Министерство науки и высшего образования Российской Федерации
ФГАОУ ВО «Северо-Восточный федеральный университет имени М.К.Аммосова»
Институт математики и информатики
Кафедра «Информационные технологии»

Технический документ

на тему:

«Обнаружение объектов с помощью RetinaNet»

Выполнили: студенты 1-го курса
группы М-ФИИТ-21
Егоров Айтал Никитич
Иннокентьев Владимир Евгеньевич
Луковцев Алексей Владимирович
Проверил: к.ф.- м.н., доцент
Григорьев Александр Виссарионович

Якутск, 2021

Введение

Чтобы получить полное понимание изображения, мы должны не только сосредоточиться на классификации различных изображений, но и попытаться точно оценить понятия и расположение объектов, содержащихся в каждом изображении. Эта задача называется обнаружением объектов и обычно состоит из различных подзадач, таких как обнаружение лиц, обнаружение пешеходов и обнаружение скелетов. Как одна из фундаментальных проблем компьютерного зрения, обнаружение объектов может предоставить ценную информацию для семантического понимания изображений и видео и связано со многими приложениями, включая классификацию изображений, анализ поведения человека, распознавание лиц и автономное вождение. Между тем, наследуя нейронные сети и связанные с ними системы обучения, прогресс в этих областях приведет к разработке алгоритмов нейронных сетей, а также окажет большое влияние на обнаружение объектов, методы которых можно рассматривать как системы обучения. Однако из-за больших различий в точках обзора, позах, окклюзиях и условиях освещения трудно идеально выполнить обнаружение объекта с помощью дополнительной задачи локализации объекта. Так много внимания было привлечено к этой области в последние годы.

Актуальность

Актуальность задач обнаружения и распознавания объектов на изображениях и их последовательностях с годами только возрастает. За последние несколько десятилетий предложено огромное количество подходов и методов обнаружения как аномалий, то есть областей изображения, характеристики которых отличаются от прогнозных, так и объектов интереса, о свойствах которых есть априорная информация, вплоть до библиотеки эталонов. В работе предпринята попытка системного анализа тенденций развития подходов и методов обнаружения, причин этого развития, а также метрик, предназначенных для оценки качества и достоверности обнаружения объектов. Рассмотрено обнаружение на основе математических моделей изображений. При этом особое внимание уделено подходам на основе моделей случайных полей и отношения правдоподобия. Проанализировано развитие сверточных нейронный сетей, направленных на задачи распознавания и обнаружения, включая ряд пред обученных архитектур, обеспечивающих высокую эффективность при решении данной задачи. В них для обучения используются уже не математические модели, а библиотеки реальных снимков. Среди характеристик оценки качества обнаружения рассмотрены вероятности ошибок первого и второго рода, точность и полнота обнаружения, пересечение по объединению, интерполированная средняя точность.

Глава 1.

Что такое распознавание объектов? Распознавание объектов — это общий термин для описания набора связанных задач компьютерного зрения, которые включают идентификацию объектов на цифровых фотографиях. Классификация изображений включает в себя предсказание класса одного объекта на изображении. Локализация объекта относится к определению местоположения одного или нескольких объектов на изображении и рисованию большого прямоугольника вокруг их границ. Обнаружение объектов объединяет эти две задачи и локализует, и классифицирует один или несколько объектов на изображении. Когда пользователь или практик говорит о «распознавании объектов», они часто имеют в виду «обнаружение объектов». Постановка задачи обнаружения объектов состоит в том, чтобы определить, где находятся объекты на данном изображении (локализация объектов) и к какой категории относится каждый объект (классификация объектов). Таким образом, конвейер традиционных моделей обнаружения объектов можно в основном разделить на три этапа: выбор информативной области, извлечение признаков и классификация.

Мы можем различать эти три задачи компьютерного зрения:

- **Классификация изображений**: предскажите тип или класс объекта на изображении.
    - Вход: изображение с одним объектом, например фотография.
    - Вывод: метка класса (например, одно или несколько целых чисел, сопоставленных с метками классов).

- **Локализация объектов**: найдите наличие объектов на изображении и укажите их местоположение с помощью ограничительной рамки.
    - Вход: изображение с одним или несколькими объектами, например фотография.
    - Выходные данные: одна или несколько ограничивающих рамок (например, определяемых точкой, шириной и высотой).

- **Обнаружение объектов**: определяйте наличие объектов с помощью ограничительной рамки и типов или классов обнаруженных объектов на изображении.
    - Вход: изображение с одним или несколькими объектами, например фотография.
    - Выходные данные: одна или несколько ограничивающих рамок (например, определяемых точкой, шириной и высотой) и метка класса для каждой ограничивающей рамки.

Архитектура нейронной сети RetinaNet

Архитектура свёрточной нейронной сети (СНС) RetinaNet состоит из 4 основных частей, каждая из которых имеет своё назначение:

a) Backbone – основная (базовая) сеть, служащая для извлечения признаков из поступающего на вход изображения. Данная часть сети является вариативной и в её основу могут входить классификационные нейросети, такие как ResNet, VGG, EfficientNet и другие.

b) Feature Pyramid Net (FPN) – свёрточная нейронная сеть, построенная в виде пирамиды, служащая для объединения достоинств карт признаков нижних и верхних уровней сети, первые имеют высокое разрешение, но низкую семантическую, обобщающую способность вторые - наоборот.

c) Classification Subnet – подсеть, извлекающая из FPN информацию о классах объектов, решая задачу классификации.

d) Regression Subnet – подсеть, извлекающая из FPN информацию о координатах объектов на изображении, решая задачу регрессии.

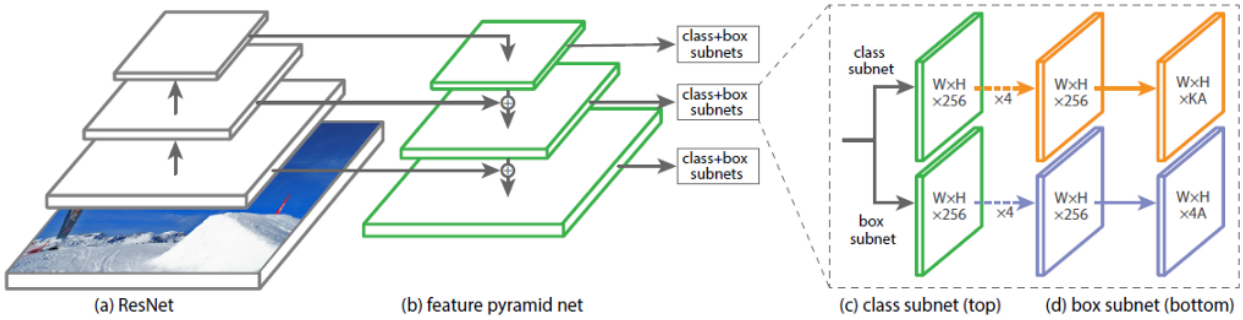

Рисунок 1 - Архитектура RetinaNet с backbone-сетью ResNet

Backbone часть сети RetinaNet

Учитывая, что часть архитектуры RetinaNet, которая принимает на вход изображение и выделяет важные признаки, является вариативной и извлеченная из этой части информация будет обрабатываться на следующих этапах, то важно выбрать подходящую backbone-сеть для лучших результатов.

Недавние исследования по оптимизации СНС позволили разработать классификационные модели, которые опередили все ранее разработанные архитектуры с лучшими показателями точности на датасете ImageNet при улучшении эффективности в 10

раз. Данные сети получили название EfficientNet-B(0-7). Показатели семейства новых сетей представлены на рис. 2.

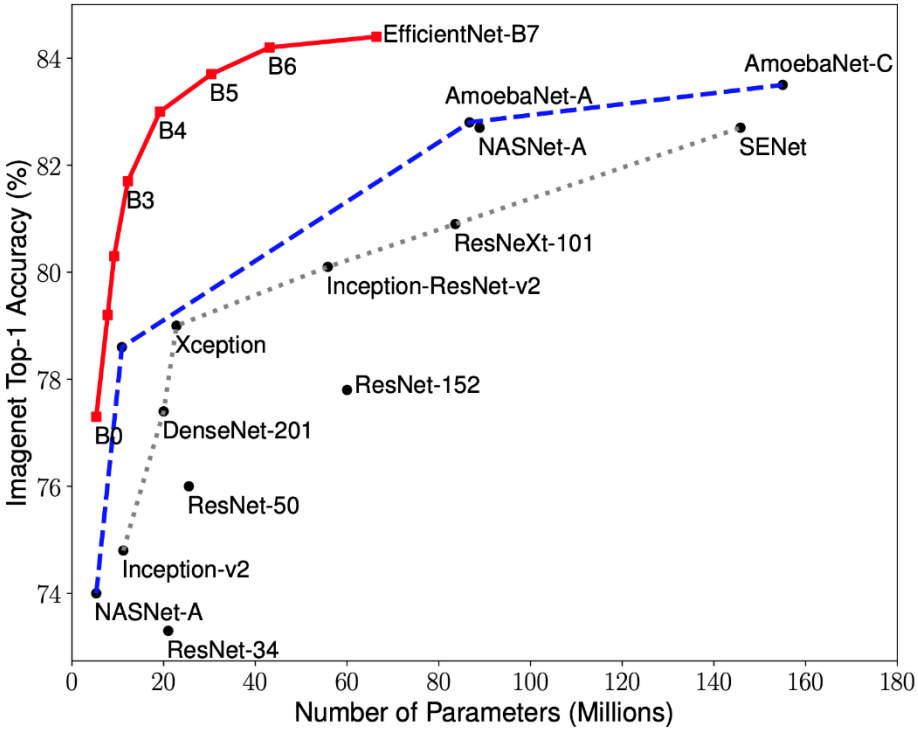

Рисунок 2 - График зависимости наибольшего показателя точности от количества весов сети для различных архитектур

Пирамида признаков

Feature Pyramid Network состоит из трёх основных частей: восходящий путь (bottom-up pathway), нисходящий путь (top-down pathway) и боковые соединения (lateral connections). Восходящий путь представляет собой некую иерархическую «пирамиду» – последовательность свёрточных слоёв с уменьшающейся размерностью, в нашем случае – backbone сеть. Верхние слои сверточной сети имеют большее семантическое значение, но меньшее разрешение, а нижние наоборот (рис. 3). Bottom-up pathway имеет уязвимость при извлечении признаков – потеря важной информации об объекте, например из-за зашумления небольшого, но значимого, объекта фоном, так как к концу сети информация сильно сжата и обобщена.

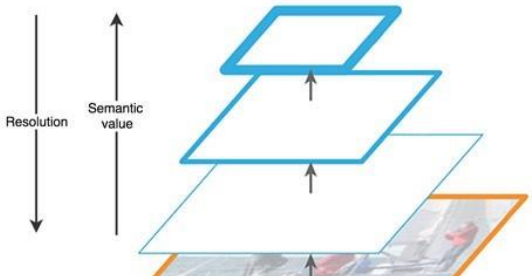

Рисунок 3 - Особенности карт признаков на разных уровнях нейросети

Нисходящий путь также представляет собой «пирамиду». Карты признаков верхнего слоя этой пирамиды имеют размер карт признаков верхнего слоя bottom-up пирамиды и увеличиваются вдвое методом ближайшего соседа (рис. 4) по направлению вниз.

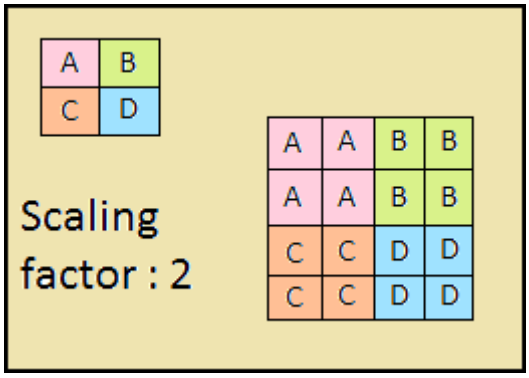

Рисунок 4 – Увеличение разрешения изображения методом ближайшего соседа

Таким образом в top-down сети каждая карта признаков вышележащего слоя увеличивается до размеров карты нижележащего. Помимо этого, в FPN присутствуют боковые соединения, это означает, что карты признаков соответствующих слоёв bottom-up и top-down пирамид поэлементно складываются, причём карты из bottom-up проходят свёртку 1*1. Этот процесс схематично представлен на рис. 5.

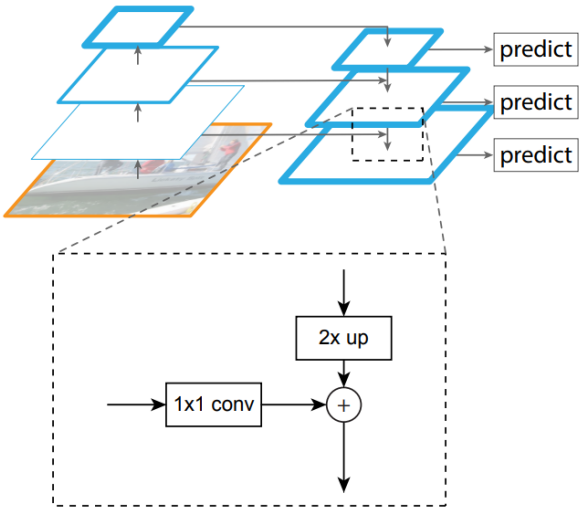

Рисунок 5 – Устройство пирамиды признаков

Боковые соединения решают проблему затухания важных сигналов в процессе прохода по слоям, совмещая семантически важную информацию, полученную к концу первой пирамиды и более детальную информацию, полученную в ней ранее.

Далее, каждый из полученных слоёв в top-down пирамиде обрабатывается двумя подсетями.

Подсети классификации и регрессии

Третей частью архитектуры RetinaNet являются две подсети: классификационная и регрессионная (рис. 6). Каждая из этих подсетей образует на выходе ответ о классе объекта и его расположении на изображении. Рассмотрим принцип работы каждой из них.

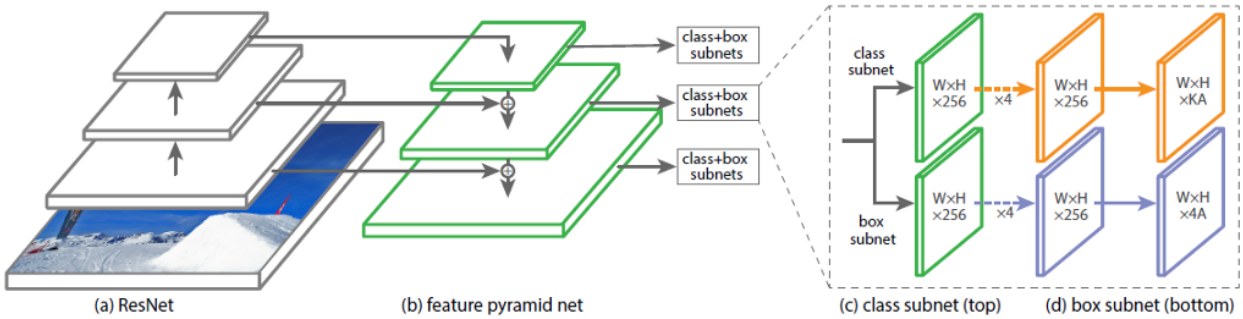

Рисунок 6 – Подсети RetinaNet

Разница в принципах работы рассматриваемых блоков (подсетей) не отличается до последнего слоя. Каждый из них состоит из 4 слоёв свёрточных сетей. В слое формируются 256 карт признаков. На пятом слое количество карт признаков изменяется: регрессионная подсеть имеет 4*A карт признаков, классификационная – K*A карт признаков, где A – количество якорных рамок (подробное описание якорных рамок в следующем подразделе), K–количество классов объектов.

В последнем, шестом, слое каждая карта признаков преобразуется в набор векторов. Регрессионная модель на выходе имеет для каждой якорной рамки вектор из 4 значений, указывающих смещение целевой рамки (англ. ground-truth box) относительно якорной. Классификационная модель имеет на выходе для каждой якорной рамки one-hot вектор длиной K, в котором индекс со значением 1 соответствует номеру класса, который нейросеть присвоила объекту.

Якорные рамки

В прошлом разделе был использован термин якорных рамок. Якорная рамка (англ. anchor box) – гиперпараметр нейросетей-детекторов, заранее определенный ограничивающий прямоугольник, относительно которого работает сеть.

Допустим, сеть имеет на выходе карту признаков размером 3*3. В RetinaNet каждая из ячеек имеет 9 якорных рамок, каждая из которых имеет разный размер и соотношение сторон (рис. 7). Во время обучения каждой целевой рамке подбираются в соответствие якорные рамки. Если их показатель IoU имеет значение от 0.5, то якорная рамка назначается целевой, если значение меньше 0.4, то она считается фоном, в других случаях якорная рамка будет проигнорирована для обучения. Классификационная сеть обучается относительно выполненного назначения (класс объекта или фон), регрессионная сеть обучается относительно координат якорной рамки (важно отметить, что ошибка вычисляется относительно якорной, но не целевой рамки).

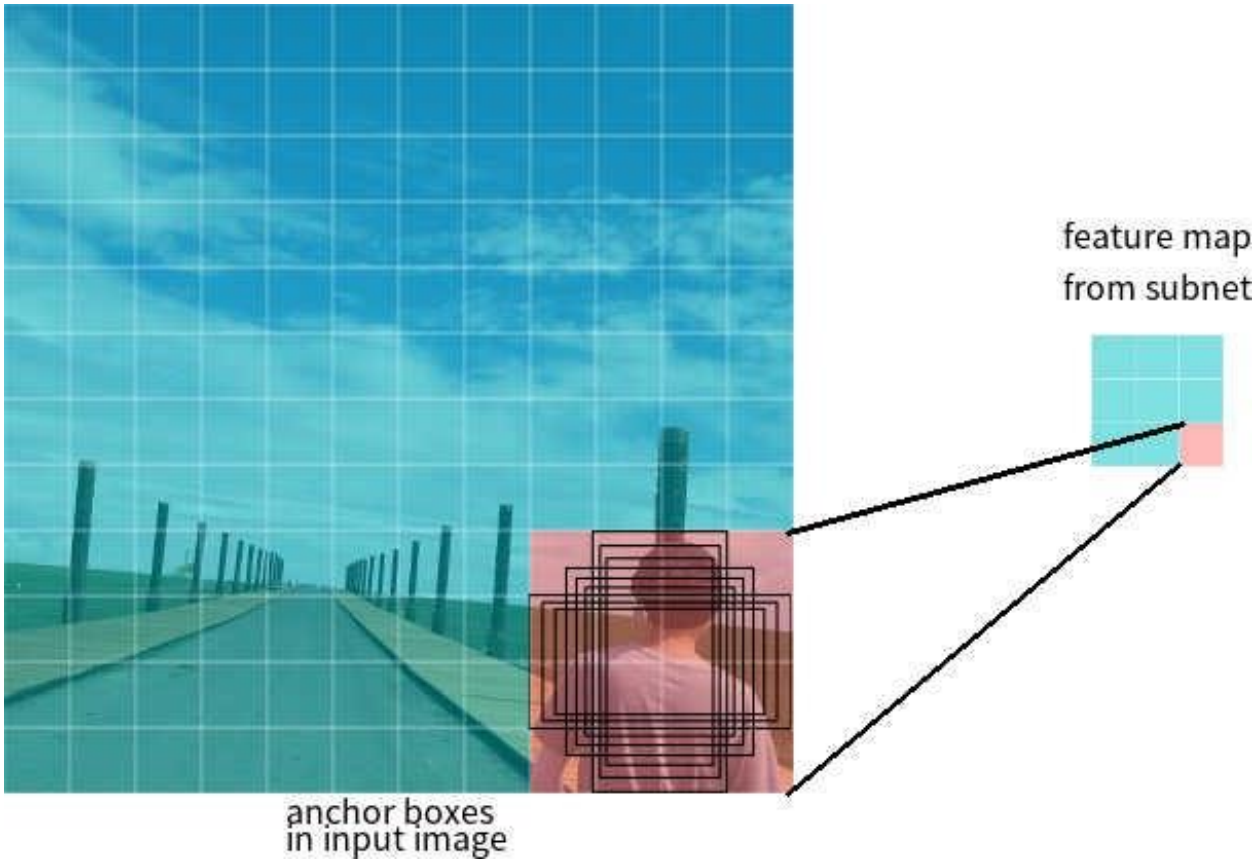

*Рисунок 7 - Якорные рамки для одной ячейки карты признаков с размером 3*3*

Функции потерь

Потери RetinaNet являются составными, их составляют два значения: ошибка регрессии, или локализации (ниже обозначено как Lloc), и ошибка классификации (ниже обозначено как Lcls). Общая функция потерь может быть записана как:

$$L = \lambda Lloc + Lcls$$

Где λ является гиперпараметром, который контролирует баланс между двумя потерями.

Рассмотрим подробнее вычисление каждой из потерь.

Как было описано ранее, каждой целевой рамке назначается якорная. Обозначим эти пары как (Ai,Gi)i=1,...N, где A представляет якорь, G – целевую рамку, а N количество сопоставленных пар.

Для каждого якоря регрессионная сеть предсказывает 4 числа, которые можно обозначить как Pi=(Pix,Piy,Piw,Pih). Первые две пары означают предсказанную разницу между координатами центров якорной Ai и целевой рамки Gi, а последние две – предсказанную разницу между их шириной и высотой. Соответственно, для каждой целевой рамки вычисляется Ti, как разница между якорной и целевой рамкой:

$$Lloc = \sum j \in \{x, y, w, h\}\, smoothL1\,(Pij - Tij)$$

Где smoothL1(x) определяется формулой ниже:

$$smooth_{\mathrm{L1}}(x) = \begin{cases} 0.5x^2, & |x| < 0 \\ |x| - 0.5, & |x| \geq 0 \end{cases}$$

Потери задачи классификации в сети RetinaNet вычисляются с помощью функции Focal loss.

$$Lcls = -i = 1Kiyilog(pi)(1 - pi)\gamma$$

где K – количество классов, yi – целевое значение класса, p – вероятность предсказания i-го класса, γ – параметр фокуса, α – коэффициент смещения. Данная функция является усовершенствованной функцией кросс-энтропии. Отличие заключается в добавлении параметра γ∈(0,+∞), который решает проблему несбалансированности классов. Во время обучения, большая часть объектов, обрабатываемых классификатором, является фоном, который является отдельным классом. Поэтому может возникнуть проблема, когда нейросеть обучится определять фон лучше, чем другие объекты. Добавление нового параметра решило данную проблему, уменьшив значение ошибки для легко классифицируемых объектов. Графики функций focal и cross entropy представлены на рис.8.

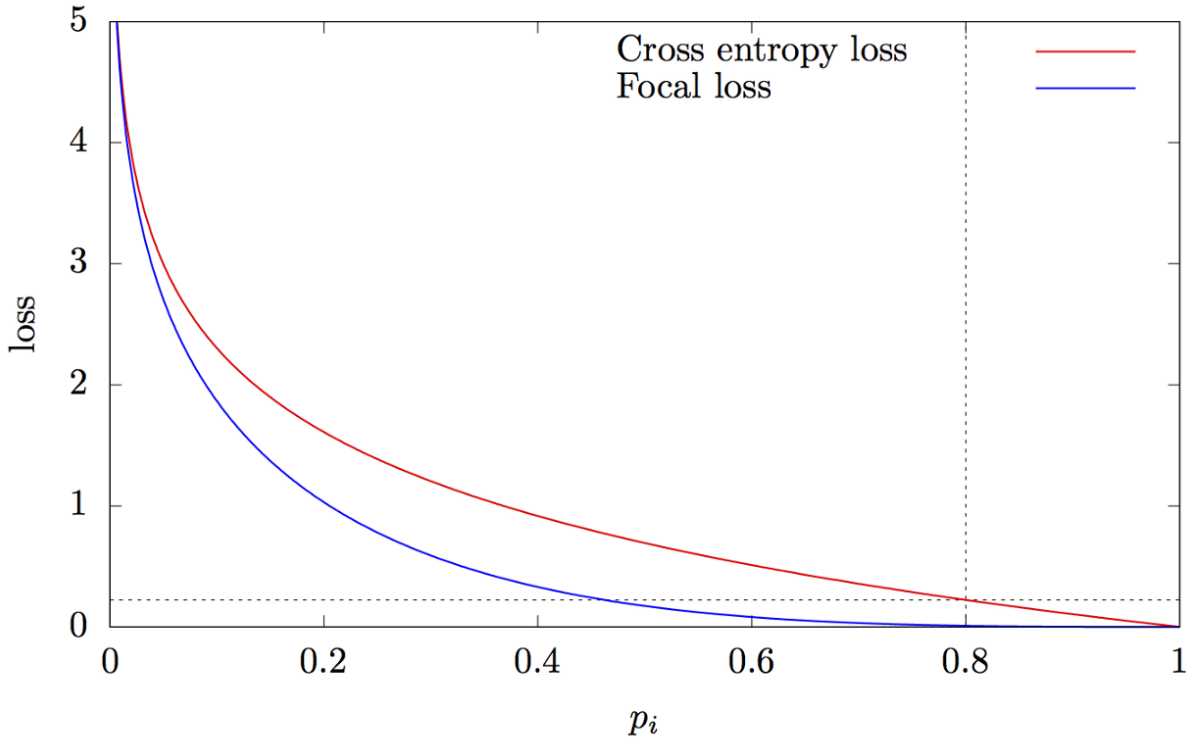

Рисунок 8 – Графики focal и cross entropy функций