация и классификация) объектов на изображениях является фундаментальной задачей компьютерного зрения, имеющей широкий спектр применений [1]. В частности, актуальным направлением является распознавание летающих объектов (ЛО), таких как беспилотные летательные аппараты (БПЛА), птицы и т. д. с целью повышения безопасности воздушного пространства.

Среди различных моделей сверточных нейронных сетей (СНС) для распознавания ЛО модели YOLO (англ. You Only Look Once) приобрели значительную популярность благодаря своей высокой скорости вычислений и приемлемой точности распознавания объектов на изображениях [2]. Так, например, в работе [3] авторы исследуют эффективность применения как стандартных моделей СНС класса YOLO, так и их модификаций в системах мониторинга воздушного пространства. Однако, несмотря на успехи моделей YOLO, детектирование ЛО на изображениях по-прежнему остается сложной задачей. Такие факторы, как малый размер ЛО и плохое качество изображений требуют разработки более совершенных моделей YOLO.

Цель данной работы заключается в создании гибридных моделей СНС на основе стандартной модели YOLOv8s класса YOLO с использованием блоков трансформеров и механизма внимания и в последующем исследовании их эффективности при решении задачи детектирования на изображениях ЛО различных размеров и ЛО малых размеров.

Используемые датасеты

Для обучения, валидации и тестирования предлагаемых моделей СНС были использованы два модифицированных нами исходных датасета с изображениями размером 416×416 пикселей: airspace и airspace-small. В первом датасете объекты на размеченных изображениях имеют разные размеры (“смешанные по размерам объекты”), причем на одном изображении может быть один или несколько объектов, при этом они могут быть разных классов. На изображениях второго датасета так же могут быть один и более объектов разных классов, но все они имеют малые размеры. ЛО представлены тремя классами: БПЛА самолетного типа (aircraft-type_uav), БПЛА вертолетного типа (helicopter-type_uav) и Птица (bird). Все объекты разделены на три категории по размерам: малые ЛО – площадь не более 32×32 пикселей); средние – от 33×33 до 96×96 пикселей и большие объекты с площадью от 97×97 пикселей и более.

Дополнение имеющихся в нашем распоряжении исходных датасетов (модернизация их) изображениями проводилось с помощью системы Roboflow. При создании этих модифицированных датасетов использовался принцип максимальной реалистичности: изображения должны были точно отражать реальные ситуации в небе (на них присутствовали один или несколько ЛО заданных классов, различный фон и т. д.) Также, для увеличения количества данных, часть изображений была подвергнута аугментации. Примеры размеченных изображений представлены на рисунке 1.

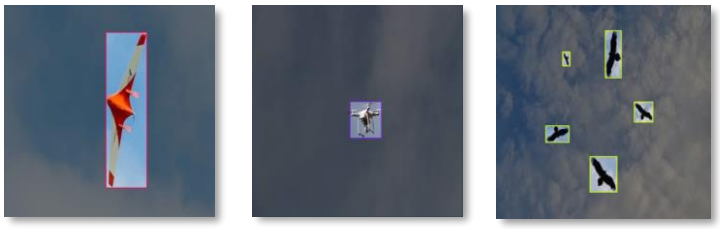

Рис. 1. Примеры аннотированных летающих объектов

В результате получено два модифицированных датасета, состоящих из 2748 изображений с объектами малых размеров и 4011 изображений с объектами “смешанных” размеров. Все изображения каждого из них были разделены на обучающую, валидационную и тестовую выборки в соотношении 70/20/10%. Распределение по классам ЛО для каждого из датасетов представлено в виде столбчатой диаграммы на рисунке 2.

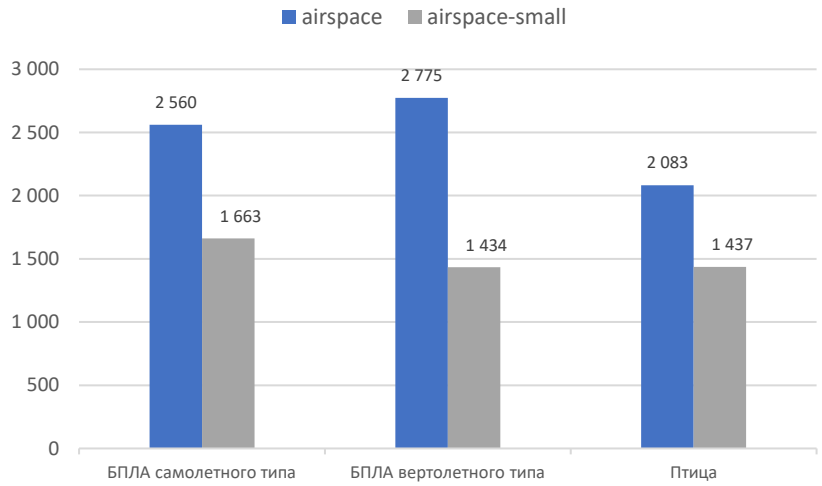

Рис. 2. Распределение летающих объектов по классам в датасетах

Стандартная модель СНС YOLOv8 и гибридные модели на ее основе

Архитектура стандартной модели СНС YOLOv8 представлена тремя базовыми компонентами [4]: “позвоночником” (англ. Backbone), “шеей” (англ. Neck) и “головой”.

Также в рамках исследования была рассмотрена гибридная модель LeYOLOs (версия small), предложенная в [5] и спроектированная на базе модели YOLOv8s. Она имеет облегченную архитектуру “позвоночника” на основе сети MobileNet, упрощенную схему “шеи” с удалением лишних ресурсоемких слоев и связей, а также разделенную сетевой структурой “голову” (англ. DNiN - Decoupled Network-in-Network) с использованием поточечных сверточных слоев (англ. pointwise convolutions). Авторы полагают, что она является более легковесной и точной по сравнению со стандартной моделью YOLOv8. Эти положения подлежат нашему исследованию при решении задачи детектирования ЛО.

Нами была предложена гибридная модель MobileViT-YOLOv8. Она использует в качестве базовой стандартную модель СНС YOLOv8s. Основным изменением является применение в качестве “позвоночника” MobileViT-блоков [6] – трансформеров,

представляющих собой облегченную и более эффективную версию стандартного визуального трансформера (англ. Visual Transformer). В таблице 1 представлены все изменения, реализованные в ходе разработки гибридной модели MobileViT-YOLOv8.

Таблица 1. Перечень изменений архитектуры модели СНС YOLOv8, представленные в MobileViT-YOLOv8

Компонент	Изменение	Пояснение
Backbone	CSPDarknet заменена на MobileViTv3 (версия XS)	MobileViTv3 является легковесным, более точным и производительным по сравнению со стандартным визуальным трансформером. Использование этого трансформера в “позвоночнике” вместо стандартного CSPDarknet [4] позволит оптимизировать и улучшить процесс извлечения признаков за счет более точного учета локального и глобального контекстов.
Neck	PANet заменена на PANet-P2	Стандартная PANet, составляющая “шею”, была дополнена связью с P2-уровнем “позвоночника”. Это позволяет учитывать в “шее” более низкоуровневые карты признаков, что, в свою очередь, обеспечивает лучшую обобщающую способность при локализации объектов малого размера.
Блоки свертки	Блок C2f заменен на C2fSimAM и C2fGhost, базовый блок свертки заменен на GhostConv	В блоке C2f был заменен модуль обычный модуль Bottleneck, состоящий из стандартных блоков свертки, на GhostBottleneck [7], основанный на использовании разделения обычной свертки на несколько шагов для того, чтобы уменьшить избыточность карт признаков, тем самым сократив количество вычислений. Также на выходе данного блока был установлен упрощенный механизм внимания, SimAM (англ. simple, parameter-free attention module) [8], позволяющий точнее выделять ключевые регионы на картах признаков.
Функция активации	Сигмоидальноая функция активации (SiLU) заменена на GELU	В блоках свертки сигмоидальная функция была заменена на GELU, которая по форме кривой представляет ту же сигмоиду, однако обладает меньшей вычислительной сложностью, что делает ее выгодной при оптимизации работы сети.

На рисунке 3 с учетом изменений и пояснений к ним из таблицы 1 схематично представлена архитектура предлагаемой гибридной модели СНС MobileViT-YOLOv8.

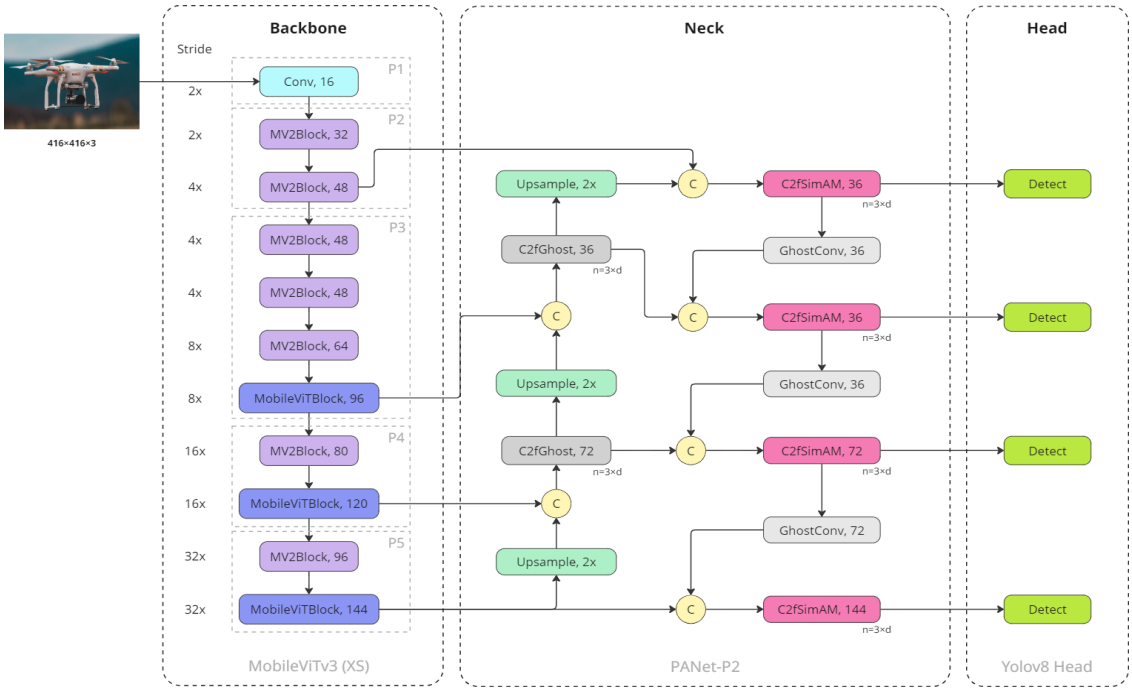

Рис. 3. Архитектура гибридной модели СНС MobileViT-YOLOv8

Обучение и валидация стандартной и гибридных моделей YOLOv8

Разработка и настройка рассматриваемых моделей СНС семейства YOLOv8 были проведены с использованием языка программирования Python 3.7 и библиотеки глубокого обучения PyTorch. Для оценки точности работы моделей использовались метрики: Precision, Recall, $AP_{0,5}$ (англ. Average Precision), $mAP_{0,5}$ (англ. mean Average Precision) и $mAP_{0,5\text{-}0,95}$. Все вычисления проводились с использованием графического процессора NVIDIA RTX A4000 с 16 ГБ видеопамяти и 24 ГБ оперативной памяти.

Обучение и валидация моделей СНС проводились при следующих настройках:

- epochs: 250 – количество эпох обучения;
- patience: 15 – допустимое количество эпох обучения без улучшения результата;
- batch: 16 – размер порции (батча) одновременно обрабатываемых изображений;
- conf_thres: 0,25 – доверительный порог;
- iou_thres: 0,6 – порог IoU (англ. Intersection over Union) для срабатывания NMS (англ. Non-Maximum Suppression);
- optimizer: Adam – оптимизатор значений гиперпараметров в ходе обучения;
- imgsz: 416 – целевой размер изображений;
- augment: True – необходимость дополнительной аугментации данных.

Результаты исследования стандартной и гибридных моделей СНС

После обучения и валидации вышеупомянутых стандартной YOLOv8s (версия small) и гибридных моделей LeYOLOs (версия small) и разработанной MobileViT-YOLOv8 были выполнены предсказания на тестовых выборках исходных датасетов. Результаты по точности детектирования по метрике $AP_{0,5}$ по каждому классу ЛО и по метрике $mAP_{0,5}$ по всем классам ЛО для каждого из датасетов представлены в таблице 2 и таблице 3.

Таблица 2. Результаты точности детектирования ЛО стандартной и гибридными моделями YOLOv8 на тестовой выборке датасета airspace

Класс	**$AP_{0,5}$, $mAP_{0,5}$**		
	YOLOv8s	**LeYOLOs**	**MobileViT-YOLOv8**
БПЛА вертолетного типа	0.941	0.940	0,979
БПЛА самолетный типа	0.930	0.960	0,944
Птица	0.901	0.913	0,928
Все	0.924	0.938	0.950

Таблица 3.

Результаты точности детектирования летающих объектов стандартной и гибридными моделями YOLOv8 на тестовой выборке датасета airspace-small

Класс	**$AP_{0,5}$, $mAP_{0,5}$**		
	YOLOv8s	**LeYOLOs**	**MobileViT-YOLOv8**
БПЛА вертолетного типа	0,932	0,890	0.974
БПЛА самолетный типа	0,919	0,914	0.927
Птица	0,862	0,843	0.873
Все	0,904	0,882	0,927

Из представленных выше таблиц 2 и 3 видно, что обученные и валидированые модели точнее работают при детектировании объектов “смешанных” размеров, причем лучше всего они определяют класс “БПЛА вертолетного типа”, а хуже всего – класс “Птица”. В каждом из двух случаев разработанной гибридной модели MobileViT-YOLOv8 удалось улучшить точность детектирования объектов класса “Птица” относительно стандартной YOLOv8s и гибридной LeYOLOs.

Также была получена усредненная статистика по всем метрикам точности: Precision, Recall, $mAP_{0.5}$, $mAP_{0.5\text{-}0,95}$. Результаты представлены в таблице 4.

Таблица 4. Сводная таблица точности детектирования ЛО с помощью стандартной и гибридных моделей СНС YOLOv8

Модель	Precision	Recall	$mAP_{0,5}$	$mAP_{0,5-0,95}$
YOLOv8s (airspace)	0.890	0.930	0.924	0.724
YOLOv8s (airspace-small)	0,877	0,843	0,904	0,643
LeYOLOs (airspace)	0.900	0.916	0.938	0.731
LeYOLOs (airspace-small)	0,860	0,820	0,882	0,670
MobileViT-YOLOv8 (airspace)	0.940	0.918	0.950	0.751
MobileViT-YOLOv8 (airspace-small)	0,900	0,861	0,927	0,680

** зеленым цветом выделены лучшие результаты для датасета airspace, синим – для датасета airspace-small*

Помимо точности детектирования ЛО с помощью представленных моделей была оценена их производительность и ресурсоемкость. В частности, были получены оценки количества параметров (в миллионах, М), GFLOPs (англ. Giga-FLoating-point OPerations per Second – миллиардов операций с плавающей точкой в секунду), FPS (англ. Frames per Second – количество кадров в секунду), а также объем в мегабайтах (МБ) дискового пространства, необходимого для хранения весовых коэффициентов и промежуточных буферов моделей. В таблице 5 представлены значения этих показателей для каждой из исследуемых моделей.

Таблица 5. Сводная таблица производительности стандартной и гибридных моделей СНС YOLOv8

Модель	FPS	Количество параметров, М	GFLOPs	Размер модели, МБ
YOLOv8s	67.32	11,2	28.4	21.5
LeYOLOs	61.61	1,9	4.36	3.9
MobileViT-YOLOv8	58.12	2,9	17.8	7,2

Анализ результатов исследования и выводы

Разработанная гибридная модель MobileViT-YOLOv8 позволила повысить точность детектирования отдельных классов ЛО. Так, например, “трудноразличимый” стандартной моделью класс “Птица” данная модель определяет точнее на 2,7% в рамках датасета со смешанными объектами и на 1,1% для датасета с малоразмерными объектами.

Результаты, представленные в таблице 4, позволяют сделать вывод о том, что гибридные модели, в частности разработанная нами MobileViT-YOLOv8, в целом дают более точные результаты детектирования по сравнению с результатами, получаемыми с помощью стандартной модели СНС YOLOv8s. Это видно из анализа значений метрик $mAP_{0.5}$ и $mAP_{0.5-0.95}$ – гибридной модели MobileViT-YOLOv8 удалось улучшить результаты по точности детектирования ЛО на величину от 2 до 4%.

Из таблицы 5 следует, что гибридные модели являются более легковесными и производительными, чем стандартная модель. Об этом говорят низкие показатели GFLOPs, малое количество параметров (в несколько раз меньше, чем у стандартной модели) и меньший объем дискового пространства, необходимого для хранения весовых коэффициентов моделей. Однако, по скорости вычислений первое место все еще занимает стандартная модель, хотя гибридные модели и имеют вполне конкурентную с ней скорость вычислений, что видно по значениям метрики FPS.

Заключение

В работе было проведено исследование эффективности стандартной модели СНС YOLOv8s и разработанных на ее основе двух гибридных моделей СНС при решении задачи объектного детектирования на изображениях ЛО трех классов. Особенностью разработанной

гибридной модели MobileViT-YOLOv8 является использование в качестве “позвоночника” MobileViT-блоков – трансформеров. Перед проведением исследований стандартная модель СНС YOLOv8s и гибридные модели СНС на ее основе были обучены и валидированы на датасете с изображениями ЛО “смешанных” размеров и на датасете с изображениями ЛО малых размеров. Исследования этих обученных моделей по точности детектирования ЛО на изображениях и по скорости вычисления моделей проводились с использованием тестовых выборок этих датасетов.

В целом исследования показали, что гибридные модели дают более точные результаты детектирования ЛО по сравнению с результатами, получаемыми с помощью стандартной модели СНС YOLOv8s, и имеют вполне конкурентную с ней скорость вычислений. Это позволяет сделать вывод о перспективности использования этих гибридных моделей в системах компьютерного зрения реального времени, предназначенных для мониторинга воздушного пространства с целью обнаружения и классификации ЛО.

Список использованных источников

1. Pathak, A. R., Pandey, M., & Rautaray, S. (2018). Application of deep learning for object detection. Procedia Computer Science, 132, 1706–1717. doi.org/10.1016/j.procs.2018.05.144.

2. Terven, J.; Córdova-Esparza, D.-M.; Romero-González, J.-A. A Comprehensive Review of YOLO Architectures in Computer Vision: From YOLOv1 to YOLOv8 and YOLO-NAS. Mach. Learn. Knowl. Extr. 2023, 5, 1680–1716. https://doi.org/10.3390/make5040083.

3. Nebaba, S. G., & Markov, N. G. (2024). Convolutional neural networks of YOLO family for mobile computer vision systems. Computer Research and Modeling, 16(3), 615–631. https://doi.org/10.20537/2076-7633-2024-16-3-615-631.

4. Yaseen, M. (2024). What is YOLOv8: An In-Depth Exploration of the Internal Features of the Next-Generation Object Detector. arXiv preprint arXiv:2408.15857.

5. Hollard, Lilian & Mohimont, Lucas & Vaillant-Gaveau, Nathalie & Steffenel, Luiz Angelo. (2024). LeYOLO, New Scalable and Efficient CNN Architecture for Object Detection. 10.48550/arXiv.2406.14239.

6. Wadekar, S.N., Chaurasia, A. (2022). MobileViTv3: Mobile-Friendly Vision Transformer with Simple and Effective Fusion of Local, Global and Input Features. arXiv preprint arXiv:2209.15159.

7. Han, K., Wang, Y., Tian, Q., Guo, J., Xu, C., Xu, C. (2019). GhostNet: More Features from Cheap Operations. arXiv preprint arXiv:1911.11907.

8. Wang, J., Wu, J., Wu, J., Wang, J., & Wang, J. (2023). YOLOv7 Optimization Model Based on Attention Mechanism Applied in Dense Scenes. In Applied Sciences (Vol. 13, Issue 16, p. 9173). MDPI AG. doi.org/10.3390/app13169173.