Модули внимания в сверточных нейронных сетях для распознавания малоразмерных объектов

Д.И. Краснов [1]
[1] *Университет ИТМО,*
197101, Россия, г. Санкт-Петербург, Кронверкский пр., д. 49

Аннотация

Задача распознавания малоразмерных объектов часто встречается в биомедицинских системах и системах безопасности. При этом обнаружение таких объектов часто осложняется наличием плотных облаков или объектов инфраструктуры. В данной работе представлены результаты использования различных механизмов внимания для повышения точности в задаче сегментации малоразмерных объектов на изображении с помощью сверточных нейронных сетей. Были рассмотрены модули внимания по каналам и по пикселам. Подобный подход позволяет эффективно подавлять менее информативные каналы и области изображения и усиливать более информативные каналы и области изображения. При этом весовые коэффициенты в модулях внимания автоматически адаптируются к обучающим данным. Проведена оценка влияния механизмов внимания в архитектуре сверточной нейронной сети на ее способность подавлять сложный фон (облака, тучи и объекты инфраструктуры) и сегментировать малоразмерные объекты. Результаты представлены в виде таблиц с тестовыми метриками, графиков precision-recall и ROC-кривых и тепловых карт, показывающих эффективность подавления фона. Полученные результаты позволяют эффективно внедрять описанные модули внимания в сверточные нейронные сети любой сложности для повышения точности распознавания объектов размером 10 – 40 пикселей на сложном фоне.

Ключевые слова: сегментация, малоразмерный объект, сверточная нейронная сеть, модуль внимания, компьютерное зрение.

Цитирование: **Краснов, Д.И.** Модули внимания в сверточных нейронных сетях для распознавания малоразмерных объектов / Д.И. Краснов // Компьютерная оптика. – 2024. – Т. 48, № 6. – С. 963-968. – DOI: 10.18287/2412-6179-CO-1468.

Citation: Krasnov DI. Attention modules in convolutional neural networks for small object recognition. Computer Optics 2024; 48(6): 963-968. DOI: 10.18287/2412-6179-CO-1468.

Введение

В настоящее время системы компьютерного зрения на основе сверточных нейронных сетей используются во многих областях человеческой деятельности. Например, в биомедицинских системах для сегментации новообразований на снимках, полученных путем магнитно-резонансной томографии [1], или в системах безопасности для определения посторонних людей или объектов в защищенной зоне [2].

В описанных выше системах часто встречается задача распознавания маленьких объектов размером около 10 – 40 пикселей. Это могут быть небольшие кровяные клетки, опухоли или удаленные от объектива камеры летающие объекты (самолеты, вертолеты или беспилотные летательные аппараты). Маленькие объекты содержат в себе очень мало семантической информации, что затрудняет их определение даже человеческим глазом. Кроме того, малоразмерные объекты часто находятся на сложном фоне, например летательные аппараты на фоне кучевых облаков или объектов инфраструктуры прилегающей зоны. Все описанные выше факторы значительно затрудняют определение таких объектов в автоматических системах компьютерного зрения.

Существует множество принципиально различных подходов к распознаванию малых объектов, которые имеют свои достоинства и недостатки.

Одним из наиболее простых подходов в системах безопасности и специального назначения является использование классических алгоритмов обработки изображения. Это позволяет подавлять фон и выделять объект с помощью различных видов фильтрации: разность фильтров Гаусса [3], локальная контрастная фильтрация [4] или вейвлет-преобразование [5]. Стоит отметить, что такой подход наиболее часто применяется в обработке изображений инфракрасного диапазона, потому что объект на таком изображении имеет вид пятна, яркость которого может значительно отличаться от яркости фона.

Другим подходом является использование последовательности кадров аналогично зрительному аппарату человека. Это позволяет применять алгоритмы вычитания фона с последующей сегментацией [6], сверточные нейронные сети с блоками долгой кратковременной памятью [7] или классификаторы для анализа траектории полета объекта [8].

Наиболее эффективным подходом является использование сверточных нейронных сетей для обра-

ботки одного кадра. В таком подходе часто используют вспомогательную сеть для повышения разрешения исследуемой области [9] или дополнительные блоки для подавления менее информативных областей и усиления более информативных областей [10]. Кроме того, возможно использование изображений инфракрасного диапазона [11 – 12], что позволяет генерировать обучающие данные в процессе обучения, помещая яркое пятно на инфракрасное изображение фона.

Следует отметить, что традиционные архитектуры нейронных сетей не способны справиться с задачей обнаружения малоразмерных объектов с достаточной точностью, поэтому данная задача остается актуальной. В работе исследованы блоки пространственного и межканального внимания для повышения точности распознавания малых объектов. Проведен эксперимент, показывающий повышение точности нейронной сети на реальных данных. Исследована способность нейронной сети подавлять сложный фон и концентрироваться на области, содержащей искомый объект. Результаты представлены в виде таблицы с тестовыми метриками (F-мера, мера Жаккара, ROC AUC и Average Precision), графиков с precision-recall кривыми и тепловых карт, полученных путем усреднения тензоров с последних слоев исследуемых нейронных сетей.

1. Набор данных

Для проведения исследования собран набор данных, содержащий 12242 изображения в видимом диапазоне с малоразмерными объектами в виде самолетов, вертолетов и беспилотных летательных аппаратов (БПЛА) размером от 5 до 70 пикселей. Набор данных был получен путем объединения трех других, находящихся в свободном доступе, с последующим исключением изображений, содержащих слишком большие объекты (больше 70 пикселей). Первый набор данных Purdue UAV Dataset [13] содержит в себе 50 размеченных видеофайлов с БПЛА, снятых в полете с установленной на одном из них камеры с разрешением 1920×1080. Эти видеофайлы были покадрово преобразованы в изображения, затем каждое изображение разделено на 9 перекрывающихся областей, для каждой из которых созданы маски, содержащие 0 на пикселах фона и 1 на пикселах объекта. Второй набор данных [14] содержит в себе видеофайлы с самолетами, вертолетами и дронами в разрешении 640×512. Видеофайлы были покадрово преобразованы в изображения, к которым были созданы маски аналогично описанному выше принципу. Третий набор данных [15] содержит размеченные видеофайлы с радиоуправляемыми самолетами и дронами, снятые с земли и воздуха, с разрешением 1280×720. Примеры полученных изображений с сегментационными масками представлены на рис. 1. Распределение объектов по их размеру в пикселях в итоговом наборе данных представлено на рис. 2.

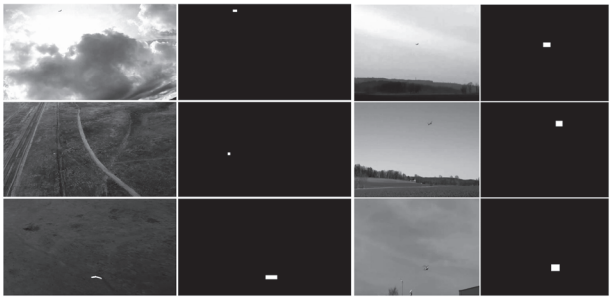

Рис. 1. Примеры используемых изображений

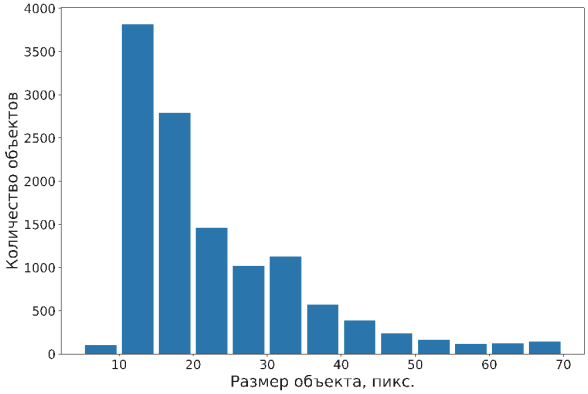

Рис. 2. Гистограмма распределения размера объектов в наборе данных

Описанные выше изображения были разделены на обучающую, валидационную и тестовую выборки по 8569, 2460 и 1213 изображений соответственно. В процессе обучения была использована аугментация изображений с помощью аффинных преобразований (повороты и отражения), моделирования погодных эффектов (дождь, снег и засветка) и моделирования помех камеры и оптической системы (шум, дефокусировка и смаз).

2. Обучаемые модели

В качестве базовой нейронной сети была выбрана симметричная сеть для семантической сегментации ESNet [16]. Архитектура базовой сети состоит из блоков (рис. 3), которые соединены последовательно. Блоки на схеме изображены условно, поскольку их структура не играет роли в рамках данного исследования. Базовая модель имеет типичную для задачи сегментации структуру энкодер-декодер, где энкодер отвечает за извлечение семантических признаков из входного изображения, а декодер за использование извлеченных признаков для сегментации объекта.

Наиболее часто применяемым методом повышения точности с минимальным изменением архитектуры нейронной сети является слияние признаков (feature fusion) [17]. Это позволяет использовать извлеченные семантические признаки с различными уровнями глубины одновременно. Выходные тензоры с нескольких последних блоков приводятся к одному разрешению с помощью интерполяции (upsampling) и соединяются вдоль каналов перед заключительным сверточным слоем (рис. 4). В данной работе для сли-

яния признаков использовались выходные тензоры от блоков 3, 4 и 5 декодера базовой сети.

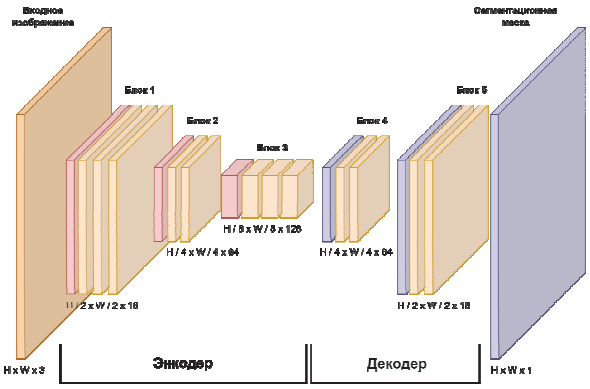

Рис. 3. Архитектура базовой модели

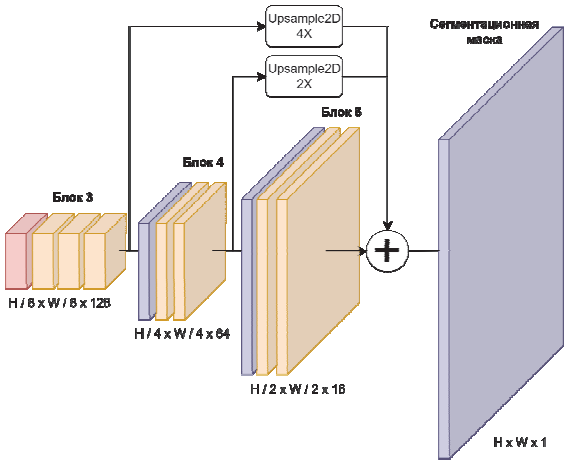

Рис. 4. Архитектура декодера модели с модулем слияния признаков

Исследуемые блоки межканального и пространственного внимания [18] имеют следующие архитектуры (рис. 5 и 6). Эти блоки являются самостоятельными единицами и могут быть встроены в нейронную сеть практически любой топологии. Принцип работы блока межканального внимания состоит в преобразовании входного тензора в одномерный вектор весов путем применения глобального пулинга (global max pooling или global average pooling). Длина вектора соответствует количеству каналов во входном тензоре. Полученный вектор весов проходит через последовательность сверточных слоев (Conv2D) с единичным ядром, что эквивалентно полносвязной сети, и нормируется с помощью сигмоидной функции (sigmoid). После этого каждый канал входного тензора умножается на соответствующий ему вес из вектора весов. Поскольку веса адаптируются к входным данным в процессе обучения, нейронная сеть будет автоматически подавлять менее информативные каналы и усиливать более информативные. Принцип работы блока пространственного внимания состоит в преобразовании входного тензора в нормированную маску весов с одним каналом и исходным разрешением с помощью последовательности сверточных слоев с функцией активации (ReLU) и сигмоидной функции. После этого полученная маска поэлементно умножается на входной тензор, в результате чего подавляются менее информативные области (фон) и усиливаются более информативные (объект). Для совместного использования двух блоков их выходные тензоры суммируются, т.к. они имеют одинаковые размерности.

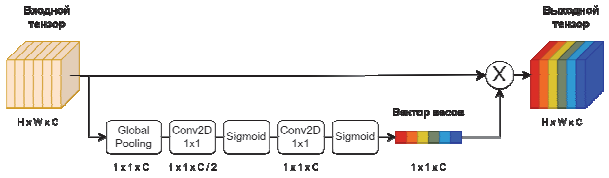

Рис. 5. Архитектура блока межканального внимания

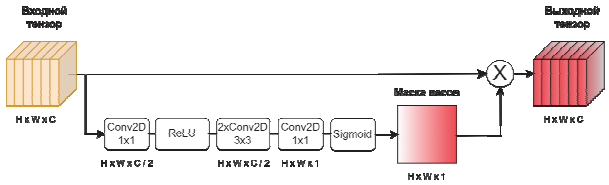

Рис. 6. Архитектура блока пространственного внимания

Для проведения эксперимента в топологию базовой сети были добавлены модуль слияния признаков, модуль межканального внимания, модуль пространственного внимания и модули пространственного и межканального внимания одновременно. Описанные выше блоки внедрены в декодер базовой модели после блоков 3, 4 и 5. Таким образом, были получены 5 архитектур сверточных нейронных сетей.

Полученные модели были обучены на выборке с использованием разрешения 512×512. Обучение проводилось на NVIDIA RTX 3060 12GB со следующими параметрами:

– размер пакета – 16;
– количество эпох – 200;
– оптимизатор Adam с learning rate 0,001 и экспоненциальным уменьшением 0,9;
– функция потерь dice loss [19].

В процессе обучения контролировались метрики по валидационной выборке: точность (precision), полнота (recall), IoU [20] (мера Жаккара) и average precision (AP).

3. Результаты

Результаты обучения пяти описанных выше моделей приведены в приложении А в табл. 1А. Значения функции потерь для обучающей выборки и метрик для валидационной выборки указаны для последней эпохи.

После обучения пяти моделей было проведено тестирование на выборке, которая не участвовала в обучении и валидации. Результаты тестирования моделей представлены в виде графиков precision-recall кривых (рис. 7) и ROC-кривых (рис. 8).

Кроме того, остальные метрики: f-мера, IoU, площадь под ROC-кривой (ROC AUC) и площадь под PR-кривой (Average Precision) для тестовой выборки были собраны в табл. 1.

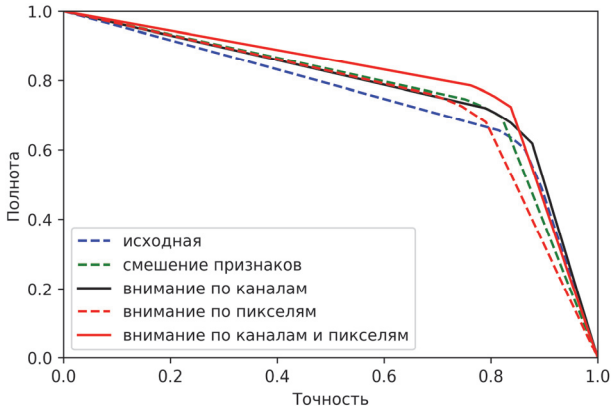

Рис. 7. Precision-recall кривые для тестовой выборки

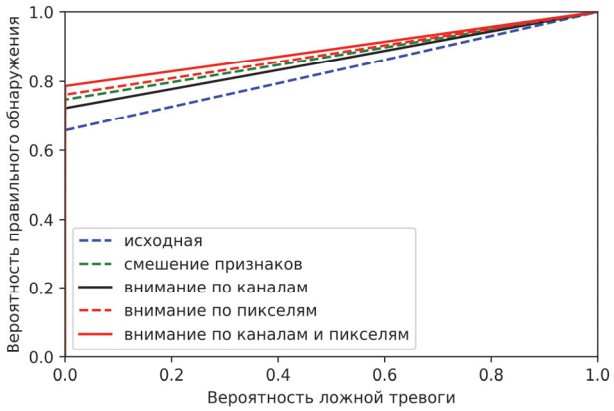

Рис. 8. ROC-кривые для тестовой выборки

Табл. 1. Метрики для тестовой выборки

Модель	F-мера	IoU	ROC AUC	AP
Исходная	0,706	0,547	0,824	0,571
Слияние признаков	0,750	0,595	0,876	0,613
Внимание по каналам	0,742	0,590	0,864	0,620
Внимание по пикселям	0,715	0,552	0,873	0,582
Внимание по каналам и пикселям	0,761	0,615	0,891	0,639

По графикам precision-recall и ROC-кривых можно заметить, что блок слияния признаков дает большое улучшение в точности сегментации объектов. Следует отметить, что такой блок практически не влияет на скорость работы модели, т.к. вносит всего лишь несколько дополнительных слоев. Блоки межканального и пространственного внимания по отдельности дают небольшой прирост точности, меньший, чем блок слияния признаков. Однако их комбинация позволяет увеличить точность модели по AP и IoU более чем на 10 % (табл. 2), что дает большее улучшение, чем модуль слияния признаков. По рис. 5 и 6 можно заметить, что описанные блоки немного усложняют архитектуру модели, однако они могут быть внедрены практически в любую топологию сверточной нейронной сети без существенного ухудшения производительности.

Для того чтобы оценить эффективность блоков внимания в задаче подавления сложного фона и локализации объекта, были построены тепловые карты, полученные путем усреднения выходного тензора с предпоследних слоев моделей (рис. 9*а* – *в*). Они позволяют визуально оценить, как модель с пространственными и межканальными блоками внимания подавляет области с фоном и уменьшает шумы, локализуя объект (рис. 9*в*) за счет назначения большего веса информативным областям. Можно заметить, что область объекта в базовой модели определяется нечетко (рис. 9*б*), в то время как в улучшенной модели границы объекта более яркие и резкие.

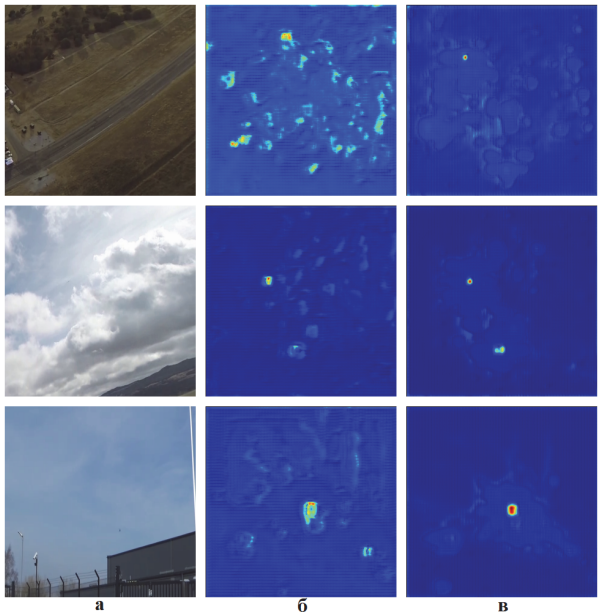

Рис. 9. Тепловые карты внимания (а – входное изображение, б – базовая модель и в – модель с межканальным и пространственным вниманием)

Однако даже в эффективных блоках внимания может быть ложное срабатывание. На рис. 9*в* для второго изображения видно, что модель с межканальным и пространственным вниманием ошибочно назначает большой вес области, не содержащей искомый объект. При этом внимание к истинному объекту значительно выше, что позволяет судить о высокой эффективности блоков внимания в задаче обнаружения малоразмерных объектов.

Заключение

В работе освещена проблема точного определения малоразмерных объектов на изображении с помощью сверточных нейронных сетей в различных системах компьютерного зрения и методы ее решения. Проведено исследование о влиянии внедрения блоков внимания в архитектуру сверточной нейронной сети на точность распознавания малоразмерных объектов (удаленных летательных аппаратов) в системах безопасности. Блоки внимания были внедрены в базовую нейронную сеть для семантической сегментации ESNet, в результате чего были получены 4 новые модели. Эти модели были обучены на смешанном наборе данных с малоразмерными летательными аппаратами. Результаты

тестирования обученных моделей приведены в виде графиков precision-recall и ROC-кривых и таблиц с метриками f-меры, IoU, ROC AUC и AP. Результаты показали, что наилучший прирост точности по Average Precision и IoU (более 10%) обеспечило внедрение блоков межканального и пространственного внимания одновременно. Блок слияния признаков позволил получить лучшую точность по сравнению с отдельным использованием блоков межканального и пространственного внимания (IoU 0,595 против 0,590 и 0,552). По полученным тепловым картам можно судить о высокой эффективности подавления фона и локализации искомых объектов. Описанные блоки могут модернизироваться в зависимости от поставленной задачи и легко встраиваться в сверточную нейронную сеть с практически любой архитектурой, а их весовые коэффициенты адаптируются к входным данным автоматически в процессе обучения.

В настоящее время исследуется возможность внедрения пространственной фильтрации с помощью разности фильтров Гаусса или вейвлет-преобразования в сверточную нейронную сеть для лучшего подавления фона.

References

[1] Chattopadhyay A, Maitra M. MRI-based brain tumour image detection using CNN based deep learning method. Neurosci Inform 2022; 2(4): 100060. DOI: 10.1016/j.neuri.2022.100060.

[2] Hashib H, Leon M, Salaque AM. Object detection based security system using machine learning algorthim and Raspberry Pi. 2019 Int Conf on Computer, Communication, Chemical, Materials and Electronic Engineering (IC4ME2) 2019: 1-4. DOI: 10.1109/IC4ME247184.2019.9036531.

[3] Wang X, Lv G, Xu L. Infrared dim target detection based on visual attention. Infrared Phys Techn 2012; 55(6): 513-521. DOI: 10.1016/j.infrared.2012.08.004.

[4] Chen CLP, Li H, Wei Y, Xia T, Tang YY. A local contrast method for small infrared target detection. IEEE Trans Geosci Remote Sens 2014; 52(1): 574-581. DOI: 10.1109/TGRS.2013.2242477.

[5] Wang H, Xin Y. Wavelet-based contourlet transform and kurtosis map for infrared small target detection in complex background. Sensors 2020; 20(3): 755. DOI: 10.3390/s20030755.

[6] Stojnić V, Risojević V, Muštra M, Jovanović V, Filipi J, Kezić N, Babić Z. A method for detection of small moving objects in UAV videos. Remote Sens 2021; 13(4): 653. DOI: 10.3390/rs13040653.

[7] Liu X, Li X, Li L, Su X, Chen F. Dim and small target detection in multi-frame sequence using Bi-Conv-LSTM and 3D-Conv structure. IEEE Access 2021; 9: 135845-135855. DOI: 10.1109/ACCESS.2021.3110395.

[8] Mazurek P. Convolutional neural network reference for track-before-detect applications. Remote Sens 2023; 15(18): 4629. DOI: 10.3390/rs15184629.

[9] Wang Z, Wang C, Chen Y, Li J. Target detection algorithm based on super- resolution color remote sensing image reconstruction. J Meas Eng 2023; 12(1): 15. DOI: 10.21595/jme.2023.23510.

[10] Liu H, Ding M, Li S, Xu Y, Gong S, Kasule AN. Small-target detection based on an attention mechanism for apron-monitoring systems. Appl Sci 2023; 13(9): 5231. DOI: 10.3390/app13095231.

[11] Fan M, Tian S, Liu K, Zhao J, Li Y. Infrared small target detection based on region proposal and CNN classifier. Signal Image Video P 2021; 15: 1927-1936. DOI: 10.1007/s11760-021-01936-z.

[12] Li B, Xiao C, Wang L, Wang Y, Lin Z, Li M, An W, Guo Y. Dense nested attention network for infrared small target detection. IEEE Trans Image Process 2023; 32: 1745-1758. DOI: 10.1109/TIP.2022.3199107.

[13] Li J, Ye DH, Kolsch M, Wachs JP, Bouman CA. Fast and robust UAV to UAV detection and tracking from video. IEEE Trans Emerg Top Comput 2022; 10(3): 1519-1531. DOI: 10.1109/TETC.2021.3104555.

[14] Svanström F, Englund C, Alonso-Fernandez F. Real-time drone detection and tracking with visible, thermal and acoustic sensors. 25th Int Conf on Pattern Recognition (ICPR) 2020: 7265-7272. DOI: 10.1109/ICPR48806.2021.9413241.

[15] Rozantsev A, Lepetit V, Fua P. Flying objects detection from a single moving camera. 2015 IEEE Conf on Computer Vision and Pattern Recognition (CVPR) 2015: 4128-4136. DOI: 10.1109/CVPR.2015.7299040.

[16] Wang Y, Zhou Q, Xiong J, Wu X, Jin X. ESNet: An efficient symmetric network for real-time semantic segmentation. In Book: Lin Z, Wang L, Yang J, Shi G, Tan T, Zheng N, Chen X, Zhang Y, eds. Pattern Recognition and Computer Vision. Cham: Springer Nature Switzerland AG; 2019: 41-52. DOI: 10.1007/978-3-030-31723-2_4.

[17] Huang L, Chen C, Yun J, Sun Y, Tian J, Hao Z, Yu H, Ma H. Multi-scale feature fusion convolutional neural network for indoor small target detection. Front Neurorobot 2022; 16: 881021. DOI: 10.3389/fnbot.2022.881021.

[18] Agac S, Durmaz Incel O. On the use of a convolutional block attention module in deep learning-based human activity recognition with motion sensors. Diagnostics 2023; 13(11): 1861. DOI: 10.3390/diagnostics13111861.

[19] Sudre CH, Li W, Vercauteren T, Ourselin S, Jorge Cardoso M. Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations. In Book: Cardoso MJ, Arbel T, Carneiro G, et al, eds. Deep learning in medical image analysis and multimodal learning for clinical decision support. Cham: Springer International Publishing AG; 2017: 240-248. DOI: 10.1007/978-3-319-67558-9_28.

[20] Rezatofighi H, Tsoi N, Gwak JY, Sadeghian A, Reid I, Savarese S. Generalized intersection over union: A metric and a loss for bounding box regression. 2019 IEEE/CVF Conf on Computer Vision and Pattern Recognition (CVPR) 2019: 658-666. DOI: 10.1109/CVPR.2019.00075.

Приложение А

Табл. 1А. Результаты обучения и валидации пяти моделей сегментации

Модель	Функция потерь		Precision	Recall	IoU	ROC AUC	Average Precision (AP)
Исходная	Обучение	0,097	0,831	0,621	0,549	0,824	0,575
	Валидация	0,278					
Слияние признаков	Обучение	0,072	0,762	0,725	0,588	0,871	0,606
	Валидация	0,246					
Внимание по каналам	Обучение	0,069	0,791	0,707	0,593	0,864	0,627
	Валидация	0,239					
Внимание по пикселям	Обучение	0,075	0,707	0,716	0,545	0,868	0,580
	Валидация	0,258					
Внимание по каналам и пикселям	Обучение	0,061	0,754	0,756	0,607	0,888	0,633
	Валидация	0,227					

Сведения об авторе

Краснов Дмитрий Игоревич, 2000 года рождения, магистрант Университета ИТМО по направлению 12.04.02 «Техническое зрение». В 2022 году с отличием окончил МГТУ имени Н.Э. Баумана по направлению 12.03.02 «Оптотехника». Область научных интересов: компьютерное зрение, обработка изображений, машинное обучение и искусственный интеллект. E-mail: *dmitriy_krasnov@outlook.com*

ГРНТИ: 20.53.19
Поступила в редакцию 5 декабря 2023 г. Окончательный вариант – 13 марта 2024 г.

Attention modules in convolutional neural networks for small object recognition

D.I. Krasnov [1]
[1] *ITMO University,*
197101, Saint Petersburg, Russia, Kronverkskiy Prospekt 49, bldg. A

Abstract

A problem of small object recognition is frequently encountered in biomedical and security systems. However, the detection of such objects is often complicated by presence of dense clouds or infrastructure objects. Results of using various attention mechanisms to improve accuracy in small objects segmentation with convolutional neural networks are presented in this paper. Modules of channel attention and spatial attention are considered. This approach allows one to effectively suppress less informative channels and image areas, while enhancing more informative channels and image areas. Meanwhile, weights of the attention modules are automatically adapted to the input data during training. An assessment of influence of the attention mechanisms in convolutional neural network architecture on the ability to suppress complex backgrounds (clouds and infrastructure objects) and segment small objects is performed. The results are presented in the form of tables with test metrics and figures with precision-recall curves, ROC curves and heatmaps showing an effectiveness of background suppression. The results obtained allow one to implement the described attention modules in the convolutional neural networks of any complexity and increase the recognition accuracy of objects of 10-40 pixels in size on a complex background.

Keywords: semantic segmentation, small object, convolutional neural network, attention modules, computer vision.

Citation: Krasnov DI. Attention modules in convolutional neural networks for small object recognition. Computer Optics 2024; 48(6): 963-968. DOI: 10.18287/2412-6179-CO-1468.

Author's information

Dmitriy Igorevich Krasnov, (b. 2000), master student of ITMO University 12.04.02 "Computer Vision" program. Graduated from BMSTU in 2022 with a bachelor degree in 12.03.02 "Optical Engineering". Research interests: computer vision, image processing, machine learning and artificial intelligence. E-mail: *dmitriy_krasnov@outlook.com*

Received December 5, 2023. The final version – March 13, 2024.