СТАТЬИ

УДК 004.855

МОДИФИКАЦИЯ АРХИТЕКТУРЫ TINY-YOLO ДЛЯ ЗАДАЧИ ОБНАРУЖЕНИЯ ОБЪЕКТОВ В РЕАЛЬНОМ ВРЕМЕНИ

Денисенко А.А.

ЧП Денисенко, Ирпень, e-mail: alexey.denisenko.work@gmail.com

Обнаружение объектов остается одним из наиболее активных направлений исследования в области машинного обучения и компьютерного зрения. Значительные успехи в этой области были получены благодаря внедрению глубоких сверточных нейронных сетей. Однако, несмотря на достигнутые результаты, одной из самых больших проблем при широкомасштабном развертывании систем, в основе которых лежат такие сети, на мобильных и периферийных устройствах являются все более растущие требования к вычислительным ресурсам и памяти. Таким образом, в последние годы наблюдается повсеместный интерес специалистов к исследованию и разработке эффективной архитектуры нейронной сети, предназначенной для повседневного использования. В рамках данной работы будет представлена модификация архитектуры Tiny-YOLO, легковесной версии YOLO – многослойной сверточной нейронной сети для решения задачи обнаружения объектов. Предлагаемая архитектура позволяет спроектировать систему, в основе которой лежит нейронная сеть размером около 4 Мб (примерно на 15% меньше оригинальной архитектуры Tiny-YOLO) и требует 4,23 млрд операций для получения результата (примерно на 30% меньше оригинальной архитектуры), при этом для используемого на этапе обучения и тестирования датасета VOC 2007 точность обнаружения на неизученных данных все еще достигает 69%, как и в оригинальной архитектуре YOLO (примерно на 10% выше, чем в Tiny-YOLO). Результаты проведенных экспериментов по скорости и точности обнаружения демонстрируют эффективность разработанной модификации.

Ключевые слова: обнаружение объектов, нейронные сети, сверточные нейронные сети, глубокое обучение, компьютерное зрение

MODIFICATION OF TINY-YOLO ARCHITECTURE FOR OBJECT DETECTION IN REAL TIME

Denysenko A.A.

PE Denysenko, Irpin, e-mail: alexey.denisenko.work@gmail.com

Object detection remains one of the most active research areas in machine learning and computer vision. Significant advances in this area have been achieved through the implementation of deep convolutional neural networks. However, despite the results achieved, one of the biggest challenges in large-scale deployments of systems that rely on such networks on mobile and peripheral devices is the ever-increasing requirements for computing resources and memory. Thus, in recent years, there has been a widespread interest of specialists in the research and development of an effective neural network architecture intended for everyday use. As part of this work, a modification of the Tiny-YOLO architecture, a lightweight version of YOLO, a multilayer convolutional neural network for solving the problem of object detection, will be presented. The proposed architecture allows you to design a system based on a neural network of about 4 MB in size (about 15% less than the original Tiny-YOLO architecture) and requires 4.23 billion operations to get the result (about 30% less than the original architecture), with at the same time, for the VOC 2007 dataset used at the stage of training and testing, the detection accuracy on unstudied data still reaches 70%, as in the original YOLO architecture (about 10% higher than in Tiny-YOLO). The results of the experiments on the speed and accuracy of detection demonstrate the effectiveness of the developed modification.

Keywords: object detection, neural networks, convolutional neural networks, deep learning, computer vision

Во всех современных системах обнаружения на основе компьютерного зрения необходимо не только локализовать местоположение объектов в пределах сцены, но и назначить ту или иную метку класса каждому из них. Самые крупные из недавних успехов в этой области связаны с последними достижениями в области глубокого обучения, в частности с использованием многослойных сверточных нейронных сетей.

До тех пор, пока основной фокус был сосредоточен на улучшении точности, появлялись все более сложные архитектуры, такие как SSD, R-CNN, Mask R-CNN, а также модификации этих сетей [1]. Несмотря на то что системы, в основе которых лежали модели этих сетей, продемонстрировали высокую производительность и качество обнаружения, их было практически невозможно развернуть на портативных и мобильных устройствах из-за ограничений памяти и недостаточных вычислительных мощностей. По сути, даже более быстрые модификации, такие как Faster R-CNN, демонстрировали низкое быстродействие при обнаружении объектов на встроенных процессорах мобильных устройств [2]. Кроме того, такие ограничения помешали широкому распространению сетей для широкого круга приложений, таких как бес-

пилотные летательные аппараты, системы видеонаблюдения, системы автономного вождения, где также требуется локальная обработка. Таким образом, исследование и разработка высокоэффективных архитектур глубоких нейронных сетей для обнаружения объектов, которые больше подходят для периферийных и мобильных устройств, являются актуальной задачей.

Целью исследования является применение принципов проектирования системы обнаружения объектов на основе семейства архитектур YOLO для создания легковесной сети с настраиваемой макро- и микроархитектурой на уровне модулей, адаптированной для задачи обнаружения.

Материалы и методы исследования

Методы глубокого обучения являются более совершенными моделями машинного обучения, обеспечивающими лучшую производительность при решении ряда задач. В то время как традиционные методы машинного обучения занимаются преимущественно ручным извлечением необходимых признаков из доступных входных данных, методы глубокого обучения иерархически извлекают эти признаки и обучаются без учителя или с частичным его участием. На рис. 1 проиллюстрирована структурная схема традиционной модели; схема модели глубокого обучения [3] представлена на рис. 2.

Глубокое обучение все еще далеко от того, чтобы быть зрелой и хорошо изученной областью, но оно уже используется многими приложениями реального мира, такими как обнаружение и распознавание на основе видения, распознавание и синтезирование речи, энергосбережение, поиск лекарств, финансы и маркетинг. С одной стороны, глубокое обучение предлагает множество возможностей для исследований и эксплуатации, с другой – в глубоком обучении много нерешенных проблем. Это также дает потенциальную возможность первым вывести на рынок разработку, публикацию или некоторый новый продукт в этой области.

YOLO (You Only Look Once) – алгоритм обнаружения объектов в реальном времени. Лежащий в основе YOLO алгоритм принимает в качестве входа исходное изображение или кадр видеопотока один раз, а на выходе возвращает фрагменты этого изображения – каждый участок кадра имеет ограничительную рамку с вероятностью нахождения внутри нее того или иного объекта [4]. Первая версия системы YOLO достигла точности более 50 % для решения задачи обнаружения объектов в реальном времени наборов данных VOC 2007 и VOC 2012, что сделало ее лучшим выбором для этой задачи – как показала практика, YOLO работает гораздо точнее и быстрее, чем любой из алгоритмов, основанных на принципе скользящего окна. Методика модели YOLO основывается только на незначительных изменениях в уже известных алгоритмах.

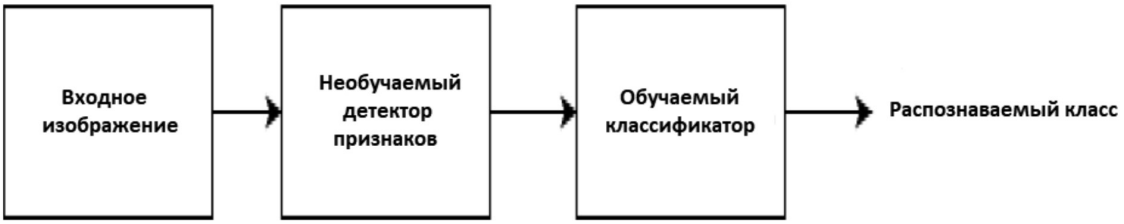

Рис. 1. Структурная схема традиционных методов машинного обучения

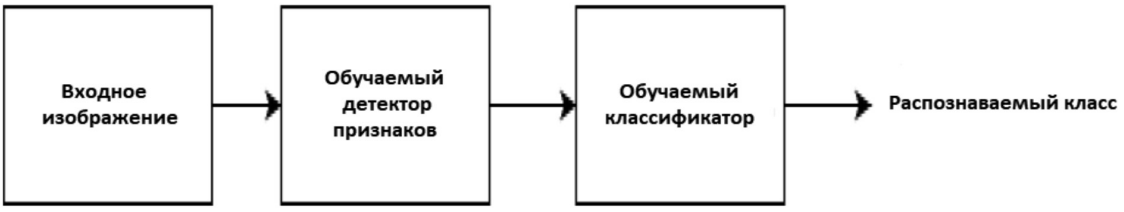

Рис. 2. Структурная схема модели глубокого обучения

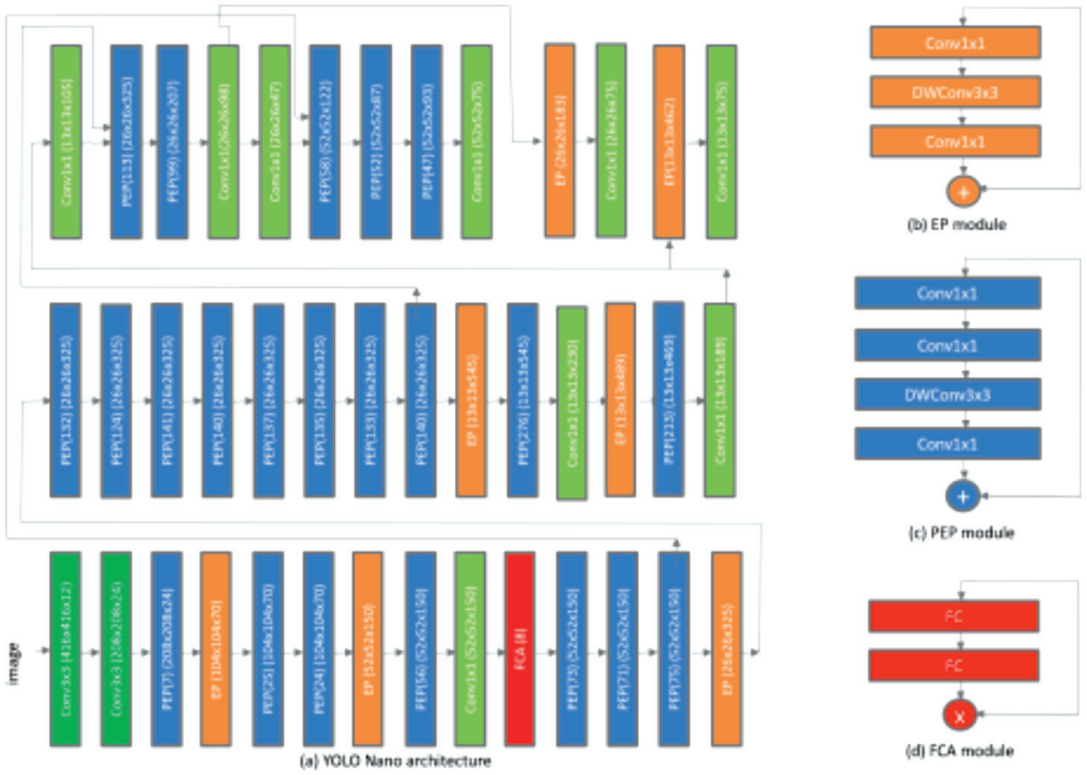

Рис. 3. Архитектура модифицированной сети Tiny-YOLO. Здесь PEP (x) указывает на x каналов в первом слое проекции остаточного модуля PEP, а FCA (x) указывает на коэффициент уменьшения x

Отличительной чертой сетей семейства YOLO является то, что, в отличие от так называемых Single-Shot сетей, которые полагаются на построение сетей, предсказывающих местоположение объектов на сцене с их последующей классификацией [4], они используют общую архитектуру для обработки входного кадра и генерации необходимых результатов. Таким образом, все прогнозы выполняются за один прямой проход по сравнению с сотнями и тысячами проходов, которые необходимо выполнить для получения окончательных результатов сетей-предшественников. Это делает архитектуры на основе YOLO значительно более быстрыми и, следовательно, более подходящими для обнаружения объектов различных размеров.

Первоначальный прототип проекта, использованный в этом исследовании, черпает вдохновение из семейства сетевых архитектур YOLO и состоит из стека модулей представления признаков с быстрыми соединениями между модулями. Модули представления признаков конфигурируются способом, подобным сетям пирамид признаков, так что они способны представлять объекты в трех различных масштабах. За этими модулями представления признаков следуют несколько сверточных слоев, выход которых представляет собой трехмерный тензор, кодирующий результат предсказания в виде ограничивающей рамки с принадлежностью к тому или иному классу для трех различных масштабов. В результате такой первоначальный прототип архитектуры позволил обнаруживать многомерные объекты.

Архитектура предлагаемой сети на основе Tiny-YOLO [2] для обнаружения объектов представлена на рис. 3.

Первая особенность архитектуры сети Tiny-YOLO, значительно отличающая ее от семейства других сетей YOLO, заключается в том, что она состоит из модулей с уникальной макроархитектурой на паттерне «проекция – расширение – проекция» (PEP) в дополнение к макроархитектурам «расширение – проекция» (EP). Остаточная макроархитектура PEP состоит из: сверточного слоя со свертками размера 1×1, который преобразует выходные каналы в тензор с более низкой размерностью, слоя расширения со свертками 1×1, который увеличивает количество каналов до более высокой размерности, смещен-

ного вглубь сверточного слоя, который выполняет пространственные свертки с различными фильтрами на каждом из отдельных выходных каналов из слоя расширения, и, наконец, проекционного слоя со свертками 1×1, который конвертирует выходные каналы в выходной тензор с более низкой размерностью. Использование остаточной макроархитектуры PEP позволяет значительно снизить вычислительную сложность при сохранении остальных параметров модели.

Второе значимое наблюдение об архитектуре сети – это введение легковесного модуляционного полносвязного слоя, который состоит из двух полносвязных слоев, изучающих динамические нелинейные взаимозависимости между слоями и создающих веса модуляции для повторного пересчета весов. Использование таких слоев облегчает повторную калибровку динамических признаков на основе глобальной информации, чтобы уделить больше внимания информативным признакам и лучше использовать доступную пропускную способность сети [5]. Это, в свою очередь, позволяет добиться прочного баланса между сниженной архитектурной, вычислительной сложностью и выразительностью модели.

Третье наблюдение касается высокой неоднородности не только макроархитектур (разнообразное сочетание модулей PEP, модулей EP, полносвязных модуляционных слоев, а также отдельных сверточных слоев со свертками 3×3 и 1×1), но также с точки зрения микроархитектур отдельных модулей и слоев представления признаков, причем каждый модуль или слой сети имеет уникальную микроархитектуру. Преимущество высокой неоднородности микроархитектуры в этой модификации Tiny-YOLO заключается в том, что она позволяет индивидуально адаптировать каждый компонент архитектуры сети для достижения сильного баланса между архитектурной и вычислительной сложностью и выразительностью модели.

Результаты исследования и их обсуждение

Чтобы изучить эффективность модифицированной сети Tiny-YOLO для обнаружения объектов в режиме реального времени, необходимо исследовать размер модели, точность обнаружения объектов и вычислительные затраты на наборах данных PASCAL VOC. Для сравнения две существующие и описанные ранее в литературе модификации Tiny-YOLO использовались в качестве базовых. Учитывался тот факт, что они являются одними из самых популярных легковесных глубоких нейронных сетей для обнаружения объектов, небольшие размеры их моделей и низкую вычислительную сложность. Наборы данных VOC2007 / 2012 состоят из полученных в естественной среде изображений с 20 различными типами (классами) объектов. Глубокие нейронные сети были обучены с использованием обучающих наборов данных VOC2007 / 2012, а средняя точность на тестовой выборке (т.е. на не изученных ранее данных) была вычислена на тестовом наборе данных VOC2007 для оценки точности обнаружения объектов глубокими нейронными сетями, что является стандартной практикой в исследовательской литературе.

В таблице показаны размеры моделей и точность обнаружения объектов предлагаемой модификации сети Tiny-YOLO, а также ее ближайших конкурентов Tiny-YOLO 2 и Tiny-YOLO 3.

Сравнение точности обнаружения объектов для основанных на Tiny-YOLO архитектур сетей на тестовом наборе VOC 2007. Размер входных кадров – 416×416 для всех протестированных сетей

Название модели	Размер, Мб	Точность на наборе VOC 2007, %	Число операций, млрд
Tiny-YOLO 2	60,5	57,1	6,87
Tiny-YOLO 3	33,4	58,4	5,52
Tiny YOLO mod.	4,0	69,1	4,57

Заключение

В ходе выполнения работы, кроме полноценного анализа предметной области, связанной с данными о компьютерном зрении, машинном и глубоком обучении, были исследованы предшествующие достижения в области задачи обнаружения объектов, в том числе по решению данной задачи другими методами, а также выполнен сравнительный анализ различных модификаций архитектуры Tiny-YOLO, предназначенной для решения задачи, в том числе представленной в работе архитектуры Tiny-YOLO mod.

Как следует из таблицы, описанная в данной работе модель Tiny-YOLO mod. демонстрирует лучшие результаты по всем параметрам.

Во-первых, можно заметить, что размер модели составил 4,0 МБ, что на 15% и 8% меньше, чем у Tiny-YOLO 2 и Tiny-YOLO 3 соответственно, что очень важно

для периферийных и мобильных устройств с учетом ограничений памяти.

Во-вторых, модифицированная модель Tiny-YOLO, несмотря на то, что она намного меньше по размеру модели в Мб, достигла результата в 69,1% точности обнаружения объекта на тестовом наборе данных VOC 2007, что приблизительно на 12% и на 10,7% больше, чем у Tiny-YOLO 2 и Tiny-YOLO 3 соответственно.

В-третьих, модификация Tiny-YOLO требует всего 4,57 млрд операций для выполнения операции обнаружения и вывода результата на кадр, что на 34% меньше, чем у Tiny-YOLO 2, и приблизительно на 17% меньше, чем у Tiny-YOLO 3.

Список литературы

1. Lin T.-Y., Dollar P., Girshick R., He K., Hariharan B., Belongie S. Feature pyramid networks for object detection. IEEE Conference on Computer Vision and Pattern Recognition (Honolulu, HI, USA, 21-26 July 2017). IEEE. 2017. P. 936–944.

2. Redmon J., Farhadi A. Yolov3: An incremental improvement. ArXiv. 2018. [Electronic resource]. URL: https://arxiv.org/pdf/1804.02767.pdf (date of access: 10.04.2021).

3. Patterson J. Deep learning: a practitioner's approach. Sebastopol, CA: O'Reilly, 2017. 532 p.

4. Shafiee M.J., Chywl B., Li F., Wong A. Fast YOLO: A fast you only look once system for real-time embedded object detection in video. ArXiv. 2017. [Electronic resource]. URL: https://arxiv.org/pdf/1709.05943.pdf (date of access: 10.04.2021).

5. Girshick R., Donahue J., Darrell T., Malik J. Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation. 2014 IEEE Conference on Computer Vision and Pattern Recognition (Columbus, OH, USA, 23-28 June 2014). IEEE, 2014. P. 580–587.