АРХИТЕКТУРА СОВРЕМЕННОЙ НЕЙРОННОЙ СЕТИ ДЛЯ СЕГМЕНТАЦИИ ОБЪЕКТОВ НА ИЗОБРАЖЕНИЯХ

Гончаров И. В., Потешкин А. С., Курулев А. П.
Кафедра теоретических основ электротехники, Белорусский государственный университет информатики и радиоэлектороники
Минск, Республика Беларусь
E-mail: {ivan98goncharov, artem.poteshkin92}@gmail.com

В статье излагается исследование одной из современной нейронной сети для сегментации объектов на изображениях на примере Mask R-CNN, разработанной Facebook Research.

Введение

Возможность видеть и распознавать объекты – естественная и привычная возможность для человека. Однако для компьютера пока что – это чрезвычайно сложная задача. На данный момент нейронные сети способны выполнять куда более сложные и интересные задания по обработке изображений. А именно, сеть с архитектурой Mask R-CNN позволяет выделять на фотографиях контуры («маски») экземпляров разных объектов, даже если таких экземпляров несколько, они имеют различный размер и частично перекрываются. Сеть так же способна к распознаванию поз людей.

I. Типы задач компьютерного зрения

Обычно современные задачи компьютерного зрения разделяют на четыре вида :

- Classification — классификация изображения по типу объекта, которое оно содержит.
- Semantic segmentation — определение всех пикселей объектов определённого класса или фона на изображении. Если несколько объектов одного класса перекрываются, их пиксели никак не отделяются друг от друга.
- Object detection — обнаружение всех объектов указанных классов и определение охватывающей рамки для каждого из них.
- Instance segmentation — определение пикселей, принадлежащих каждому объекту каждого класса по отдельности.

На примере изображения с воздушными шарами из (рис. 1) это можно проиллюстрировать так:

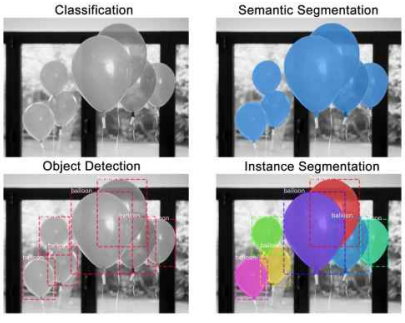

Рис. 1 – Типы задач компьютерного зрения

II. Концепции, лежащие в основе в Mask R-CNN

Концепции, лежащие в основе в Mask R-CNN прошли поэтапное развитие через архитектуры нескольких промежуточных нейросетей, решавших разные задачи из приведённого выше списка.

Convolution Neural Networks состоит из сверхточных слоев, которые позволяют объединять значения расположенных рядом пикселей и выделять более обобщённые признаки изображения. Для этого по картинке последовательно скользят квадратным окном небольшого размера (3x3, 5x5 пикселей и т.п.) называемым ядром. Каждый элемент ядра имеет свой весовой коэффициент, умножаемый на значение того пикселя изображения, на который в данный момент наложен элемент ядра. Затем полученные для всего окна числа складываются, и эта взвешенная сумма даёт значение очередного признака. Для получения матрицы («карты») признаков всего изображения, ядро последовательно сдвигается по горизонтали и вертикали. В следующих слоях операция свертки применяется уже к картам признаков, полученным из предыдущих слоёв (рис.2).

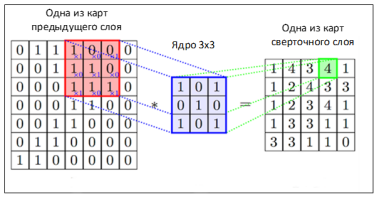

Рис. 2 – Графический процесс получения "карты"признаков изображения

Архитектура сети R-CNN (Regions With CNNs) была разработана командой из UC Berkley для применения Convolution Neural Networks к задаче object detection. Процедуру детектирования объектов сетью R-CNN можно разделить на следующие шаги:

- Выделение регионов-кандидатов при помощи Selective Search.
- Преобразование региона в размер, принимаемый CNN CaffeNet.

- Получение при помощи CNN 4096-размерного вектора признаков.
- Проведение N бинарных классификаций каждого вектора признаков при помощи N линейных SVM.
- Линейная регрессия параметров рамки региона для более точного охвата объекта.

Fast R-CNN предложили ускорила процесс R-CNN за счёт пары модификаций: пропускать через CNN не каждый из 2000 регионов-кандидатов по отдельности, а всё изображение целиком. Предложенные регионы потом накладываются на полученную общую карту признаков. Преобразование признаков, попавших в разные регионы, к фиксированному размеру производилось при помощи процедуры RoIPooling. Окно региона шириной w и высотой h делилось на сетку, имеющую HW ячеек размером h/H w/W. По каждой такой ячейке проводился Max Pooling для выбора только одного значения, давая результирующую матрицу признаков HW.

После улучшений, сделанных в Fast R-CNN, самым узким местом нейросети оказался механизм генерации регионов-кандидатов. В 2015 команда из Microsoft Research смогла сделать этот этап значительно более быстрым. Они предложили вычислять регионы не по изначальному изображению, а по карте признаков, полученных из CNN (рис.3).Для этого был добавлен модуль Region Proposal Network (RPN). В рамках RPN по извлечённым CNN признакам скользят «мини-нейросетью» с 3x3 окном. Полученные с её помощью значения передаются в два параллельных полносвязанных слоя: box-regression layer (reg) и box-classification layer (cls).

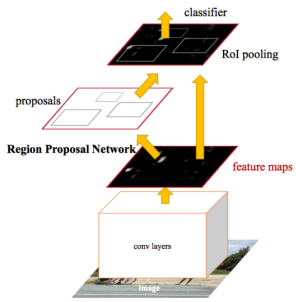

Рис. 3 – Архитектура Faster R-CNN

III. Mask R-CNN

Mask R-CNN развивает архитектуру Faster R-CNN путём добавления ещё одной ветки, которая предсказывает положение маски, покрывающей найденный объект, и решает уже задачу instance segmentation (рис.4).

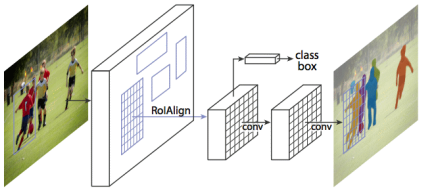

Рис. 4 – Архитектура Mask R-CNN

Маска представляет собой просто прямоугольную матрицу, в которой 1 на некоторой позиции означает принадлежность соответствующего пикселя объекту заданного класса, 0 — что пиксель объекту не принадлежит. Выделение маски происходит в class-agnostic стиле: маски предсказываются отдельно для каждого класса, без предварительного знания, что изображено в регионе, и потом просто выбирается маска класса, победившего в независимом классификаторе. Утверждается, что такой подход более эффективен. Одна из основных модификаций, возникших из-за необходимости предсказывать маску — изменение процедуры RoIPool(вычисляющей матрицу признаков для региона-кандидата) на так называемую RoIAlign. Дело в том, что карта признаков, полученная из CNN, имеет меньший размер, чем исходное изображение, и регион, охватывающий на изображении целочисленное количество пикселей, не получается отобразить в пропорциональный регион карты с целочисленным количеством признаков (рис.5).

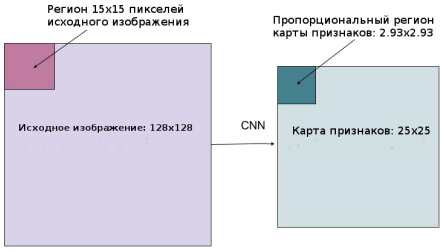

Рис. 5 – Работа процедуры RoIAlign

RoIPool проблема решалась просто округлением дробных значений до целых. Такой подход нормально работает при выделении охватывающей рамки, но вычисленная на основе таких данных маска получается слишком неточной. В противоположность этому, в RoIAlign не используется округление, все числа остаются действительными, а для вычисления значений признаков используется билинейная интерполяция по четырём ближайшим целочисленным точкам.

IV. Заключение

Сегментация объектов - одна из перспективных сфер для развития нейронных сетей. Данную технологию можно использовать на дорогах, в аэропортах, на улицах и т.д. Mask R-CNN стала новым уровнем искусства с точки зрения сегментации экземпляров.

Список Литературы

1. R. Beith: Computer Vision: Algorithms and Applications, -2010. -979c
2. https://blog.athelas.com/a-brief-history-of-cnns-in-image-segmentation-from-r-cnn-to-mask-r-cnn-34ea83205de4
3. https://engineering.matterport.com/splash-of-color-instance-segmentation-with-mask-r-cnn-and-tensorflow-7c761e238b46