Исследование оптимальной конфигурации сверточной нейронной сети для идентификации объектов в режиме реального времени

М.А. Исаев[1], Д.А. Савельев[1,2]

[1]Самарский национальный исследовательский университет им. академика С.П. Королева, Московское шоссе 34А, Самара, Россия, 443086
[2]Институт систем обработки изображений РАН - филиал ФНИЦ «Кристаллография и фотоника» РАН, Молодогвардейская 151, Самара, Россия, 443001

Аннотация. Данная статья содержит сравнение различных сверточных нейронных сетей, которые являются основой многих актуальных решений в области компьютерного зрения. Исследование включает сравнение данных решений в области компьютерного зрения по таким критериям как mAP, FPS, т.е. возможность их использования в реальном времени. В заключении, делается вывод об оптимальной модели свёрточной нейронной сети и методе глубокого обучения в рамках задачи обработки изображений в реальном времени.

1. Введение

В настоящее время активно развивается сфера применения компьютерного зрения, особенно с момента появления свёрточных нейронных сетей (convolutional neural network – CNN) [1, 2] и беспилотных устройств [3]. Другой неотъемлемой частью компьютерного зрения является детектирование объектов. Обнаружение объектов успешно применяется в трекинге транспортных средств, определении поз, наблюдении [4]. Разница между алгоритмами классификации и обнаружения объектов заключается в том, что в алгоритмах детектирования находятся границы интересующей области (объекта) и определяются на изображении.

Для задач детектирования объектов не стоит использовать обычную нейронную сеть с полносвязным слоем в конце. Это связано с тем, что как правило длина выходного слоя динамическая, что связано с не фиксированным количеством появляющихся объектов.

Одним из подходов для решения этой задачи является получение различных областей с изображения (RoI) и использование CNN для определения наличия объекта в рамках этой области. Данное решение не учитывает возможность различного расположения объекта и различных пропорций сторон. Следовательно, необходимо будет обрабатывать огромное количество таких областей, что затратно с точки зрения вычислительной мощности. Другой вариант решения – это специальные алгоритмы, которые были разработаны для задачи обнаружения объектов в реальном времени [5].

Решения в области распознавания изображений в режиме реального времени делятся на два основных семейства: Region Proposes (поочередно предлагаются и классифицируются регионы кадра) и Single Shot (на полученном изображении сразу детектируются все объекты). К первому семейству относятся такие нейронные сети, как R-CNN, Fast R-CNN, Faster R-CNN [6 – 8]. Ко

второму же семейству относятся YOLO, SSD [4, 9]. Нейронные сети, использующие распознавание по регионам, имеют достаточно медленное время распознавания при качественном определении объектов.
В данной работе проводится исследование определения оптимального решения для задачи обнаружения объектов в реальном времени на основе тестирования решений R-CNN, R-FCN, SSD(VGG-16), YOLOv3(Darknet-53), опираясь на такие метрики, как mAP и FPS.

2. Использование сверточной нейронной сети для идентификации объектов
Для исследования использовались следующие параметры: набор данных – PASCAL VOC 2012 [5], тестировались решения Faster R-CNN, R-FCN, SSD(VGG-16), YOLOv3(Darknet-53).
Faster R-CNN вместо медленного алгоритма селективного поиска использует RPN (Region Proposal Network). RPN является полной заменой селективному алгоритму, и работает следующим образом: на последнем уровне изначальной CNN, скользящее окно размерностью 3x3 обходит карту признаков и уменьшает ее размерность и для каждого положения скользящего окна, RPN генерирует множество возможных областей, основанных на k границах возможного объекта.
R-FCN, или Region-based Fully Convolutional Net является полносвязной сетью и поднимает одну из главных проблем в проектировании нейронных сетей. С одной стороны, выполняя классификацию объекта, необходимо обучить модель свойству инвариантности расположения объекта: несмотря на то, где объект появляется на изображении, объект должен определяться однозначно. С другой стороны, необходимо, чтобы обученная модель выделяла границы объекта в том месте изображения, где он появляется (локальная вариантность). Компромисс между вариантностью и инвариантностью расположения: карты оценок, чувствительные к позиции.
Входное изображение обрабатывается CNN, добавляя полносвязный слой для создания хранилища карт оценок, чувствительных к позиции, генерирует RoI. Далее для каждого региона проверяется хранилище оценок на факт того, является ли этот регион соответствующей позицией некоторого объекта.
SSD, в отличии от Faster R-CNN, которая использует алгоритмы областной классификации и генерации областей предсказаний, одновременно определяет рамку объекта, а также его класс в момент обработки изображения. SSD передает изображение на обработку через серию сверточных слоев, получая несколько наборов карт признаков, для каждого положения в каждой из этих карт признаков используется 3x3 сверточный фильтр для получения набора эталонных координат границ изображения, где для каждого набора координат одновременно рассчитывается смещение и вероятность нахождения в рамках границ этих координат объекта.
YOLOv3, как и SSD относится к семейству Single Shot, а также использует softmax с независимыми логистическими классификаторами для вычисления схожести входных данных с конкретным классом. Вместо использования MSE (mean squared error) для вычисления ошибки классификации, YOLOv3 использует бинарную кросс-энтропию для каждой метки класса. Для определения координат границ объекта, YOLOv3 применяет алгоритм k-means кластеризации.
Обобщенный результат исследований приведен в таблице 1.

Таблица 1. Сравнение результатов идентификации объектов рассматриваемыми нейронными сетями.

Решение	Набор обучающих данных	mAP, %
R-FCN	COCO + VOC 12	59.8
Faster R-CNN	COCO + VOC 12	60.15
SSD	PASCAL VOC 12	64.00
YOLOv3	COCO	63.35

На рисунке 1 показано число размеченных данных по классам, на рисунке 2 приведены результаты оценки средней точности обнаружения для некоторых классов объектов при использовании решения YOLOv3.

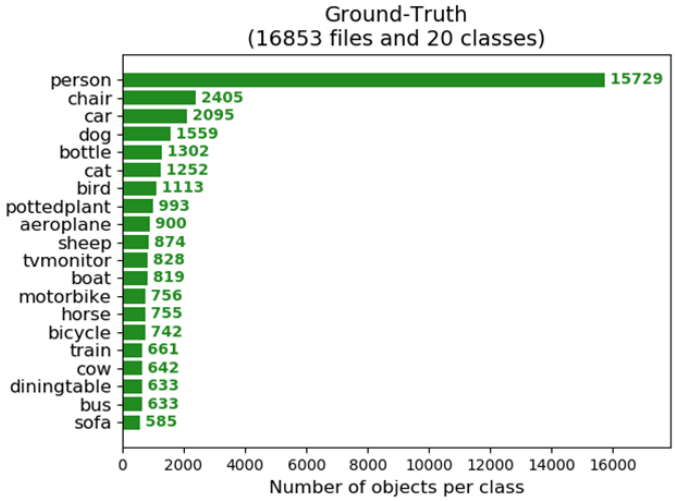

Рисунок 1. Число размеченных данных по классам.

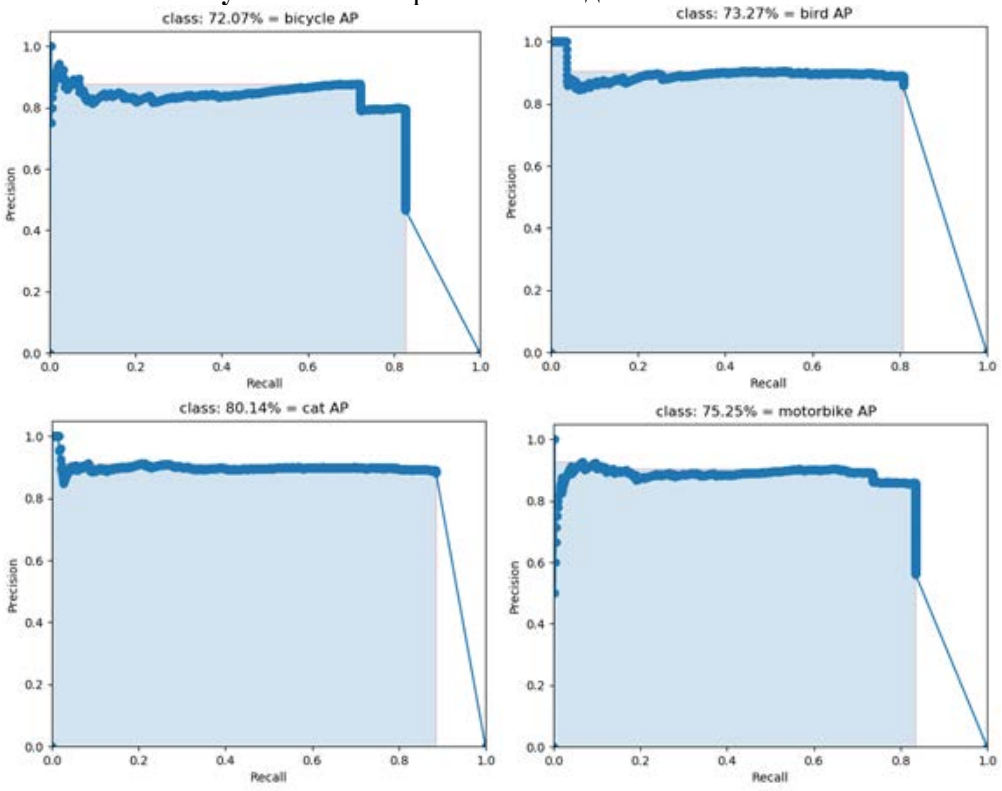

Рисунок 2. Результаты оценки AP для некоторых классов объектов.

3. Заключение

В данной работе было проведено сравнение различных сверточных нейронных сетей. Исследование включает сравнение данных решений в области компьютерного зрения по таким критериям как mAP, FPS, т.е. возможность их использования в реальном времени. На основании проведенного исследования показано, что наиболее подходящим решением является YOLOv3. Несмотря на не самую большую оценку mAP, YOLOv3 имеет высокую скорость обработки видеопотока. Поэтому YOLOv3 имеет большие перспективы, как инструмент трекинга и обнаружения объектов в видеопотоке.

4. Литература

[1] Shi, W. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network / W. Shi, J. Caballero, F. Huszar, J. Totz, A.P. Aitken, R. Bishop, D. Rueckert, Z. Wang // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016. – P. 1874-1883.
[2] Szegedy, C. Rethinking the inception architecture for computer vision / C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, Z. Wojna // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016. – P. 2818-2826.
[3] Pestana, J. Computer vision based general object following for gps-denied multirotor unmanned vehicles / J. Pestana, J.L. Sanchez-Lopez, S. Saripalli, P. Campoy // American Control Conference (ACC), 2014. – P. 1886-1891.
[4] Redmon, J. You only look once: Unified, real-time object detection / J. Redmon, S. Divvala, R. Girshick, A. Farhadi // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016. – P. 779-788.
[5] Girshick, R. Rich feature hierarchies for accurate object detection and semantic segmentation / R. Girshick, J. Donahue, T. Darrell, J. Malik // Proceedings of the IEEE Conference on computer Vision and Pattern Recognition, 2014. – P. 580-587.
[6] Dai, J. R-fcn: Object detection via region-based fully convolutional networks / J. Dai, Y. Li, K. He, J. Sun //Advances in neural information processing systems, 2016. – P. 379-387.
[7] Girshick, R. Fast r-cnn // Proceedings of the IEEE International Conference on Computer Vision, 2015. – P. 1440-1448.
[8] Ren, S. Faster r-cnn: Towards real-time object detection with region proposal networks / S. Ren, K. He, R. Girshick, J. Sun // Advances in Neural Information Processing Systems, 2015. – P. 91-99.
[9] Liu, W. Ssd: Single shot multibox detector / W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.Y. Fu, A.C. Berg // European Conference on Computer Vision. – Springer, Cham, 2016. – P. 21-37.

Investigation of optimal configurations of a convolutional neural network for the identification of objects in real-time

M.A. Isayev[1], D.A. Savelyev[1,2]

[1]Samara National Research University, Moskovskoe Shosse 34A, Samara, Russia, 443086
[2]Image Processing Systems Institute of RAS - Branch of the FSRC "Crystallography and Photonics" RAS, Molodogvardejskaya street 151, Samara, Russia, 443001

Abstract. The article considers the comparison of different convolutional neural networks which are the core of the most actual solutions in the computer vision area. The study includes benchmarks of this state-of-the-art solutions by some criteria, such as mAP (mean average precision), FPS (frames per seconds), i.e, the possibility of real-time usability. It is concluded on the best convolutional neural network model and deep learning methods that were used at particular solution.