DOI: 10.15514/ISPRAS-2022-34(2)-8

Функции потерь для обучения моделей сегментации изображений документов

[1,2] *А.И. Перминов, ORCID: 0000-0001-8047-0114 <perminov@ispras.ru>*
[1,2] *Д.Ю. Турдаков, ORCID: 0000-0001-8745-0984 <turdakov@ispras.ru>*
[1] *О.В. Беляева, ORCID: 0000-0002-6008-9671 <belyaeva@ispras.ru>*
[1] *Институт системного программирования им. В.П. Иванникова РАН, 109004, Россия, г. Москва, ул. А. Солженицына, д. 25*
[2] *Московский государственный университет имени М.В. Ломоносова, 119991, Россия, Москва, Ленинские горы, д. 1*

Аннотация. Работа посвящена повышению качества результатов сегментации изображений документов различных научных статей и нормативно-правовых актов нейросетевыми моделями путём обучения с использованием модифицированных функций потерь, учитывающих особенности изображений выбранной предметной области. Проводится анализ существующих функций потерь, а также разработка новых функций, оперирующих, как только координатами ограничивающих прямоугольников, так и использующих информацию о пикселях входного изображения. Для оценки качества выполняется обучение нейросетевой модели сегментации с модифицированными функциями потерь, а также проводится теоретическая оценка с помощью симуляционного эксперимента, показывающего скорость сходимости и ошибку сегментации. В результате исследования созданы быстро сходящиеся функции потерь, улучшающие качество сегментации изображений документов с использованием дополнительной информации о входных данных.

Ключевые слова: сегментация изображений документов; функции потерь; модификация функции потерь

Для цитирования: Перминов А.И., Турдаков Д. Ю., Беляева О.В. Функции потерь для обучения моделей сегментации изображений документов. Труды ИСП РАН, том 34, вып. 2, 2022 г., стр. 89-110. DOI: 10.15514/ISPRAS–2022–34(2)–8

Loss functions for train document image segmentation models

[1,2] *A.I. Perminov, ORCID: 0000-0001-8047-0114 <perminov@ispras.ru>*
[1,2] *D.Y. Turdakov, ORCID: 0000-0001-8745-0984 <turdakov@ispras.ru>*
[1] *O.V. Belyaeva, ORCID: 0000-0002-6008-9671 <belyaeva@ispras.ru>*
[1] *Ivannikov Institute for System Programming of the Russian Academy of Sciences, 25, Alexander Solzhenitsyn st., Moscow, 109004, Russia*
[2] *Lomonosov Moscow State University, GSP-1, Leninskie Gory, Moscow, 119991, Russia*

Abstract. The work is devoted to improving the quality of the results of image segmentation of documents of various scientific articles and legal acts by neural network models by learning using modified loss functions that take into account the features of images of the selected subject area. The analysis of existing loss functions is carried out, as well as the development of new functions that operate both with the coordinates of the bounding boxes and using information about the pixels of the input image. To assess the quality, a neural network segmentation model with modified loss functions is trained, and a theoretical assessment is carried out using a simulation experiment showing the convergence rate and segmentation error. As a result of the study,

rapidly converging loss functions were created that improve the quality of document image segmentation using additional information about the input data.

Keywords: document image segmentation; loss functions; loss function modifications

For citation: Perminov A.I., Turdakov D. Y., Belyaeva O.V. Loss functions for train document image segmentation models. Trudy ISP RAN/Proc. ISP RAS, vol. 34, issue 2, 2022, pp. 89-110 (in Russian). DOI: 10.15514/ISPRAS–2022–34(2)–8

1. Введение

Огромное количество различных документов хранятся в электронном виде. Некоторые из них содержат копируемый текст и предоставляют возможность навигации по нему с помощью оглавления и ссылок. Однако есть документы (например, книги, отчеты, квитанции и т.д.), которые представляют собой сканированные копии бумажных документов. Такие документы не имеют текстового слоя. Таким образом, документы не упорядочены, в них сложно искать информацию и невозможно напрямую извлечь из них текст.

В связи с тем, что существует необходимость систематизировать и анализировать информацию в документах, предполагается, что документы имеют определенную структуру. Это означает, что существуют определенные шаблоны, по которым создаются документы. Это предположение позволяет говорить о возможности автоматического анализа электронного документа и извлечения логической структуры.

Практически для всех документов общим шаблоном можно считать геометрическую организацию – шрифты, отступы, линии, более крупные блоки – абзацы текста, заголовки, изображения и т.д. Для выделения таких признаков используется сегментация страницы документа.

В настоящее время для сегментации используются нейронные сети, поскольку они позволяют производить высококачественную обработку изображений. Для обучения нейросетей используются алгоритмы минимизации некоторой функции, способной оценить качество сети. Такие функции называются функциями потерь. Обычно, к ним предъявляют требование на дифференцируемость и непрерывность в некоторой области, однако существуют функции, для которых это свойство не выполняется в некоторых точках. В силу работы нейронных сетей с компьютерными числами, не являющимися алгебраически точным математическим представлением, это оказывается вполне допустимым.

При разработке функций потерь зачастую преследуется цель получить функцию, оценивающую качество работы оптимизируемой модели, однако далеко не всегда уделяется должное внимание анализу прочих свойств получаемых функций, как, например, градиенты, гладкость, специфические свойства данных предметной области и т.д. Функция потерь может иметь высокую скорость сходимости и показывать близкие к нулю значения после оптимизации, однако в результате оптимизации модель может работать недостаточно хорошо из-за того, что не были учтены какие-то специфические особенности входных данных. Разработка и исследование функций, специфических для конкретной задачи может существенно повысить качество получаемой модели. При этом, возможно, и существующие функции могут помочь достичь такого же качества, но, например, за большее время, из-за чего целесообразнее будет применить модифицированный вариант.

Для сегментации изображений документов в большинстве решений используются те же нейросетевые модели, что и для сегментации естественных изображений - объектов реального мира. Поскольку естественные изображения лишены общей структуры, то и модели сегментации не учитывают дополнительных свойств, присущих изображениям текстовых документов. Помимо этого, такие модели плохо и долго обучаются, а результаты их обучения на изображениях документов приводят к необходимости использования дополнительной обработки.

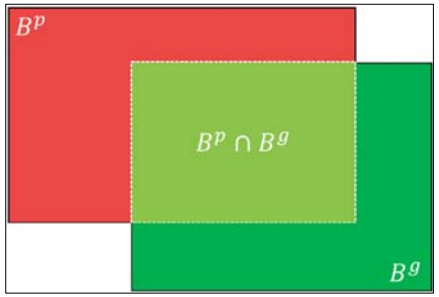
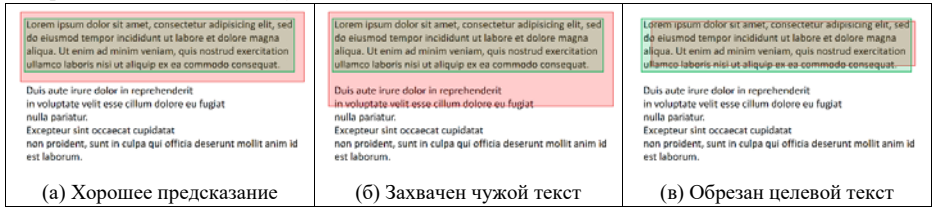


Рис. 1. Пример двух ограничивающих прямоугольников
Fig. 1. Two bounding boxes example

Модели сегментации оперируют координатами двух ограничивающих прямоугольников (рис. 1): $B^g = (x_1^g, y_1^g, x_2^g, y_2^g)$ – ожидаемым и $B^p = (x_1^p, y_1^p, x_2^p, y_2^p)$ – предсказанным, с помощью которых находятся площади самих прямоугольников и их пересечения. Основной задачей сегментации является регрессия координат предсказанной области к целевой. Оценка качества сегментации выполняется путём анализа значений получаемых площадей. Для сегментации произвольных изображений этого вполне достаточно, однако сегментация изображений документов накладывает такие ограничения, как отсутствие обрезки текста (рис. 2 (в)) или наложение на другой блок текста (рис. 2 (б)). Ситуация, в которой текст выделен с учётом этих ограничений, но при этом взято больше фона (рис. 2 (а)), также считается хорошей, в отличие от распространённых функций потерь. В то же время координатные функции вынуждают модель сегментации подстраиваться под геометрию того домена, на котором производится обучение. Из-за этого обработка документов с другим расположением элементов (но всё ещё являющихся манхэттенскими) приводит к серьёзной потере качества вплоть до нулевого.



(а) Хорошее предсказание (б) Захвачен чужой текст (в) Обрезан целевой текст

Рис. 2. Пример предсказаний с учётом перечисленных свойств
Fig. 2. An example of predictions taking into account the listed properties

По этой причине далее разрабатываются функции потерь, учитывающие не только геометрию областей, но и особенности сегментации изображений документов – возможность захватить больше фона, запрет на обрезку текста и захват чужого текста.

2. Существующие решения

В работах [1], [2], [3] к функции потерь добавляют коэффициенты, дающие ненулевые градиенты при отсутствии пересечения, чтобы избежать основной проблемы распространённой IoU функции.

В работе [3] предлагается новая функция потерь SCA для замены классической функции потерь – IoU. Разработанная функция, вместо оценки отношения площади пересечения прямоугольников к площади объединения, как это делает IoU, независимо оценивает отношения размеров пересечения и минимального покрывающего прямоугольников. Исследователи обучают различные нейросетевые модели с её использованием на нескольких

наборах данных изображений общего назначения и сравнивают результаты на метриках усреднённой точности (mean average precision) с различными порогами. По результатам оценки качества большинство моделей показывают более высокие результаты с модифицированной функцией потерь.

Исследователи в [4] активно исследуют проблемы IoU функции и скорость её сходимости и в результате предлагают семейство функций, называемое "α-IoU". Привычная IoU функция возводится в степень α и в результате проводимых экспериментов демонстрируется более высокая сходимость при определённых значениях подбора введённого коэффициента.

В работе [5] авторы исследуют гладкие версии функции IoU, используемые совместно с перекрёстной энтропией с целью повышения качества классификации объекта. Анализируя градиенты получаемой функции, исследователям удаётся уменьшить количество выбросов и разрывов, а также повысить точность сегментации на классических наборах данных.

Авторы работы [6] анализируют градиенты модификаций IoU и, жертвуя инвариантностью относительно масштабирования и используя разность площадей двух прямоугольников, получают функцию, способную хорошо различать очень похожие и расположенные близко ограничивающие прямоугольники. При этом градиенты такой функции получаются более гладкими, что приводит к высокой скорости сходимости моделей.

Описанные выше модификации функций потерь опираются лишь на геометрические признаки или математические свойства. Для учёта особенностей входных данных их использование невозможно, а потому необходимо разработать собственные функции потерь.

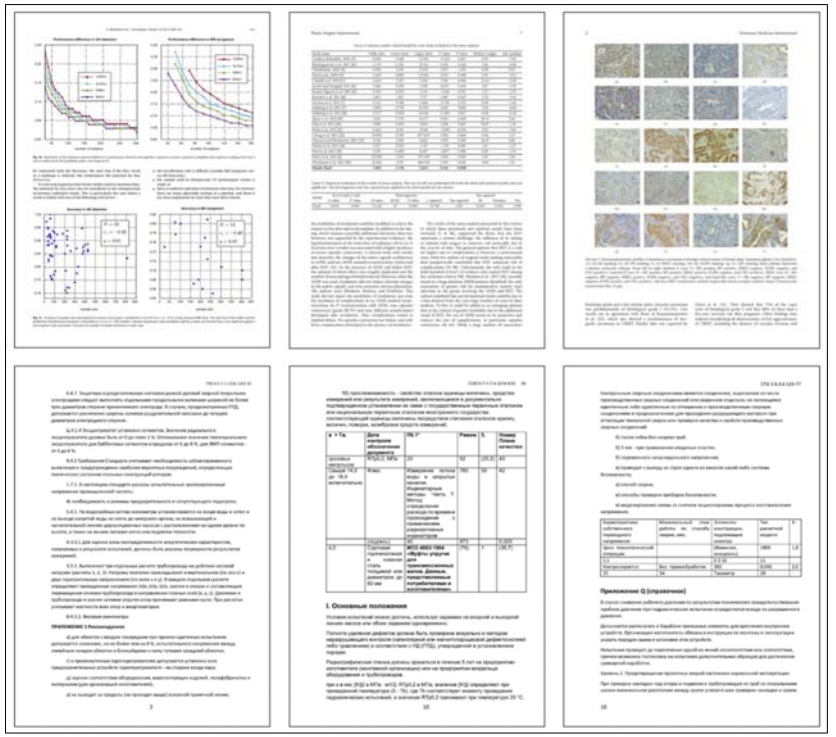


Рис. 3. Примеры реальных документов
Fig. 3. Real documents examples

3. Изображения документов для сегментации

В качестве входных данных рассматриваются отсканированные изображения различных статей и нормативно-правовых документов с тёмным текстом и светлым однородным фоном (рис. 3).

Сканированные документы, как правило, характеризуются высоким качеством символов, белым фоном, низким уровнем шума и манхэттенским стилем оформления. Манхэттенский документ – это документ, в котором границы всех блоков прямые (каждый блок или прямоугольный, или представляет собой несколько прямоугольников, у которых некоторые вершины и части сторон общие).

4. Исследование функций потерь

В настоящее время наиболее качественной моделью сегментации изображений является архитектура YOLO [7] пятой версии. Данная модель использует комбинированную функцию потерь – для регрессии используется CIoU [1], а для классификации используется перекрёстная энтропия. Каждую из составляющих легко модифицировать для собственных экспериментов, чего нельзя сказать о большинстве других используемых моделей. Также эта модель сегментации довольно быстро обучается. По вышеперечисленным причинам была выбрана модель архитектуры YOLO.

4.1 Разработка метрик, учитывающих особенности документов

Свойства метрик, учитывающие особенности изображений документов:

- содержимое целевой области должно полностью находиться внутри предсказанной области – обрезка текста должна быть сведена к минимуму;
- содержимое, относящееся к другому объекту не должно входить в выделенную область – чужой текст не должен быть захвачен;
- размер предсказанной области может быть, как меньше, так и больше определённой, если выполнены указанные выше правила – это позволяет выделять больше фоновых пикселей, что не является критичным, и в то же время позволяет выделять область ровно по контенту, если фоновых пикселей было слишком много в целевой области;
- в случае полного совпадения областей значение метрики равно единице, во всех остальных случаях значение не превышает единицы.

4.1.1 PIoU – Pixel intersection over union

В качестве самой простой функции потерь, удовлетворяющей всем условиям можно взять следующую:

$$PIoU = \frac{Black(img, |B^g \cap B^p|)}{Black(img, |B^g \cup B^p|)}. \quad (1)$$

$Black(img, v)$ – количество тёмных пикселей в области v изображения img. Под тёмным пикселем будем считать тот, чья яркость меньше некоторого порога threshold. Поскольку сегментация производится на изображениях в RGB пространстве, то в качестве яркости можно использовать усреднение по трём каналам:

$$Brightness(pixel) = \frac{r + g + b}{3} \quad (2)$$

Областью будем считать прямоугольник, определяемый координатами левого верхнего и правого нижнего углов – (x_1, y_1, x_2, y_2).

Таким образом, функция подсчёта количества тёмных пикселей может быть описана следующим образом:

$$Black = \sum_{x=x_1}^{x_2} \sum_{y=y_1}^{y_2} \delta(x, y), \quad (3)$$

$$\delta(x, y) = \begin{cases} 1, \text{если } Brightness(x, y) < threshold \\ 0, \text{иначе} \end{cases}. \quad (4)$$

4.1.2 BWIoU – Black and White intersection over union

Полезным свойством PIoU (1) является выполнение всех перечисленных условий, однако при исследовании её свойств было установлено (табл. 1), что значения метрики при небольшом наложении на чужой текст или малое обрезание целевой области лишь немного уменьшает значение метрики. Поэтому было поставлено дополнительное свойство – более резкое изменение значения при описанных ситуациях и была придумана новая метрика, учитывающая не только соотношение тёмных пикселей, но и светлых (не тёмных), а также штрафующая за тёмные пиксели вне пересечения:

$$BWIoU = \frac{w \cdot Black_\cap + (1 - w) \cdot White_\cap}{w \cdot Black_\cup + (1 - w) \cdot White_\cup + \lambda(Black_\cup - Black_\cap)}. \quad (5)$$

В качестве значений весовых коэффициентов выбраны $w = 0.7$, $\lambda = 10$. $Black_\cap$ и $Black_\cup$ – количество тёмных пикселей в пересечении и объединении областей соответственно. Аналогично, $White_\cap$ и $White_\cup$ – количество светлых пикселей в пересечении и объединении областей (рис. 4).

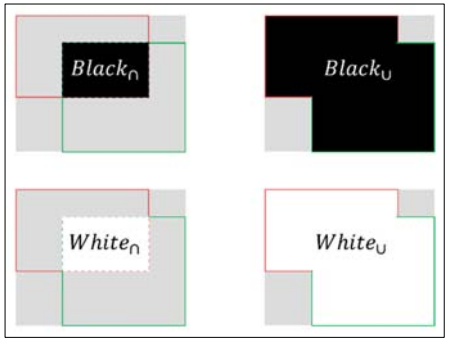


Рис. 4. Визуализация пиксельных метрик
Fig. 4. Pixel metrics visualization

4.1.3 Weighted BWIoU

Обычная функция BWIoU (5) значительно лучше учитывает обрезку текста и наложение на другой текст, однако при увеличении фона довольно сильно уменьшает своё значение (табл. 1). Это происходит в основном из-за подобранных значений коэффициента w. Величину весового коэффициента стоит выбирать адаптивно в зависимости от параметров выделенных и ожидаемых областей, а потому была получена новая метрика:

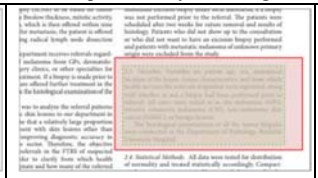
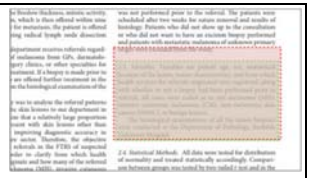
$$Weighted\ BWIoU = \frac{w_\cap \cdot Black_\cap + (1 - w_\cap) \cdot White_\cap}{w_\cup \cdot Black_\cup + (1 - w_\cup) \cdot White_\cup + \lambda(Black_\cup - Black_\cap)}. \quad (6)$$

В этой метрике весовые коэффициенты для пересечения и для объединения вычисляются следующим образом:

$$w_\cap = \frac{White_\cap}{Area_\cap}, w_\cup = \frac{White_\cup}{Area_\cup}. \quad (7)$$

Использование адаптивных коэффициентов позволило практически полностью не реагировать на изменения области лишь в фоновых пикселях, а также более резко реагировать на обрезку или наложение.

Табл. 1. Пример пиксельных метрик при различных положениях областей
Table 1. Example of a pixel metric for various positions of bounding boxes

		
IoU = 0.7906 PIoU = 0.9299 BWIoU = 0.7083 Weighted BWIoU = 0.6675	IoU = 0.7403 PIoU = 1 BWIoU = 0.7581 Weighted BWIoU = 0.9792	IoU = 0.8177 PIoU = 0.9329 BWIoU = 0.7274 Weighted BWIoU = 0.6793

4.2 Метрика, основанная на оптическом распознавании символов

Поскольку одной из основных целей сегментации изображений документов является анализ содержимого, который практически невозможен без средств оптического распознавания символов (OCR), необходимо выполнять сегментацию таким образом, чтобы результат получения текста средствами OCR максимально соответствовал тексту на анализируемом изображении. Для оценки качества можно получить текст, находящийся в целевом и предсказанном ограничивающих прямоугольниках, сопоставить области друг с другом и сравнить похожесть текстов с помощью редакционного расстояния, например, с помощью расстояния Левенштейна:

$$ratio(str_1, str_2) = 1 - \frac{D_{Levenshtein}(str_1, str_2)}{len(str_1) + len(str_2)}. \quad (8)$$

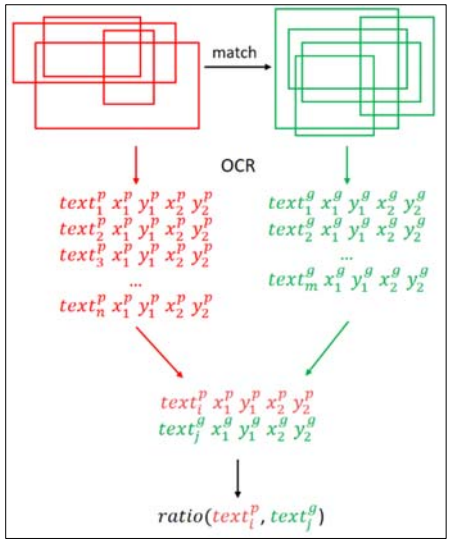


Рис. 5. Схема работы метрики OCR
Fig. 5. OCR metric flow diagram

В качестве метрики для набора данных выбирается среднее значение $ratio$ (формула 8):

$$Metric_{OCR} = \frac{1}{N_p} \cdot \sum_{i=1}^{N_p} ratio(OCR\left(Nearest(bboxes_g, bbox_i^p), OCR(bbox_i^p)\right) \quad (9)$$

Здесь $OCR(bbox)$ - получение текста средствами оптического распознавания символов в области, соответствующей прямоугольнику $bbox$. $Nearest(bboxes_g, bbox_i^p)$ – наиболее близкий прямоугольник среди целевых $bboxes_g$ к предсказанному прямоугольнику $bbox_i^p$. В качестве меры близости выбирается прямоугольник, имеющий наибольшую площадь пересечения. N_p – количество предсказанных прямоугольников. Схематично процесс вычисления данной метрики приведён на рис. 5.

4.3 Использование метрик в качестве функций потерь

Чтобы метрику можно было использовать как функцию потерь, необходимо иметь возможность вычислить градиент по входным параметрам такой функции. Входными параметрами являются координаты ограничивающих прямоугольников. Для оптимизации модели сегментации необходимо вычислять градиент по координатам предсказанной области. К сожалению, предложенные функции не имеют градиента по входным параметрам, поскольку для получения оценки качества вместо значений координат, используется взятие подобласти изображения – срез – не дифференцируемая операция, поскольку координаты ограничивающих прямоугольников являются вещественными числами (хранятся нормализованные координаты относительно размеров изображения), а для получения области на изображении необходимо использовать лишь целое число.

В результате метрика даёт некоторое число без градиента – константу, характеризующую предсказанную область. Поскольку получаемое число для каждой области своё, то его можно использовать как весовой коэффициент к дифференцируемой функции, например, к IoU. Таким образом, итоговая функция потерь будет выглядеть следующим образом:

$$L_{pixel} = 1 - f_{diff} \cdot f_{pixel} \quad (10)$$

где f_{diff} – дифференцируемая координатная метрика, f_{pixel} – значение пиксельной метрики.

Таким образом, пиксельный коэффициент будет являться регуляризатором – для областей без особенностей практически ничего не будет меняться, а для областей с наложением или обрезкой значение функции потерь будет значительно больше.

4.4 Оценка качества функций потерь

Для оценки качества работы функций потерь до непосредственно обучения модели сегментации исследователи проводят эксперимент-симуляцию (алгоритм 1): выбирается некоторое количество якорей (ограничивающих рамок, относительно которых будет проводиться оптимизация смещения) и сетка на области единичного квадрата. Случайным образом создаются целевые ограничивающие прямоугольники. Аналогично случайным образом для каждого якоря создаются случайные предсказанные прямоугольники в каждой точке выбранной сетки и запускается итерационный процесс: для каждой пары предсказанного и целевого прямоугольников вычисляется значение функции потерь, градиенты по координатам предсказанного прямоугольника и ошибка регрессии (например, сумма модулей разности координат или пиксельная метрика, описанная выше), после чего координаты обновляются методом градиентного спуска и процесс запускается сначала.

Результатом работы такой симуляции является оценка скорости сходимости функции потерь, по которой можно сделать выводы о целесообразности применения такой функции в обучении реальной модели. Основным достоинством такого эксперимента, что он не занимает столько времени как обучение полноценной модели на (обычно) большом наборе данных.

Input:
N_p - количество предсказанных прямоугольников,
N_g - количество целевых прямоугольников,
$B^p = \{B_i^p\}_{i=1}^{N_p}$ - предсказанные прямоугольники,
$B^g = \{B_j^g\}_{j=1}^{N_g}$ - целевые прямоугольники,
T - количество итераций повторения,
η - скорость градиентного спуска, по умолчанию 0.01
Output: средняя ошибка регрессии всех предсказанных прямоугольников ко всем целевым

1: **for** t= 1...T **do**
2: **for** i= 1...N_p **do**
3: **for** j= 1...N_g **do**
4: $Loss = L(B_{i,t}^p, B_j^g)$
5: $\nabla B_{i,t}^p = \frac{\partial Loss}{\partial B_{i,t}^p}$
6: $B_{i,t+1}^p = B_{i,t}^p - \eta \nabla B_{i,t}^p$
7: $RegressionError_t = |B_i^p - B_j^g|$
8: **end for**
9: **end for**
10: **end for**
11: **return** $mean(RegressionError_t)$

Алгоритм 1. Симуляционный эксперимент
Algorithm 1. Simulation experiment

Для построенных функций был проведён описанный эксперимент. В качестве дифференцируемой функции потерь была использована модификация IoU – CIoU [1].

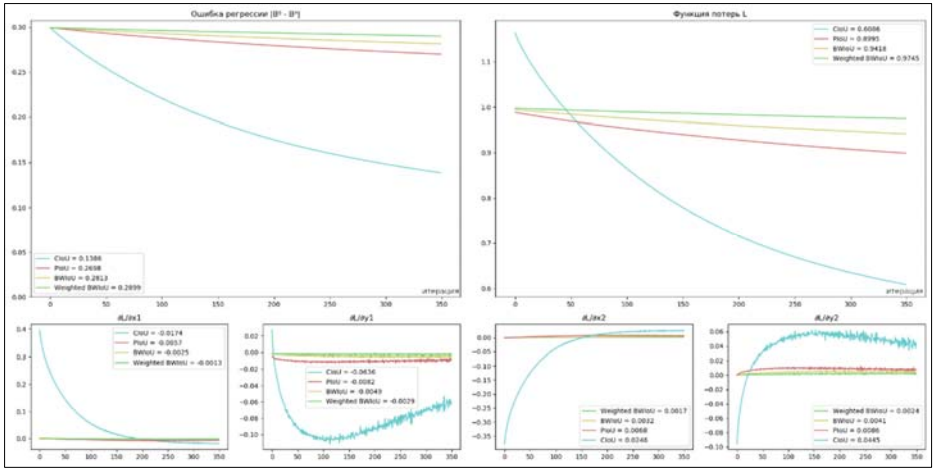


Рис. 6. Результаты симуляционного эксперимента пиксельных функций
Fig. 6. Results of simulation experiment of pixel functions

Симуляция для построенных функций (рис. 6) показала очень медленную сходимость, из-за чего, несмотря на качество оценки, в данном виде использовать их невыгодно. При этом нет никакого смысла использовать пиксельную оценку в то время, когда целевая и предсказанная области находятся далеко друг от друга – имеют маленькое пересечение. Решением данной

проблемы являлась бы функция, которая включала пиксельную метрику лишь ближе к концу, а в остальное время работала бы обычная дифференцируемая функция без изменений.

4.5 Усовершенствованная пиксельная функция потерь

Чтобы избежать проблемы медленной сходимости, но оставить преимущества пиксельного коэффициента, была введена новая функция потерь:

$$L_{fast} = 1 - f_{diff} \cdot \left(f_{pixel} + 1 - f_{diff}^*\right). \quad (11)$$

Здесь f_{diff}^* – значение функции f_{diff}, не имеющее градиента. Это позволяет не менять направление градиента, а лишь его амплитуду, аналогично изначальному варианту. Когда два ограничивающих прямоугольника далеко друг от друга или же практически не пересекаются, коэффициент $f_{pixel} + 1 - f_{diff}^*$ практически равен 1, поскольку значение f_{pixel} в такой области либо равно нулю либо ничтожно мало, как и f_{diff}^*. Когда же области находятся близко друг к другу, $1 - f_{diff}^*$ стремится к нулю, благодаря чему основной вклад вносит f_{pixel}. Таким образом большую часть времени сходится обычная функция потерь и лишь под конец пиксельная. Результаты сходимости модифицированных функций потерь представлены на рис.7.

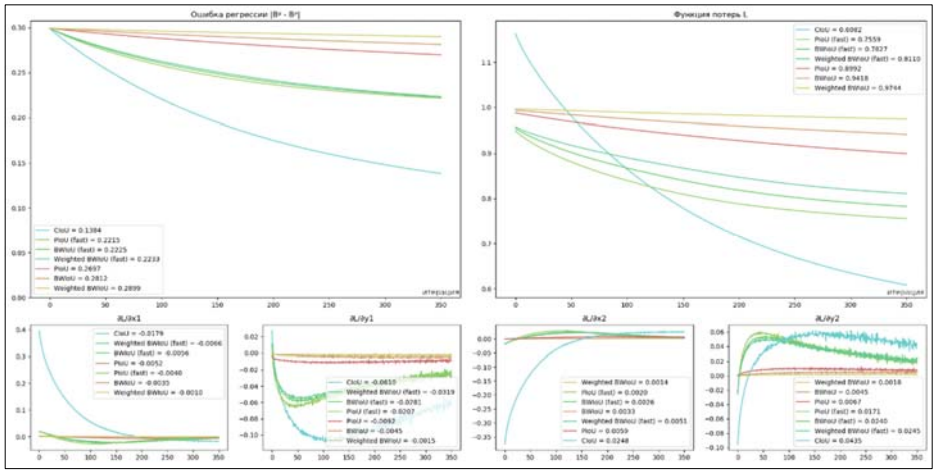


Рис. 7. Результаты симуляционного эксперимента усовершенствованных пиксельных функций
Fig. 7. Results of the simulation experiment of advanced pixel functions

Скорость сходимости модифицированных функций потерь ($PIoU$ ($fast$), $BWIoU$ ($fast$) и $Weighted\ BWIoU$ ($fast$)) выше по сравнению с исходными версиями ($PIoU$ (1), $BWIo$U (5) и $Weighted\ BWIoU$ (6)), однако всё ещё довольно низкая. Обучение сегментационной модели показало, что всё ещё нужно довольно много итераций обучения для достижения качественных результатов.

4.6 Покоординатная пиксельная функция потерь

Одним из главных недостатков описанных выше пиксельных функций является независимость коэффициента масштабирования от координат. Если одна из координат уже находится в оптимальном, с точки зрения пиксельного окружения, положении, то, вероятнее всего, её уже не нужно менять. Чтобы этого избежать, но при этом всё ещё иметь возможность вычислять градиент по координатам, можно добавить к координатной функции потерь следующую функцию:

$$L_{pixel} = k_{x_1} \cdot x_1^p + k_{y_1} \cdot y_1^p + k_{x_2} \cdot x_2^p + k_{y_2} \cdot y_2^p. \quad (12)$$

Несложно убедиться в следующих равенствах:

$$\frac{\partial L_{pixel}}{\partial x_1^p} = k_{x_1}, \quad \frac{\partial L_{pixel}}{\partial y_1^p} = k_{y_1}, \quad \frac{\partial L_{pixel}}{\partial x_2^p} = k_{x_2}, \quad \frac{\partial L_{pixel}}{\partial y_2^p} = k_{y_2}. \quad (13)$$

В качестве значений коэффициентов k_{x_1}, k_{y_1}, k_{x_2} и k_{y_2} можно вычислять отношение количества тёмных пикселей в областях, получающихся при сдвиге координат на Δ пикселей в ту или иную сторону, например, так:

$$x_{1_{left}} = black(x_1 - \Delta, y_1, x_2, y_2), \quad (14) \qquad y_{1_{top}} = black(x_1, y_1 - \Delta, x_2, y_2), \quad (15)$$

$$x_{1_{right}} = black(x_1 + \Delta, y_1, x_2, y_2), \quad (16) \qquad y_{1_{bottom}} = black(x_1, y_1 + \Delta, x_2, y_2), \quad (17)$$

$$x_{2_{left}} = black(x_1, y_1, x_2 - \Delta, y_2), \quad (18) \qquad y_{2_{top}} = black(x_1, y_1, x_2, y_2 - \Delta), \quad (19)$$

$$x_{2_{right}} = black(x_1, y_1, x_2 + \Delta, y_2), \quad (20) \qquad y_{2_{bottom}} = black(x_1, y_1, x_2, y_2 + \Delta), \quad (21)$$

$$k_{x_1} = k \cdot \frac{x_{1_{left}} - x_{1_{right}}}{black(x_1, y_1, x_2, y_2)}, \quad (22) \qquad k_{y_1} = k \cdot \frac{y_{1_{top}} - y_{1_{bottom}}}{black(x_1, y_1, x_2, y_2)}, \quad (23)$$

$$k_{x_2} = k \cdot \frac{x_{2_{left}} - x_{2_{right}}}{black(x_1, y_1, x_2, y_2)}, \quad (24) \qquad k_{y_2} = k \cdot \frac{y_{2_{top}} - y_{2_{bottom}}}{black(x_1, y_1, x_2, y_2)}. \quad (25)$$

В качестве значений Δ берётся небольшое значение – примерно 3-7 пикселей. Коэффициент k позволяет ограничивать влияние пиксельной составляющей. Эксперименты показали, что для документов оптимальнее всего использовать область в 3 пикселя и $k = 0.5$. Схематично данная функция представлена на рис. 8.

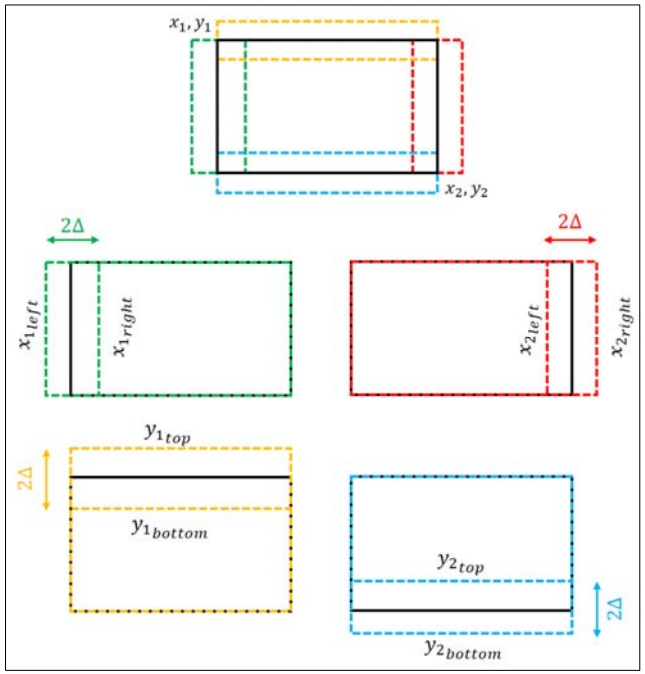


Рис. 8. Схема работы покоординатной пиксельной функции потерь
Fig. 8. Scheme of operation of the coordinate pixel loss function

4.7 Анализ координатных функций потерь

При анализе таких координатных функций потерь, как GIoU [3], DIoU [1], SCA [5] и т.д., было обнаружено, что в подавляющем большинстве случаев в первую очередь выполняется сходимость одного из углов ограничивающего прямоугольника (то есть координаты x_1,y_1 или x_2,y_2) и лишь затем начинает сходиться другой угол. Данный эффект объясняется малыми значениями градиента анализируемых функций потерь и негативно сказывается на сходимости в целом. При этом у функции SCA [5] данный эффект выражен менее сильно, что служит сигналом к более подробному её анализу.

Также большинство функций потерь имеет наибольшее значение градиента при очень близком расположении прямоугольников, из-за чего выбор большей скорости обучения приводит к осцилляциям градиента и не даёт сойтись областям.

Немало проблем дополнительно доставляет ненулевой градиент для уже сошедшихся координат – если прямоугольник совпал по каким-то (но не всем) координатам, градиент анализируемых функций потерь по этим координатам отличен от нуля, причём зачастую довольно сильно, что также приводит к осцилляциям и негативно сказывается на сходимости.

Проведённый анализ приводит к поискам оптимальной координатной функции потерь (рис. 9), отвечающей следующим свойствам:

- максимальная амплитуда градиента при большом расстоянии между областями;
- нелинейно уменьшающаяся амплитуда градиента по мере приближения к целевой области;
- максимально близкий к нулю или равный ему градиент для совпадающих координат;
- максимально похожая форма (с точки зрения размеров).

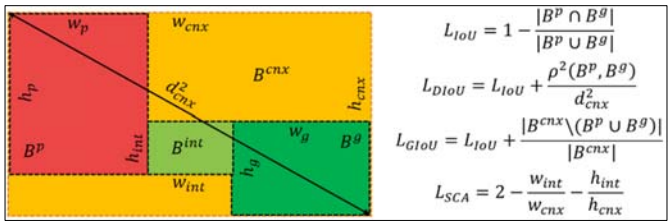


Рис. 9. Используемые координатные функции
Fig. 9. Used coordinate functions

4.8 Усовершенствованные координатные функции потерь

Были разработаны две функции потерь, основанные на SCA. Общим для них является добавление нормализованного квадрата расстояния между центрами, как это сделано в DIoU [1]:

$$L_{center} = \frac{\left(\left(x_c^g - x_c^p\right)^2 + \left(y_c^g - y_c^p\right)^2\right)}{d_{cnx}^2} \quad (26)$$

$\left(x_c^g, y_c^g\right)$, $\left(x_c^p, y_c^p\right)$ – координаты центров целевого и предсказанного прямоугольника, а d_{cnx}^2 – квадрат диагонали наименьшего покрывающего прямоугольника для целевого и предсказанного прямоугольников. Данный коэффициент позволяет быстрее сдвигать все 4 координаты предсказанного прямоугольника к целевому.

4.8.1 Степенная функция потерь отношения размеров

В SCA используется слагаемое $L_{SO} = 2 - SO$, показывающее сумму отношений ширин и высот прямоугольника пересечения и покрывающим прямоугольником. Возведение данного

коэффициента в степень больше единицы значительно увеличивает сходимость получающейся функции. Таким образом получается функция ISCA:

$$L_{ISCA} = L_{SO}^{n} + L_{center}. \quad (27)$$

4.8.2 Функция потерь, оптимизирующая форму

Проблема сходимости в первую очередь лишь одного из углов решается подстраиванием формы предсказанной области к целевой и дальнейшее её перемещение. Для получения такой же формы, как и у целевого прямоугольника, необходимо сравнивать ширины и высоты между собой. Однако необходимо получать значения от 0 до 1 для нормировки, из-за чего деление предсказанных размеров на ожидаемые невозможно, так как получающиеся значения могут быть гораздо больше единицы. Этой проблемы можно избежать, если делить минимальный из размеров на максимальный. Таким образом для сохранения формы можно ввести такой коэффициент:

$$k_{form} = \frac{min(w^g, w^p)}{max(w^g, w^p)} + \frac{min(h^g, h^p)}{max(h^g, h^p)}, \quad (28)$$

w^g, w^p – ширины целевого и предсказанного прямоугольников, а h^g, h^p – высоты соответственно.

Каждое из слагаемых гарантированно не превышает единицы, а потому их сумма не превышает двух. Таким образом для получения функции потерь из него достаточно вычесть его из двух:

$$L_{form} = 2 - k_{form} = 2 - \frac{min(w^g, w^p)}{max(w^g, w^p)} - \frac{min(h^g, h^p)}{max(h^g, h^p)}. \quad (29)$$

Итоговая функция потерь примет следующий вид:

$$L_{FM} = L_{SO} + L_{form} + L_{center}. \quad (30)$$

4.8.3 Анализ сходимости полученных координатных функций

Для полученных функций потерь был проведён описанный выше симуляционный эксперимент (рис. 10), показавший эффективность новых функций

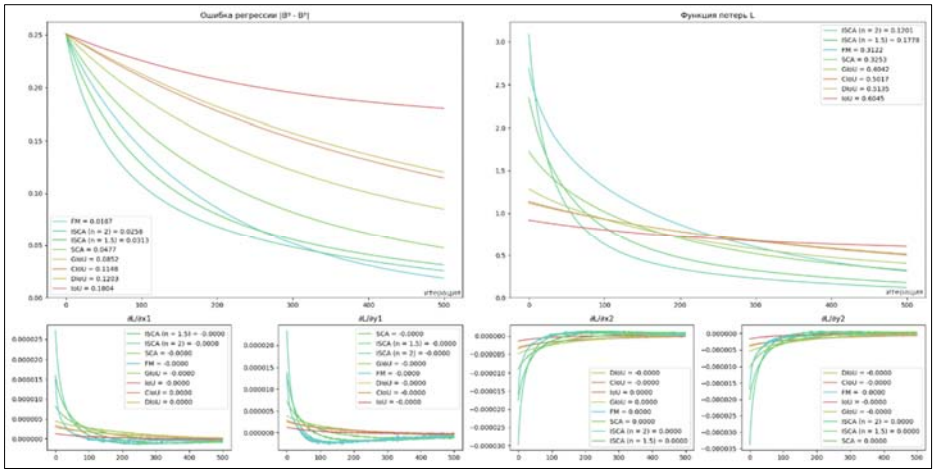


Рис. 10. Анализ сходимости новых координатных функций
Fig. 10. Analysis of the convergence of new coordinate functions

При большом расстоянии между областями градиенты полученных функций ($ISCA$ (формула 27), FM (формула 30)) по-прежнему не максимальны и изменяются нелинейно по мере приближения к целевой области, а также имеют ненулевой градиент при совпадении координат, однако в теории позволяют выполнять регрессию двух ограничивающих прямоугольников заметно быстрее и при этом позволяют сохранять форму (функция L_{FM}) по сравнению с SCA, GIoU [3], CIoU [1], DIoU [1] и IoU.

4.9 Анализ градиентов существующих координатных функций потерь

Для поиска функции потерь, способной удовлетворять описанным требованиям к градиентам, следует проанализировать форму градиентов имеющихся функций потерь. Для этого продифференцируем по x_1^p, y_1^p, x_2^p, y_2^p переменным составляющие следующих функций:

- $L_{IoU} = 1 - IoU$;
- $L_{SCA} = L_{SO} + \alpha \cdot L_{cd}$;
- $L_{FM} = L_{SO} + L_{form} + L_{center}$.

4.9.1 Градиенты L_{IoU}

$$L_{IoU} = 1 - \frac{S_\cap}{S_\cup}; \quad (31)$$

$$\frac{\partial L_{IoU}}{\partial x_1^p} = -\frac{1}{S_\cup}\left(h_p \cdot IoU - \frac{1 - sgn(x_1^g - x_1^p)}{2} \cdot H(w_\cap) \cdot max(h_\cap, 0) \cdot (1 + IoU)\right); \quad (32)$$

$$\frac{\partial L_{IoU}}{\partial y_1^p} = -\frac{1}{S_\cup}\left(w_p \cdot IoU - \frac{1 - sgn(y_1^g - y_1^p)}{2} \cdot H(h_\cap) \cdot max(w_\cap, 0) \cdot (1 + IoU)\right); \quad (33)$$

$$\frac{\partial L_{IoU}}{\partial x_2^p} = \frac{1}{S_\cup}\left(h_p \cdot IoU - \frac{1 + sgn(x_2^g - x_2^p)}{2} \cdot H(w_\cap) \cdot max(h_\cap, 0) \cdot (1 + IoU)\right); \quad (34)$$

$$\frac{\partial L_{IoU}}{\partial y_2^p} = \frac{1}{S_\cup}\left(w_p \cdot IoU - \frac{1 + sgn(y_2^g - y_2^p)}{2} \cdot H(h_\cap) \cdot max(w_\cap, 0) \cdot (1 + IoU)\right). \quad (35)$$

4.9.2 Градиенты L_{SO}

$$L_{SO} = 2 - \frac{w_\cap}{w_{cnx}} - \frac{h_\cap}{h_{cnx}}; \quad (36)$$

$$\frac{\partial L_{SO}}{\partial x_1^p} = \frac{1}{w_{cnx}}\left(\frac{1 + sgn(x_1^g - x_1^p)}{2} \cdot \frac{w_\cap}{w_{cnx}} - \frac{1 - sgn(x_1^g - x_1^p)}{2}\right); \quad (37)$$

$$\frac{\partial L_{SO}}{\partial y_1^p} = \frac{1}{h_{cnx}}\left(\frac{1 + sgn(y_1^g - y_1^p)}{2} \cdot \frac{h_\cap}{h_{cnx}} - \frac{1 - sgn(y_1^g - y_1^p)}{2}\right); \quad (38)$$

$$\frac{\partial L_{SO}}{\partial x_2^p} = \frac{1}{w_{cnx}}\left(\frac{1 - sgn(x_2^g - x_2^p)}{2} \cdot \frac{w_\cap}{w_{cnx}} - \frac{1 + sgn(x_2^g - x_2^p)}{2}\right); \quad (39)$$

$$\frac{\partial L_{SO}}{\partial y_2^p} = \frac{1}{h_{cnx}}\left(\frac{1 - sgn(y_2^g - y_2^p)}{2} \cdot \frac{h_\cap}{h_{cnx}} - \frac{1 + sgn(y_2^g - y_2^p)}{2}\right). \quad (40)$$

4.9.3 Градиенты L_{cd}

$$L_{cd} = \frac{\left(x_1^g - x_1^p\right)^2 + \left(y_1^g - y_1^p\right)^2 + \left(x_2^g - x_2^p\right)^2 + \left(y_2^g - y_2^p\right)^2}{d_{cnx}^2}; \quad (41)$$

$$\frac{\partial L_{cd}}{\partial x_1^p} = -\frac{2}{d_{cnx}^2}\left(x_1^g - x_1^p - \frac{1 + sgn\left(x_1^g - x_1^p\right)}{2} \cdot w_{cnx} \cdot L_{cd}\right); \quad (42)$$

$$\frac{\partial L_{cd}}{\partial y_1^p} = -\frac{2}{d_{cnx}^2}\left(y_1^g - y_1^p - \frac{1 + sgn\left(y_1^g - y_1^p\right)}{2} \cdot h_{cnx} \cdot L_{cd}\right); \quad (43)$$

$$\frac{\partial L_{cd}}{\partial x_2^p} = -\frac{2}{d_{cnx}^2}\left(x_2^g - x_2^p + \frac{1 - sgn\left(x_2^g - x_2^p\right)}{2} \cdot w_{cnx} \cdot L_{cd}\right); \quad (44)$$

$$\frac{\partial L_{cd}}{\partial y_2^p} = -\frac{2}{d_{cnx}^2}\left(y_2^g - y_2^p + \frac{1 - sgn\left(y_2^g - y_2^p\right)}{2} \cdot h_{cnx} \cdot L_{cd}\right). \quad (45)$$

4.9.4. Градиенты L_{form}

$$L_{form} = 2 - \frac{min(w^g, w^p)}{max(w^g, w^p)} + \frac{min(h^g, h^p)}{max(h^g, h^p)}; \quad (46)$$

$$\frac{\partial L_{form}}{\partial x_1^p} = \frac{1}{w_{max}}\left(\frac{1 + sgn\left(w_g - w_p\right)}{2} - \frac{1 - sgn\left(w_g - w_p\right)}{2} \cdot \frac{w_{min}}{w_{max}}\right); \quad (47)$$

$$\frac{\partial L_{form}}{\partial y_1^p} = \frac{1}{h_{max}}\left(\frac{1 + sgn\left(h_g - h_p\right)}{2} - \frac{1 - sgn\left(h_g - h_p\right)}{2} \cdot \frac{h_{min}}{h_{max}}\right); \quad (48)$$

$$\frac{\partial L_{form}}{\partial x_2^p} = -\frac{1}{w_{max}}\left(\frac{1 + sgn\left(w_g - w_p\right)}{2} - \frac{1 - sgn\left(w_g - w_p\right)}{2} \cdot \frac{w_{min}}{w_{max}}\right); \quad (49)$$

$$\frac{\partial L_{form}}{\partial y_2^p} = -\frac{1}{h_{max}}\left(\frac{1 + sgn\left(h_g - h_p\right)}{2} - \frac{1 - sgn\left(h_g - h_p\right)}{2} \cdot \frac{h_{min}}{h_{max}}\right). \quad (50)$$

4.9.5 Выводы

Анализируя градиенты имеющихся функций, можно заметить общую закономерность – градиенты нормируются на некоторое общее число (высоту/ширину/площадь некоторого прямоугольника – наименьшего ограничивающего, наименьшего из двух) и т.д.). При этом в некоторых условиях градиенты зависят только от нормирующего множителя, а не от близости координат.

4.10 Построение координатных функций потерь по градиенту

После проведённого анализа был выполнен поиск функций градиента, которые бы также имели нормирующий множитель, но при этом градиенты всегда зависели от координат и приводили к более высокой сходимости по сравнению с имеющимися. Одной из таких функций стала следующая:

$$\frac{\partial L_{grad}}{\partial x_1^p} = -\frac{sgn\left(x_1^g - x_1^p\right) + x_1^g - x_1^p}{w_g}; \quad (51)$$

$$\frac{\partial L_{grad}}{\partial y_1^p} = -\frac{sgn\left(y_1^g - y_1^p\right) + y_1^g - y_1^p}{h_g}; \quad (52)$$

$$\frac{\partial L_{grad}}{\partial x_2^p} = -\frac{sgn\left(x_2^g - x_2^p\right) + x_2^g - x_2^p}{w_g}; \quad (53)$$

$$\frac{\partial L_{grad}}{\partial y_2^p} = -\frac{sgn\left(y_2^g - y_2^p\right) + y_2^g - y_2^p}{h_g}. \quad (54)$$

4.10.1 Получение функции L_{grad}

Для восстановления функции L_{grad} достаточно проинтегрировать каждую из четырёх функций и сложить результат, так как каждое из слагаемых не зависит от переменных других слагаемых. При этом в качестве константы стоит выбрать ту, при которой значение функции при идеальном совпадении предсказанных и целевых координат будет равно нулю.

$$L_{grad} = L_{grad_{x_1}} + L_{grad_{y_1}} + L_{grad_{x_2}} + L_{grad_{y_2}}; \quad (55)$$

$$L_{grad_{x_1}} = -\int \frac{sgn\left(x_1^g - x_1^p\right) + x_1^g - x_1^p}{w_g}\, dx_1^p = \frac{\left|x_1^g - x_1^p\right| + 0.5\left(x_1^g - x_1^p\right)^2}{w_g}; \quad (56)$$

$$L_{grad_{y_1}} = -\int \frac{sgn\left(y_1^g - y_1^p\right) + y_1^g - y_1^p}{h_g}\, dy_1^p = \frac{\left|y_1^g - y_1^p\right| + 0.5\left(y_1^g - y_1^p\right)^2}{h_g}; \quad (57)$$

$$L_{grad_{x_2}} = -\int \frac{sgn\left(x_2^g - x_2^p\right) + x_2^g - x_2^p}{w_g}\, dx_2^p = \frac{\left|x_2^g - x_2^p\right| + 0.5\left(x_2^g - x_2^p\right)^2}{w_g}; \quad (58)$$

$$L_{grad_{y_2}} = -\int \frac{sgn\left(y_2^g - y_2^p\right) + y_2^g - y_2^p}{h_g}\, dy_2^p = \frac{\left|y_2^g - y_2^p\right| + 0.5\left(y_2^g - y_2^p\right)^2}{h_g}. \quad (59)$$

Достоинства функции градиента:

- градиенты нормализованы относительно наименьшего прямоугольника;
- каждая координата сходится независимо;
- нулевой градиент при совпадении координат;
- с отдалением от целевой координаты градиент увеличивается в обе стороны.

Недостатки:

- имеет неустранимый разрыв в окрестности равенства координат

4.10.2 Устранение проблемы разрывности функции градиента

Выражение $f(x) = sgn(x) + x$ имеет разрыв в точке 0. Решение – заменить непрерывной функцией, например, $f(x) = tanh(kx) + x$ или $f(x) = \frac{2}{\pi} \cdot atan(kx) + x$ (рис. 11):

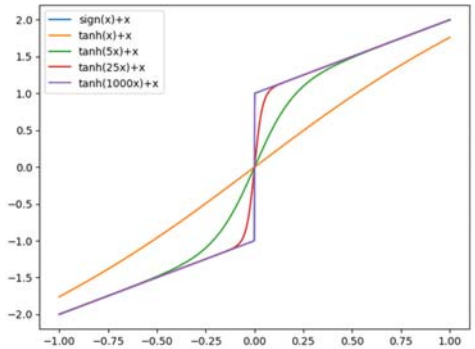


Рис. 11. Сравнение функций
Fig. 11. Functions comparison

Замена разрывной функции $sgn(x)$ на $tanh(kx) + x$ или $\frac{2}{\pi} \cdot atan(kx) + x$ даёт две новые функции градиентов (Δv обозначает $v^g - v^p$):

$$\frac{\partial L_{tanh}}{\partial x_1^p} = -\frac{tanh(k \cdot \Delta x_1) + \Delta x_1}{w_g}; \quad (60) \qquad \frac{\partial L_{atan}}{\partial x_1^p} = -\frac{\left(\frac{2}{\pi} \cdot tanh(k \cdot \Delta x_1) + \Delta x_1\right)}{w_g}; \quad (61)$$

$$\frac{\partial L_{tanh}}{\partial y_1^p} = -\frac{tanh(k \cdot \Delta y_1) + \Delta y_1}{h_g}; \quad (62) \qquad \frac{\partial L_{atan}}{\partial y_1^p} = -\frac{\left(\frac{2}{\pi} \cdot tanh(k \cdot \Delta y_1) + \Delta y_1\right)}{h_g}; \quad (63)$$

$$\frac{\partial L_{tanh}}{\partial x_2^p} = -\frac{tanh(k \cdot \Delta x_2) + \Delta x_2}{w_g}; \quad (64) \qquad \frac{\partial L_{atan}}{\partial x_2^p} = -\frac{\left(\frac{2}{\pi} \cdot tanh(k \cdot \Delta x_2) + \Delta x_2\right)}{w_g}; \quad (65)$$

$$\frac{\partial L_{tanh}}{\partial y_2^p} = -\frac{tanh(k \cdot \Delta y_2) + \Delta y_2}{h_g}; \quad (66) \qquad \frac{\partial L_{atan}}{\partial y_2^p} = -\frac{\left(\frac{2}{\pi} \cdot tanh(k \cdot \Delta y_2) + \Delta y_2\right)}{h_g}. \quad (67)$$

Интегрирование данных функций градиентов даёт следующие функции:

$$L_{tanh} = L_{tanh_{x_1}} + L_{tanh_{y_1}} + L_{tanh_{x_2}} + L_{tanh_{y_2}}; \quad (68)$$

$$L_{tanh_{x_1}} = \frac{1}{w_g}\left(\frac{\Delta x_1^{\,2}}{2} + \frac{ln\left(cosh(k \cdot \Delta x_1)\right)}{k}\right); \quad (69)$$

$$L_{tanh_{y_1}} = \frac{1}{h_g}\left(\frac{\Delta y_1^{\,2}}{2} + \frac{ln\left(cosh(k \cdot \Delta y_1)\right)}{k}\right); \quad (70)$$

$$L_{tanh_{x_2}} = \frac{1}{w_g}\left(\frac{\Delta x_2^{\,2}}{2} + \frac{ln\left(cosh(k \cdot \Delta x_2)\right)}{k}\right); \quad (71)$$

$$L_{tanh_{y_2}} = \frac{1}{h_g}\left(\frac{\Delta y_2^{\,2}}{2} + \frac{ln\left(cosh(k \cdot \Delta y_2)\right)}{k}\right); \quad (72)$$

$$L_{atan} = L_{atan_{x_1}} + L_{atan_{y_1}} + L_{atan_{x_2}} + L_{atan_{y_2}}; \quad (73)$$

$$L_{atan_{x_1}} = \frac{1}{w_g}\left(\frac{\Delta x_1^{\,2}}{2} + \frac{2}{\pi}\left(\Delta x_1 \cdot atan(k \cdot \Delta x_1) - \frac{ln\left(1 + k^2 \cdot \Delta x_1^{\,2}\right)}{2k}\right)\right); \quad (74)$$

$$L_{atan_{y_1}} = \frac{1}{h_g}\left(\frac{\Delta y_1^{\,2}}{2} + \frac{2}{\pi}\left(\Delta y_1 \cdot atan(k \cdot \Delta y_1) - \frac{ln\left(1 + k^2 \cdot \Delta y_1^{\,2}\right)}{2k}\right)\right); \quad (75)$$

$$L_{atan_{x_2}} = \frac{1}{w_g}\left(\frac{\Delta x_2^{\,2}}{2} + \frac{2}{\pi}\left(\Delta x_2 \cdot atan(k \cdot \Delta x_2) - \frac{ln\left(1 + k^2 \cdot \Delta x_2^{\,2}\right)}{2k}\right)\right); \quad (76)$$

$$L_{atan_{y_2}} = \frac{1}{h_g}\left(\frac{\Delta y_2^{\,2}}{2} + \frac{2}{\pi}\left(\Delta y_2 \cdot atan(k \cdot \Delta y_2) - \frac{ln\left(1 + k^2 \cdot \Delta y_2^{\,2}\right)}{2k}\right)\right). \quad (77)$$

4.10.3 Сходимость полученных функций

Результаты эксперимента-симуляции представлены на рис. 12. Как и ожидалось, градиенты полученных функций (L_{grad} (формула 52), L_{tanh} (формула 65) и L_{atan} (формула 70)) схожи как по форме, так и по значению, однако сглаженные функции (L_{tanh} (формула 65) и L_{atan} (формула 70)) сходятся быстрее L_{grad} (формула 52) и тем более L_{SCA} [5].

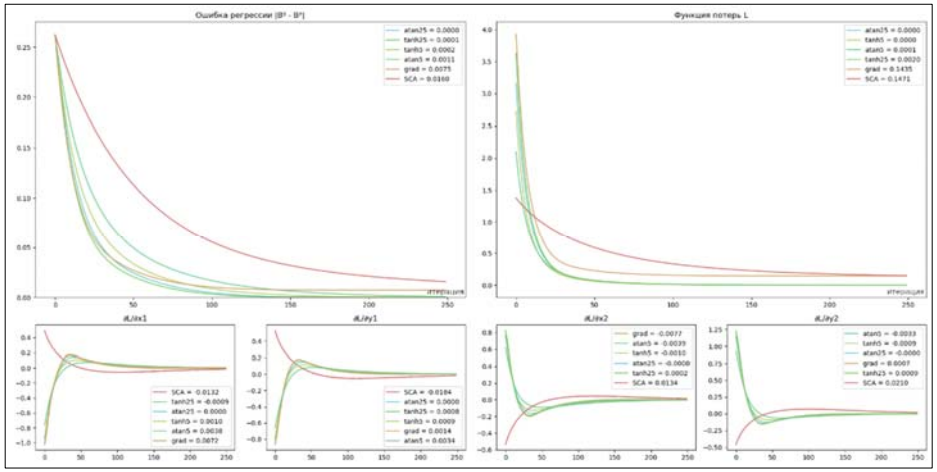


Рис. 12. Результаты симуляционного эксперимента градиентных функций
Fig. 12. Results of the simulation experiment of gradient functions

5. Выбор набора данных для обучения

Для экспериментов был выбран набор данных PubLayNet [8], представляющий собой коллекцию документов медицинских научных статей. Он содержит 335 тысяч обучающих изображений и 11.5 тысяч проверочных изображений. Поскольку обучение на таком большом наборе требует много времени, а экспериментов с функциями потерь проводить нужно большое количество раз, то для экспериментов была выбрана небольшая часть – 10 тысяч обучающих изображений и 1000 проверочных.

Помимо него также проводились эксперименты на наборе сгенерированных изображений документов различных договоров и технических заданий [9], содержащем 20 тысяч обучающих и 769 проверочных изображений.

6. Оценка качества

Для более полной оценки качества сегментации, в том числе учитывающей особенности сегментации изображений документов, были взяты три типа метрик:

- Стандартные метрики Precision, Recall, mAP (mean average precision) с порогом 0.5, mAP с интервалом порога 0.5...0.95, основанные на IoU функции;
- Перечисленные выше метрики с заменой IoU функции на пиксельную BWIoU (5) функцию;
- OCR метрика (9).

7. Результаты экспериментов

В табл.2 и 3 приведены результаты тестирования выбранной модели сегментации на наборе данных PubLayNet [8] с помощью координатной и пиксельной метрик соответственно. Наилучшие результаты по метрике mAP 0.5 и mAp 0.5…0.95 показывают функции Pixel-delta (12) и PIoU (1).

Табл. 2. Результаты экспериментов на наборе PubLayNet (координатная метрика)
Table 2. Results of experiments on the PubLayNet dataset (coordinate metric)

Функция потерь	Precision	Recall	mAP 0.5	mAP 0.5…0.95
Pixel-delta	0.931	0.906	**0.934**	**0.817**
IoU	0.928	0.901	0.923	0.801
SCA	0.94	0.895	0.926	0.803
PIoU	**0.932**	**0.909**	**0.934**	**0.817**
BWIoU	0.936	0.894	0.929	0.804
Weigted BWIoU	0.939	0.9	0.928	0.812
FM	0.953	0.899	0.932	0.813
atan	0.947	0.884	0.925	0.799
tanh	0.953	0.886	0.926	0.799

Табл. 3. Результаты экспериментов на наборе PubLayNet (пиксельная метрика)
Table 3. Results of experiments on the PubLayNet dataset (pixel metric)

Функция потерь	Precision	Recall	mAP 0.5	mAP 0.5…0.95
Pixel-delta	0.936	0.898	**0.924**	**0.742**
IoU	0.927	0.895	0.916	0.721
SCA	0.931	0.888	0.919	0.723
PIoU	0.925	**0.902**	**0.924**	0.727
BWIoU	0.931	0.884	0.913	0.716
Weigted BWIoU	0.93	0.891	0.915	0.711
FM	**0.947**	0.893	0.922	0.733
atan	0.938	0.876	0.913	0.714
tanh	0.941	0.879	0.913	0.714

В табл. 4 и 5 приведены результаты тестирования выбранной модели сегментации на наборе синтетических документов с помощью координатной и пиксельной метрик соответственно. Функция потерь Pixel-delta показывает наилучшее значения в метрике mAp 0.5…0.95, а по метрике mAp 0.5 наилучший результат показывает чисто пиксельная функция PIoU.

Табл. 4. Результаты экспериментов на наборе Generated (координатная метрика)
Table 4. Results of experiments on the Generated dataset (coordinate metric)

Функция потерь	Precision	Recall	mAP 0.5	mAP 0.5…0.95
Pixel-delta	**0.992**	**0.988**	0.994	**0.874**
IoU	**0.992**	0.983	0.986	0.858
SCA	0.987	0.981	0.987	0.861
PIoU	**0.992**	**0.988**	**0.994**	0.862
BWIoU	0.99	0.982	0.993	0.861
Weigted BWIoU	0.99	0.982	0.991	0.858
FM	0.984	0.972	0.991	0.81
atan	0.989	0.969	0.986	0.856
tanh	0.989	0.975	0.989	0.844

Табл. 5. Результаты экспериментов на наборе Generated (пиксельная метрика)
Table 5. Results of experiments on the Generated dataset (pixel metric)

Функция потерь	Precision	Recall	mAP 0.5	mAP 0.5…0.95
Pixel-delta	**0.989**	**0.982**	**0.989**	**0.736**

IoU	0.982	0.973	0.98	0.708
SCA	0.978	0.972	0.98	0.67
PIoU	**0.989**	0.977	0.983	0.697
BWIoU	0.983	0.975	0.987	0.72
Weigted BWIoU	0.98	0.973	0.981	0.733
FM	0.981	0.955	0.977	0.586
atan	0.98	0.96	0.979	0.718
tanh	0.978	0.964	0.978	0.689

В табл. 6 приведены результаты тестирования модели сегментации на наборах PubLayNet и Generated с помощью OCR метрики. Функция Pixel-delta для всех наборов данных показывает наилучшее качество, что говорит о наиболее точном выделении целевых областей.

Табл. 6. Результаты экспериментов (метрика OCR)
Table 6. Results of experiments (OCR metric)

Функция потерь	OCR metric (PubLayNet)	OCR metric (Generated)
Pixel-delta	**0.903**	**0.914**
IoU	0.891	0.903
SCA	0.892	0.904
PIoU	0.899	0.912
BWIoU	0.898	0.909
Weigted BWIoU	0.89	0.91
FM	0.891	0.902
atan	0.893	0.903
tanh	0.895	0.901

Использование функций потерь, учитывающих особенности изображений документов (PIoU, BWIoU, Weighted BWIoU), демонстрируют более высокое качество сегментации по сравнению со стандартными функциями. Разработанные координатные функции потерь (FM, tanh, atan) показывают качество на уровне существующих функций, но имеют более высокую скорость сходимости. Функции, комбинирующие координатный и пиксельный подходы (Pixel-delta), имеют более высокое качество сегментации и скорость сходимости по сравнению с функциями с одним подходом. Разработанные функции потерь, учитывающие особенности изображений документов, улучшают качество сегментации на 3-5%. Разработанные координатные функции потерь повышают скорость сходимости.

7. Выводы

В работе была выбрана модель сегментации и разработаны функции потерь, как учитывающие особенности изображений документов, так и использующие классический чисто координатный подход.

Исходя из полученных результатов, можно утверждать, что использование функций потерь, использующих особенности сегментации изображений документов позволяет достичь высокой точности сегментации, а использование координатных функций позволяет быстрее обучать модель. Комбинируя подходы, можно получать функции потерь, сочетающие в себе оба свойства.

Список литературы / References

[1] Zheng Z., Wang P. et al. Distance-IoU loss: Faster and better learning for bounding box regression. Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, no. 07, 2020, pp. 12993-13000.

[2] Rezatofighi H., Tsoi N. et al. Generalized intersection over union: A metric and a loss for bounding box regression. In Proc. of the IEEE/CVF conference on computer vision and pattern recognition, 2019, pp. 658-666.
[3] Zheng T., Zhao S. et al. SCALoss: Side and Corner Aligned Loss for Bounding Box Regression. arXiv preprint arXiv:2104.00462, 2021, 9 p.
[4] He J., Erfani S. et al. α-IoU: A Family of Power Intersection over Union Losses for Bounding Box Regression. Advances in Neural Information Processing Systems, vol. 34, 2021, 13 p.
[5] Wu S., Yang J. et al. Iou-balanced loss functions for single-stage object detection. Pattern Recognition Letters, vol. 156, 2022, pp. 96-103.
[6] Du S., Zhang B., Zhang P. Scale-Sensitive IOU Loss: An Improved Regression Loss Function in Remote Sensing Object Detection. IEEE Access, vol. 9, 2021, pp. 141258-141272.
[7] Redmon J., Farhadi A. Yolov3: An incremental improvement. arXiv preprint arXiv:1804.02767, 2018, 6 p.
[8] Zhong X., Tang J., Yepes A.J. Publaynet: largest dataset ever for document layout analysis. In Proc. of the 2019 International Conference on Document Analysis and Recognition (ICDAR), 2019, pp. 1015-1022.
[9] Беляева О.В., Перминов А.И., Козлов И.С. Использование синтетических данных для тонкой настройки моделей сегментации документов. Труды ИСП РАН, том 32, вып. 4, 2020 г., стр. 189-202. DOI: 10.15514/ISPRAS–2020–32(4)–14 / Belyaeva O.V., Perminov A.I., Kozlov I.S. Synthetic data usage for document segmentation models fine-tuning. Trudy ISP RAN/Proc. ISP RAS, vol. 32, issue 4, 2020. pp. 189-202 (in Russian).

Информация об авторах / Information about authors

Андрей Игоревич ПЕРМИНОВ является студентом магистратуры кафедры системного программирования. Научные интересы: цифровая обработка сигналов, нейросетевая обработка данных, создание искусственных данных, цифровая обработка изображений.

Andrey Igorevich PERMINOV – master's student of the Department of System Programming. Research interests: digital signal processing, neural network data processing, generation of artificial data, digital image processing.

Денис Юрьевич ТУРДАКОВ – кандидат физико-математических наук, заведующий отделом ИСП РАН, доцент кафедры системного программирования ф-та ВМК МГУ. Научные интересы: анализ естественного языка, извлечение информации, обработка больших данных, анализ социальных сетей.

Denis Yurievich TURDAKOV – PhD, Head of Department at ISP RAS, associate professor of the Department of System Programming at MSU. Research interests: natural language processing, information extraction, big data analysis, social network analysis.

Оксана Владимировна БЕЛЯЕВА – аспирант. Научные интересы: распознавание структуры документов, цифровая обработка изображений, нейросетевая обработка данных, распознавание образов.

Oksana Vladimirovna BELYAEVA – PhD Student. Research interests: document layout analysis, digital image processing, neural network data processing, image pattern recognition.