# Доклад YOLO

Павел Кошкин

Декабрь 2018

# 1 Вступление

You only look once (YOLO) - современный детектор объектов на изображениях и видео, показывающий хорошее качество в сочетании с отличной скоростью работы. Благодаря новому подходу - использование для детекции всего одного прохода сверточной нейросети, метод способен детектировать объекты на видео в реальном времени. Ниже приведены результаты замеров качества и скорости рыботы YOLO в сравнении с другими известными методами, взятые с официального сайта фреймворка.

| Model | Train | Test | mAP | FLOPS | FPS |
|---|---|---|---|---|---|
| SSD300 | COCO trainval | test-dev | 41.2 | - | 46 |
| SSD500 | COCO trainval | test-dev | 46.5 | - | 19 |
| YOLOv2 608x608 | COCO trainval | test-dev | 48.1 | 62.94 Bn | 40 |
| Tiny YOLO | COCO trainval | test-dev | 23.7 | 5.41 Bn | 244 |
| SSD321 | COCO trainval | test-dev | 45.4 | - | 16 |
| DSSD321 | COCO trainval | test-dev | 46.1 | - | 12 |
| R-FCN | COCO trainval | test-dev | 51.9 | - | 12 |
| SSD513 | COCO trainval | test-dev | 50.4 | - | 8 |
| DSSD513 | COCO trainval | test-dev | 53.3 | - | 6 |
| FPN FRCN | COCO trainval | test-dev | 59.1 | - | 6 |
| Retinanet-50-500 | COCO trainval | test-dev | 50.9 | - | 14 |
| Retinanet-101-500 | COCO trainval | test-dev | 53.1 | - | 11 |
| Retinanet-101-800 | COCO trainval | test-dev | 57.5 | - | 5 |
| YOLOv3-320 | COCO trainval | test-dev | 51.5 | 38.97 Bn | 45 |
| YOLOv3-416 | COCO trainval | test-dev | 55.3 | 65.86 Bn | 35 |
| YOLOv3-608 | COCO trainval | test-dev | 57.9 | 140.69 Bn | 20 |
| YOLOv3-tiny | COCO trainval | test-dev | 33.1 | 5.56 Bn | 220 |
| YOLOv3-spp | COCO trainval | test-dev | 60.6 | 141.45 Bn | 20 |

# 2 Используемые ранее подходы и преимущества YOLO

YOLO - далеко не первый предложенный метод для детекции объектов на изображениях. До начала широкого использования нейросетей, были известны в основном методы, заточенные под детекцию объектов определенного класса. Например вышедший в 2001 году детектор Виолы-Джонса использует в ручную подобранный набор признаков для для детекции лиц. Эти признаки состоят из фильтров разных размеров, расположений и форм. Они вычисляются по входному изображению и используются для классификатора SVM, который определяет является ли входное изображение изображением лица.

Более поздняя работа 2005 года HOG (гистограмма направленных градиентов) использовалась для детекции более широкого числа объектов: лиц, пешеходов, дорожных знаков и других. По сути это улучшение детектора Виолы-Джонса по средством использования более совершенных признаков для классификации. В качестве таких признаков берется гистограмма направленных градиентов. Для их вычисления для каждого пикселя изображения считается направление самого сильного увеличения яркости изображения, при этом число направлений фиксировано. Далее изображение разбивается на куски не большого размера и для каждого из этих кусков вычисляются гистограммы направлений с предыдущего шага. Наборы таких гистограмм и являются признаками для классификации.

Существует большое количество улучшений методов, описанных выше, например с использованием пирамиды изображений разных разрешений.

Нейросетевые методы появились в 2012 году. Первой идеей было использовать предобученную на задаче классификации сверточную нейросеть для вычисления признаков различных регионов на изображении и выбора в качестве области детекции объекта смеси регионов, для которых классификатор имеет наибольшую уверенность в принадлежности к детектируемому классу. Для выбора регионов использовался метод скользящего окна, который перебирает куски картинки разных размеров с каким-то фиксированным шагом. Очевидным минусом подхода является очень большое время работы, поскольку число окон на изображении может достигать очень больших значений.

До выхода YOLO практически все нейросетевые детекторы основывались на работе R-CNN, идея которой заключается в улучшении метода скользящего окна. Авторы статьи предлагают запускать классификатор только на множестве подизображений фиксированного размера. Для генерации этих подизображений используется специальная процедура, названная "селективный выбор". Эта процедура генерирует подизображения со случайными расположением, формой и размером и пытается объединить их, основываясь на сходствах их цветов, текстур и интенсивности.

Существует большое количество улучшений R-CNN, многие которые появились совсем недавно (например Mask R-CNN). Однако YOLO предлагает принципиально другой подход. Основная идея (как видно из названия) - единственный проход сверточной нейросети по всему изображению. Такой подход позволяет не только существенно уменьшить время работы алгоритма, но и повысить качество, поскольку единая нейросеть может скомбинировать информацию из разных областей изображения, что не возможно для R-CNN. Например, нейросеть, получающая на вход все изображение может учесть факт, что по дорогам чаще всего движутся автомобили, а если на человеке есть футболка, то вероятно на этом же изображении можно найти шорты.

# 3 Метод работы

## 3.1 Общее описание

Авторы YOLO предлагают разбить исходное изображение на SxS (7x7 или 13x13 в зависимости от версии YOLO) ячеек.

Для каждой ячейки модель предсказывает $B$ (2 или 5) регионов, которые кодируются векторами из 4 чисел: $(x, y, h, w)$. Здесь $x$ и $y$ - сдвинутые относительно расположения ячейки координаты левого верхнего угла региона, а $h$ и $w$ - нормированные на размеры изображения высота и ширина. Для каждого из таких векторов предсказывается также вероятность число, передающее уверенность модели в том, что рассматриваемый регион содержит некоторый объект. Оно предствляется как вероятность обнаружить в регионе объект $P(object)$, умноженная на $IoU$ - размер пересечения выделенного региона с реальным деленный на размер их объединения. Важный момент заключается в том, что эта матрика не содержит никакой информации о том, к какому классу этот объект принадлежит. Для определения класса региона предсказывается другой вектор, размера количества классов $C$ (20 или 25). Этот вектор содержит условные вероятности $P(class_i|object)$ принадлежности каждому из классов при условии, что в регионе есть некоторый объект.

Итоговый score региона вычисляется по формуле

$$score_i = P(class_i) \cdot IoU = P(class_i|object) \cdot P(object) \cdot IoU$$

Таким образом, предсказания модели YOLO представляют из себя вектора трехмерные тензоры размерности $(S, S, B \cdot 5 + C)$. Поскольку каждая из ячеек может предсказывать регионы только одного класса, модель YOLO имеет проблемы с предсказаниями объектов, которые находятся очень близко дург к другу.

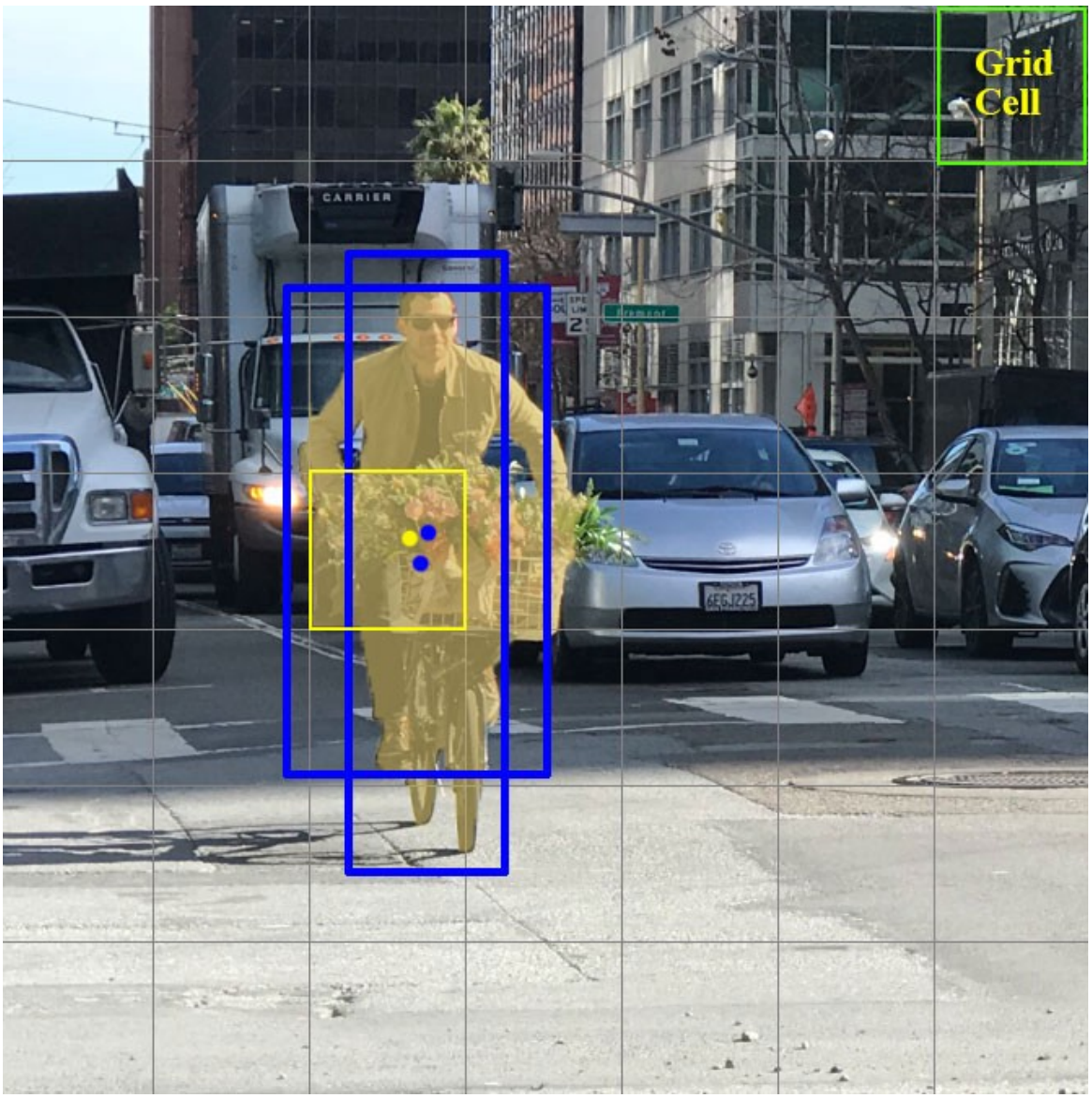


Рис. 1: Каждая ячейка предсказывает фиксированное число регионов.

## 3.2 Архитектура сети

Архитектура сети, которая предсказывает тензоры размерности $(S, S, B \cdot 5 + C)$ изображена на рисунке 2 (для первой версии YOLO). Источник - https://arxiv.org/pdf/1506.02640.pdf. Она представляет из себя большое количество сверточных и пулинговых слоев.

## 3.3 Функция потерь

Функция потерь для обучения YOLO состоит из трех составляющих.

- Classification loss - отвечает за совпадение распределения вероятностей классов $P(class_i|object)$ внутри одной ячейки с истинным. Представляет из себя квадрат разности реальной и предсказанной вероятности.
- Localization loss - отвечает правильное расположение предсказываемого региона. Включяает в себя слагаемые с квадратами разностей соответствующих координат предсказанного и реального региона, а также квадратичных корней из их ширины и высоты.

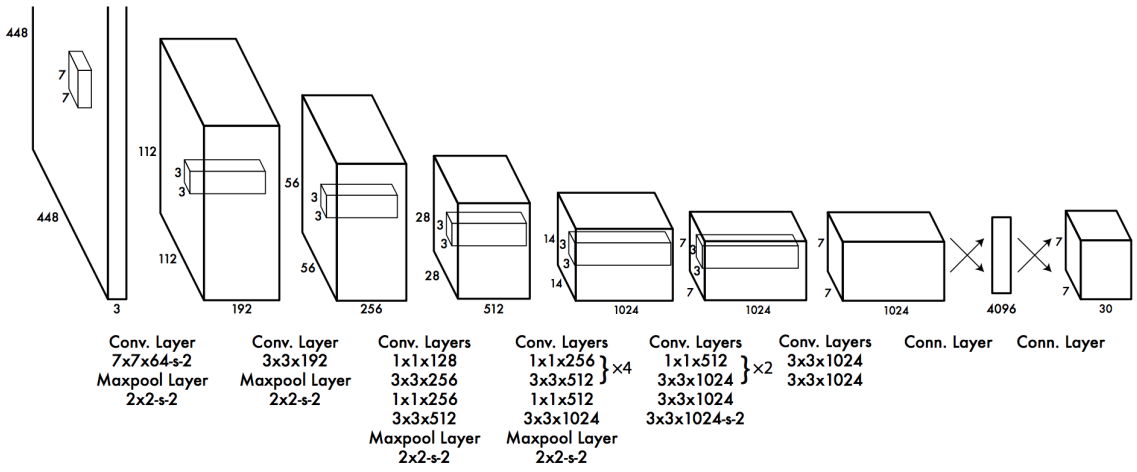


Рис. 2: Архитектура YOLO.

- Confidence loss - отвечает за то, чтобы уверенность сети в том, что в конерктной ячейке присутствует объект совпадала с реальной. Как и в предыдущих пунктах для оценки различия вероятностей используется квадрат разности.

# 4 Вместо вывода

Новый подход, предложенный авторами YOLO позволил создать метод детекции объектов на изображении, который сильно отличается по своим характеристикам от всех предложенных ранее. Перечислим преимущества и недостатки YOLO.

Преимущества YOLO

- Высокая скорость работы благодаря одному проходу сверточной нейросети
- Обучение модели может выполняться end-to-end
- Имеет хорошую обобщающую способность. Обходит старые методы на задаче domain adaptation (Например - тестирование сети, обученной на фотография, на живописи)
- Нет зависимости от метода выделения регионов для дальнейшего ранжирования (как в R-CNN)
- Детектирование по сетке ячеек позволяет позволяет достигать хороших показателей разнообразности предсказаний.

Недостатки YOLO

- Ограничения на количество регионов и классов для одной ячейки не позволяют YOLO делать точные предсказании в случае большого количества объектов, находящихся рядом друг с другом (с очень сильным перекрытием).

Метод породил серию исследований в области детектирования объектов YOLO-подобными архитектурами. Уже сейчас известно несколько улучшений стандартной модели - YOLO2, YOLO9000, YOLO3.