УДК 004.032.26

ИССЛЕДОВАНИЕ ПРИМЕНЕНИЯ СВЕРТОЧНЫХ НЕЙРОННЫХ СЕТЕЙ ДЛЯ РЕШЕНИЯ ЗАДАЧ СЕМАНТИЧЕСКОЙ СЕГМЕНТАЦИИ ИЗОБРАЖЕНИЙ

Широкова Вера Николаевна[1], Ушанкова Мария Юрьевна[2]

[1]Студент;
Государственный университет «Дубна»;
Россия, 141980, Московская обл., г. Дубна, ул. Университетская, 19;
e-mail: shvn.20@uni-dubna.ru.

[2]Старший преподаватель;
Государственный университет «Дубна»;
Россия, 141980, Московская обл., г. Дубна, ул. Университетская, 19;
e-mail: ushankova.m.ju@uni-dubna.ru.

Семантическая сегментация – операция в компьютерном зрении, заключающаяся в классификации и попиксельной локализации объектов на цифровом изображении. Данная статья содержит в себе обзор существующих модификаций классической архитектуры сверточной нейронной сети, направленных на решение проблемы искажения информации с исходного изображения. Проведено сравнение эффективности рассмотренных моделей в условиях бинарной и множественной семантической сегментации. Статья может быть полезной для ML/DL-разработчиков, желающих изучить проблематику сегментации изображений в рамках своей предметной области.

Ключевые слова: нейронные сети, сверточные нейронные сети, сегментация изображений, семантическая сегментация, U-Net, SegNet, DeepLabV3.

Для цитирования:
Широкова В. Н., Ушанкова М. Ю. Исследование применения сверточных нейронных сетей для решения задач семантической сегментации изображений // Системный анализ в науке и образовании: сетевое научное издание. 2024. № 2. С. 21-29. EDN: DPJCBC. URL : https://sanse.ru/index.php/sanse/article/view/616.

THE APPLICATION OF CONVOLUTIONAL NEURAL NETWORKS FOR SOLVING SEMANTIC IMAGE SEGMENTATION PROBLEMS

Shirokova Vera N.[1], Ushankova Maria Yu.[2]

[1]Student;
Dubna State University;
19 Universitetskaya Str., Dubna, Moscow region, 141980, Russia;
e-mail: shvn.20@uni-dubna.ru.

[2]Senior teacher;
Dubna State University;
19 Universitetskaya Str., Dubna, Moscow region, 141980, Russia;
e-mail: ushankova.m.ju@uni-dubna.ru.

Semantic segmentation is an operation in computer vision that involves the classification and localization of objects in an image. The article provides an overview of different modifications to the classical convolutional neural network architecture designed to solve the problem of distortion in image data The efficiency of the considered models in conditions of binary and multiple semantic segmentation is compared. The article may be useful for ML/DL-developers who wish to study the problem of image segmentation within their subject area.

Keywords: convolutional neural networks, neural networks, image segmentation, semantic segmentation, UNet, SegNet, DeepLabV3.

For citation:
Shirokova V. N., Ushankova M. Yu. The application of convolutional neural networks for solving semantic image segmentation problems. *System analysis in science and education*, 2024;(2):21-29 (in Russ). EDN: DPJCBC. Available from: https://sanse.ru/index.php/sanse/article/view/616.

Введение

В современном мире технологии компьютерного зрения становятся все более популярными. Они востребованы во многих областях, будь то автономное вождение, виртуальная реальность, обработка медицинских или географических снимков и пр. Одной из ключевых задач в сфере компьютерного зрения является семантическая сегментация, заключающаяся в классификации и локализации объектов на изображении. И зачастую, основным инструментом для ее решения становятся сверточные нейронные сети.

В статье анализируются различные подходы к созданию сверточных нейронных сетей для решения задачи семантической сегментации с целью определения наиболее эффективных архитектур.

Проблема сегментации изображений

Сегментация изображений – одна из самых важных операций в компьютерном зрении. Ее цель заключается в разделении цифрового изображения на несколько сегментов (см. рис. 1). Это позволяет изменить представление об поступивших визуальных данных, упрощая их анализ и дальнейшую обработку. Результатом является изображение, каждому пикселю которого присвоена метка, относящая его к одной из искомых категорий [1].

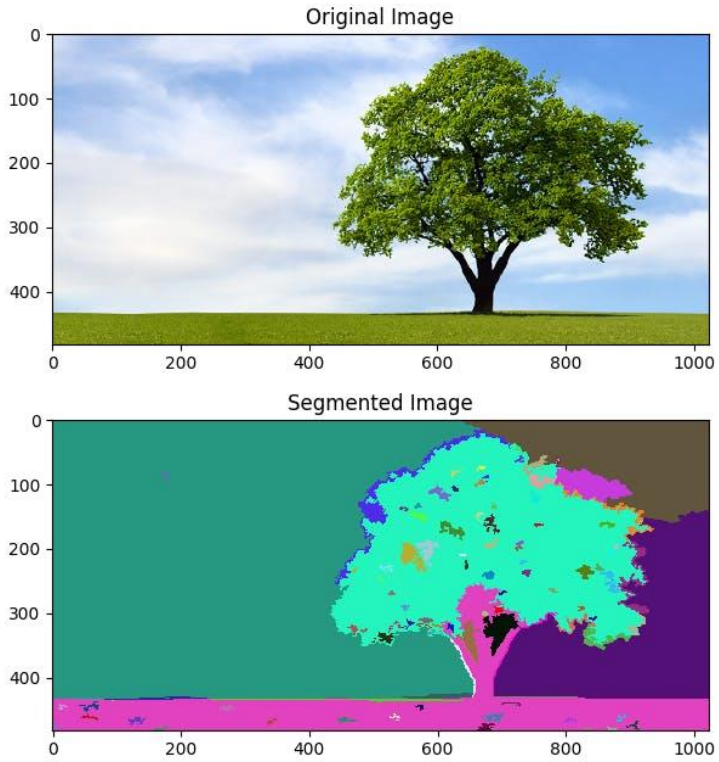

Рис. 1. Пример работы сегментации изображения

Задачи сегментации изображения можно разделить на несколько групп в зависимости от количества и типа информации, которую они выделяют. Далее будет рассмотрена задача семантической сегментации, заключающаяся в выделении на изображении значимых по смыслу объектов (классов) (см. рис. 2). Для сравнения, классическая сегментация изображений характеризуется разделением изображения на несколько частей, без попытки понять, что из себя представляют эти данные. По сути, задачу семантической сегментации можно охарактеризовать как совмещение задачи классификации и локализации объектов на изображении.

Рис. 2. Пример семантической сегментации изображения

Наиболее качественной маской сегментации будет та, чьи итоговые разметки классов будут максимально точно совпадать с их реальным расположением. Для вычисления меры сходства этих множеств существует две метрики: коэффициент Жаккара или «Пересечение через объединение» (англ. *Intersection over Union*) и коэффициент Сёренсена-Дайса, известный также как *F1*-мера [2].

Эти метрики имеют много общего, поэтому на практике для оценки проекта часто используется только одна из них, на выбор разработчиков.

Коэффициент Жаккара представляет из себя отношение пересечения множеств к их объединению. Чем больше общих элементов между множествами, тем большее число выведет метрика.

$$Jaccard\ (A,B) = \frac{|A \cap B|}{|A \cup B|}$$

Интерпретация коэффициента Сёренсена-Дайса несколько сложнее, поэтому эта метрика не так популярна. Коэффициент равен удвоенному числу элементов, общих для обоих наборов, делённому на сумму количества элементов в каждом наборе.

$$Dice(A,B) = \frac{2|A \cap B|}{|A| + |B|}$$

Аналогично коэффициенту Жаккара большее значение метрики в диапазоне от 0 до 1 соответствует более высокому качеству работы модели.

Также для данных формул предусмотрен ряд усовершенствований:

- если модель решает задачу множественной классификации, среднее значение метрики вычисляется как взвешенное среднее по каждому классу, где весами является частота встречаемости пикселей каждого класса;
- в зависимости от условий задачи пиксели на границах объектов могут учитываться с меньшим весом по сравнению с внутренними пикселями.

На сегодняшний день основным инструментом для решения задач семантической сегментации стали сверточные нейронные сети, благодаря своей способности извлекать сложные признаки из визуальных данных [3].

Сверточная нейронная сеть

Нейронная сеть – математическая модель, имитирующая работу нервной системы человека. Получая на вход какой-либо объект, она анализирует значение его признаков путем перемножения их на

коэффициенты значимости, и после ряда подобных преобразований делает предсказание в рамках решаемой задачи.

Одним из главных преимуществ нейронных сетей перед традиционными алгоритмами является способность к обучению. В процессе обучения нейронная сеть изменяет упомянутые выше коэффициенты значимости и тем самым способна выявлять сложные зависимости между входными данными и выходными, а также выполнять обобщение. Это значит, что в случае успешного обучения сеть сможет вернуть верный результат на основании неполных, «зашумленных» и частично искажённых данных.

Сверточная нейронная сеть (*convolutional neural network, CNN*) — специальная архитектура нейронных сетей, предложенная Яном Лекуном, нацеленная на работу с изображениями. В отличие от классической архитектуры многослойного персептрона, здесь учитывается их двухмерная топология. Т.е. вместо разложения пикселей входного изображения в вектор, как в полносвязной сети, мы обрабатываем изображение в виде матрицы, что позволяет нам находить больше признаков принадлежности к искомым классам объектов.

В данной архитектуре вместо нейронов с персональными весовыми коэффициентами, используется матрица весов небольшого размера, которую «двигают» по всему обрабатываемому слою. Это ядро свертки умножает каждый пиксель входного изображения на соответствующие значения своего окна, формируя таким образом выходную карту признаков. Его работу можно интерпретировать как графическое кодирование какого-либо признака, например, наличие наклонной линии под определённым углом. При этом, веса ядер свертки не закладываются исследователем заранее, а формируются самостоятельно путём обучения сети классическим методом обратного распространения ошибки. Подобно нейронам, несколько параллельно применяемых сверток образуют собой слой, называемый сверточным.

Также в *CNN* существуют слои субдискретизации (*pooling*), чьей целью является уменьшение размерности карт предыдущего слоя. Если на предыдущей операции свертки уже были выявлены некоторые признаки, то для дальнейшей обработки настолько подробное изображение уже не нужно и оно уплотняется. Подобная фильтрация ненужных деталей также помогает сети не переобучаться (*overfitting*). Чаще всего используется функция максимума (*max pooling*) или среднего значения (*average pooling*). Обучаемых параметров у этого слоя нет.

Между сверточными и подвыборочными слоями к картам признаков применяются функции активации, аналогичные тем, что используются в других архитектурах нейросетей. Последние несколько слоев *CNN* обычно представляют из себя классический персептрон.

Использование CNN для семантической сегментации

Как было определено ранее, задача семантической сегментации изображений также включает в себя локализацию объектов, т.е. попиксельное выделение их границ. Чтобы *CNN* смогла выполнить эту задачу, необходимо внести в ее базовую архитектуру ряд изменений.

Основными элементами в сети для сегментации являются сверточные слои, поскольку именно они отвечают за выделение признаков на изображении. С помощью последовательности этих слоев можно получить вывод, который объединяет информацию со всех карт признаков, относящихся к искомым классам задачи. Тем самым мы получим матрицу размером $X \times X \times Y$, где $X \times X$ – размер ядер сверток предпоследнего слоя, равный размеру входного изображения, а Y – количество искомых классов.

Однако, такой подход имеет свои недостатки. Особенности устройства сверточных слоев приводят к тому, что после обработки карты признаков становятся менее информативными по сравнению с оригинальным изображением, слои субдискретизации также приводят к уменьшению изображения, что заметно ухудшает итоговый результат работы сети в рамках рассматриваемой задачи.

Существующие решения

В данной статье мы рассмотрим существующие улучшения базовой архитектуры *CNN*, направленных на решение проблемы искажения информации с оригинального изображения.

U-Net

U-Net – сверточная нейронная сеть для сегментации биомедицинских изображений, созданная в 2015 году [3]. Архитектура *U-Net* состоит из двух частей, как называют сами авторы — сжимающийся (*contracting*) путь и расширяющийся (*expansive*) (см. рис. 3). Сжимающийся путь повторяет устройство обычной сверточной сети и состоит из двойных слоев сверток 3x3, за которыми следует функция активации *ReLU* и слой субдискретизации (*maxpooling*) с ядром 2x2. На каждом этапе количество каналов изображения удваивается. Расширяющийся путь представляет из себя комбинацию из deconvolution-слоя, объединяющего в себе операцию повышения дискретизации (*maxunpooling*) и свертки с ядром 2x2, а также двойной последовательности сверток 3x3, за которыми следует *ReLU*. Недостаток информации, возникающий при реализации *unpooling* компенсируется за счет передачи карты признаков с симметричного слоя сворачивающей части нейросети. Данный метод принято называть *skip-connection*. На последнем слое используется свертка 1x1 для сопоставления 64-канальной карты признаков с искомым количеством классов.

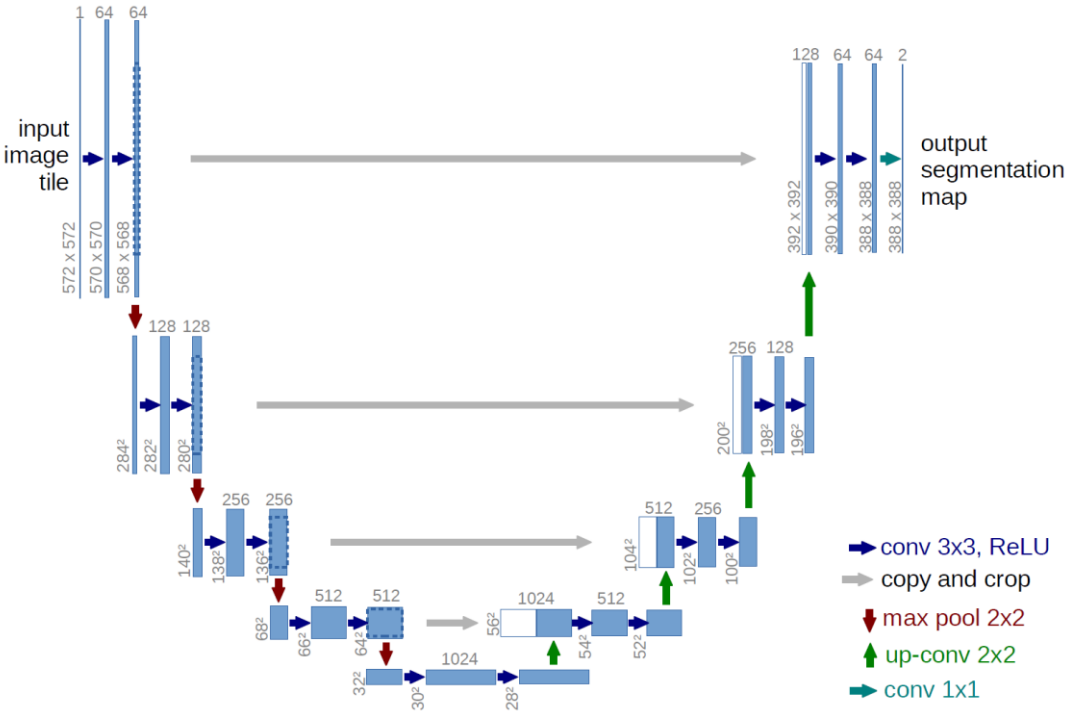

Рис. 1. Архитектура сети U-Net

SegNet

Общая идея архитектуры сети *SegNet* похожа на сеть *U-Net*, она также включает в себя две симметричные части, кодирующую и декодирующую (см. рис. 4) [4].

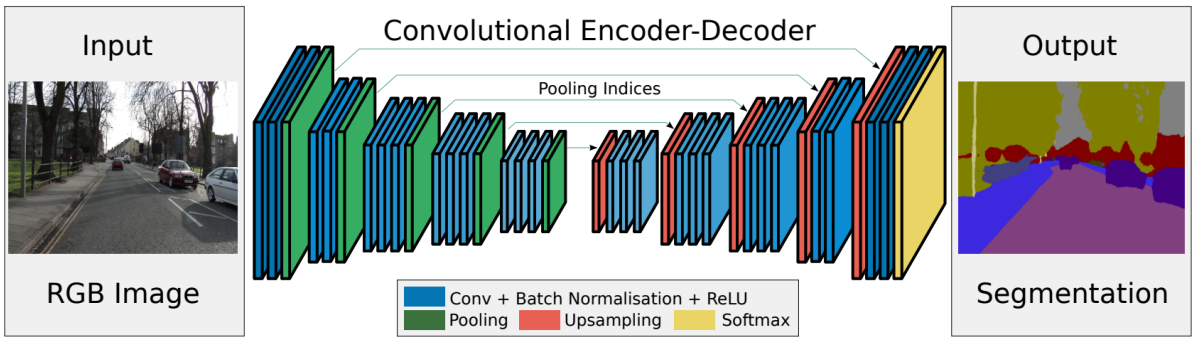

Рис. 2. Архитектура сети SegNet

Кодирующая часть повторяет устройство сверточной сети для классификации изображений *VGG16*, состоит из 5 крупных сегментов, содержащих в себе сверточные слои с ядром 3x3, пакетную нормализацию, *ReLU* и операцию *MaxPooling* с ядром 2x2. Декодирующая часть симметрично повторяет кодирующую, заменяя операцию *MaxPooling* на *MaxUnooling*. Восстановление данных происходит за счет передачи индексов, сохраненных на этапе субдискретизации (см. рис. 5). Последний сверточный слой преобразует многоканальную карту признаков таким образом, чтобы количество каналов соответствовало числу искомых классов. На выходе из сети используется функция *softmax*, которая принимает на вход вектор из действительных чисел и нормализует его в распределение вероятностей.

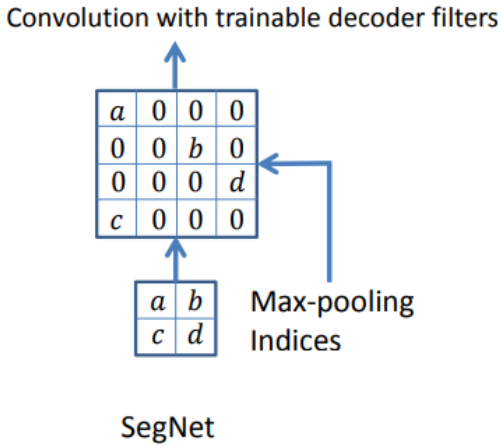

Рис. 3. Операция MaxUnpooling с использованием сохраненных индексов

DeepLabV3

Архитектура сети *DeepLabV3* заметно отличается от рассмотренных выше *U-Net* и *SegNet* (см. рис.6) [5]. В первую очередь из-за использования слоев расширенной свертки (*Dilated/Atrous convolution*) для решения проблемы сегментации объектов разных масштабов и большего учета контекста. Расширенная свертка позволяет увеличить обрабатываемую область без потери качества изображения и роста числа параметров.

Рассмотренные сверточные слои образуют собой *Atrous Spatial Pyramid Pooling* — комбинацию из четырех сверток с разными отступами и слоя субдисктретизации, параллельно обрабатывающих одну карту признаков, что позволяет проводить классификацию объектов произвольного масштаба.

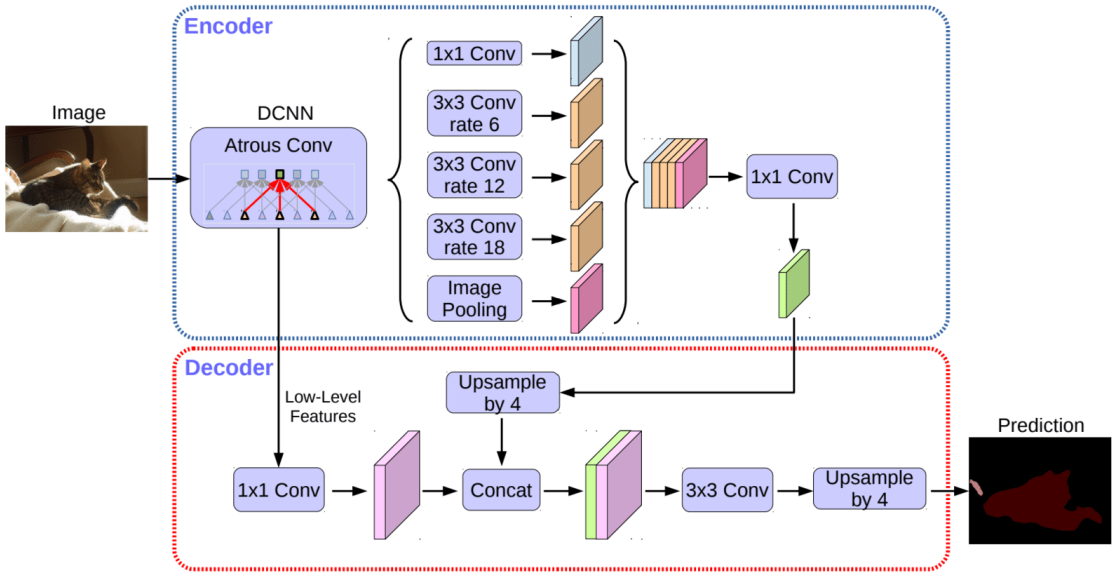

Рис. 4. Архитектура DeepLabV3

Исследование применимости

Проведено исследование с целью определения наиболее эффективной архитектуры из рассмотренных выше. В качестве задач для анализа взяты ситуации бинарной сегментации, подразумевающей разделение пикселей по категориям объект – фон, и множественной семантической сегментации, предполагающей обнаружение и локализацию на изображении нескольких классов объектов.

В ходе эксперимента модели обучались на одном наборе данных в течение равного количества эпох с одинаковыми гиперпараметрами, такими как: функция потерь, оптимизатор и его темп обучения (*learning rate*). Основными показателями для сравнения служили: время прохождения одной эпохи и качество обработки изображений из тестовой выборки. Метрикой оценки качества выбран коэффициент Жаккара или «Пересечение через объединение» (англ. *Intersection over Union*).

Для рассмотрения задачи бинарной классификации был выбран датасет *Caltech-UCSD Birds-200-2011 (CUB-200-2011)* [6], содержащий в себе 11 788 изображений птиц (см. рис. 7)

Рис. 5. Пример объекта из датасета CUB-200-2011

Для рассмотрения задачи множественной семантической сегментации выбран набор данных *CamVid* (*Cambridge-Driving Labeled Video Database*) [7]. Он состоит из 700 пар изображений с автомобильных видеорегистраторов и их сегментированными масками, содержащими разметки 32 классов объектов (см. рис. 8). Для упрощения процесса обучения сетей 32 класса были сгруппированы в 9 крупных категорий, включающих в себя движимые и недвижимые объекты, дорогу, тротуар, растения, средства организации дорожного движения и т.д.

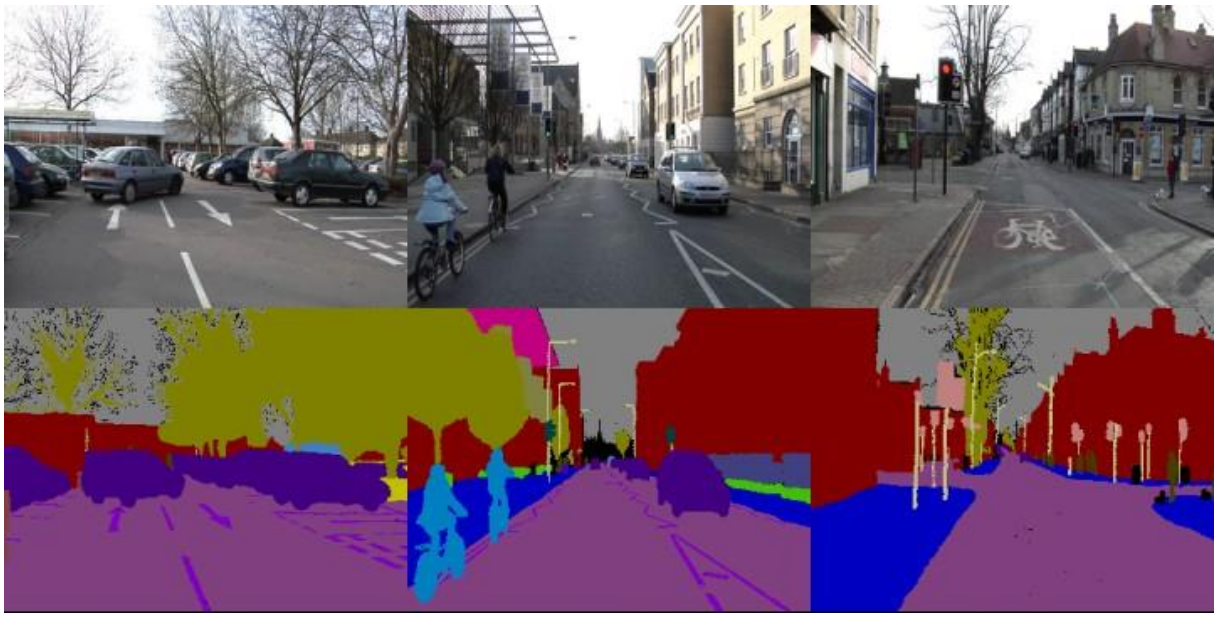

Рис. 6. Пример объектов из датасета CamVid

Программные модули исследования реализованы на языке Python. Данный язык активно используется для машинного обучения так как имеет большое количество специализированных фреймворков и библиотек. В частности, в этой работе активно использовался фреймворк *PyTorch* [8], предоставля-

ющий широкий инструментарий для глубокого обучения (*deep learning*), также были подключены созависимые библиотеки *torchvision, torchmetrics*, упрощающие обработку изображений и подсчет метрик качества соответственно.

Модели сетей *U-Net* и *SegNet* были собраны нами самостоятельно, согласно описанным выше архитектурам. Модель *DeepLabV3* была импортирована из указанного выше фреймворка *PyTorch*. Обучение сетей проходило с помощью графического ускорителя (*GPU*) на видеокарте *GeForce RTX 4070 Ti*.

Результаты, полученные в ходе исследования приведены в табл. 1.

Табл. 1. Сравнительные характеристики моделей

Модель сети	**Бинарная сегментация**			**Множественная сегментация**		
	Количество эпох	**Длительность эпохи (сек)**	**Средний IoU**	**Количество эпох**	**Длительность эпохи (сек)**	**Средний IoU**
U-Net	20	362	0.75	60	150	0,53
SegNet	20	248	0.76	60	137	0,58
DeepLabV3	10	225	0.84	60	130	0,68

Основываясь на них, можно сделать следующие выводы:

- особенности устройства *U-Net*, предполагающие сохранение полных карт сегментации в процессе анализа изображения, сделали модель самой затратной по памяти и времени в процессе обучения;
- *SegNet*, также относясь к архитектурам к типа *Encoder-Decoder*, оказалась более оптимизированной за счет сохранения только индексов слоев субдискретизации (*pooling*);
- *DeepLabV3*, построенная с помощью расширенных сверток и *Atrous Spatial Pyramid Pooling*, оказалась самой результативной и быстрообучаемой из всех рассматриваемых моделей. (В задаче бинарной сегментации для достижения качественных результатов ей потребовалось пройти вдвое меньше эпох, чем *U-Net* и *SegNet*.)

В рамках поставленного эксперимента показано, что все сети справляются с задачей сегментации, но наибольшую эффективность продемонстрировала модель *DeepLabV3*.

Заключение

Задача семантической сегментации изображений сложнее задачи классификации, для которой изначально создавалась сверточная нейронная сеть. Поэтому классическая архитектура *CNN* требует изменений, позволяющих избежать потери информации в процессе обработки визуальных данных. Все рассмотренные модели усовершенствования *CNN* продемонстрировали способность справиться с задачами бинарной и множественной сегментации, однако, использование архитектуры *DeepLabV3* — наиболее эффективно. Особенно при работе с изображениями, где объекты могут быть представлены в разных масштабах.

Список источников

1. Лукашик Д. В. Анализ современных методов сегментации изображений // Экономика и качество систем связи. – 2022. – №2. – С. 57-65.

2. Monteux A. Metrics for semantic segmentation // Excursions in data: [сайт]. – Angelo Monteux, 2019 – Дата публикации: 10.05.2019. – URL: https://ilmonteux.github.io/2019/05/10/segmentation-metrics.html (дата обращения: 23.05.2024).

3. Ronneberger O., Fischer P., Brox T. U-net: Convolutional networks for biomedical image segmentation // Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international

conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18. – Springer International Publishing, 2015. – P. 234-241.

4. Badrinarayanan V., Kendall A., Cipolla R. Segnet: A deep convolutional encoder-decoder architecture for image segmentation // IEEE transactions on pattern analysis and machine intelligence. – 2017. – Vol. 39, No. 12. – P. 2481-2495.

5. Rethinking atrous convolution for semantic image segmentation / L. Chen, G. Papandreou, F. Schroff, H. Adam // arXiv preprint. [2017]. – arXiv:1706.05587.

6. Caltech-UCSD Birds-200-2011 (CUB-200-2011) [: Dataset] / C. Wah, S. Branson, P. Welinder [et al.]; California Institute of Technology // Perona Lab. – URL: https://www.vision.caltech.edu/datasets/cub_200_2011/ (дата обращения: 23.05.2024).

7. Brostow G. J., Fauqueur J., Cipolla R. Semantic object classes in video: A high-definition ground truth database // Pattern recognition letters, 2009. – Vol. 30. – №. 2. – P. 88-97.

8. PyTorch documentation // pytorch.org. – URL: https://pytorch.org/docs/stable/index.html (дата обращения: 23.05.2024).