РАЗРАБОТКА МОДЕЛИ СЕМАНТИЧЕСКОЙ СЕГМЕНТАЦИИ ОБЛАКОВ ТОЧЕК ОБЪЕКТОВ В ПОМЕЩЕНИЯХ

Штайн В.А.
Томский политехнический университет, ИШИТР, студент гр. 8И12,
e-mail: vas136@tpu.ru

Аннотация

В данной работе описываются результаты процесса разработки модели для решения задачи семантической сегментации объектов помещений и интерьеров. Приведены результаты численных экспериментов на наборе данных ScanNet по метрикам Accuracy, IoU.

Ключевые слова: семантическая сегментация, облака точек, нейронные сети, PointNet, трёхмерное зрение.

Введение

Семантическая сегментация – одна из задач в области компьютерного зрения, целью которой является классификация отдельных пикселей (в случае изображений) или точек (в случае облаков точек).

Облака точек – данные, представляющие собой описание множества точек в трёхмерном пространстве. Для каждой точки, помимо самих координат, могут храниться метаданные – это может быть цвет, материал, класс, и другие признаки. Данные, как правило, получаются с использованием лазерных сканеров и характеризуются пространственной разреженностью.

Семантическая сегментация облаков точек актуальна для задач робототехники, навигации, инспекции зданий и планирования интерьеров, так как является их составной частью. Актуальность работы обусловлена практической потребностью в разработке моделей семантической сегментации облаков точек для анализа сцен, снятых внутри помещений при помощи LiDAR [1].

Целью работы является оценка эффективности применения архитектуры PointNet [2] для решения задачи семантической сегментации при обучении на выбранном наборе данных. На основе результатов будут предложены возможные модификации модели, а также оценена применимость для задач сегментации текущей её версии. Стоит отметить, что работа находится в процессе, и текущие результаты являются промежуточными.

Набор данных

Большинство работ, связанных с исследованием моделей для облаков точек, используют простые наборы данных с одиночными объектами, к примеру, ShapeNet [3], и ModelNet [4]. В некоторых сценариях один объект может содержать несколько классов – составных частей. Такие наборы данных приемлемы для базовых исследований, однако недостаточны для моделирования реальных сценариев.

В рамках данной работы была поставлена цель изучить применимость моделей для семантической сегментации облаков точек в условиях, приближенных к реальным. Следовательно, как и в случае с прикладными задачами для семантической сегментации изображений, стоит рассматривать для ответа на вопрос о применимости моделей составные сцены, включающие в себя множество объектов. Существует ряд наборов данных для таких задач - для улиц (дорог, деревьев, пешеходов, зданий), к примеру, KITTI [5], и для внутренних помещений (столы, стулья, стены), например, S3DIS [6] и ScanNet [7]. В рамках сравнения последних двух, был выбран ScanNet и его подмножество из 40 классов, так как он имеет большее количество сцен, комнат, и семантических категорий.

Предварительный анализ выбранного набора данных

На этапе общего обзора полученных данных было выяснено, что набор данных ScanNet имеет 40 классов, соответствующих разметке NYU40[8], и содержит один дополнительный класс для неразмеченных данных, приводя к 41 результирующему классу, для которого был произведён маппинг цветов для соответствия каждого класса своему цвету. Всего в наборе содержится 1513 сцен. Некоторые сцены представляют собой вариации 707 уникальных помещений, снятых в других условиях, или являющихся частью основных помещений (например, отдельно отсканированная ванная комната). Следовательно, в среднем, каждая сцена имеет две вариации. Координаты в наборе данных не нормализованы, каждая единица расстояния соответствует одному реальному метру.

В результате проведения статистического анализа по всем предоставленным в наборе сценам, было выяснено, что в среднем на сцену приходится около 150 тысяч точек. Средний размер помещений в выборке – приблизительно 5x5x2,4м. На основе этих результатов было принято решение о необходимости предобработки данных перед обучением модели. В целях повышения скорости обработки данных моделью для каждой сцены было проведено прореживание до десяти тысяч точек, выбранных случайным образом. Также была проведена нормализация координат, так как это предполагается оригинальной архитектурой PointNet.

Примеры проекций всех точек на плоскостях приведены на рис. 1.

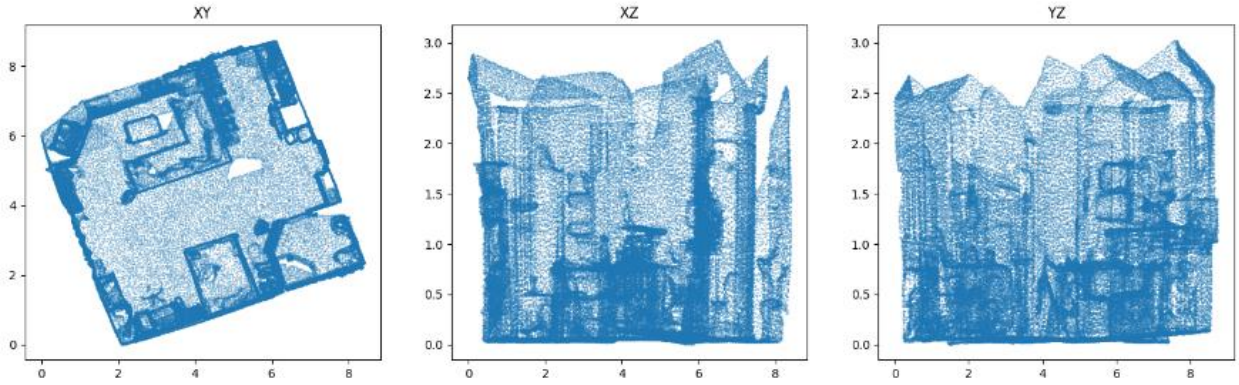

Рис. 1. Проекции точек в трёхмерном пространстве для сцены в ScanNet без нормализации координат

Разработка модели для семантической сегментации

В качестве основы для решения задачи семантической сегментации облака точек была выбрана оригинальная архитектура PointNet с блоком для сегментации, учитывающим помимо 1024 глобальных признаков, 64 локальных. Общая архитектура приведена на рис. 2.

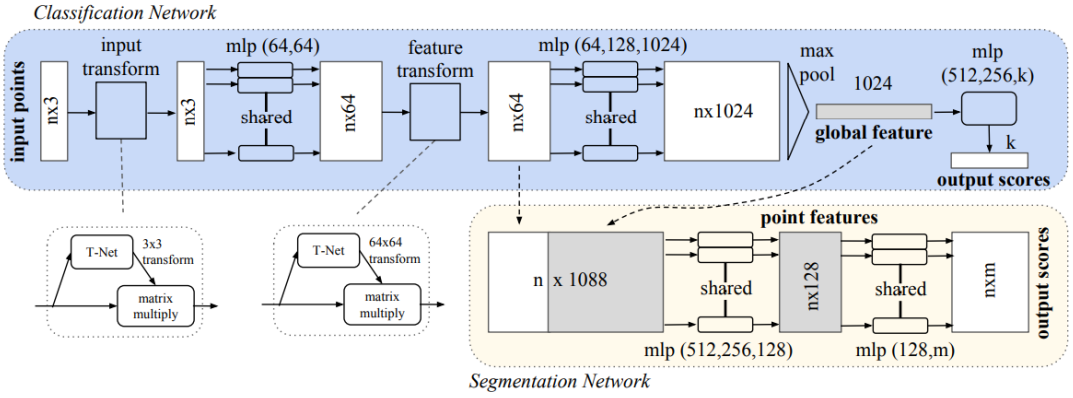

Рис. 2. Архитектура модели PointNet с блоком для семантической сегментации

В работе для выходных данных использовалась только ветка, отвечающая за семантическую сегментацию. Блок классификации в итоговой архитектуре не использовался.

Можно также отметить вспомогательный модуль преобразования T-Net, предназначенный для обеспечения инвариантности к аффинным преобразованиям. В результате её работы обучается матрица преобразования, обеспечивающая устойчивость к изменениям геометрических свойств объектов.

После того, как входные данные проходят через T-Net, они попадают в многослойный перцептрон, из которого при помощи max pooling создаётся глобальный вектор признаков. Этот глобальный вектор в нашем случае проходит также с объединёнными локальными признаками через блок семантической сегментации. Признаки извлекаются серией свёрток в полносвязных слоях, выдавая на выходе вероятности для k классов.

В ряде экспериментов [9-10] эмпирически доказана эффективность функций активации ACON и Swish сравнительно с ReLU и её модификациями – наблюдаемая точность на большинстве приведённых наборах данных выше без увеличения вычислительных затрат. На основании этого было предложено реализовать данные функции активации в качестве возможных улучшений оригинальной архитектуры.

Численные эксперименты

Для входных данных при обучении и валидации была выполнена предобработка – их координаты нормализованы, а каждая сцена была ограничена случайными подвыборками из десяти тысяч точек. Параметры обучения модели: размер батча равен 16, использовался оптимизатор Adam (lr=10e-4) с планировщиком, изменяющим параметр learning rate при отсутствии изменений метрики mIoU между эпохами. Основная функция потерь, используемая во время обучения – кросс-энтропия, для которой описана формула 1 ниже.

$$H(p,q) = -\sum_{x} p(x) * \log q(x), \quad (1)$$

где
x – класс
p(x) – истинное распределение вероятности для класса
q(x) – предсказанное распределение вероятности для класса

Помимо метрики mIoU и общей точности (overall accuracy), учитывалась средняя точность по классам (Mean Accuracy).

Результаты обучения оценивались с разными функциями активации, упомянутыми в предыдущем разделе – ReLU, Swish, ACON-C, формулы для которых определены формулами 2-4 соответственно.

$$ReLU(x) = \begin{cases} x, & x \geq 0 \\ 0, & x < 0' \end{cases} \quad (2)$$

$$Swish(x) = x * \sigma(x) = x * \frac{1}{1 + e^{-x}}, \quad (3)$$

$$ACON\ C(x) = (p1 - p2) * x * \sigma(\beta x) + p2 * x, \quad (4)$$

где
p1 – нелинейная “Swish”-составляющая функции
p2 – линейная “ReLU”-составляющая функции
β – обучаемый параметр ACON-блока

Результаты обучения модели на 20 эпохах с разными функциями активации в слоях и соответствующие метрики на проверочном наборе данных приведены в таблице 1.

Таблица 1. Метрики сегментации сцен из тестовой выборки набора данных ScanNet

Функция активации	Overall Accuracy	Mean Accuracy	mIoU
ReLU	0.457	0.073	0.041
ACON-C	0.465	0.077	0.043
Swish	0.476	0.091	0.055

Лучшие показатели можно отметить для функции активации Swish. Работа находится в процессе, и подбор оптимальных гиперпараметров при этом не выполнялся. Пример предсказания лучшей модели с функцией активации Swish для одной из проверочных сцен продемонстрирован на рис. 3.

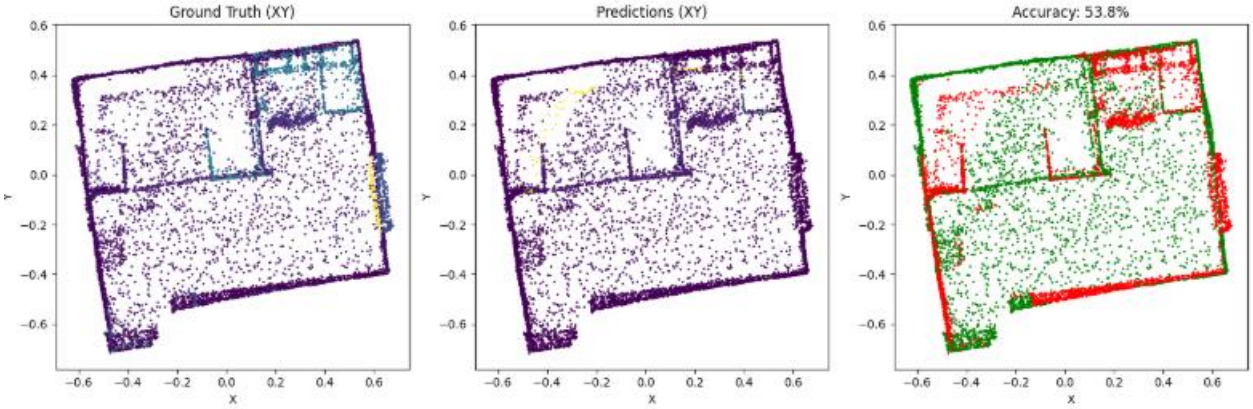

Рис. 3. Пример предсказания обученной модели:
слева – проекция с оригинальной разметкой сцены в цветах хранимых метаданных;
по центру – проекция предсказаний модели в цветах соответствующих классов;
справа – проекция с разметкой корректности предсказания, где красный – неверно предсказанный класс, зелёный – правильно предсказанный

Результаты и их обсуждение

В сужении до 10 лучших классов можно сделать вывод о том, какие модель определяет лучше всего – результаты продемонстрированы в таблице 2.

Таблица 2. Значения метрик для наиболее хорошо определяемых классов

Класс	IoU	Accuracy
Пол	0.740	0.987
Потолок	0.463	0.647
Стена	0.389	0.882
Стул	0.276	0.523
Кровать	0.093	0.175
Шкаф	0.093	0.128
Другой объект без класса (otherprop)	0.082	0.145
Душевая занавеска	0.057	0.058
Туалет	0.049	0.052
Ванна	0.037	0.041
Средние метрики	**0.228**	**0.364**

Классы в наборе данных были представлены несбалансированно, и модель обучилась определять лучше всего самые распространённые – пол, потолок, стены, стулья. Стены и пол

занимают около 45 % точек набора данных, в то время как ещё 10 % не размечены. Около 24 классов, по отдельности, представляют менее одного процента. Однако можно сделать вывод и об определении моделью признаков объектов – классы туалета и ванной распространены меньше, чем, например, классы дверей и окон, но предсказываются лучше.

Время инференса для одной сцены без учёта предобработки (нормализации) порядка одной секунды на примере среднего времени обработки на этапе валидации. На данном этапе разработки лучшее применение модели – семантическая сегментация отобранных классов, которые модель определяет с наилучшей точностью – к примеру, пола, стен и стульев.

Заключение

Для решения задачи семантической сегментации объектов в помещениях на примере модели PointNet с модифицированной архитектурой были проведены численные эксперименты для набора данных ScanNet.

Обученная модель подходит для решения задачи семантической сегментации конкретных классов, которые наиболее точно предсказываются – пола, стен, потолков, стульев. Для других классов точность недостаточно высока для однозначного определения объектов. Время инференса, если допустить возможность незначительных затрат на нормализацию, приемлемо для некоторых систем реального времени, не требующих быстрого реагирования, для сцен, сходных со средними параметрами сцен из набора данных ScanNet.

Возможными улучшениями для обучения модели могут быть улучшение архитектуры, балансировка классов, подбор оптимальных гиперпараметров при обучении.

Список использованных источников

1. The Basics of LiDAR - Light Detection and Ranging - Remote Sensing // NSF NEON: сайт. – URL: neonscience.org/resources/learning-hub/tutorials/lidar-basics (дата обращения 28.03.2025).

2. PointNet: Deep Learning on Point Sets for 3D Classification and Segmentation / Charles R. Qi, Hao Su, Kaichun Mo, Leonidas J. Guibas // arXiv:1612.00593v2.

3. ShapeNet: сайт. – URL: shapenet.org/ (дата обращения 28.03.2025).

4. Princeton ModelNet: сайт. –URL: modelnet.cs.princeton.edu (дата обращения: 28.03.2025).

5. The KITTI Vision Benchmark Suite // svlibs.net: сайт. – URL: cvlibs.net/datasets/kitti/ (дата обращения 28.03.2025).

6. Stanford Large-Scale 3D Indoor Spaces Dataset (S3DIS) // Stanford Data Farm: сайт. – URL: redivis.com/datasets/9q3m-9w5pa1a2h (дата обращения 28.03.2025).

7. ScanNet: Richly-annotated 3D Reconstructions of Indoor Scenes // ScanNet: сайт. – URL: scan-net.org/ (дата обращения 28.03.2025).

8. NYU40 Semantic Labels Description // GitHub.com: сайт. – URL: github.com/apple/ml-hypersim/blob/main/code/cpp/tools/scene_annotation_tool/semantic_label_descs.csv (дата обращения 28.03.2025).

9. Activate or Not: Learning Customized Activation / Ningning Ma, Xiangyu Zhang, Ming Liu, Jian Sun // Computer Vision Foundation open access: сайт. – URL: openaccess.thecvf.com/content/CVPR2021/papers/Ma_Activate_or_Not_Learning_Customized_Activation_CVPR_2021_paper.pdf (дата обращения 28.03.2025).

10. Swish: A Self-Gated Activation Function / Prajit Ramachandran, Barret Zoph, Quoc V. Le // arXiv: 1710.05941v1.