Об ускорении архитектуры сверточной нейронной сети на базе ResNet в задаче распознавания объектов дорожной сцены

М. Г. Лобанов [I], Д. Л. Шоломов [II,III]

[I] Московский государственный университет им. М.В. Ломоносова, г. Москва, Россия
[II] Институт проблем передачи информации им. А.А. Харкевича Российской академии наук, г. Москва, Россия
[III] Национальный исследовательский технологический университет "МИСиС", г. Москва, Россия

Аннотация. В настоящее время подходы к детектированию объектов дорожной сцены на основе свёрточных нейронных сетей достигли приемлемого уровня для использования их в задачах автономного управления транспортным средством и ADAS системах. Однако как правило, лучшие современные сетевые архитектуры достаточно тяжеловесны и не могут быть использованы в системах реального времени. В связи с этим наиболее остро стоит проблема ускорения сетей и нахождения оптимального баланса между скоростью и качеством их работы. В данной работе предложен метод облегчения архитектуры Deformable Convolutional Network с базовой сетью ResNet, дающий трехкратное увеличение скорости прямого прохода. При этом качество детектирования объектов дорожной сцены уменьшается не столь существенно. Кроме того, в работе приведено сравнение качества работы сети данной архитектуры при обучении на различных открытых наборах данных – BDD и MS-COCO.

Ключевые слова: детектор объектов, объекты дорожной сцены, deformable convolutional network, ResNet, ADAS системы, ускорение сверточной сети, BDD, MS-COCO, распознавание пешеходов, распознавание машин.

DOI 10.14357/20718632190305

Введение

На текущий момент одной из наиболее важных и крайне востребованных задач компьютерного зрения является задача детектирования объектов на изображении. В последнее время благодаря развитию глубоких нейронных сетей в решении данной задачи был совершен принципиальный прорыв.

Уже сегодня качество детектирования объектов на основе свёрточных нейронных сетей достигло приемлемого уровня для использования их в задачах автономного управления транспортным средством [1] (Рис 1). Но при этом лучшие результаты все равно показывают те архитектуры сетей, скоростные характеристики которых, даже на современном оборудовании, не укладываются в требования для их использования в системах реального времени. Минимальным требованием по скорости детектирования, как правило, является обработка не менее 15 fps (кадров в секунду) в fullHD разрешении, при этом лучшие на данный момент детектирующие сети имеют скорость, не превосходящую 5 fps на высокопроизводительной видеокарте типа nVidia GeForce GTX 1080Ti. Следовательно, одной из первоочередных задач является ускорение архитектуры сетей и поиск оптимального баланса между скоростью и качеством их работы.

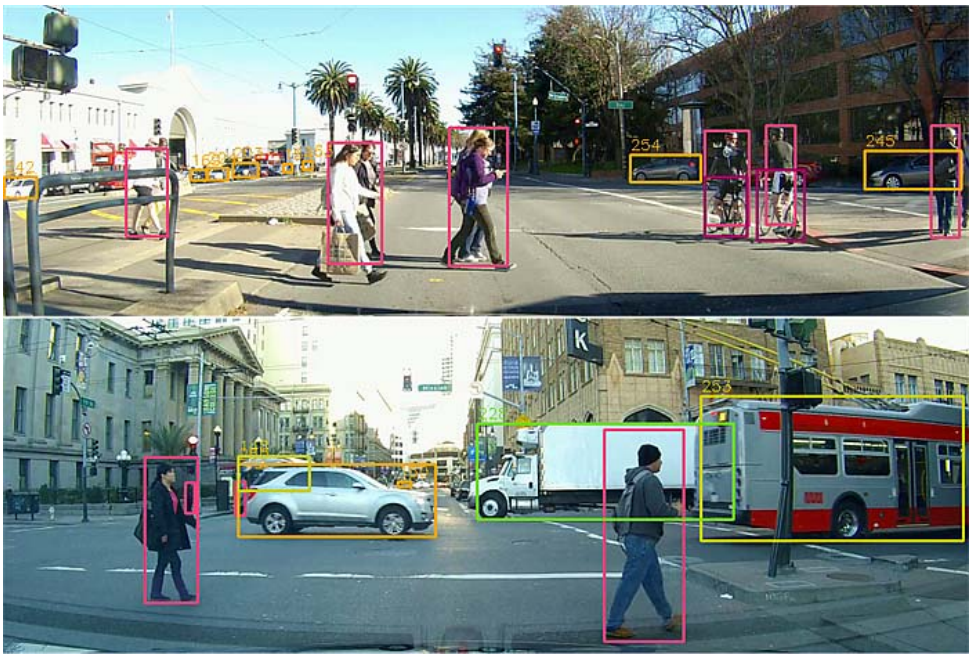

Рис. 1. Примеры детектирования объектов дорожной сцены сверточной нейронной сетью Deformable RFCN ResNet-50 v2 Light, представленной в данной статье

Исторически первые шаги по использованию нейронных сетей для детектирования объектов были сделаны Гиршеком в работе [2]. Дальнейшие улучшения модели [3-5] позволили работать с изображениями произвольного размера, при этом признаки вычисляются для изображения только один раз, а предсказание регионов, в которых могут содержаться объекты, также производится нейронной сетью.

В работе [6] была предложена новая архитектура сверточной нейронной сети Deformable RFCN, использующая так называемые deformable-свёртки, позволяющие адаптировать ядро свёртки под форму объектов каждого класса. В задаче распознавания объектов дорожной сцены данная архитектура достигает хороших результатов на наборе данных MS-COCO [7] с использованием одного масштаба для входного изображения при прямом проходе сети.

В настоящее время для обучения и тестирования моделей сетей существует несколько открытых наборов данных (датасетов). Недавно опубликованным и наиболее обширным датасетом является Berkley Deep Drive (BDD) [8]. Он содержит 70 тысяч изображений для обучения, 10 тысяч для валидации и 20 тысяч для тестирования детектирования объектов дорожной сцены. Кроме того, он включает аннотацию для обучения семантической сегментации экземпляров объектов (instance segmentation), дорожного полотна (driveable area) и линий дорожной разметки. Другим датасетом, широко используемым для задачи детектирования, является Microsoft Common Objects In Context (MS-COCO) [7]. Он содержит как дорожные сцены, так и прочие графические данные по различной тематике. Большим его плюсом является наличие огромного количества людей в различных положениях и снятых с разного расстояния. Объем MS-COCO составляет около 120 тысяч изображений, но большая часть из них не относится к дорожной сцене, что является минусом для его использования в решаемой задаче. Также следует упомянуть датасеты Cityscapes [9] и Mapillary Vistas [10], однако они содержат разметку только для семантической сегментации и не содержат разметки описывающими прямоугольниками. Автоматическое конвертирование разметки не позволяет получить данные необходимого качества, поэтому в рамках текущих исследований эти наборы было решено не использовать.

Рассмотренные выше детектирующие сети могут включать различные архитектуры базо-

вых сетей (backbones). Обычно для этого берутся модели, хорошо проявившие себя в задачах классификации объектов.

Используемая в работе [6] базовая сеть ResNet впервые была описана в [11] и совершила серьезный прорыв в конкурсе классификации изображений большого разрешения ImageNet [12]. Благодаря предложенному в работе [11] разностному обучению, стало возможным применять на практике более глубокие сети (глубиной до 1202 слоев). Широко используемая ранее базовая сеть VGG [13] имеет лишь 19 слоёв, но при этом, несмотря на увеличение глубины, модели на базе ResNet имеют меньшую вычислительную сложность. Например, модель VGG-19 имеет сложность 19,6 гигафлоп, а модель ResNet-152 при ее глубине – только 11,3 гигафлоп.

1. Ускорение базовой сети ResNet

Целью проводимых экспериментов является получение модели с качеством распознавания объектов дорожной сцены сравнимым с моделью, полученной авторами статьи [6]. При этом модель должна работать существенно быстрее, практически в реальном времени. Время обработки одного кадра должно составлять не более 60 ms на одном вычислителе. К основным объектам дорожной сцены относятся автомобили с подтипами (легковой, грузовой, автобус), пешеходы, велосипедисты и мотоциклисты. Важным является, например, относить велосипедистов и мотоциклистов к разным классам, т.к. динамика этих объектов существенно отличается, и неправильная типизация может привести к роковым ошибкам на этапе принятия решений ADAS комплексом.

В качестве первого шага исследования был проведен детальный анализ времени исполнения слоев детектирующей сети Deformable RFCN (схема сети представлена на Рис. 2). Оказалось, что большую часть времени (75%) занимает базовая сеть (стадии 1-3), поэтому было принято решение, прежде всего, ускорить ее.

Известно, что обучение глубоких нейронных сетей сложно из-за проблемы затухания градиентов [14, 15]. Чем дальше по глубине слой находится от функции ошибки, тем меньше становятся градиенты для его весов при использовании алгоритма обратного распространения. Для решения этой проблемы в работе [11] было предложено внести в сеть дополнительные связи (shortcuts, skip-connections), что позволило существенно увеличить эффективную глубину сетей.

Данная связь применяется для нескольких последовательных свёрток и осуществляется при помощи простого поэлементного сложения входа блока и выхода последнего слоя блока. ResNet состоит из нескольких разностных блоков (residual units). Блок может быть представлен формулами:

$$y_l = x_l + F(x_l, W_l),$$
$$x_{l+1} = f(y_l),$$

где x_l и x_{l+1} - вход и выход l-го блока, F- разностная функция, W_l- веса разностного блока, f- функция активации ReLU [16].

Таким образом, если $F(x_l, W_l) = 0$, разностный блок является единичным отображением входа. Авторы предположили, что такое реше-

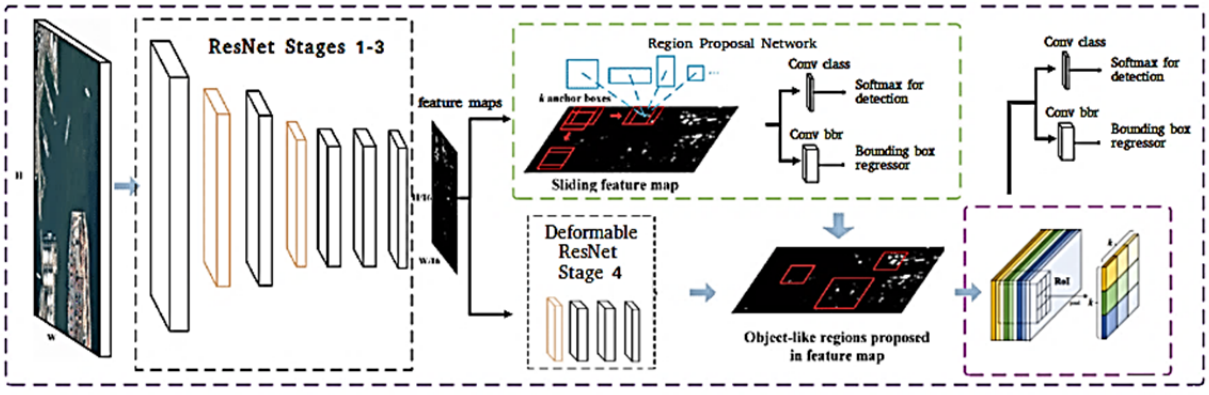

Рис. 2. Схема Deformable RFCN

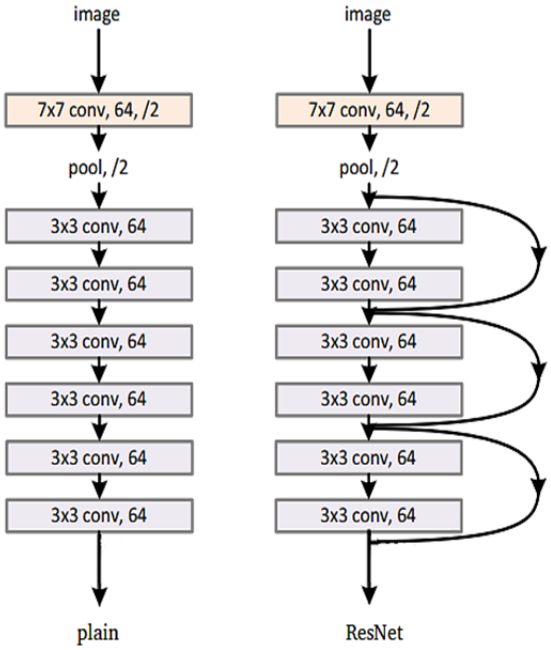

Рис. 3. Сравнение простой сверточной сети с разностной сетью ResNet

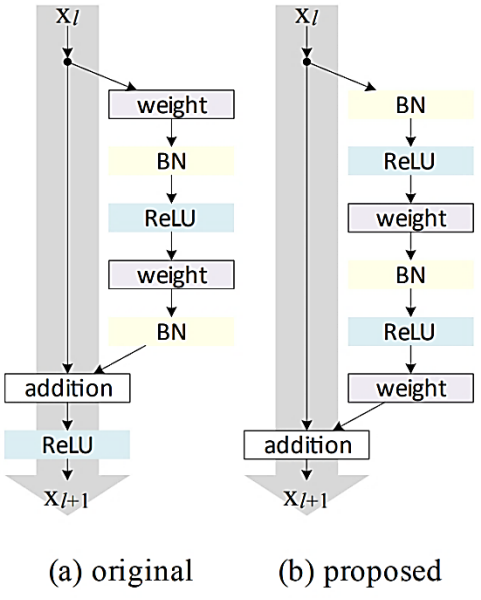

Рис. 4. Сравнение оригинального разностного блока и его второй версии

ние позволит лучше обучаться нейронной сети, поскольку если единичное отображение является в определенном смысле оптимальным, то при обучении разностного блока его веса должны просто стремиться к нулю, тогда как без введенных связей получить единичное отображение сложнее. Данное предположение успешно подтвердилось экспериментами. Сравнение архитектуры ResNet с простой сверточной сетью, использовавшейся ранее, представлено на Рис. 3.

В работе [17] был проведен анализ различных архитектур разностных блоков, в результате чего было получена улучшенная архитектура сети ResNet v2. Сравнение оригинального разностного блока и его второй версией представлено на Рис. 4.

Полученное улучшение основано на том, что сигнал напрямую проходит через всю сеть, как при прямом, так и обратном проходе. В новой версии оригинальный сигнал в каждом блоке получает некоторую добавку, при этом он не теряет существенную часть сигнала при использовании ReLU активации между блоками, как в первой версии.

Оригинальная модель сети Deformable RFCN использует старую версию разностных блоков, в текущем же исследовании для улучшения результатов было решено использовать более новую версию - ResNet v2.

Разница базовых сетей ResNet-34 и ResNet-50 заключается в том, что в ResNet-50 используются так называемые bottleneck-блоки, позволяющие увеличить глубину, а как следствие, и качество распознавания, при этом добавляя не слишком много параметров и вычислительной сложности, в версии же ResNet-34 используются обычные разностные блоки, имеющие меньшую производительность. Архитектура этих блоков представлена на Рис. 5.

ResNet-50 v2 позволила получить существенное ускорение модели со 180 мс до 120 мс на прямом проходе сети с входным разрешением 1920x1080 пикселей на видеокарте NVIDIA GeForce 1080Ti.

Архитектура ResNet состоит из нескольких стадий, на которых происходит уменьшение тензора и изменяется количество фильтров в свёрточных слоях. Для ускорения ResNet-50 было решено уменьшить количество фильтров на стадиях 1-4 в 2 раза. Данное решение позволило сократить время прямого прохода до 60мс, что уже является 3-кратным ускорением исходной сети ResNet-101.

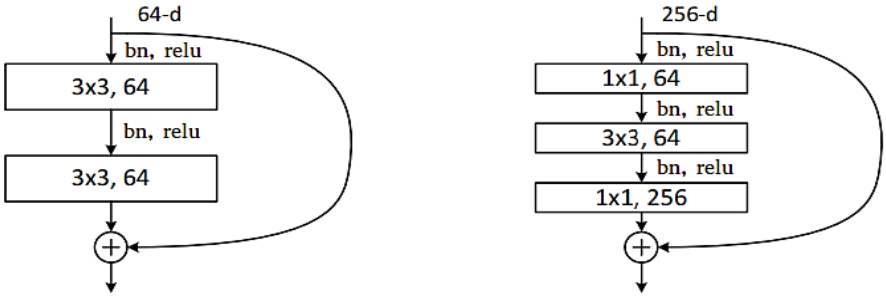

Рис. 5. Обычный разностный блок (слева) и bottleneck блок (справа)

Табл. 1. Сравнение архитектур базовых сетей

Базовая сеть		Стадия 1	Стадия 2	Стадия 3	Стадия 4
ResNet-101	Количество блоков	3	4	23	3
	Выходное разрешение	270x480	135x240	68x120	68x120
	Выходное количество фильтров	256	512	1024	2048
ResNet-50 v2	Количество блоков	3	4	6	3
	Выходное разрешение	270x480	135x240	68x120	68x120
	Выходное количество фильтров	256	512	1024	2048
ResNet-50 v2 Light	Количество блоков	3	4	6	3
	Выходное разрешение	270x480	135x240	68x120	68x120
	Выходное количество фильтров	128	256	512	1024

Сравнение архитектур представлено в Табл. 1, предложенная в работе облегченная сеть обозначена как ResNet-50 v2 Light.

В соответствии со статьей [6] на 4-ой стадии в блоках ResNet используется слой с deformable сверткой.

2. Эксперименты

Авторами было проведено несколько экспериментов по обучению нейронной сети для детектирования объектов дорожной сцены на изображении. В качестве первичной архитектуры была взята сеть Deformable RFCN [6] с базовой сетью ResNet-101 [11], обученная на наборе данных MS-COCO [7]. Данная сеть была протестирована на наборе данных, состоящем из 20 тысяч изображений разрешения 1920x1080, содержащем 23 тысячи размеченных автомобилей, 4,5 тысячи пешеходов и 5,8 тысяч велосипедистов и мотоциклистов. Набор включал в себя как сцены с хорошими погодными условиями, так и с плохими (дождь, снегопад, низкая освещенность). Показатели детектирования каждого из классов (автомобилей, пешеходов и мотоциклистов с велосипедистами) измерялись отдельно. В качестве целевых метрик использовались точность (precision), полнота (recall), производная от них метрика f-score (F1), а также метрика fppi - количество ложноположительных детекций на кадр. Оценка проводилась для относительно хорошо различимых объектов (размер объекта не менее 40 пикселей для машин, мотоциклов и велосипедов и от 60 пикселей для пешеходов, а заслонение не более 30% для автомобилей и не более 20% для пешеходов, велосипедов и мотоциклов).

Эксперименты проводились с несколькими вариантами базовой сети. Обучение велось на двух наборах данных: MS-COCO 2014 [7] (120 тысяч изображений) и Berkley Deep Drive [8] (80 тысяч изображений). Основные параметры

обучения были взяты в соответствии со статьей [6]. В качестве аугментации использовалось отражение изображения по вертикальной оси. Длительность обучения составила 8 эпох для MS-COCO, т.е. около 1,2 миллиона итераций после аугментации и фильтрации изображений, не содержащих нужные классы (примерно 150 тысяч изображений на эпоху). Для BDD длительность обучения составила 15 эпох (2 миллиона итераций). При обучении моделей с базовыми сетями ResNet 101 v1 и ResNet 50 v2 [17] использовались веса моделей, предобученных на наборе данных ImageNet [12]. При обучении модели с базовой сетью ResNet 50 v2 Light в качестве начальных весов базовой сети бралась часть весов ResNet 50 v2. Однако данную модель можно считать не предобученной, так как маловероятно, что часть весов, взятая из ResNet 50 v2, оптимальна. В связи с этим количество эпох при обучении данной модели было увеличено до 25. Входное изображение приводилось к разрешению с меньшей стороной в 720 px и с большей стороной - не более 1280 px. Ограничением в проведении экспериментов являлась длительность обучения модели при использовании двух видеокарт NVIDIA GeForce 1080 Ti (Табл. 2).

В результате проведенных экспериментов были получены следующие качественные результаты (Табл. 3).

Зависимость качества распознавания от времени прямого прохода модели представлена на Рис. 6.

Табл. 2. Длительность обучения моделей с различными базовыми сетями данных BDD

Базовая сеть	**Кол-во эпох**	**Итераций в секунду**	**Длительность, час**	**Разрешение**	
				target	**max**
ResNet 101	15	5.6	120	720	1280
ResNet 50 v2	15	6.2	108	720	1280
ResNet 50 v2 Light	25	9.5	118	720	1280

Табл. 3. Сравнение моделей

Базовая сеть	**Обучающий датасет**	**Количество кадров**	**Количество объектов**	**Точность (precision)**	**Полнота (recall)**	**f-score**	**fppi**	**Скорость**
Автомобили								
ResNet 101 v1	MS-COCO	11671	23473	0.976	0.955	0.965	0.0480	180 мс
ResNet 101 v1	BDD	11671	23473	0.971	0.983	0.977	0.0589	180 мс
ResNet 50 v2	BDD	11671	23473	0.974	0.964	0.969	0.0516	120 мс
ResNet 50 v2 L	BDD	11671	23473	0.986	0.955	0.970	0.0284	60 мс
Пешеходы								
ResNet 101 v1	MS-COCO	2753	4599	0.954	0.927	0.941	0.0751	180 мс
ResNet 101 v1	BDD	2753	4599	0.951	0.961	0.956	0.0846	180 мс
ResNet 50 v2	BDD	2753	4599	0.961	0.923	0.941	0.0632	120 мс
ResNet 50 v2 L	BDD	2753	4599	0.972	0.884	0.926	0.0418	60 мс
Велосипеды и Мотоциклы								
ResNet 101 v1	MS-COCO	5676	5809	0.960	0.857	0.906	0.0152	180 мс
ResNet 101 v1	BDD	5676	5809	0.931	0.966	0.948	0.0308	180 мс
ResNet 50 v2	BDD	5676	5809	0.943	0.916	0.930	0.0238	120 мс
ResNet 50 v2 L	BDD	5676	5809	0.941	0.918	0.930	0.0247	60 мс

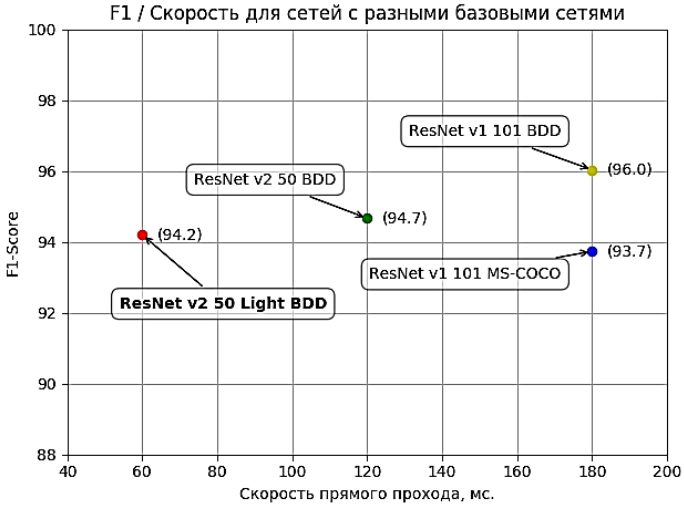

Рис. 6. Зависимость f-score от времени прямого прохода модели усредненная по всем типам объектов

Из результатов экспериментов видно, что модель с представленной в работе базовой сетью ResNet 50 v2 Light, имея скорость в три раза большую, чем у первичной модели, представленной в работе [6], даже немного превосходит её по качеству распознавания. Это, прежде всего, вызвано тем, что модель обучалась на более специализированном наборе данных, но также и связано с удачной архитектурой сети.

Если сравнивать модели, обученные на одном наборе данных, то ResNet 50 v2 Light не сильно уступает по качеству распознавания более тяжеловесным моделям с базовыми сетями ResNet 101 и ResNet 50 v2, обученными на том же наборе данных BDD. Наиболее заметным ухудшением является уменьшение полноты детектирования пешеходов, но в целом сеть ResNet 50 v2 Light натренирована в сторону точности и по показателю f-score ухудшение уже является допустимым. По другим типам объектов облегченная сеть ResNet 50 v2 Light лишь незначительно уступает тяжеловесной сети ResNet 101, обученной на BDD.

При сравнении одной архитектуры сети, обученной на различных наборах данных, получается, что сеть с базовой архитектурой ResNet 101, обученная на данных BDD, имеет лучшие показатели, чем при обучении на данных MS-COCO.

Заключение

В задаче детектирования объектов дорожной сцены современные архитектуры сверточных нейронных сетей достигли весьма хорошего уровня. Примером такой архитектуры может служить Deformable RFCN с базовой сетью ResNet-101. Но при этом скоростные характеристики данной сети невысоки и составляют порядка 5 кадров в секунду. В работе предложен метод ускорения данной сверточной архитектуры путем облегчения базовой сети ResNet. Результаты показывают, что данный подход ускоряет сверточную сеть в три раза со 180 мс до 60 мс, что уже позволяет использовать ее для управления транспортным средством в режиме реального времени. Качественные же характеристики сети при этом изменяются не столь существенно. Кроме того, в работе показано, что сеть с архитектурой Deformable RFCN ResNet-101, обученная на наборе данных BDD, превосходит сеть, обученную на MS-COCO, по всем типам объектов дорожной сцены. Замеры проводились для четырех типов объектов для пяти географических локаций и различных погодных условий.

Литература

1. Prun V.E., Postnikov V.V., Sadekov R.N., Sholomov D.L. “Development of Active Safety Software of Road Freight Transport, Aimed at Improving Inter-City Road Safety, Based on Stereo Vision Technologies and Road Scene Analysis” // Proceedings of the Scientific-Practical Conference “Research and Development – 2016”, Springer, Cham, pp.209-218. – 2017
2. Girshick, R., Donahue, J., Darrell, T., Malik, J.: Rich feature hierarchies for accurate object detection and semantic segmentation. In: CVPR, 2014.

3. R. Girshick, “Fast R-CNN,” in ICCV, 2015
4. S. Ren, K. He, R. Girshick, and J. Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. TPAMI, 2016.
5. J. Dai, Y. Li, K. He, and J. Sun. R-fcn: Object detection via region-based fully convolutional networks. In NIPS, 2016.
6. J. Dai, H. Qi, Y. Xiong, Y. Li, G. Zhang, H. Hu, and Y. Wei. Deformable convolutional networks. ICCV, 2017.
7. T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollar, and C. L. Zitnick. Microsoft COCO: Common objects in context. In ECCV. 2014.
8. F. Yu, W. Xian, Y. Chen, F. Liu, M. Liao, V. Madhavan, and T. Darrell. BDD100K: A Diverse Driving Video Database with Scalable Annotation Tooling. ArXiv e-prints, 2018.
9. M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, “The cityscapes dataset for semantic urban scene understanding,” in Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016.
10. G. Neuhold, T. Ollmann, S. R. Bulo, and P. Kontschieder, “The mapillary vistas dataset for semantic understanding of street scenes,” in Proceedings of the International Conference on Computer Vision (ICCV), Venice, Italy, 2017, pp. 22–29.
11. K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In CVPR, 2016.
12. J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In IEEE Conference on Computer Vision and Pattern Recognition, 2009.
13. K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In ICLR, 2015.
14. Y. Bengio, P. Simard, and P. Frasconi. Learning long-term dependencies with gradient descent is difficult. IEEE Transactions on Neural Networks, 5(2):157–166, 1994
15. X. Glorot and Y. Bengio. Understanding the difficulty of training deep feedforward neural networks. In AISTATS, 2010.
16. Nair, V., Hinton, G.E.: Rectified linear units improve restricted boltzmann machines. In: ICML. (2010)
17. K. He, X. Zhang, S. Ren, and J. Sun. Identity mappings in deep residual networks. In European Conference on Computer Vision, pages 630–645. Springer, 2016.

Лобанов Михаил Генрихович. Московский государственный университет им. М. В. Ломоносова, г. Москва, Россия. Студент магистратуры. Область научных интересов: нейронные сети, глубокое обучение, компьютерное зрение.
E-mail: m.lobanov1994@yandex.ru

Шоломов Дмитрий Львович. Институт проблем передачи информации им. А.А. Харкевича Российской академии наук г. Москва, Россия. Старший научный сотрудник, кандидат технических наук. Национальный исследовательский технологический университет ”МИСиС”, г. Москва, Россия. Старший преподаватель. Количество печатных работ: 25. Область научных интересов: цифровая обработка изображений, компьютерное зрение, искусственный интеллект.
E-mail: sholomov@list.ru

On the Acceleration of the Convolutional Neural Network Architecture Based on Resnet in the Task of Road Scene Objects Recognition

M. G. Lobanov[I], D. L. Sholomov[II,III]

[I] Lomonosov Moscow State University, Moscow, Russia
[II] The Institute for Information Transmission Problems of Russian Academy of Sciences, Moscow, Russia
[III] National University of Science and technology ”MISiS”, Moscow, Russia

Abstract. Recent approaches to road scene objects detection based on convolutional neural networks have reached an acceptable level to be used in autonomous vehicle control and ADAS systems. However, the best modern network architectures are rather heavy and cannot be integrated in real-time systems. Thus the most actual problem is to accelerate networks and to find the optimal balance between their quality and performance. This paper proposes a method to facilitate the architecture of the Deformable Convolutional Network based on ResNet backbone that provides a threefold increase in the inference performance. At the same time, the quality of detection of road scene objects is reduced not so significantly. In addition, the paper compares the quality of the network of this architecture trained on different open datasets – BDD and MS-COCO.

Keywords: object detection, road scene objects, deformable convolutional network, ResNet, ADAS, convolutional network acceleration, BDD, MS-COCO, pedestrian detection, vehicle detection.

DOI 10.14357/20718632190305

Reference

1. Prun V.E., Postnikov V.V., Sadekov R.N., Sholomov D.L. “Development of Active Safety Software of Road Freight Transport, Aimed at Improving Inter-City Road Safety, Based on Stereo Vision Technologies and Road Scene Analysis” // Proceedings of the Scientific-Practical Conference “Research and Development – 2016”, Springer, Cham, pp.209-218. – 2017
2. Girshick, R., Donahue, J., Darrell, T., Malik, J.: Rich feature hierarchies for accurate object detection and semantic segmentation. In: CVPR, 2014.
3. R. Girshick, “Fast R-CNN,” in ICCV, 2015
4. S. Ren, K. He, R. Girshick, and J. Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. TPAMI, 2016.
5. J. Dai, Y. Li, K. He, and J. Sun. R-fcn: Object detection via region-based fully convolutional networks. In NIPS, 2016.
6. J. Dai, H. Qi, Y. Xiong, Y. Li, G. Zhang, H. Hu, and Y. Wei. Deformable convolutional networks. ICCV, 2017.
7. T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollar, and C. L. Zitnick. Microsoft COCO: Common objects in context. In ECCV. 2014.
8. F. Yu, W. Xian, Y. Chen, F. Liu, M. Liao, V. Madhavan, and T. Darrell. BDD100K: A Diverse Driving Video Database with Scalable Annotation Tooling. ArXiv e-prints, 2018.
9. M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, “The cityscapes dataset for semantic urban scene understanding,” in Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016.
10. G. Neuhold, T. Ollmann, S. R. Bulo, and P. Kontschieder, “The mapillary vistas dataset for semantic understanding of street scenes,” in Proceedings of the International Conference on Computer Vision (ICCV), Venice, Italy, 2017, pp. 22–29.
11. K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In CVPR, 2016.
12. J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In IEEE Conference on Computer Vision and Pattern Recognition, 2009.
13. K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In ICLR, 2015.
14. Y. Bengio, P. Simard, and P. Frasconi. Learning long-term dependencies with gradient descent is difficult. IEEE Transactions on Neural Networks, 5(2):157–166, 1994
15. X. Glorot and Y. Bengio. Understanding the difficulty of training deep feedforward neural networks. In AISTATS, 2010.
16. Nair, V., Hinton, G.E.: Rectified linear units improve restricted boltzmann machines. In: ICML. (2010)
17. K. He, X. Zhang, S. Ren, and J. Sun. Identity mappings in deep residual networks. In European Conference on Computer Vision, pages 630–645. Springer, 2016.

Lobanov M. G. Lomonosov Moscow State University, Moscow, Russia. Graduate student. Scientific interests: neural networks, deep learning, computer vision, e-mail: m.lobanov1994@yandex.ru

Sholomov D. L. Ph.D. The Institute for Information Transmission Problems of Russian Academy of Sciences, Moscow, Russia. Senior researcher. National university of science and technology ”MISiS”, Moscow, Russia, Senior lecturer. Number of the printed works: 25. Scientific interests: image processing, computer vision, artificial intelligence, e-mail: sholomov@list.ru