УДК 528.9
DOI 10.33764/2618-981X-2024-1-260-268

М. А. Федорочев[1]*

Выбор оптимальной архитектуры нейронной сети для распознавания объектов городской застройки

[1] Пермский государственный национальный исследовательский университет, г. Пермь, Российская Федерация
* e-mail: fedorochev.m.a@yandex.ru

Аннотация. В контексте ускорения урбанизационных процессов, данное исследование сосредоточено на использовании нейронных сетей для автоматического распознавания объектов градостроительства. Современные методы компьютерного зрения, основанные на применении нейронных сетей, предоставляют новые возможности для создания автоматизированных систем мониторинга и управления городской средой. В ходе исследования были проанализированы одни из наиболее распространенных моделей: U-Net и Mask R-CNN. Исходя из анализа метрик и характеристик моделей, U-Net, благодаря своей простоте и быстроте обучения, может быть предпочтительной при работе с большим объемом данных. Однако, Mask R-CNN, несмотря на более медленное обучение, показывает лучшие результаты при распознавании городских объектов и не требует дополнительной постобработки. Для оптимизации производительности модели и повышения точности распознавания объектов, необходимо расширение датасета за счет увеличения объема данных и их детализации.

Ключевые слова: глубокое обучение, сегментация, метрики, городская застройка, нейронная сеть

M. A. Fedorochev[1]*

Choosing the optimal neural network architecture for recognizing urban development features

[1] Perm State University, Perm, Russian Federation
* e-mail: fedorochev.m.a@yandex.ru

Abstract. In the context of accelerating urbanization processes, this research focuses on the use of neural networks for automatic recognition of urban development objects. Modern computer vision methods based on the use of neural networks provide new opportunities for creating automated systems for monitoring and managing the urban environment. In this research, some of the most common models were analyzed: U-Net and Mask R-CNN. Based on the analysis of metrics and model characteristics, U-Net, due to its simplicity and speed of learning, may be preferable when working with a large amount of data. However, Mask R-CNN, despite slower learning, shows better results in recognizing urban objects and does not require additional postprocessing. To optimize the performance of the model and improve the accuracy of object recognition, it is necessary to expand the dataset by increasing the amount of data and their detail.

Keywords: deep learning, segmentation, metrics, urban development, neural network

Введение

Выбор оптимальной нейронной сети является ключевым этапом для достижения высокой точности распознавания объектов градостроительства, включая здания, дороги, зеленые зоны и другие элементы. Эффективная нейронная сеть должна обладать способностью адаптации к различным условиям среды и обеспечивать высокую производительность даже при ограниченных вычислительных ресурсах. В отличие от многих других методов машинного обучения, нейронные сети могут автоматически извлекать и использовать признаки из данных, что упрощает процесс подготовки данных и улучшает производительность модели.

Подготовка набора данных для обучения

В связи с уникальностью каждого региона и глобальным характером доступных наборов данных, возникает необходимость создания специализированных наборов данных для более точного отображения объектов на территории России. Открытые наборы спутниковых данных могут быть оптимальными для определенных задач, однако для создания цифрового двойника города требуется более высокая точность, которую могут обеспечить ортофотопланы. Для обучения нейронной сети создан набор обучающих данных на основании ортофотоплана с разрешением 0,07 м. Набор включает три класса сцен с различным типом городских объектов – здания, растительность и дороги [1]. Несбалансированные обучающие выборки напрямую влияют на конечный обучающий эффект модели. При этом количество объектов в разных классах не обязательно идентично. После извлечения выборок из изображения данные следует дополнить, чтобы гарантировать хорошее представление признаков типов с небольшими размерами выборки. Набор выборок может быть расширен с помощью таких процессов, как ротация, перевод и аугументация [2].

В процессе подготовки данных для обучения модели, ортофотоплан разделяется на отдельные квадратные сегменты, каждый из которых содержит истинную маску интересующего объекта по заданным классам. Эти сегменты затем используются в качестве обучающих выборок. Архитектуры моделей могут иметь различные требования к входным данным, однако, ключевым требованием является то, что объект исследования должен занимать значительную часть входного изображения.

Учитывая разрешение ортофотоплана и размеры объектов на местности, можно предположить, что оптимальным размером для тайлов будет 512x512 пикселей. Это соответствует примерно 36x36 метрам на местности. Такой размер позволит полностью захватить большую часть объектов, не сильно замедляя скорость обучения и увеличивая относительное присутствие объектов интереса на снимке [2].

Подбор необходимой архитектуры нейронной сети

В области глубокого обучения и анализа пространственных данных существует множество предварительно обученных моделей и инструментов, которые

могут значительно упростить процесс разработки и реализации нейронных сетей. Предварительно обученные модели, такие как те, которые доступны в библиотеках глубокого обучения, таких как TensorFlow и PyTorch, уже обучены на больших наборах данных и могут быть адаптированы для конкретной задачи с помощью техник, таких как передача обучения. Это может значительно ускорить процесс обучения и улучшить производительность модели, особенно если доступных данных мало. Однако модели, первоначально обученные на данных спутниковых данных, не предназначены к работе с данными с разрешением 10 см, так как они предоставляют более детализированную информацию, которая может включать в себя новые признаки и паттерны, влияющие на качество работы модели.

Выбор архитектуры нейронной сети определяется техническими требованиями, спецификой задачи и размером обучающего набора данных. Доступные вычислительные ресурсы могут ограничивать выбор архитектуры, особенно для задач, требующих высокой скорости выполнения. Специализированные архитектуры, способные создавать детализированные маски объектов, могут быть необходимы для задач пиксельной сегментации. Кроме того, размер обучающего набора данных влияет на выбор архитектуры, где сложные модели могут переобучиться при недостатке данных, в то время как большие наборы данных могут обеспечить лучшую производительность с более сложными моделями.

Для задач, связанных с сегментацией объектов, две архитектуры часто выделяются как особенно эффективные: Mask R-CNN и U-Net.

Mask R-CNN, расширяющая Faster R-CNN, и U-Net являются архитектурами нейронных сетей, применяемыми для сегментации изображений. Mask R-CNN, включающая двухэтапный процесс с использованием сети предложения регионов и детектора R-CNN, эффективна в задачах распознавания объектов городской застройки. U-Net, полностью сверточная сеть с кодировщиком и декодером, широко используется в анализе спутниковых изображений и способна генерировать высококачественные маски сегментации [7, 8].

Для выбора между архитектурами нейронных сетей, такими как Mask R-CNN и U-Net, важно провести сравнительный анализ их производительности на конкретной задаче. Это требует создания обучающей выборки и оценки результатов по нескольким метрикам. Обучающая выборка создана на основе данных о городской застройке Первоуральска, полученной с помощью аэрофотосъемки. Это позволит моделям обучиться на конкретных примерах объектов городской застройки, которые встречаются в этом городе и типичны для большей части городов Российской Федерации.

Сравнительный анализ архитектур

После обучения моделей на выборке, можно оценить их производительность, используя различные метрики, такие как точность, полнота, F-мера и средняя точность [11, 12, 13]. Это поможет определить, какая модель лучше справляется с задачей выделения объектов городской застройки. В зависимости от результатов, можно решить, какую модель предпочтительнее использовать выделения объектов в городской застройке. Если разница в производительности

между моделями незначительна, выбор может зависеть от других факторов, таких как скорость обучения и инференции, а также удобство использования и интеграции с другими системами.

Для обучения данных использовался набор, который включал в себя небольшие обучающие выборки на территорию города, захватывая как частный сектор, так и многоэтажную застройку. При проведении сравнительного анализа эффективности Mask R-CNN и U-Net, были использованы одинаковые обучающие выборки, содержащие данные о растительности, зданиях и дорогах на одной и той же территории. Обе модели были обучены на этих данных и протестированы на соответствующих тестовых выборках.

В данном эксперименте наборы данных были случайным образом разделены на два непересекающихся набора: 80 % выборок с 2470 изображениями были приняты в качестве обучающих наборов, а по 10 % выборок с 308 изображениями были использованы в качестве проверочных и валидационных наборов.

Для сравнения качества и скорости обучения различных архитектур необходимо анализировать параметры обучения, включая скорость обучения (learning rate), время одной эпохи, потери на этапах обучения, измеряющую разницу между предсказаниями модели и истинными значениями, валидации, а также итоговую точность классификации. Скорость обучения подбирается экспериментально, начиная, например, с 0,001, и далее корректируется для достижения оптимальной эффективности и стабильности обучения. Автоматические методы, такие как learning rate finders, могут динамически адаптировать скорость обучения в процессе, предоставляя оптимальный способ поиска этого параметра. Этот метод включает проведение теста с постепенным увеличением скорости обучения после каждого обработанного набора данных (batch), с последующим анализом графика зависимости потерь от скорости обучения, поэтому он был использован для нахождения оптимальных значений.

На графиках (рис.1) видно, что функция потерь (Loss) уменьшается при увеличении количества обработанных пакетов (Batches processed), это означает, что модель обучается и адаптируется к обучающим данным. В процессе обучения модель постепенно корректирует свои параметры (веса), чтобы минимизировать функцию потерь. Это происходит на каждом шаге обучения, и с каждым новым обработанным пакетом модель получает возможность уточнить свои предсказания.

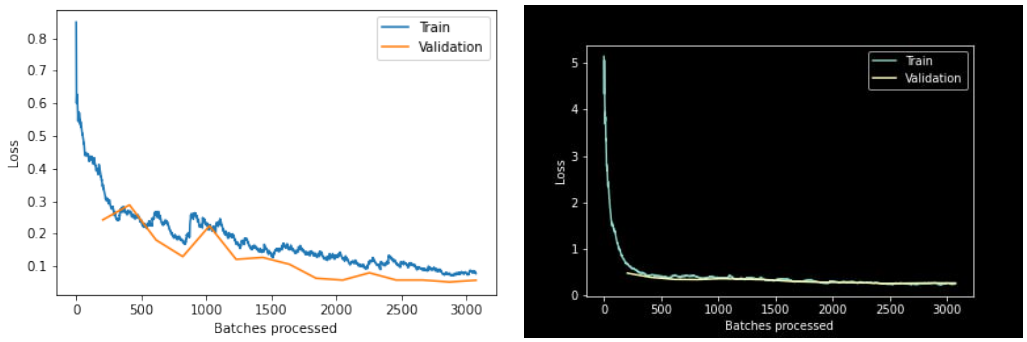

Рис. 1. Графики обучения моделей U-Net (слева) и Mask R-CNN (справа)

Уменьшение функции потерь на обучающих данных не всегда означает, что модель будет хорошо работать на новых, ранее использованных в обучении данных. Это связано с тем, что модель может переобучиться, то есть слишком хорошо адаптироваться к обучающим данным, потеряв при этом способность к обобщению. В случае, когда функция потерь на обучающих данных продолжает снижаться, в то время как на валидационных данных наблюдается ее рост, это может указывать на переобучение модели. Однако, как видно из (таб.1), функция потерь на валидационных данных также продолжает уменьшаться с каждым новым пакетом, что свидетельствует о продолжающемся обучении модели без признаков переобучения.

Таблица 1

Результаты первого обучения моделей
на основе архитектур U-Net и Mask-RCNN

Эпоха	Потери на обучающей выборке (U-Net)	Потери на валидационной выборке (U-Net)	Время обучения, мин. (U-Net)	Потери на обучающей выборке (Mask R-CNN)	Потери на валидационной выборке (Mask R-CNN)	Время обучения, мин. (Mask R-CNN)
0	0.344409	0.242502	31:09	0.676037	0.480075	47:01
1	0.262926	0.288320	30:35	0.438373	0.389590	45:33
2	0.261236	0.180606	30:42	0.426197	0.345558	45:54
3	0.176830	0.129805	30:48	0.403556	0.341466	46:06
4	0.218454	0.224050	30:41	0.376633	0.360927	51:09
5	0.172256	0.121447	30:42	0.339450	0.350274	50:15
6	0.146376	0.127197	30:39	0.323808	0.336017	46:22
7	0.153877	0.106253	30:19	0.319094	0.302142	47:09
8	0.133059	0.063350	31:17	0.303575	0.287357	47:48
9	0.122241	0.057727	31:00	0.289599	0.281104	52:12
10	0.110619	0.080129	30:40	0.281095	0.273731	48:40
11	0.111581	0.057894	35:52	0.247082	0.257894	43:18
12	0.088422	0.057956	30:08	0.252228	0.263275	49:41
13	0.078393	0.051883	28:55	0.258187	0.268252	45:01
14	0.078336	0.057338	30:03	0.254529	0.266271	50:04

В результате анализа процесса обучения и оценки среднего показателя точности (Average Precision Score), было определено, что U-Net обеспечивает более эффективное и быстрое обучение в сравнении с Mask R-CNN. Это различие в производительности в основном обусловлено отличиями в архитектуре этих сетей.

U-Net, который отличается эффективным использованием пространственной информации и агрегацией контекста на разных уровнях, продемонстрировал высокую способность к извлечению и интерпретации признаков на уровне пикселей. Этот аспект делает U-Net применимым для задач сегментации, где важно учитывать локальные детали и структуру объектов. Mask R-CNN, в силу особенностей его архитектуры, включающей в себя детекцию и сегментацию объектов, приводит к более сложному процессу обучения и более длительному времени конвергенции [10]. Тем не менее, высокий уровень точности Mask R-CNN в сегментации объектов среди известных архитектур делает его ценным инструментом в задачах, где требуется высокая точность, детализированное разделение объектов и создание ограничивающих полигонов.

Таблица 2

Результаты метрик после трех обучений

Архитектура	Класс	Precision	Recall	F1 Score	Average Precision (AP)	Количество объектов (features)
Mask R-CNN	Здания	0,894345	0,859414	0,881509	0,877664	3211
	Дороги	0,833123	0,821551	0,821095	0,850950	2255
	Растительность	0,866650	0,816217	0,841999	0,863740	4856
	Все классы	0,864706	0,832394	0,848201	0,864118	10322
U-Net	Здания	0,682158	0,871579	0,818510	0,791590	3211
	Дороги	0,641251	0,851081	0,731795	0,761515	2255
	Растительность	0,646694	0,859029	0,678266	0,742772	4856
	Все классы	0,656701	0,860563	0,742857	0,761959	10322

Для объективного и более полного сравнения моделей, обученных на идентичных наборах данных, необходимо вычислить ключевые метрики, такие как Точность (Precision), Полнота (Recall), F1 Score и Средняя Точность (Average Precision, AP). Эти показатели являются стандартными метриками для оценки эффективности моделей машинного обучения, в частности, в задачах классификации, детекции и сегментации объектов. Метрики вычисляются на основании матрицы ошибок (confusion matrix), которая включает в себя истинно положительные (TP), истинно отрицательные (TN), ложно положительные (FP) и ложно отрицательные (FN) результаты. Эти метрики помогают оценить эффективность модели, учитывая различные аспекты ее производительности, они могут быть особенно полезны при сравнении различных моделей.

Точность (Precision) отражает способность модели правильно идентифицировать только релевантные экземпляры. Это отношение числа истинно поло-

жительных результатов к общему числу положительных результатов, предсказанных моделью. Чем выше точность, тем меньше ложных срабатываний [9].

Полнота (Recall) измеряет способность модели идентифицировать все релевантные экземпляры в данных. Это отношение числа истинно положительных результатов к общему числу реальных положительных случаев. Чем выше полнота, тем меньше вероятность пропустить положительный случай [9].

F1 Score является гармоническим средним между точностью и полнотой. Он стремится уравновесить оба показателя и является полезной метрикой, когда классы несбалансированы. Чем выше F1 Score, тем более эффективной является модель при одновременном учете точности и полноты [6, 9].

Средняя Точность (Average Precision, AP) является обобщением понятия точности для случая, когда порог классификации варьируется. Она вычисляется как средневзвешенное значение точности при каждом изменении значения отзыва. Чем выше AP, тем лучше модель справляется с задачей классификации при различных пороговых значениях. Итоговое значение указывает на реальные прогнозы данных для определенного класса [9].

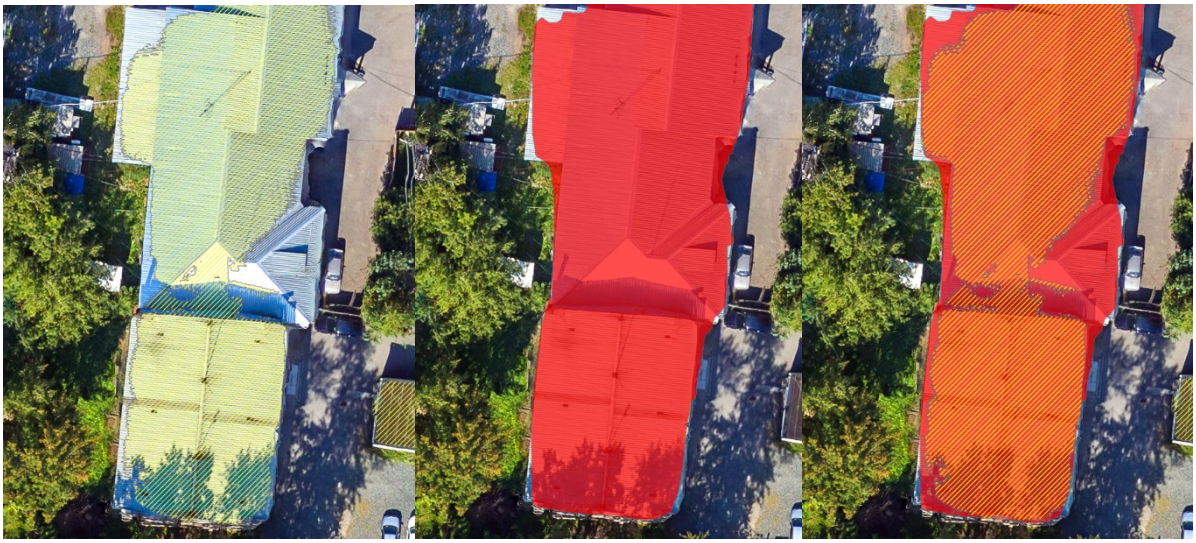

Рис. 2. Результаты определения зданий, красный фон (Mask R-CNN) и желтая штриховка штриховка (U-Net)

Исходя из результатов значений по метрикам, представленных в таб.2, выделение объектов на основании модели, обученной на архитектуре Mask-RCNN, показывает лучшие результаты практически по всем метрикам, за исключением Recall, где значения практически идентичные.

Согласно полученным результатам - U-Net может проявлять несогласованность в сегментации, приводя к неровным границам и разрозненным данным внутри выделяемого объекта. Для достижения точной сегментации объектов, соответствующей реальным формам, требуется последующая обработка данных, включающая заполнение пропусков и геометрическую коррекцию. Однако, эффективность такого подхода не гарантирована и может варьироваться в зависи-

мости от параметров постобработки и характеристик исходных данных. В то же время маски, полученные с использованием Mask R-CNN, обычно обладают большей целостностью и легче поддаются обработке, что делает данную модель более предпочтительной для задач, требующих высокой точности сегментации и предварительной детекции объектов.

Заключение

Современные тенденции обуславливают повышение популярности использования нейронных сетей, особенно для long-term задач, где они показывают свою максимальную пользу в виду специфики подготовки данных и последующего обучения. В данной работе были рассмотрены две наиболее популярные архитектуры глубокого обучения, выполняющие сегментацию изображения и получая на выходе маску требуемых объектов. По итогу при лучших значениях обученной модели U-Net она оказалась хуже, чем Mask R-CNN практически по всем параметрам при выделении объектов на новой территории, не входящей в исходную выборку обучающих данных.

В контексте анализа городской застройки, результаты, полученные с использованием различных архитектур глубокого обучения, могут в значительной степени зависеть от специфических характеристик каждой архитектуры. Архитектура U-Net, благодаря своей относительной простоте и быстроте обучения, может оказаться предпочтительной при работе с большим объемом данных и ограниченными вычислительными ресурсами. С другой стороны, архитектура Mask R-CNN может быть более подходящим выбором для задач, требующих более точного определения форм зданий и других объектов. Это обусловлено ее способностью предварительно создавать маски объектов перед их сегментацией, что может привести к более точным и детализированным результатам.

Учитывая, что для эффективного обучения моделей требуется значительное количество данных и времени, параллельное обучение двух моделей представляется нецелесообразным. Вместо этого следует выбрать наиболее оптимальную модель, которой, как показывает анализ, является Mask R-CNN. Несмотря на то, что скорость обучения этой модели в среднем меньше на 55 %, она демонстрирует значительно лучшие результаты при идентификации объектов городской среды, увеличивая показатели метрик F1 Score и AP на 10 % в сравнении с U-Net. Кроме того, Mask R-CNN не требует дополнительной постобработки, что в конечном итоге практически компенсирует разницу в скорости обучения.

Полученные результаты выявления объектов являются не вполне удовлетворительными в контексте выделения по данным ортофотопланов, что подчеркивает необходимость дополнительного обучения модели. Существует необходимость оптимизации датасета для повышения эффективности модели и точности идентификации объектов градостроительства. Расширение объема данных и углубление их детализации обеспечит модели возможность обучения на более широком спектре примеров и улучшит ее способность к обобщению. Кроме того, интеграция новых данных в датасет обеспечит более полное представление о городской среде.

БИБЛИОГРАФИЧЕСКИЙ СПИСОК

1. Экспорт обучающих данных для глубокого обучения [Электронный ресурс]. URL: https://desktop.arcgis.com/ru/arcmap/latest/tools/spatial-analyst-toolbox/export-training-data-for-deep-learning.htm

2. Hensman, P., & Masko, D. (2015). The Impact of Imbalanced Training Data for Convolutional Neural Networks. Degree Project in Computer Science, DD143X. Supervisor: Pawel Herman. Examinator: Örjan Ekeberg. May 8, 2015.

3. Aarno Oskar Vuola, Saad Ullah Akram, Juho Kannala Mask-RCNN and U-net Ensembled for Nuclei Segmentation, January 2019, 6 страниц

4. Everything about Mask R-CNN: A Beginner's Guide [Electronic resource]: https://viso.ai/deep-learning/mask-r-cnn/

5. Mask R-CNN for Object Detection and Segmentation [Electronic resource]: https://github.com/matterport/Mask_RCNN

6. Takahashi, K., Yamamoto, K., Kuchiba, A., & Koyama, T. (2022). Confidence interval for micro-averaged F1 and macro-averaged F1 scores. Applied Intelligence, 52, 4961–4972.

7. How Mask R-CNN works [Electronic resource]: https://developers.arcgis.com/python/guide/how-maskrcnn-works/.

8. Wang, Y., Li, S., Teng, F., Lin, Y., Wang, M., & Cai, H. (n.d.). Improved Mask R-CNN for Rural Building Roof Type Recognition from UAV High-Resolution Images: A Case Study in Hunan Province, China.

9. Метрики классификации и регрессии [Электронный ресурс]: https://education.yandex.ru/handbook/ml/article/metriki-klassifikacii-i-regressii.

10. Shamir, O. (2018). Exponential Convergence Time of Gradient Descent for One-Dimensional Deep Linear Neural Networks. arXiv preprint arXiv:1809.08587v1 [cs.LG].

11. Model Evaluation [Electronic resource]: https://scikit-learn.org/stable/modules/model_evaluation.html.

12. How to Calculate Precision, Recall, F1, and More for Deep Learning Models [Electronic resource]: https://machinelearningmastery.com/how-to-calculate-precision-recall-f1-and-more-for-deep-learning-models.

13. How to Benchmark the Performance of Machine Learning Platforms [Electronic resource]: https://www.neuraldesigner.com/blog/how-to-benchmark-the-performance-of-machine-learning-platforms.