УДК 004.932

АНАЛИЗ АРХИТЕКТУР СВЕРТОЧНЫХ НЕЙРОННЫХ СЕТЕЙ ДЛЯ ОБНАРУЖЕНИЯ И КЛАССИФИКАЦИИ ОБЪЕКТОВ НА ВИДЕОПОСЛЕДОВАТЕЛЬНОСТЯХ БОЛЬШОГО РАЗРЕШЕНИЯ

И.Ю. ЗАХАРОВА, Д. ВОРОБЬЕВ
(Представлено: канд. техн. наук, доц. Р.П. БОГУШ)

В приложении к задаче обнаружения объектов на видео большого разрешения рассмотрены архитектуры современных сверточных нейронных сетей. Выявлено, что применение масштабирования входного кадра большого разрешения к относительно малым размерам входного слоя СНС будет приводить к потере объектов небольшого размера, что ухудшает процедуру обнаружения в целом.

Последовательности видеоизображений высокого разрешения характеризуются тем, что содержат значительный объем информации, но при этом на обработку и извлечение необходимых данных требуют огромных вычислительных затрат. Однако, значительное повышение производительности вычислительной техники за последнее десятилетие предопределило развитие и использование ресурсоемких методов обработки информации, возможности которых ранее были ограничены аппаратными средствами, что привело к расширению спектра решаемых практических задач.

Обнаружение объектов на видеопоследовательностях и дальнейшее их распознавание являются актуальными задачами для автоматизированных систем управления и принятия решений, использующих техническое зрение различного назначения, а также видеонаблюдение. Эффективным инструментом для этого являются сверточные нейронные сети (СНС), включающие кроме традиционных полносвязных слоев, сверточные и слои субдискретизации. К первой СНС относят свёрточную нейронную сеть LeNet-5 [1]. С учетом перспективы применения СНС и большого интереса к ним ученых, в последнее время предложено ряд архитектур, которые направлены на обнаружение и классификацию объектов на изображениях и могут быть использованы для обработки видеопоследовательностей большого разрешения.

Модель СНС AlexNet [2] включает восемь взвешенных слоев, из которых первые пять являются сверточными, а последние три представляют собой полносвязные слои. Данная модель была обучена на базе данных ImageNet LSVRC-2010, содержащей 1.2 миллиона изображений, которые разбиты на 1000 классов. При этом, использовались искусственные расширения обучающей выборки, такие как сдвиг, вращение и удаление областей изображений. При тестировании сети на метриках top-1 и top-5 коэффициенты ошибок составили 67.4% и 40.9%, соответственно. Данная модель сети имеет ограничения при работе с многомасштабным анализом, подвержена переобучению из-за пропуска объектов на боковых выбросах, требует значительных вычислительных затрат.

Модель СНС R-CNN предполагает выделение областей интереса на основе предположений о местоположении объектов с использованием метода выборочного поиска. Далее размер регионов масштабируется к размеру входного слоя и обрабатывается СНС AlexNet. На последнем этапе выполняется бинарная классификация с использованием метода опорных векторов, которые были получены для каждого класса объектов. Повышение устойчивости модели к ошибочному делению объекта на фрагменты достигается за счет применения подавление не максимумов. При этом, совпадающие границы областей интереса на соседних фрагментах одного объекта удаляются из-за отсутствия существенного изменения величины градиента. Значение коэффициента ошибок при тестировании определяется величиной 15,3% в метрики ошибок top-5 [3]. Недостатком данной СНС является значительное время, затрачиваемое на обработку потока данных.

Для уменьшения временных затрат в 2015 году предложена модификация данного метода, Fast R-CNN [4], в которой используется усеченный метод опорных векторов. Особенностью Fast R-CNN является также представление регионов интереса в виде сверточной карты признаков. При этом используется 4-х размерный вектор {r, c, h, w}, где r, c - координаты верхнего левого угла, а h и w - высота и ширина региона соответственно. Далее карта признаков подается на слой субдискретизации с размером окна 7×7, выходом которого является максимальное значение для каждого положения окна. Данный метод быстрее осуществляет классификацию в выбранных областях интереса, но не учитывает существенные временные затраты на выделение этих областей перед обработкой в СНС.

Дальнейшее уменьшение временных затрат было предложено в модели Faster R-CNN[5]. Следует отметить, что в отличие от Fast R-CNN данная модель учитывает временные затраты на выделение

объектов. В Faster R-CNN используется нейронная сеть для предположения о нахождении регионов интереса (Region Proposal Network, PRN). Вход сети RPN выделяет сверточные признаки, которые затем попадают на структуру, состоящую из двух полностью соединенных сверточных слоев. Первый слой является регрессионным и предсказывает прямоугольную область, ограничивающую объект. Второй слой является классификационным. Данная структура не одинаково хорошо показывает свои результаты работы с некоторыми группами объектов. Также Faster R-CNN не устойчива к мелким объектам и к зашумленным изображениям.

Модель GoogLeNet [6] состоит более чем из 100 слоев, однако полностью подключенные слои не используются совсем. Также, по сравнению с AlexNet, количество обнаруживаемых параметров уменьшено в 12 раз и сверточные ядра большого размера заменены на последовательность ядер меньшего размера. В модели данной СНС выделяют блок Inception, который включает операции свертки с ядрами размером 5×5, 3×3, 1×1 и субдискретизации для окна размером 3×3. Данный блок последовательно повторяется в архитектуре сети девять раз. Применение указанных выше ядер свертки позволяет извлекать признаки различных размеров в одном блоке. В данной модели по сравнению с R-CNN на первом этапе улучшен подход к выделению регионов интереса за счет включения в метод выборочного поиска алгоритма генерации множества ограничивающих рамок (multibox) [7]. В алгоритме применяется СНС AlexNet, которая генерирует заданное количество ограничивающих рамок для каждой области интереса, далее используется подавление не максимумов для исключения наименее совпадающих ограничивающих рамок с действительной областью интереса. При обучении СНС GoogLeNet выборка искусственно расширялась путем многократного использования входных изображений, после чего результирующие значения многопеременной логической функции (softmax) усреднялись для идентичных изображений, что позволило улучшить результат классификации. Для данной модели СНС значение коэффициента ошибок в метрике top-5 составляет 6,7%. Однако, CoogleNet не учитывает контекстную информацию обо всем изображении, так как на вход СНС подаются лишь области интереса. Также, выделение ограничивающих рамок расширенным методом выборочного поиска значительно замедляет обработку изображения.

В работе [8] предложена модификация блока Inception v2 для CoogleNet путем факторизации сверточного слоя 5×5 на 2 слоя с размером 3×3, что позволило увеличить скорость вычислений в 9,36 раза. Дальнейшее применение факторизации (разложения) для всех сверточных слоев n×n в два слоя с размерами n × 1 и 1× n позволило уменьшить вычислительные затраты на 33%. Для исключения резкого уменьшения размерности карты признаков и снижения вычислительных затрат половина признаков, вычисленных на предыдущих этапах, подается на слой субдискретизации. Другая часть признаков поступает на последующих сверточный слой. Данная модель также предполагает начальное задание минимальных весовых коэффициентов, равномерно распределенных относительно каждого класса.

Модификация Inception v3 [9] характеризуется добавлением нормализации выборки из [11] на последних слоях вместо технологии отсева (Dropout). Модель СНС состоит из 11 чередующихся блоков Inception v3 и при тестировании достигает коэффициента ошибки 4,2% для метрики top-5. Однако, по сравнению со своим ранним аналогом, затрачивает в 2.5 раз больше вычислительной мощности.

Рассмотренная в [10] архитектура СНС ResNet содержит начальный сверточный слой с размерностью ядра [7×7], чередующиеся сверточные слои с ядрами [3×3] и [1×1], а также обеспечивает возможность соединения по технологии быстрого доступа между входами чередующихся слоев и их выходами. В работе отмечено, что использование технологии быстрого доступа позволяет исключить ухудшение качества работы детектора при увеличении количества слоев СНС. Коэффициент ошибки составляет 3,57% в метрике top-5.

В 2016 году компания Google представила четвертую версию блока Inseption v4, а также его модификации Inseption-ResNet [11]. Применение технологии TensorFlow позволило обучать модель GoogleNet на одном устройстве, в отличие от предыдущих версий данной СНС, обучение которых выполнялось по частям с дальнейшим их соединением в полностью обученную модель.

Представленная в 2016 модель СНС Yolo (You Only Look Once) [12] использует сверточные слои с ядрами размером [7×7] и [3×3] для выделения признаков, а третьи слой с ядром размером [1×1] для понижения размерности пространства предположений. Также используются слои субдискретизации размером [2×2]. За каждым сверточным слоем [3×3] размещен полносвязный слой [1×1], передающий координаты ограничивающих рамок и выходные вероятности о нахождении объекта и классовой принадлежности этого объекта. Два конечных полносвязных слоя выполняют задачу классификации. Размеры областей интереса во время обучения выбираются вручную.

Модель делит входное изображение на сетку размером [S × S] частей (ячеек). Для каждой ячейки сетки предсказываются ограничивающие рамки и их весовые коэффициенты, которые характеризуют вероятность наличия объекта в ограничивающей рамке и величину этой вероятности (Pr(Object)). Если центр объекта попадает в ячейку сетки, то для нее устанавливается максимальная вероятность наличия рассматриваемого объекта. Каждая ограничивающая рамка характеризуется 5 параметрами: координатами центра относительно границ ячейки сетки (x, y), шириной (w) и высотой (h) ячейки, и (Pr(Object)). После разбиения каждая ячейка сети поступает на входной слой сети, масштабируясь под его размер. Во время тестирования вероятность наличия объекта в ограничивающей рамке, а также вероятность принадлежности объекта к конкретному классу умножаются, и их произведение дает классификационную оценку для каждого региона. Эти оценки кодируют как вероятность того, что этот класс появится в поле, так и насколько хорошо предсказанный регион подходит для объекта. По метрике Top-5 точность данной модели составляет 88%. Недостатком модели является ручной выбор размера рамок, множество ошибок локализации для объектов малого размера, а также при перекрывающихся рамках проявляется нестабильность модели.

Для улучшения сегментации и классификации были предложены следующие модификации Yolo v2, Yolo9000[13], в которых предусмотрены:

– нормализация данных, позволяющая не использовать технологию «отсева» без опасения возникновения переобучения;

– повышение размерности классификатора для Yolo v2 до [448x448] для 10 эпох ImageNet;

– использование сети, выносящей предположение о нахождении регионов интереса (RPN) по аналогии с моделью Faster R-CNN;

– применение метода k-средних (k=5) для уточнения размера ограничивающих рамок и предварительной сегментации объектов в каждой области интереса;

– более точное предсказание местоположения осуществляется за счет передачи на последующие слои не собственных координат ограничивающих рамок, а их смещения относительно верхнего левого угла исходного изображения;

– вычисление детализированных признаков с разрешением карты признаков [26x26];

– многомасштабное обучение, которое позволяет сети сегментировать и классифицировать объекты при разных разрешениях, исключающее полносвязные слои в архитектуре СНС;

– использование новой классификационной модели Darknet-19, которая включает 19 сверточных слоя и 5 слоев субдискретизации.

Yolo9000 имеет ту же архитектуру, что и Yolo v2, однако количество выходных гипотез ограничивается 3. Модификация архитектуры улучшенной версии Yolo заключается в замене 2 последних полносвязных слоев исходной архитектуры на сеть, выносящую предположения о наличии регионов интереса (Region Proposal Network, RPN) и применении метода k-средних (причем, k=5) для каждой ячейки. Таким образом, каждая ячейка может включать 5 объектов, в то время как первая версия Yolo ограничивалась одним объектом для ячейки. Далее выделенные области интереса поступают на вход классификатора Darknet-19, в котором используется 19 сверточных слоев размером [3x3] и [1x1], а также 5 слоев субдискретизации [2×1]. Обучение Yolo 9000 выполнено на совместной базе данных (ImageNet и COCO), состоящей из 9000 классов. Вероятность правильного обнаружения для данной модели достигает значения 91,2% в метрике top-5, однако при этом вычислительные затраты значительно больше, чем у СНС Yolo. Для некоторых классов объектов, например, «человек», «одежда» вероятность правильной классификации уменьшается.

Представленный анализ современных архитектур СНС показал, что применение данных моделей для обработки видеоизображений большого разрешения требует применения масштабирования входного изображения к размерам входного слоя СНС, что будет приводить к потере объектов небольшого размера в кадре. Поэтому, для устранения данного недостатка необходимо использование специальных алгоритмов обработки [14].

ЛИТЕРАТУРА

1. LeCun, Y. Gradient-based learning applied to document recognition. Proceedings of IEEE / Y. LeCun [et al.]. – 1998. – 86(11). – P. 2278–2324.
2. Krizhevsky, I. Hinton. ImageNet classification with deep convolutional neural networks», Advances in Neural Information Processing Systems 25: Proc. of 26th Annual Conference on Neural Information Processing Systems / I. Krizhevsky, G.E. Sutskever. – 2012., Lake Tahoe, Nevada, United States. – P. 1106–1114.

3. R-CNN for Object Detection [Electronic resource]. – Mode of acess: https://courses.cs.washington.edu/courses/cse590v/14au/cse590v_wk1_rcnn.pdf. – Date of access: 02.12.2017.
4. Ross B. Girshick, «Fast R-CNN», 2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December. – 2015. – P. 1440–1448.
5. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks [Electronic resource]. – Mode of access: http://papers.nips.cc/paper/5638-faster-r-cnn-towards-real-time-object-detection-with-region-proposal-networks.pdf. – Date of access: 20.12.17.
6. Christian Szegedy, Wei Liu, Yangqing Jia, Sermanet, Scott E. Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, Andrew Rabinovich «Going Deeper with Convolutions [Electronic resource]. – Mode of access: https://arxiv.org/pdf/1409.4842.pdf.
7. D. Erhan, C. Szegedy, A. Toshev, and D. Anguelov. «Scalable object detection using deep neural networks», 2014 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2014, Columbus, OH, USA. – 2014. – P. 2155–2162.
8. Rethinking the Inception Architecture for Computer Vision [Electronic resource]. – Mode of access: https://arxiv.org/pdf/1512.00567.pdf. – Date of access: 10.03.17.
9. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift [Electronic resource]. – Mode of access: https://arxiv.org/pdf/1502.03167.pdf. – Date of access: 10.03.17
10. He, K. Deep Residual Learning for Image Recognition / K. He, X. Zhang, S. Ren, J. Sun // Proc. of IEEE Conf. on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA. – 2016. – P. 770–778.
11. Szegedy, C. Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning / C. Szegedy, S. Ioffe, V. Vanhoucke // Proc. of the Thirty-First AAAI Conference on Artificial Intelligence (AAAI-17). – 2017. – 4278–4284.
12. Redmon, J. You Only Look Once: Unified, Real-Time Object Detection / J. Redmon, S.K. Divvala, R.B. Girshick, A. Farhadi // Proc. of IEEE Conf. on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA. – 2016. – P. 779–788.
13. YOLO9000 Better, Faster, Stronger [Electronic resource]. – Mode of access: https://arxiv.org/pdf/1612.08242.pdf. – Date of access: 15.06.17.
14. Vorobjov, D. An effective object detection algorithm for high resolution video by using convolutional neural network / D. Vorobjov, I. Zakharava, R. Bohush, S. Ablameyko // Proc. of the 15th Int. Symposium on Neural Networks (ISNN 2018), June 25–28 2018, Minsk. – 2018. – P. 503–508.