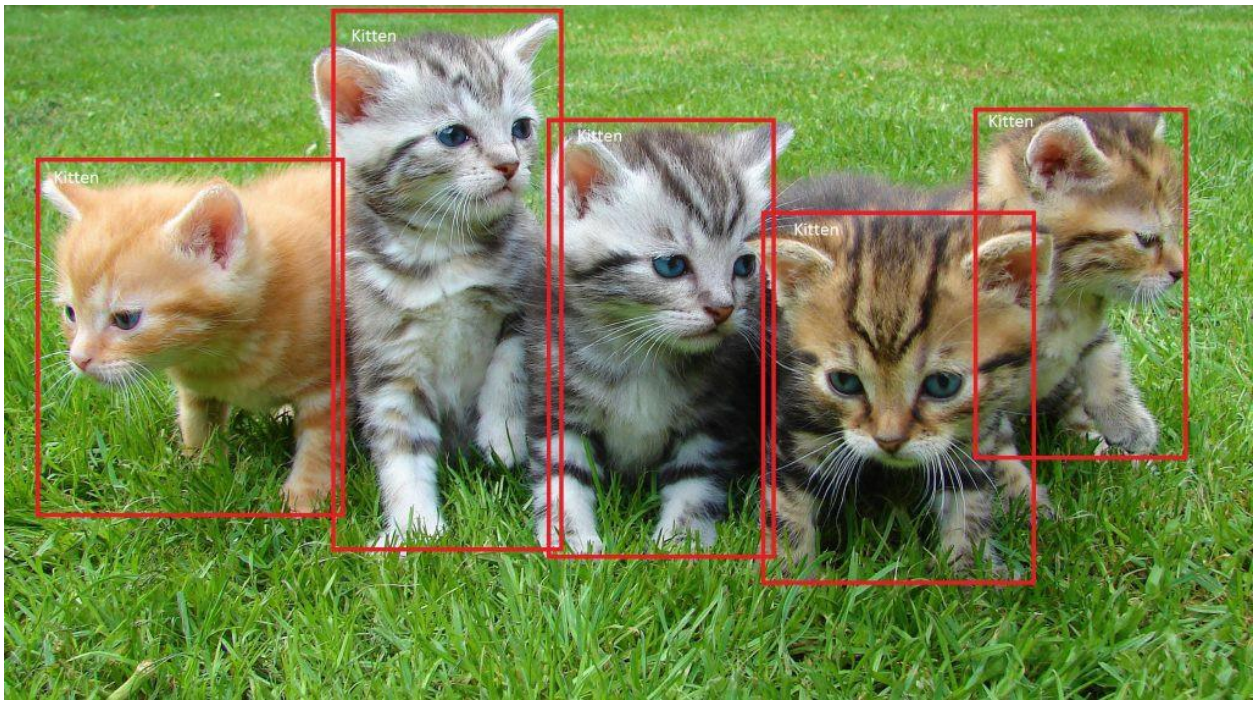
ОБЗОР SSD: детектирование объектов в реальном времени в глубоком обучении



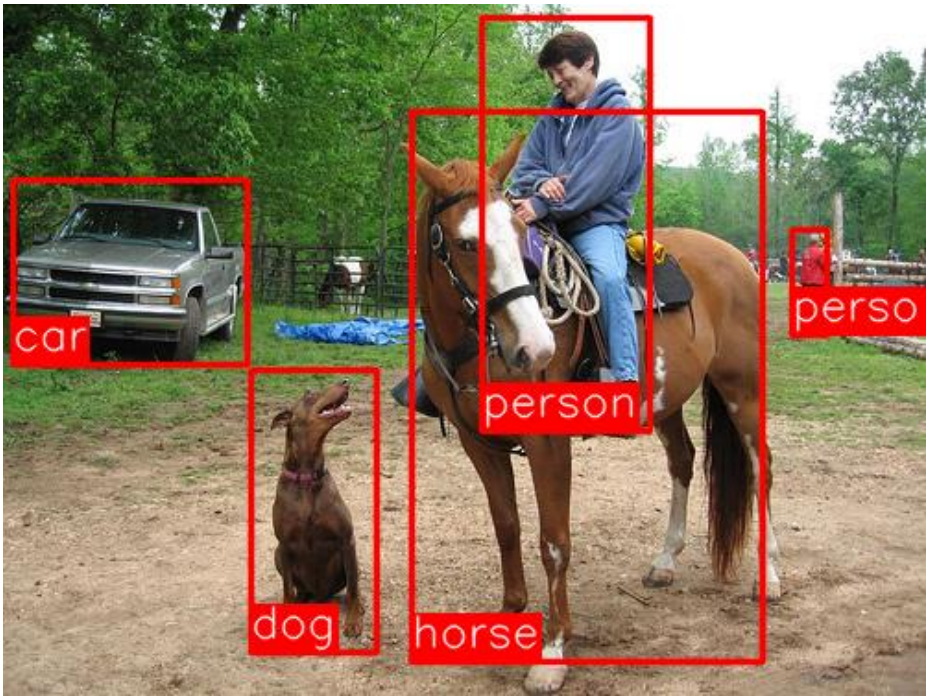
С тех пор, как AlexNet в 2012г. стремительно взял на абордаж исследовательский мир в сфере машинного обучения (МО) и нейронных сетей (НС), глубокое обучение (ГО) стало методом перехода к задачам распознавания изображений, намного превосходящим более традиционные методы компьютерного зрения, состоящие из категоризации изображений по заданному набору классов (например, кошка, собака) и сети, определяющей самый сильный класс, присутствующий в изображении.

В настоящее время есть алгоритмы ГО, которые лучше классифицируют изображения, чем люди. Однако с нами, людьми, не все потеряно. Мы все еще можем гораздо больше, чем просто классифицировать картинки при наблюдении и взаимодействии с миром. Например, мы локализуем и классифицируем каждый элемент в пределах нашего поля

зрения. Это гораздо более сложные задачи, которые машины все еще пытаются выполнить так же, как и люди.

R-CNN

Несколько лет назад, используя некоторые из достижений, которые стали возможными в компьютерном зрении с помощью CNN, исследователи разработали R-CNN для решения задач обнаружения, локализации и классификации объектов. Вообще говоря, R-CNN - это особый тип CNN, который способен обнаруживать и детектировать (определять, узнавать) объекты на изображениях: выходные данные, как правило, представляют собой набор ограничивающих рамок, которые близко соответствуют каждому из обнаруженных объектов, а также выходной класс для каждого обнаруженного объекта. Изображение ниже показывает, что выводит типичная R-CNN:



Исследователи и разработчики не остановились на достигнутом и улучшали R-CNN, создав Fast-R-CNN и Faster-R-CNN. Как можно догадаться, на каждом следующем этапе совершенствовалась основная работа, проделываемая R-CNN, так, чтобы сеть быстрее обнаруживала объекты в режиме реального времени. Тем не менее, несмотря на потрясающие достижения, ни одна из этих архитектур не может быть детектором объектов действительно в реальном времени. Были выявлены следующие основные проблемы с вышеуказанными сетями:

- алгоритм обучения громоздкий и весьма дорогой

- обучение проходит в несколько этапов (например, определение региона/области обучения относительно классификатора)

- сеть слишком медленно выдает результаты при работе с новыми входными данными (т.е. на которых она не обучалась и которые ‘в глаза не видела’).

К счастью, человеческая мысль не стоит на месте, и в последние несколько лет были созданы новые архитектуры для устранения узких мест R-CNN и ее преемников/потомков, которые позволили вплотную приблизиться к обнаружению объектов в режиме реального времени. Наиболее известными из них являются YOLO и SSD.

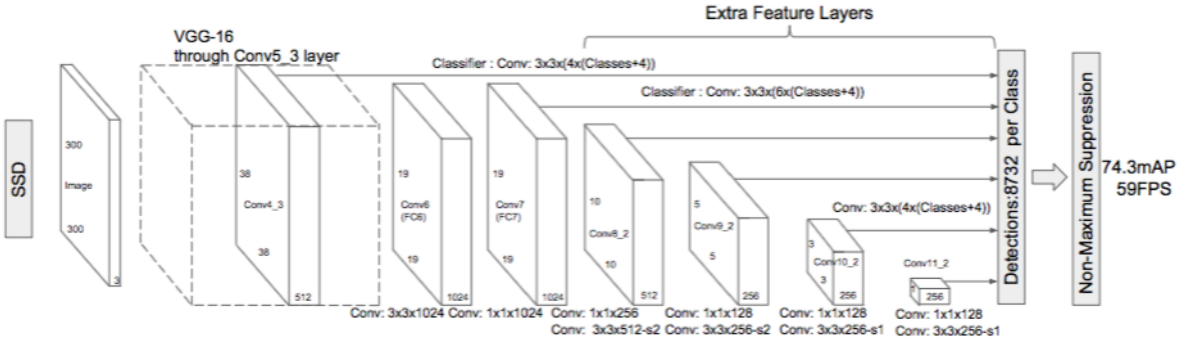
SSD

Первая статья о SSD (автор: C. Szegedy и др.) была опубликована в конце ноября 2016 года. В ней говорилось, что SSD достигла новых рекордов с точки зрения производительности и точности для задач обнаружения объектов, набрав более 74% mAP (mean Average Precision) на скорости 59 кадров в секунду на стандартных наборах данных, таких как PascalVOC и COCO. Попробуем понять что это за диво дивное, SSD (Single Shot Detector):

- Single Shot означает, что задачи локализации и классификации объектов выполняются за один проход сети с использованием методики регрессии ограничивающего прямоугольника

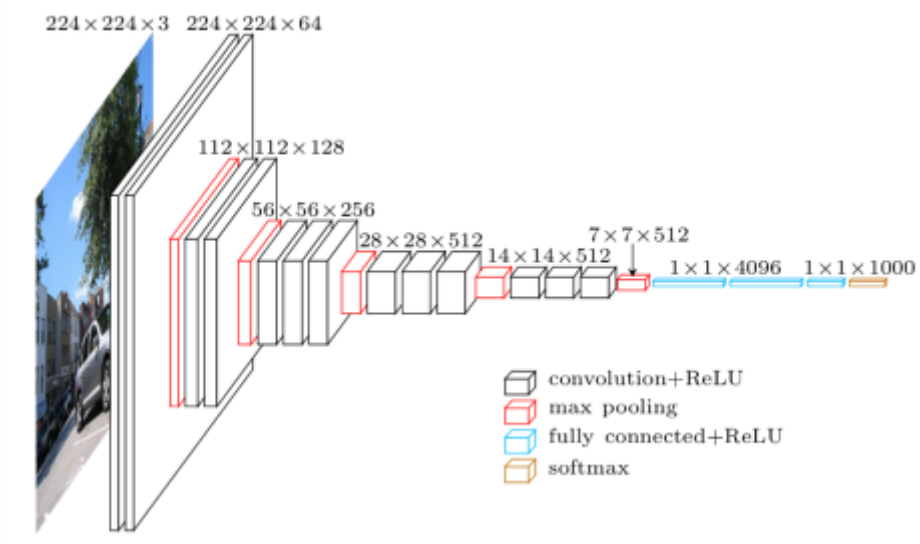
- Detector является поисковиком/определителем/обнаружителем объектов, который также классифицирует эти обнаруженные объекты.

Архитектура:



Как видно, архитектура SSD основана на архитектуре VGG-16, оставляя некоторые уровни несвязанными. Причина, по которой VGG-16 был использован, заключается в его высокой производительности в задачах классификации изображений высокого качества и его

популярности для задач, где обучение с помощью переноса помогает улучшить результаты. Вместо оригинальных полностью связанных слоев был добавлен набор вспомогательных сверточных слоев (начиная с 6), что позволило извлекать элементы в нескольких масштабах и постепенно уменьшать размер входных данных для каждого последующего слоя.



Не вдаваясь далее в технические детали архитектуры SSD, подытожим:

- чем больше контуров/прямоугольников (boxes) по умолчанию, тем точнее обнаружение, хотя это влияет на скорость

- наличие регрессионого модуля на нескольких слоях также приводит к лучшему обнаружению, благодаря тому, что детектор работает с функциями с несколькими разрешениями

- 80% времени уходит на базовую сеть VGG-16: это означает, что при более быстрой и одинаково точной сети производительность SSD может быть еще лучше

- SSD-500 (вариант с наивысшим разрешением, использующий входные изображения 512x512) достигает наилучшего значения mAP на PascalVOC2007 при 76,8%, но принося в жертву скорость, когда частота кадров падает до 22 кадров в секунду; SSD-300, таким образом, является гораздо лучшим компромиссом с 74,3 mAP при 59 кадрах в секунду

- SSD дает худшую производительность для небольших объектов, так как они могут отображаться не на всех картах объектов; увеличение разрешения входного изображения облегчает эту проблему, но не решает ее полностью