РАСПОЗНАВАНИЕ ПРЕПЯТСТВИЙ С ИСПОЛЬЗОВАНИЕМ СЕМАНТИЧЕСКОЙ СЕГМЕНТАЦИИ ПРИ ДВИЖЕНИИ АВТОНОМНОГО МОБИЛЬНОГО РОБОТА[1*]

В.В. Маносин[1], М.И. Кумсков[1], В.М. Буданов[2]

[1] Механико-математический факультет МГУ имени М.В. Ломоносова,
[2] Научно-исследовательский институт механики МГУ имени М.В. Ломоносова

В данной работе рассмотрена задача навигации автономного мобильного робота по незнакомой пересеченной местности с использованием современных методов семантической сегментации. Более конкретно исследуется задача поиска и распознавания препятствий, возникающих по ходу движения робота. Проводится обзор и сравнение современных нейросетевых архитектур, предназначенных для семантической сегментации, таких как FCN, SegNet, U-Net, применительно к поставленной задаче. Вычислительные эксперименты осуществлялись на Robot Unstructured Ground Driving (RUGD) Dataset с измененным числом классов. С помощью архитектуры U-Net на данном наборе данных удалось достичь качества сегментации 94.84% по метрике IoU.

Ключевые слова: мобильный робот, распознавание препятствий, нейронные сети, семантическая сегментация.

1. Введение

1.1. Мобильный робот

Мобильный робот (ровер) представляет собой высокотехнологичное устройство, специально разработанное и предназначенное для автономного перемещения при различных сценариях. Высокая маневренность ровера обеспечивается использованием колесного шасси. Однако необходимо отметить, что колесная система является одновременно и недостатком, так как позволяет преодолевать лишь небольшие неровности ландшафта. Информация об окружающей среде поступает с помощью высококачественной камеры и датчика LiDAR (Light Detection and Ranging), установленных на мобильном роботе. Сочетание данных с этих двух устройств позволяет как идентифицировать объекты, так и с высокой точностью определять расстояние до них.

2. Задача распознавания препятствий

2.1. Описание задачи

Задача распознавания препятствий направлена на создание алгоритмов, способных выявлять преграды, мешающие продвижению ровера, в окружающей среде. Эти преграды могут быть как статическими (камни, деревья, здания), так и динамическими (люди, животные). Главная цель распознавания препятствий заключается в том, чтобы предоставить роверу информацию о рельефе окружающей местности. С помощью полученной информации автономный робот может эффективно ориентироваться в окружающей среде, корректировать свой маршрут и избегать столкновений с различными потенциально опасными объектами.

[1*] Работа выполнена при частичной поддержке проекта 23-Ш01-11 «Исследование интеллектуальной робототехнической системы, оснащенной ветроэнергетической установкой» НОШ МГУ «Космос».

2.2. Семантическая сегментация в задаче распознавания препятствий

Рассмотрим более подробно способ решения задачи распознавания препятствий с помощью семантической сегментации изображения. Семантическая сегментация – это задача компьютерного зрения, которая заключается в классификации каждого пикселя изображения, то есть отнесении его к определенному классу. В отличие от обычной классификации, где изображение зачастую помечено всего одной меткой, семантическая сегментация обеспечивает более подробную информацию о содержимом изображения, разделяя объекты попиксельно.

Решение задачи семантической сегментации подразумевает создание детализированной попиксельной цветовой маски изображения, где каждому пикселю присваивается определённый цвет в зависимости от того, к какому типу объектов он относится. В частности, объекты, идентифицируемые как преграды, будут отмечены специфическим цветом, что позволит как определять их местоположение, так и понимать их форму и размеры. Также существует возможность разделения большого класса преград на более узкие (деревья, кусты, камни, здания, ямы) для более точного распознавания объектов. Однако решение такой задачи может быть связано с усложнением модели и существенным увеличением вычислений, что не всегда допустимо в условиях ограниченности вычислительных ресурсов, поэтому данная задача требует дальнейших исследований.

2.3. Сочетание камеры и LiDAR

Камера и LiDAR могут быть связаны для решения задачи распознавания препятствий с помощью семантической сегментации изображения. Эти два датчика отлично дополняют друг друга: камера предоставляет информацию о цвете и текстуре объектов, что полезно для идентификации и сегментации различных классов препятствий. В то время как с помощью LiDAR можно определить расстояние до объектов и построить их 3D-модель, позволяя более точно идентифицировать местоположение препятствий и их форму. Также объединение информации разной природы способствует улучшению точности предсказания за счет компенсации недостатков друг друга. Например, в условиях низкой освещенности камера выдает нечеткую зашумленную картинку, из которой может быть сложно извлечь какие-либо признаки о препятствиях, в свою очередь, LiDAR работает в любое время суток и при любой освещенности без ограничений. Таким образом, модели, использующие комбинированные данные, могут быть более эффективны для решения задачи распознавания препятствий.

2.4. Использование нейронных сетей в задаче распознавания препятствий

В настоящее время методы глубокого обучения, представляющие собой мощные архитектуры нейронных сетей, стали основой для решения множества задач в области компьютерного зрения, в частности, задачи семантической сегментации изображения. Эти подходы обеспечивают значительные улучшения в качестве и точности сегментации по сравнению с традиционными методами, позволяя эффективно справляться со сложными изображениями с большим количеством объектов на них.

3. Архитектуры нейронных сетей для семантической сегментации

3.1. Fully Convolution Network

FCN, или Fully Convolutional Network, представляет собой архитектуру полностью сверточной нейронной сети, которая впервые продемонстрировала возможность успешного применения таких нейросетей для решения задачи семантической сегментации изображения, значительно превосходя по метрикам классические методы [1]. Основной принцип работы FCN заключается в переосмыслении классических классификационных архитектур путём адаптации их структуры, а именно замены полносвязных слоев на слои свертки, что позволило нейросети обрабатывать входные данные произвольного размера.

Большинство современных архитектур для сегментации изображений разработаны на основе концепции Fully Convolutional Networks. То есть они используют исключительно свёрточные слои, которые позволяют извлекать признаки из изображений, а также слои пулинга, уменьшающие пространственное разрешение для упрощения вычислений и выделения наиболее значимых характеристик. В дополнение к этому используется апсемплинг, необходимый для увеличения размерности изображения и восстановления его исходного разрешения для построения корректной сегментационной маски.

3.2. U-Net

U-Net – это архитектура сверточной нейронной сети, разработанная для решения задачи семантической сегментации биомедицинских изображений [2]. U-Net в настоящее время считается одной из самых популярных и широко используемых моделей для семантической сегментации благодаря своей способности достигать хороших результатов при сравнительно небольшом объеме вычислительных ресурсов и минимальных требованиях к размерам обучающих наборов данных. Это делает её универсальным инструментом для решения многих задач семантической сегментации далеко за пределами биомедицинских изображений.

Архитектура U-Net состоит из двух основных компонент: кодировщика и декодировщика, связанных между собой. Кодировщик выполняет функцию сжатия входного изображения и извлечения признаков. Процесс начинается с применения двух сверток размера 3x3, после каждой из которых используется функция активации ReLU. Затем применяется операция пулинга, которая снижает пространственные размеры входных данных, что уменьшает количество вычислений и выделяет наиболее существенные признаки изображения. Данная последовательность операций (свертка + ReLU + свертка + ReLU + пулинг) повторяется четыре раза, каждый раз уменьшая размерность и увеличивая количество каналов, что обеспечивает извлечение всё более сложных признаков из исходного изображения.

Декодировщик осуществляет обратное преобразование сжатых данных, однако в оригинальной архитектуре не восстанавливает размер изображения до исходного. Он также использует две последовательные свертки 3x3 с активацией ReLU после каждого применения. В дополнение к этому применяется транспонированная свертка, увеличивающая пространственные размеры входных данных. Как и в кодировщике, данный блок операций (свертка + ReLU + свертка + ReLU + транспонированная свертка) также повторяется четыре раза.

Финальный блок модели включает в себя две свертки 3x3 с активацией ReLU после каждой свертки. Для окончательной интерпретации выходных данных используется свертка размером 1x1, которая отвечает за сопоставление каждого вектора признаков с желаемым количеством классов, что позволяет получить желаемую сегментационную маску.

Одной из ключевых особенностей архитектуры U-Net является использование skip connections [3]. Они обеспечивают объединение слоев кодировщика с соответствующими слоями декодировщика. Эти соединения позволяют использовать как низкоуровневые, так и высокоуровневые признаки, что значительно повышает качество сегментации. Благодаря этому подходу, модель способна восстанавливать мелкие детали объектов, которые могли бы быть утеряны во время прохода через слои кодировщика.

3.3. SegNet

SegNet – это архитектура сверточной нейронной сети, разработанная для семантической сегментации дорожных сцен и сцен в помещениях [4]. Основной акцент при проектировании этой архитектуры был сделан на обеспечение высокой производительности нейронной сети как с точки зрения времени вычислений, так и с точки зрения экономии используемой памяти. Это особенно важно, поскольку в реальных приложениях, таких как автономные автомобили, требуется проводить вычисления в реальном времени. Также обычно вычислительные ресурсы на мобильных устройствах существенно ограничены, что является дополнительным плюсом в пользу SegNet.

Архитектура SegNet, как и U-Net, состоит из кодировщика и декодировщика. Кодировщик в SegNet состоит из 13 последовательных применений свертки, после каждой из которых применяется процедура нормализации, которая улучшает сходимость и стабильность обучения модели, а также функция активации ReLU, вводящая нелинейность в модель. Переход от одного блока сверток к другому осуществляется с помощью операции пулинга, что позволяет уменьшать размерность и выделять наиболее значимые признаки изображения. Для инициализации процесса обучения SegNet можно использовать заранее подготовленные веса, обученные на больших наборах данных, например, на известном наборе данных ImageNet [5]. Важной особенностью SegNet является то, что после каждого блока свертки запоминаются индексы, полученные в результате пулинга, которые затем передаются соответствующему блоку в декодировщике. Это позволяет декодировщику эффективно восстанавливать информацию, потерянную в процессе сжатия данных.

Ключевым отличием SegNet от архитектуры U-Net является метод использования информации о пулинге. Вместо того, чтобы передавать полные карты признаков, как это делает U-Net, SegNet использует только индексы пулинга, что позволяет существенно сократить объем памяти, необходимый для обучения модели и ее реализации. Каждый уровень декодировщика соотносится с соответствующим уровнем кодировщика, и таким образом декодировщик также состоит из 13 блоков сверток с процедурой нормализации и функцией активации ReLU. Переход между блоками декодировщика реализуется посредством апсемплинга, что позволяет восстанавливать разрешение изображения, постепенно увеличивая размеры выходных данных на каждом уровне. Последним слоем применяется многопеременная логистическая регрессия, которая представляет результаты в форме сегментационных масок, позволяя выделять различные классы объектов на изображении.

4. Вычислительные эксперименты

4.1. Формирование набора данных

Для проведения сравнения представленных архитектур необходимо сначала сформировать набор данных, на котором будут обучаться модели. В дальнейшем на этом же наборе будет замеряться время работы уже обученных нейросетей. Для этих целей был использован RUGD Dataset, ориентированный на семантическое понимание неструктурированной внешней среды с целью применения в приложениях автономной навигации по сложным и труднопроходимым участкам (например, лесные массивы) [6].

Набор данных состоит из видеопоследовательностей, которые были сняты с камеры, установленной на мобильном колесном роботе. Данный робот был спроектирован так, чтобы обладать компактными размерами, что позволяет ему маневрировать в загроможденной среде и успешно проходить через узкие пространства. При этом он также обладает достаточной прочностью и устойчивостью, чтобы преодолевать сложную местность, состоящую из разных типов рельефа и различных природных препятствий.

Плотные попиксельные аннотации, предоставляемые с данным набором данных, были созданы для каждого пятого кадра видеоряда, что позволяет обеспечить высокую детализацию информации о каждом элементе среды. В общей сложности в аннотациях можно встретить 24 смысловые категории, которые описывают окружающую среду и объекты в ней.

Для удобства анализа смысловые категории были разделены на две основные группы. В первую группу входят те категории, которые не представляют значительных препятствий для перемещения робота. Эти категории, такие как void, dirt, sand, grass, sky, asphalt, gravel, mulch и rock_bed, характеризуют доступные для проезда участки, где условия не оказывают серьезного влияния на способность ровера передвигаться. Во вторую группу попадают категории, представляющие собой препятствия, которые могут затруднять или полностью блокировать движение ровера. К этим категориям относятся tree, pole, water, vehicle, container, building, log, bicycle, person, fence, bush, sign, rock, bridge и picnic_table. Эти аннотации позволяют моделям обучаться различать проходимые и непроходимые участки, что крайне важно для разработки эффективных алгоритмов автономной навигации и обеспечения безопасности роботов при их перемещении по сложным ландшафтам.

4.2. Эксперименты с U-Net

Проведем вычислительные эксперименты для U-Net с предобученными на ImageNet энкодерами, такими как Resnet34 [3] и различными модификациями Mobilenetv3_small [7], (время работы замеряется как среднее время генерации масок для батчей с 32 изображениями в облачной среде Kaggle).

Таблица 1. Результаты экспериментов

Предобученный энкодер	Число параметров	Время работы	IoU
Resnet34	21 млн.	0.30 с.	94.00 %
Mobilenetv3_small_100	0.93 млн.	0.28 с.	94.84 %
Mobilenetv3_small_minimal_100	0.43 млн.	0.25 с.	93.80 %

Предсказания, полученные с помощью Resnet34 + U-Net достаточно точны, что подтверждается метрикой IoU. Однако из-за большого числа параметров обученная модель занимает много места в хранилище. Объем модели можно уменьшить за счет использования энкодера, предназначенного для мобильных устройств.

Результаты сегментации с использованием Mobilenetv3_small_100 + U-Net получаются ничуть не хуже, а даже немного лучше предыдущей архитектуры (94.84 % IoU против 94.00 % IoU). Число параметров же уменьшается больше, чем в 22.5 раза, что существенно экономит пространство на жестком диске, занимаемое обученной моделью. Также немного уменьшается время работы (на 0.02 секунды).

Последняя архитектура Mobilenetv3_small_minimal_100 + U-Net показывает худшее из 3 моделей качество сегментации, всего 93.8 % IoU, однако отставание от лучшей модели не так велико (1.04 %), выигрыш же в числе параметров почти в 49 раз по сравнению с самой тяжелой моделью. Также получено ускорение на 0.05 секунд по сравнению с Resnet34 или на 0.03 секунды по сравнению с mobilenetv3_small_100.

5. Выводы

Таким образом, архитектуры семантической сегментации полностью применимы к решению задачи распознавания препятствий при движении ровера. Полученные результаты в 94.84 % по метрике IoU с высокой скоростью работы гарантируют качественное построение сегментационных масок. Это в дальнейшем позволит строить точную карту местности, необходимую для полноценной навигации мобильного автономного робота. Для построения карты местности возможно использование сочетания таких методов как SLAM и глубокого обучения с подкреплением [8], однако это выходит за рамки представленного доклада.

Литература

1. Jonathan Long, Evan Shelhamer, Trevor Darrell Fully Convolutional Networks for Semantic Segmentation. URL: arxiv.org/abs/1411.4038 (дата обращения: 14.06.2024).
2. Olaf Ronneberger, Philipp Fischer, Thomas Brox U–Net Convolutional Networks for Biomedical Image Segmentation. URL: arxiv.org/abs/1505.04597 (дата обращения: 23.07.2024).
3. Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun Deep Residual Learning for Image Recognition. URL: arxiv.org/abs/1512.03385 (дата обращения: 24.07.2024).
4. Vijay Badrinarayanan, Alex Kendall, Roberto Cipolla SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation. URL: arxiv.org/abs/1511.00561 (дата обращения: 02.08.2024).

5. Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, Li Fei-Fei ImageNet Large Scale Visual Recognition Challenge. URL: arxiv.org/abs/1409.0575 (дата обращения: 02.08.2024).
6. Maggie Wigness, Sungmin Eum, John G. Rogers III, David Han, Heesung Kwon A RUGD Dataset for Autonomous Navigation and Visual Perception in Unstructured Outdoor Environments. URL: rugd.vision/pdfs/RUGD_IROS2019.pdf (дата обращения: 21.08.2024).
7. Andrew Howard, Mark Sandler, Grace Chu, Liang-Chieh Chen, Bo Chen, Mingxing Tan, Weijun Wang, Yukun Zhu, Ruoming Pang, Vijay Vasudevan, Quoc V. Le, Hartwig Adam Searching for MobileNetV3. URL: arxiv.org/abs/1905.02244 (дата обращения: 04.09.2024).
8. Seunghyeop Nam, Changseok Woo, Sinkyu Kang, Tuan Anh Nguyen, Dugki Min SLAM_DRLnav: A SLAM-Enhanced Deep Reinforcement Learning Navigation Framework for Indoor Self-driving. URL: ieeexplore.ieee.org/document/10181042 (дата обращения: 15.09.2024).