Градиент как основа для построения функции потерь

И.А. Килбас[1], Р.А. Парингер[1,2]

[1]Самарский национальный исследовательский университет им. академика С.П. Королева, Московское шоссе 34А, Самара, Россия, 443086
[2]Институт систем обработки изображений РАН - филиал ФНИЦ «Кристаллография и фотоника» РАН, Молодогвардейская 151, Самара, Россия, 443001

Аннотация. Глубокие нейронные сети достигли большого успеха в различных областях, но их обучение всё еще представляет трудность. Точность нейронной сети зависит не только от её архитектуры, но также от выбора функции потерь. Перекрестная энтропия является наиболее частым выбором для решения задач классификации. Данная функция потерь имеет недостаток – она плохо работает с несбалансированными данными. Для того, чтобы решить данную проблему, была предложена функция потерь Focal Loss. В данной работе мы исследуем причины хорошей работы данной функции потерь в условиях несбалансированных данных. Мы находим ряд свойств градиента Focal Loss, которые могут быть использованы для построения новых функций потерь, а также предлагаем некоторые из них. Состоятельность предложенных функций потерь показана экспериментально.

1. Введение

С введением свёрточных нейронных сетей глубокое обучение стало основным инструментом для решения многих задач в сфере компьютерного зрения. В частности, свёрточные нейронные сети стали главным инстументом в задаче детектирования объектов. Детектирование объектов на изображении состоит из двух подзадач: локализации и классификации. Первая заключается в регресии координат прямоугольной рамки, в которую обводится объект, а вторая – в ответе на вопрос, к какому классу относится объект в данной рамке.

К последним достижениям в области детектирования объектов относятся двуступенчатые и одноступенчатые детекторы объектов, основанные на нейронных сетях. К первым относятся детекторы, использующие Region Proposal подход, такие как R-CNN [1], Fast R-CNN [2] и Faster R-CNN [3]. Данный подход подразумевает два этапа детектирования: выявление областей интереса и последующую регрессию координат рамок с их классификацией уже предобученной свёрточной нейронной сетью. Данные детекторы обладают высокой точностью, но из-за специфики подхода имеют низкую скорость детектирования. Одноступенчатые детекторы в свою очередь имеют высокую скорость детектировании, но имеют более низкую точность по сравнению с Region Proposal подходами. В данной работе будут рассматриваться именно одноступенчатые детекторы объектов.

К одноступенчатым детекторам объектов относятся такие, как YOLO [4] и SSD [5]. В отличие от Region Proposal подхода, пространство входного изображения изначально дискретизировано множеством областей интереса (рамок) различного размера, называемыми «якорями» (anchors). Таким образом, одноступенчатые детекторы выполняют классификацию

«якорей» в один проход вместо итеративной классификации выделенных областей отдельной нейронной сетью. Результат классификации в дальнейшем фильтруется с помощью алгоритма Non-Maximum Supression.

В силу плотного покрытия пространства картинки «якорями», большинство из них принадлежат к классу фона. Превалирование «фоновых якорей» представляет собой проблему, поскольку во время обучения детектора класс фона преобладает над другими. Обучающие данные для детекторов представляют собой массивы меток для каждого «якоря». Поскольку абсолютное большинство меток занимают «фоновые якоря», обучающие данные являются крайне несбалансированными. В результате детектор склонен классифицировать все объекты как фон, либо просто не обучается.

Для решения вышеозвученной проблемы было предложено во время обучения искусственно ограничивать количество «фоновых якорей», результаты классификации которых включаются в функцию потерь. Данный подход называется Online Hard Example Mining (OHEM) [6]. В терминах данного подхода «фоновые якоря» называются лёгкими примерами, а «якоря», содержащие объекты, тяжелыми примерами. Суть подхода состоит в том, чтобы искусственно контролировать отношение количества тяжелых примеров к количеству лёгких примеров. Для обучения SSD авторы использовали соотношение 1 к 3. Для сравнения, без использования OHEM данное соотношение может быть 1 к 100 или даже 1 к 1000, что является сильным дисбалансом в сторону лёгких примеров.

В качестве иного подхода для решения данной проблемы было предложено использовать адаптивную функцию потерь. Авторы [7] предложили новую функцию потерь для задачи классификации и назвали её Focal Loss. Данная функция представляет собой модификацию перекрестной энтропии. Помимо логарифма в формуле также появился множитель, который дополнительно уменьшает значение функции при стремлении аргумента к 1.

$$Crossentropy\ Loss = -\log(p_t)$$
$$Focal\ Loss = -(1-p_t)^{\gamma}\log(p_t)$$

где p_t – вероятность принадлежности объекта классу t; γ – гиперпараметр.

С введением данного множителя функция потерь может принимать значения близкие к нулю при p=0.6, что уже является верным ответом классификации. В силу данного обстоятельства во время обучения суммарное значение функции потерь по лёгким примерам становится меньше ошибки по немногочисленным тяжелым примерам. Из-за этого детектор во время обучения начинает фокусироваться на тяжелых примерах, что сильно повышает итоговую точность.

В данной работе мы попытаемся ответить на вопрос, какие свойства Focal Loss позволяют ей обучать детекторы объектов в условиях сильного дисбаланса и возможно ли построить собственную функцию потерь, удовлетворяющую данным свойствам и также позволяющую обучать детекторы объектов в условиях дисбаланса.

2. Целевая функция потерь для обучения детекторов объектов

Минимизируемая целевая функция потерь во время обучения детектора объектов представляет собой сумму функции потерь классификации и функции потерь локализации. Функция для локализации количественно измеряет ошибку регрессии координат «якорей», функция для классификации количественно измеряет ошибку классификации данных «якорей».

$$Loss = CL + \lambda LL$$

где CL – функция потерь классификации; LL – функция потерь локализации; λ – множитель функции потерь локализации, обычно равен 1.

Функция потерь классификации обычно представляет собой сумму значений перекрестной энтропии результатов классификации «якорей». Функция потерь локализации аналогично представляет собой сумму значений потерь локализации, функции Хьюбера или Smooth L1 Loss, для результатов регрессии «якорей».

Данная функция минимизируется посредством алгоритма градиентного спуска [8]. Данный алгоритм является итеративным и может быть представлен следующей формулой:

$$W_{t+1} = W_t - \theta \frac{dLoss}{dW_t}$$

где W_t – значения весов нейронной сети (детектора объектов) на шаге t; θ – коэффициент, определяющий размер вектора изменения весов $\frac{dLoss}{dW_t}$; $\frac{dLoss}{dW_t}$ – вектор градиента функции потерь по весам нейронной сети.

2.1. Градиент перекрестной энтропии, Focal Loss и коэффициент масштаба

То, как изменяется градиент целевой функции ошибки, играет ключевую роль в процессе обучения детектора объектов. В общем случае градиент функции потерь классификации для одного примера представим в следующем виде:

$$\frac{dCL}{dW} = \frac{dCL}{dp}\frac{dp}{dW}$$

где CL – функция потерь классификации; p – значение вероятности принадлежности примера тому или иному классу; W – веса нейронной сети.

Из формулы градиента видно, что его можно разбить на два множителя. Множитель $\frac{dp}{dW}$ не зависит от вида функции потерь классификации, поэтому в дальнешем учитываться не будет. Следовательно то, как изменяется множитель $\frac{dCL}{dp}$, является ключевым фактором, оказывающим влияние на процесс обучения детектора объектов. Данный множитель будем в дальнейшем называть коэффициентом масштаба.

Коэффициент масштаба перекрестной энтропии имеет вид:

$$\frac{dCE}{dp} = -\frac{1}{p}$$

Коэффициент масштаба Focal Loss имеет вид:

$$\frac{dFL}{dp} = \gamma(1-p)^{\gamma-1}\log(p) - \frac{(1-p)^{\gamma}}{p}$$

Из вида коэффициента масштаба Focal Loss можно выделить его следующие свойства:

- монотонно увеличивается при $p \to 1$;
- стремится к нулю при $p \to 1$.

Коэффициент масштаба перекрестной энтропии не удовлетворяет второму свойству. Именно данное свойство позволяет Focal Loss «сконцентрировать внимание» на тяжелых примерах в процессе обучения детектора, когда точность на лёгких примерах уже достигла некоторого порога. Для иллюстрации данного феномена смоделируем следующую ситуацию. Пусть дано 1000 лёгких примеров, для которых p=0.9, и 10 тяжелых примеров, для которых p=0.3. Таким образом, суммарное значение коэффициента масштаба для перекрёстной энтропии равно

$$1000\frac{dCE}{dp}|_{p=0.9} = -1111$$

$$10\frac{dCE}{dp}|_{p=0.3} = -33$$

Суммарное значение коэффициента масштаба для Focal Loss равно

$$1000\frac{dFL}{dp}|_{p=0.9}^{\gamma=2.0} = -32$$

$$10\frac{dFL}{dp}|_{p=0.3}^{\gamma=2.0} = -33$$

Как видно из данного примера, при обучении детектора объектов используя перекрестную энтропию градиент по лёгким примерам «затмевает» градиент по тяжелым примерам, из-за чего детектор учится классифицировать все «якори» как фон. В то же время Focal Loss гасит градиент по лёгким примерам, что позволяет детектору фокусироваться на тяжелых примерах во время обучения.

3. Модификация коэффициента масштаба

Коэффициент масштаба перекрестной энтропии стремится к -1 при $p \to 1$, что не удовлетворяет второму свойству коэффициента масштаба Focal Loss. Простой модификацией, которая позволит исправить это, будет прибавление единицы. После данной модификации коэффициент масштаба перекрестной энтропии будет иметь вид:

$$\frac{dMCE}{dp} = -\frac{1}{p} + 1 = -\frac{1-p}{p}$$

3.1. Maki Loss

Подобно виду Focal Loss можно сделать обобщение для данной модификации:

$$\frac{dMCE}{dp} = -\frac{(1-p)^{\gamma}}{p}$$

где γ – гиперпараметр.

Недостатком данного обобщения является то, что выражение справа не интегрируется в квадратурах для произвольного γ. Несмотря на это, можно проинтегрировать данное выражение для $\gamma \in \aleph_0$.

$$Maki\ Loss = -(\log(p) + \sum_{k=1}^{\gamma} \frac{C_{\gamma}^{k}(-p)^{k}}{k} - \sum_{k=1}^{\gamma} \frac{C_{\gamma}^{k}(-1)^{k}}{k})$$

где C_{γ}^{k} – число сочетаний из γ по k.

Вторая сумма в формуле, где p заменено на 1, необходима для того, чтобы $Maki\ Loss = 0$ при $p = 1$. Название Maki Loss входновлено названием фреймворка MakiFlow [9], в котором данная функция потерь была реализована.

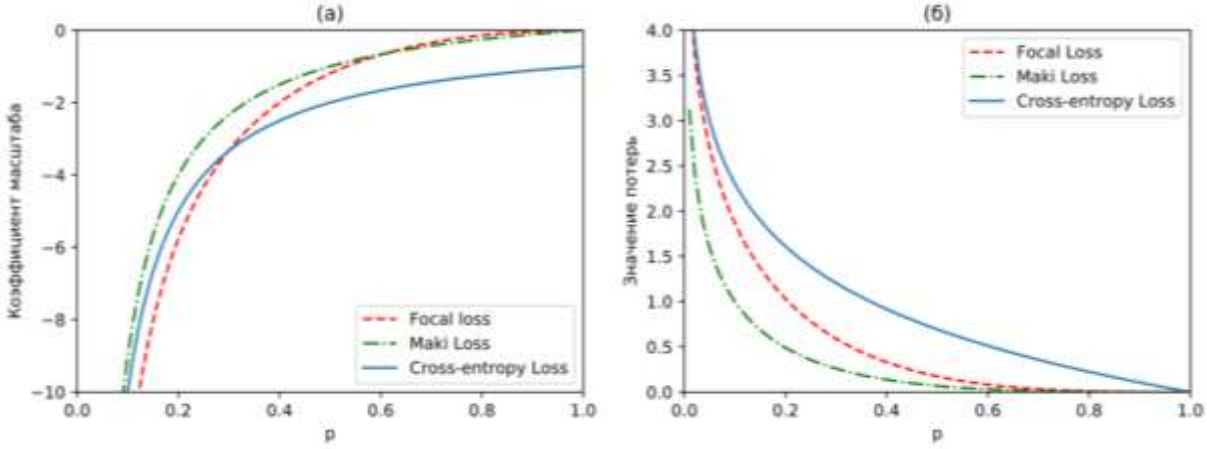

Рисунок 1. Графики значения коэффициента масштаба (а) для перекрестной энтропии, Focal Loss и Maki Loss, а также непосредственно самих значений функций потерь (б). Значения для Focal Loss вычислялись при γ=2.0. Значения для Maki Loss вычислялись при $\gamma = 1$.

3.2. Quadratic Cross-Entropy Loss

Продолжая тему модификации перекрестной энтропии можно предлагать произвольные функции, которые будут удовлетворять свойствам коэффициента масштаба Focal Loss. Примером такой функции может выступить натуральный логарифм. Умножив коэффициент масштаба перекрестной энтропии на натуральный логарифм получим:

$$\frac{dQCE}{dp} = \frac{\log(p)}{p}$$

Минус в данной формуле отсутствует, поскольку значения логарифма уже являются отрицательными. Проинтегрировав данное выражение получим формулу для новой функции потерь:

$$QCE = \frac{\log^2(p)}{2}$$

Данная функция названа квадратичной перекрестной энтропией (Quardratic Cross-Entropy или QCE), поскольку в сущности она представляет собой возведённую в квадрат перекрестную энтропию. Данная функция отличается от всех ранее рассмотренных крайней простотой реализации.

3.3. Детали реализации Maki Loss и QCE Loss

Нормализация значений Maki Loss происходит подобно Focal Loss и OHEM – значения функции потерь делятся на количество тяжелых примеров в обучающем батче. QCE Loss, в свою очередь, нормализуется общим числом примеров в батче (количество «якорей», умноженное на размер батча). Было решено нормализовать данную функцию потерь именно таким образом, чтобы сделать её наиболее простой как аналитическом смысле, так и в отношении реализации в коде.

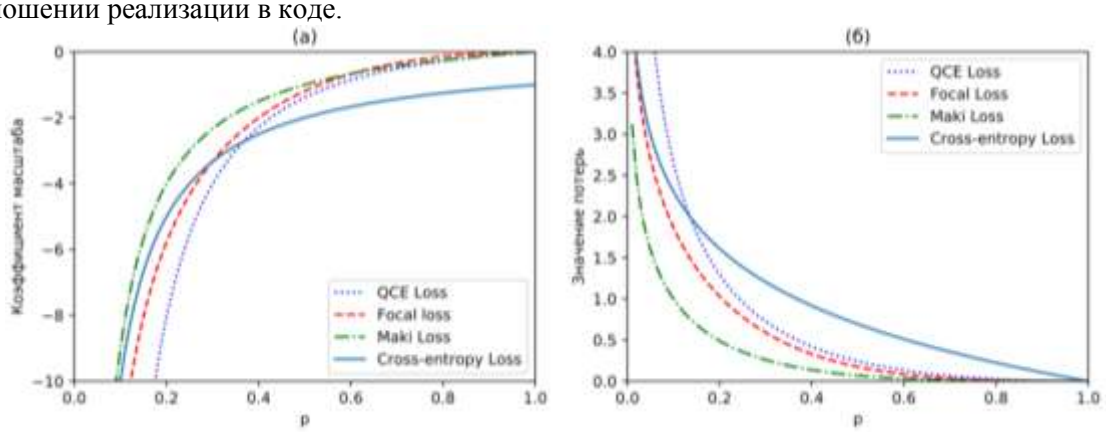

Рисунок 2. Графики значения коэффициента масштаба (а) для перекрестной энтропии, Focal Loss, Maki Loss и QCE Loss, а также непосредственно самих значений функций потерь (б). Значения для Focal Loss вычислялись при γ=2.0. Значения для Maki Loss вычислялись при γ=1.

4. Постановка эксперимента

Для проверки предложенных функций потерь на каждой из них был обучен детектор объектов на основе предобученной свёрточной нейронной сети MobileNetV2 [10]. Детектор объектов был реализован и обучен средствами фреймворка MakiFlow. В качестве обучающих данных использовались датасеты Pascal2012 и Pascal2007. Точность обученных детекторов объектов измерялась метрикой mAP.

4.1. Гиперпараметры обучения

Гиперпараметры для каждой из функций потерь (θ , λ , γ) подбирались индивидуально. Детекторы объектов обучались в течение 37000 итераций. Размер батча равен 128. Для обучения использовался оптимизатор Адам [11].

4.2. Подготовка данных для обучения и тестирования

Все изображения нормализовывались делением на 255. Для увеличения количества обучающих данных использовалась аугментация: добавление гауссового шума, горизонтальный и вертикальный повороты, изменение контраста и яркости. Для тестирования обученных детекторов объектов из PascalVOC2012 было взято 2125 изображений, из PascalVOC2007 – 700 изображений.

4.3. Подробности гиперпараметров обучения

В таблице ниже представлены гиперпараметры, подобранные для каждой из функций потерь. Параметры подбирались посредством многочисленных тестов и наблюдений за изменением точности и значений функций потерь в ходе обучения детекторов объектов.

Таблица 1. Значения гиперпараметров обучения для каждой функции потерь.

Название гиперпараметра	Перекрестная энтропия с OHEM	Focal Loss	Maki Loss	QCE Loss
Learning rate θ	1e-3	4e-4	1e-3	1e-3
Localization loss term λ	1.0	1.0	0.2	0.1
Отношение количества тяжелых примеров к лёгким	1 : 1	-	-	-
Гамма γ	-	2.0	1	-

5. Результаты эксперимента

На рисунке 4 представлены графики точности детекторов объектов по ходу обучения. Результаты обучения не являются высокими, поэтому вопрос о подборе оптимальных гиперпараметров остаётся открытым. Несмотря на это, данные графики показывают, что предложенные функции потерь имеют место быть и могут быть использованы для обучения детекторов объектов.

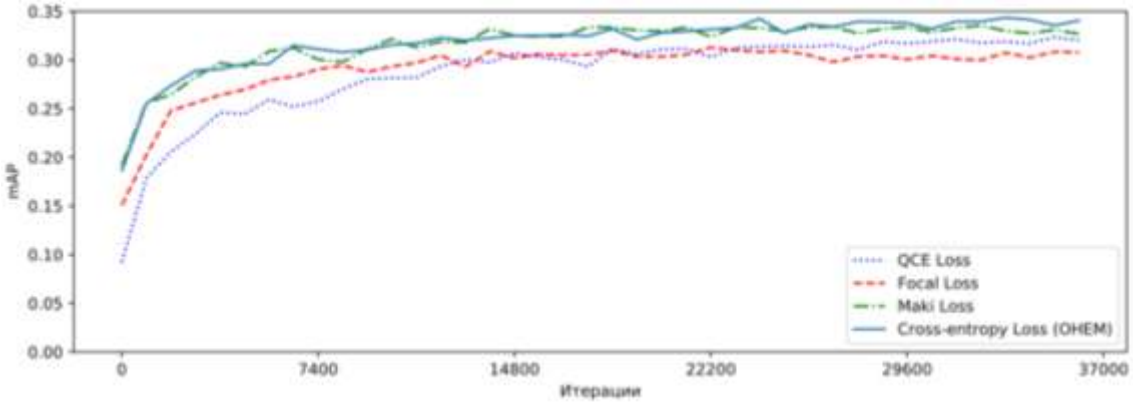

Рисунок 3. Графики изменения mAP во время обучения детекторов объектов.

6. Обсуждение результатов

Результаты эксперимента показали, что функции потерь, удовлетворяющие свойствам коэффициента масштаба Focal Loss, могут быть использованы для обучения детекторов объектов в условиях дисбаланса тяжелых и лёгких примеров. Было показано, что простая модификация градиента перекрестной энтропии позволяет сделать её устойчивой к данной проблеме дисбаланса.

Стоит заметить, что обучение с QCE Loss происходит несколько медленнее. Связать данный феномен можно с типом нормализации данной функции потерь. Нормализация числом тяжелых примеров в батче может ускорить процесс обучения.

7. Развитие идеи модификации градиента, градиентный инжиниринг

Хотя предложенные функции потерь следуют двум вышеозвученным свойствам градиента Focal Loss, мы предполагаем, что решающее влияние на обучение оказывает именно второе свойство. Примером дальнейшей модификации градиента функций потерь может быть инкапсуляция стратегий обучения нейронных сетей в коэффициент масштаба.

Так, авторы [12] использовали малое θ в начале обучения нейронной сети, что повышало стабильность дальнейшего обучения при больших значениях θ. Такой способ повышения

стабильности обучения они назвали warm up. Объяснить данный феномен можно тем, что значения p в начале обучения близки к 0.1, вследствие чего величина градиента высока, что в свою очередь негативно сказывается на процессе обучения. Для решения данной проблемы можно сконструировать такой коэффициент масштаба, величина которого не будет превышать заранее заданного порога для малых p.

В [13] авторы предлагают изменять значения θ периодически, используя функцию косинус. Данная стратегия обучения нейронных сетей также может быть инкапсулирована в градиент функции потерь посредством построения синусообразного коэффициента масштаба.

В общем случае коэффициент масштаба может строиться с учетом индивидуальных особенностей нейронных сетей и обучающих данных. Например, среднее соотношение количества тяжелых и легких примеров в батче зависит как от способа дискретизации пространства входного изображения «якорями», так и от самого изображения. Данное соотношение может быть использовано для подбора оптимальной формы коэффициента масштаба. Проверке вышеозвученных гипотез будут посвящены последующие работы.

8. Заключение

В данной работе была подробно рассмотрена функция потерь Focal Loss и выявлен ряд свойств её градиента: монотонное увеличение и стремление к нулю при $p \to 1$. Предложено две новые функций потерь, построенные с учетом данных свойств: QCE Loss и Maki Loss. Было экспериментально показано, что данные функции потерь позволяют обучать нейронные сети в условиях дисбаланса данных. Также был предложен новый подход к обучению нейронных сетей – построение собственной функций потерь путём дизайна коэффициента масштаба, подходящего для конкретного случая.

9. Благодарности

Работа выполнена в рамках государственного задания по теме FSSS-2020-0017, при частичной финансовой поддержке Российского фонда фундаментальных исследований № 19-29-01135.

Работа выполнена с использованием рабочей станции NVIDIA DGX Station, входящей в состав оборудования ИЦ "Большие данные" Самарского университета.

10. Литература

[1] Girshick, R. Rich feature hierarchies for accurate object detection and semantic segmentation / R. Girshick, J. Donahue, T. Darrell, J. Malik // Proceedings of the IEEE conference on computer vision and pattern recognition. – 2014. – P. 580-587.

[2] Girshick, R. Fast r-cnn // Proceedings of the IEEE international conference on computer vision. – 2015. – P. 1440-1448.

[3] Ren, S. Faster r-cnn: Towards real-time object detection with region proposal networks / S. Ren, K. He, R. Girshick, J. Sun // Advances in neural information processing systems. – 2015. – P. 91-99.

[4] Redmon, J. You only look once: Unified, real-time object detection / J. Redmon, S. Divvala, R. Girshick, A. Farhadi // Proceedings of the IEEE conference on computer vision and pattern recognition. – 2016. – P. 779-788.

[5] Liu, W. Ssd: Single shot multibox detector / W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.Y. Fu, A.C. Berg // European conference on computer vision. – 2016. – P. 21-37.

[6] Shrivastava, A. Training region-based object detectors with online hard example mining / A. Shrivastava, A. Gupta, R. Girshick // Proceedings of the IEEE conference on computer vision and pattern recognition. – 2016. – P. 761-769.

[7] Lin, T.Y. Focal loss for dense object detection / T.Y. Lin, P. Goyal, R. Girshick, K. He, P. Dollár // Proceedings of the IEEE international conference on computer vision. – 2017. – P. 2980-2988.

[8] Ruder, S. An overview of gradient descent optimization algorithms [Электронный ресурс]. – Режим доступа: https://arxiv.org/pdf/1609.04747.pdf (дата обращения 01.12.2019).

[9] Фреймворк MakiFlow [Электронный ресурс]. – Режим доступа: https://github.com/MakiResearchTeam/MakiFlow.
[10] Sandler, M. Mobilenetv2: Inverted residuals and linear bottlenecks / M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, L.C. Chen // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. – 2018. – P. 4510-4520.
[11] Kingma, A. A method for stochastic optimization / A. Kingma, P. Diederik, J. Ba [Электронный ресурс]. – Режим доступа: https://arxiv.org/pdf/1412.6980.pdf (дата обращения 01.12.2019).
[12] He, K. Deep residual learning for image recognition / K. He, X. Zhang, S. Ren, J. Sun // Proceedings of the IEEE conference on computer vision and pattern recognition. – 2017. – P. 770-778.
[13] Loshchilov, I. Sgdr: Stochastic gradient descent with warm restarts / I. Loshchilov, F. Hutter – [Электронный ресурс]. – Режим доступа: https://arxiv.org/pdf/1608.03983.pdf (дата обращения 01.12.2019).

Gradient as a foundation for building a loss function

I.A. Kilbas[1], R.A. Paringer[1,2]

[1]Samara National Research University, Moskovskoe Shosse 34A, Samara, Russia, 443086
[2]Image Processing Systems Institute of RAS - Branch of the FSRC "Crystallography and Photonics" RAS, Molodogvardejskaya street 151, Samara, Russia, 443001

Abstract. Deep neural networks have achieved a tremendous success in a variety of applications across many disciplines. Yet, training a neural network comes with several challenges that have to be solved. The performance of a deep learning models rely not only on the network architecture but also on the choice of a loss function. Cross-entropy loss found to be the most common choice for the classification problem. But its main downside is that it can't handle data with huge class imbalance. In order to tackle this problem, the Focal loss has been proposed. In this paper we investigate the reasons behind its good performance. We find several properties of the Focal loss' gradient that can be applied for building new loss functions and propose a few of them. We also show an experimental evidence of the validity of the proposed functions.