Применение метода градиентного спуска для балансировки данных в задачах анализа диагностических изображений

А.В. Мухин[1], И.А. Килбас[1], Р.А. Парингер[1,2], Н.Ю. Ильясова[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. В статье предложен и исследован алгоритм балансировки данных, основанный на методе градиентного спуска. Описанный в работе алгоритм позволяет частично нивелировать проблему несбалансированности данных, возникающую нередко при анализе биомедицинских изображений. Так, было проведено экспериментальное исследование точности сверточных нейронных сетей в задаче семантической сегментации изображений глазного дна без балансировки и с применением предложенного алгоритма балансировки. Так же в работе сформулированы рекомендации по использованию разработанного алгоритма.

1. Введение

Задача семантической сегментации[1][2] является важной и востребованной в настоящее время. Её суть заключается в попиксельной классификации изображений. Наиболее сложной и специфичной является задача семантической сегментации биомедицинских данных [3][4]. Так, чаще всего для её решения используют либо нейронные сети [5][6], либо текстурные признаки [7][8]. Для решения задачи семантической сегментации были использованы полносверточные нейронные сети [9], которые принимают на вход изображение и возвращают маску того же размера на выходе с размеченными в соответствии с предсказаниями пикселями. Также в работе решались проблемы, присущие задаче семантической сегментации.

Первой из них является проблема дисбаланса данных [10][11][12].

Данная проблема заключается в том, что некоторые классы среди исходных данных встречаются намного реже остальных, в силу чего, точность классификации данных классов мала.

Существующие методы балансировки [13] данных так или иначе сводятся к двум техникам, а именно: undersampling – удаление объектов, содержащих доминирующий класс в выборке и oversampling – дублирование объектов, содержащих редкий класс в выборке.

Эти техники можно без труда применить в бинарных задачах, например в бинарной классификации, или в задачах, где каждое изображение из исходного набора данных может содержать только один класс [13]. Однако балансирование данных становится нетривиальной задачей, если изображения содержат более одного класса. В подобных случаях исследователи предлагают новые алгоритмы балансировки данных.

Например, одним из способов борьбы с проблемой дисбаланса классов в задаче семантической сегментации является использование весовых карт [14]. Однако, в силу серьезного дисбаланса, в нашем случае весовые маски не оказали пользы.

Так, при решении проблемы дисбаланса данных в нашей задаче, нами был разработан новый алгоритм балансирования данных, основанный на методе градиентного спуска.

Второй проблемой, с которой мы столкнулись – является проблема оценки качества обученных нейронных сетей. Так в задачах семантической сегментации чаще всего применяют метрики “intersection over union” (IoU) [15] и коэффициента Дайса [16]. В задачах семантической сегментации биомедицинских данных чаще применяют именно коэффициент Дайса. Но в силу того, что эта метрика является бинарной, а поставленная задача не является бинарной, нами было разработано её обобщение на случай множества классов – vDice.

2. Биомедицинские данные

Исходные биомедицинские данные представлены 115-ю изображениями глазного дна (Рисунок 1). Изображения размечены попиксельно экспертным методом на 9 классов (Рисунок 2): диск зрительного нерва, макула, сосуды, твердый экссудат, мягкий экссудат, свежие коагуляты, пигментированные коагуляты, ретинальные геморрагии.

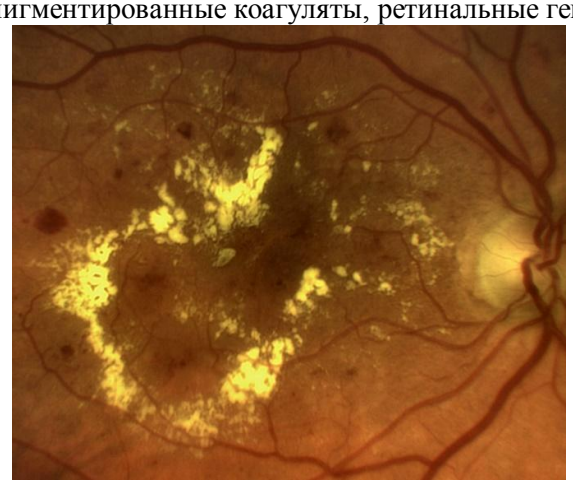

Рисунок 1. Пример изображения глазного дна.

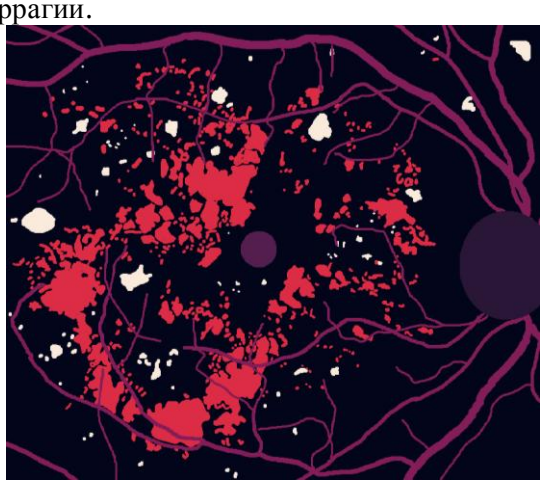

Рисунок 2. Пример разметки глазного дна.

Не все классы содержатся на изображениях в одинаковой мере, присутствует ярко выраженный дисбаланс (Рисунок 3). Так классы 6, 7 являются самыми редкими – присутствуют менее чем в 20% изображений (Рисунок 4).

3. Алгоритм балансировки данных

Разработанный алгоритм сводится к полуавтоматическому балансированию исходных данных при помощи техник oversampling и undersampling с использованием метода градиентного спуска. Идея создать подобную реализацию пришла после осознания того, что вручную балансировать данные, в которых имеется более 3-ех классов одновременно – нетривиальная задача. Попытки сделать это вручную приводили к усугублению проблемы дисбаланса классов. Поэтому было решено сформулировать новую, более простую и общую задачу (такая формулировка позволит использовать наш алгоритм для балансировки любых данных).

Каждое исходное изображение может быть описано с помощью вектора λ состоящего из нулей и единиц. Данный вектор построен так, что его i-ая компонента равна 1, если соответствующее изображение содержит i-ый класс, иначе компонента равен 0. Таким образом исходные изображение представляются в виде набора векторов, этот набор обозначим как Ω.

Обозначим $\sum(\Omega)$ как сумму всех векторов из множества Ω. Тогда i-ый элемент этой суммы будет обозначать количество изображений содержащих i-ый класс.

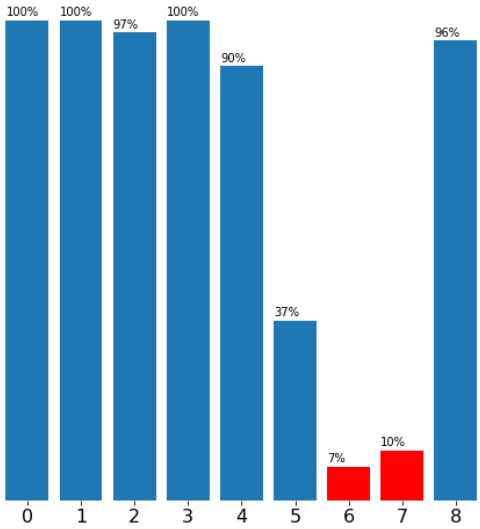

Рисунок 3. Процент картинок содержащих i-ый класс.

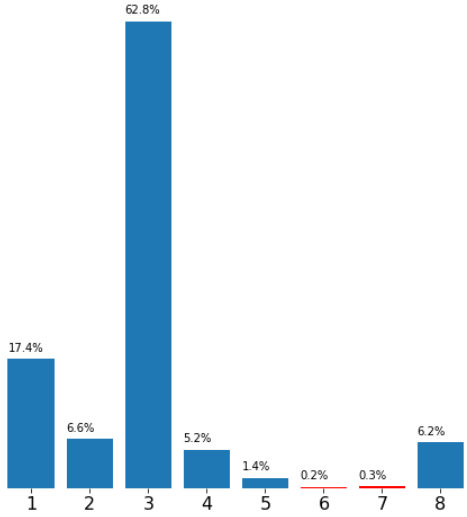

Рисунок 4. Отношение количества пикселей i-ого класса к общему количеству пикселей (на графике отсутствует класс нормы).

Пусть π – вектор распределения классов (или целевой вектор), где на i-ой позиции находится отношение количества изображений, имеющих класс i, к общему числу изображений. Этот вектор необходимо задавать вручную, т.к. он описывает распределение, к которому мы стремимся и которое хотим получить в последствии.

Теперь мы можем сформулировать задачу: нам нужно из набора векторов Ω создать набор векторов $\Omega`$ такой, что выполняется отношение $\frac{\sum(\Omega`)}{|\Omega`|} = \pi$, где $|\Omega`|$ - общее число элементов из $\Omega`$.

Задачу, описанную выше, можно упростить. Пусть $\psi(\Omega)$ — это набор уникальных векторов λ из Ω. Тогда $\Lambda[\psi(\Omega)] = \sum_i \alpha_i \lambda_i$, где α_i - количество копий вектора λ_i из набора $\psi(\Omega)$. Набор $\Omega`$ будет состоять из векторов λ_i, продублированных α_i раз.

Для балансировки набора $\Omega`$ определим следующую функцию:

$$L = \left| \frac{\Lambda[\psi(\Omega`)]}{|\Omega`|} - \pi \right|$$

Данная функция измеряет расстояние от вектора $\Lambda[\psi(\Omega`)]$ до вектора $\pi$. Минимизируя это расстояние с помощью подбора параметров $\alpha_i$ методом градиентного спуска, мы получим сбалансированный набор $\Omega`$, что будет является решением поставленной задачи.

Такая функция позволяет найти решение задачи, однако такое решение далеко не всегда является оптимальным и порождает чрезмерное количество дубликатов данных, что может отрицательно сказаться на качестве обучения нейронной сети. Поэтому нами была сформулирована следующая функция регуляризации.

$$RL = |P_\Omega - P_\alpha| = \sum_{i=1}^{N} \left(\frac{|\Omega_i|}{|\Omega|} - \frac{\alpha_i}{\sum_{j=1}^{N} \alpha_j} \right)^2, \text{ где}$$

$|\Omega|$ - общее число элементов из Ω, $|\Omega_i|$- число элементов из Ω содержащих i-ый класс, α_i - мощность (вес) -го класса.

Данная функция регуляризации позволяет предотвратить сильное отклонение параметров α_i от изначальных значений весов, что позволит найти более оптимальное решения и обойтись меньшим количество дубликатов.

4. Применение алгоритма балансировки данных

Реализацию данного алгоритма в виде исходного кода на языке программирования Python можно найти в нашем фреймворке MakiFlow [17]. Алгоритм не является полностью автоматическим, поэтому необходимо контролировать процесс балансировки, изменяя

параметры градиентного спуска и вектора распределения для получения необходимого результата.

Рекомендации по выбору целевого вектора π. Одной из идей в качестве целевого вектора может быть вектор единиц. В ходе экспериментов было замечено, что учет специфики данных позволяет улучшить качество результатов. Так, для ряда классов, процентное содержание которых в исходных данных было около 10%, нельзя было получить значения 90% или даже 40% без большого количество дубликатов в сбалансированной выборке. Несмотря на это, данное процентное соотношение можно было увеличить на 10-15%, при этом избежав создания лишних дубликатов, сделав сбалансированную выборку качественной. Поэтому рекомендуется:

- Для редких классов:
    - в случае, если используется регуляризация, ставить значения близкие к единице
    - в случае, если регуляризация не используется, инициализировать значения близкими к исходному распределению классов (такие значение можно подобрать эмпирически)
- Остальные классы стоит инициализировать их процентным содержанием в выборке.

Также рекомендуется использовать оптимизатор Adam [18] для градиентного спуска. Learning Rate необходимо подбирать эмпирически.

В случае, если используется регуляризация – функция, минимизируемая градиентным спуском является суммой двух функций:

$$ObjectiveFunction = L + \gamma RL$$

Влияние функции регуляризации контролируется параметром γ, который выбирается также эмпирически.

Что касается инициализации весов, в работе было исследовано несколько подходов. Изначально было опробовано инициализировать веса константой, затем веса задавались мощностью классов, умноженных на константу. Оба этих подхода позволили получить схожие результаты. Поэтому вопрос о лучшей инициализации весов остаётся открытым и является объектом будущих исследований.

5. Метрика

В настоящем времени в доминирующем большинстве случаев исследователи в задачах детектирования объектов и семантической сегментации использую такую метрику, как IoU (Intersection Over Union). Однако, исходя из наших наблюдений в задаче семантической сегментации биомедицинских данных – очень часто использую коэффициент Дайса для оценки точности предсказаний нейронной сети.

Коэффициент Дайса используется для оценки качества бинарной сегментации. В нашей задаче производится многоклассовая сегментация, в силу этого обстоятельства нами было предложено обобщение коэффициента Дайса - vDice. Её реализация заключается в подсчете коэффициента Дайса для каждого из классов в отдельности посредством сведения задачи к бинарной: необходимый класс в маске помечается как класс 1, все остальные – как класс 0. vDice равен среднему значению коэффициентов Дайса, подсчитанных для всех классов.

6. Постановка эксперимента

Для экспериментальной проверки предложенного алгоритма было сформировано три выборки: несбалансированная (оригинальная), сбалансированная без применения регуляризации, сбалансированная с применением регуляризации. В каждой использовались все 115 изображений.

Далее, на основе каждой выборки было составлено соответственно три набора для кросс-валидации (Таблица 1). Наборы различаются составом тренировочного и тестового множеств. Причем тестовые множества составлены без пересечений. При дублировании изображения подвергались аугментации (эластичная трансформация, отражения). В каждом наборе было около 5000 изображений.

Для решения задачи семантической сегментации биомедицинских данных была построена и обучена сверточная нейронная сеть с помощью разработанного нами фреймворка MakiFlow. В

качестве архитектурной основы для неё, выступила нейронная сеть U-Net [14]. В качестве экстрактора признаков была выбрана предобученная свёрточная нейронная сеть Xception-65 [19]. Нейронная сеть обучалась с использованием оптимизатора Adam, на learning rate = 8e-3, с функцией ошибки FocalLoss [20]. Обучение длилось 10 эпох, а размер батча был равен 8.

Построенная нейронная сеть была обучена на всех составленных наборах с нуля. В конце каждой обучающей эпохи производилось тестирование, подсчитывались коэффициенты Дайса для каждого класса и метрика vDice. Достоверность эксперимента обеспечена кросс-валидацией.

Таблица 1. Пример итогового распределения данных.

	Класс							
	1	2	3	4	5	6	7	8
Несбалансированная	100%	97%	100%	90%	37%	7%	10%	96%
Сбалансированная без регуляризации	100%	98%	100%	81%	56%	31%	44%	95%
Сбалансированная с регуляризацией	100%	98%	100%	93%	36%	16%	20%	97%

7. Результаты

Из таблицы 2 можно увидеть, что обучение нейронной сети на сбалансированной выборке даёт результат лучше, чем обучение на несбалансированной выборке. Значение коэффициента Дайса для класса 5 на выборке сбалансированной без использования регуляризации можно объяснить переобучением сети [21] вследствие создания большого количества дубликатов изображений, содержащих данный класс. Кроме того, результаты подтверждают наши опасения о влиянии большого числа дубликатов на результат обучения нейронной сети. Так, меньшее количество дубликатов, полученное балансировкой с использованием регуляризации, положительно сказывается на качестве обучения.

Таблица 2. Усредненные значения максимально достигнутой точности по выборкам.

	Не сбалансирована	Сбалансирована (без применения регуляризации)	Сбалансирована (с применением регуляризации)
Класс 1 (Диск зрительного нерва)	0.9130	0.9198	**0.9266**
Класс 2 (Макула)	0.6287	**0.7364**	0.7049
Класс 3 (Сосуды)	0.7346	0.7416	**0.7425**
Класс 4 (Твердый экссудат)	0.3786	0.5639	**0.5892**
Класс 5 (Мягкий экссудат)	0.3060	0.1899	**0.3132**
Класс 6 (Свежие коагулянты)	0.0801	0.0857	**0.1031**
Класс 7 (Пигментированные коагулянты)	0.0929	0.0895	**0.1808**
Класс 8 (ретинальные геморрагии)	0.5728	0.6182	**0.6491**
vDice	0.4511	0.4762	**0.5050**

8. Выводы

В работе представлен алгоритм балансировки данных, основанный на методе градиентного спуска. Были даны рекомендации по применению предложенного алгоритма. Для оценки точности нейронной сети был использован коэффициент Дайса и его обобщение на случай множества классов - метрику vDice. Для обеспечения достоверности полученных результатов была использована кросс-валидация.

С использованием предложенного алгоритма балансировки данных, значение коэффициента Дайса на редких классах, а именно на 6 и 7, увеличился на 2.3% и 8.8% соответственно. В свою очередь, коэффициент Дайса на остальных классах не подвергся уменьшению. Также

увеличилось значение vDice на 5.4%. Исходя из полученных результатов можно сказать, что предложенный в работе алгоритм позволяет уменьшить влияние проблемы несбалансированности данных на точность сверточной нейронной сети.

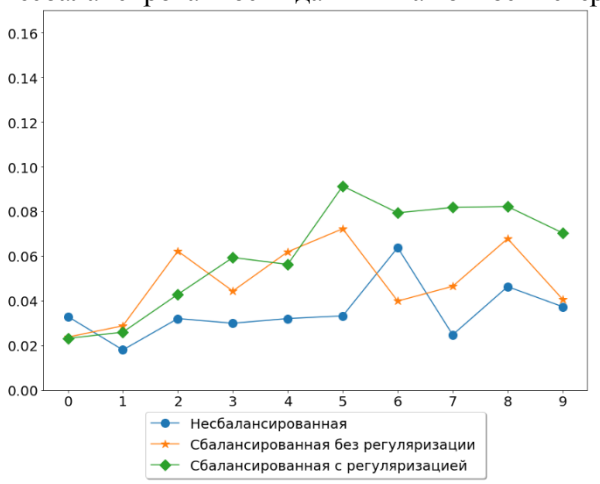

Рисунок 5. График зависимости значения коэффициента Дайса для 6-го класса от количества эпох во время обучения для разных выборок.

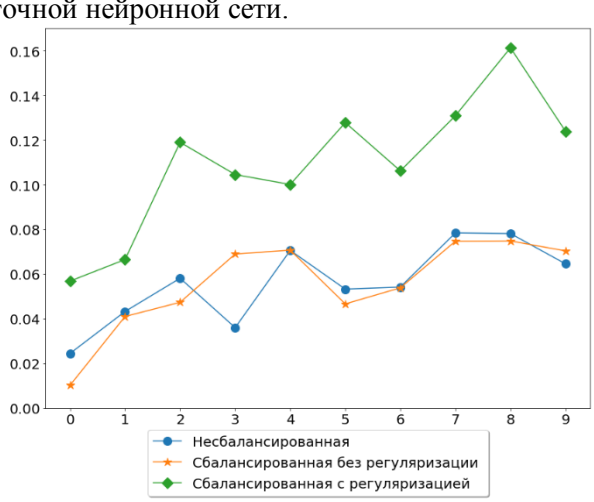

Рисунок 6. График зависимости значения коэффициента Дайса для 7-го класса от количества эпох во время обучения для разных выборок.

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] Noh, H. Learning deconvolution network for semantic segmentation / H. Noh, S. Hong, B. Han // Proceedings of the IEEE international conference on computer vision. – 2015. – P. 1520-1528.

[3] Zhang, Y. Deep Adversarial Networks for Biomedical Image Segmentation Utilizing Unannotated Images / Y. Zhang, L. Yang, J. Chen, M. Fredericksen, D.P. Hughes, D.Z. Chen // Medical Image Computing and Computer-Assisted Intervention – MICCAI. – 2017. – P. 408-416. DOI:10.1007/978-3-319-66179-7_47.

[4] Yang, L. Suggestive Annotation: A Deep Active Learning Framework for Biomedical Image Segmentation / L. Yang, Y. Zhang, J. Chen, S. Zhang, D.Z. Chen // Medical Image Computing and Computer-Assisted Intervention – MICCAI. – 2017. – P. 399-407. DOI: 10.1007/978-3-319-66179-7_46.

[5] Milletari, F. V-Net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation / F. Milletari, N. Navab, S.A. Ahmadi // Fourth International Conference on 3D Vision (3DV), 2016. DOI:10.1109/3dv.2016.79.

[6] Havaei, M. Brain tumor segmentation with deep neural networks / M. Havaei, A. Davy, D. Warde-Farley, A. Biard, A. Courville, Y. Bengio, H. Larochelle // Medical image analysis. – 2017. – Vol. 35. – P. 18-31.

[7] Ilyasova, N. Intelligent feature selection technique for segmentation of fundus images / N. Ilyasova, R. Paringer, A. Kupriyanov, D. Kirsh // Seventh International Conference on Innovative Computing Technology (INTECH), 2017. DOI:10.1109/intech.2017.8102433.

[8] Ilyasova, N.Y. A modified technique for smart textural feature selection to extract retinal regions of interest using image pre-processing / N.Y. Ilyasova, A.S. Shirokanev, R.A. Paringer, A.V. Kupriyanov, A.V. Zolotarev // Journal of Physics: Conference Series. – 2018. – Vol. 1096. – P. 012095. DOI:10.1088/1742-6596/1096/1/012095.
[9] Shelhamer, E. Fully Convolutional Networks for Semantic Segmentation / E. Shelhamer, J. Long, T. Darrell // IEEE Transactions on Pattern Analysis and Machine Intelligence. – 2017. – Vol. 39(4). – P. 640-651. DOI:10.1109/tpami.2016.2572683.
[10] Haibo, H. Learning from Imbalanced Data / H. Haibo, E.A. Garcia // IEEE Transactions on Knowledge and Data Engineering. – 2009. – Vol. 21(9). – P. 1263-1284. DOI: 10.1109/ tkde.2008.239.
[11] Mostafizur Rahman, M. Addressing the Class Imbalance Problem in Medical Datasets / M. Mostafizur Rahman, D.N. Davis // International Journal of Machine Learning and Computing. – 2013. – Vol. 3(2). – P. 224-228. DOI:10.7763/IJMLC.2013.V3.307.
[12] Longadge, R. Class imbalance problem in data mining review / R. Longadge, S. Dongre // arXiv preprint arXiv:1305.1707, 2013.
[13] Yap, B.W. An application of oversampling, undersampling, bagging and boosting in handling imbalanced datasets / B.W. Yap, K.A. Rani, H.A.A. Rahman, S. Fong, Z. Khairudin, N.N. Abdullah // Proceedings of the first international conference on advanced data and information engineering (DaEng-2013) – Springer, Singapore, 2014. – P. 13-22. DOI: 10.1007/978-981-4585-18-7_2.
[14] Ronneberger, O. U-net: Convolutional networks for biomedical image segmentation / O. Ronneberger, P. Fischer, T. Brox // International Conference on Medical image computing and computer-assisted intervention – Springer, Cham, 2015. – P. 234-241. DOI: 10.1007/978-3-319-24574-4_28.
[15] Nowozin, S. Optimal decisions from probabilistic models: the intersection-over-union case / S. Nowozin // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. – 2014. – P. 548-555.
[16] Shamir, R.R. Continuous dice coefficient: a method for evaluating probabilistic segmentations / R.R. Shamir, Y. Duchin, J. Kim, G. Sapiro, N. Harel // arXiv preprint arXiv:1906.11031, 2019.
[17] Фреймворк MakiFlow [Электронный ресурс]. – Режим доступа: https://github.com/MakiResearchTeam/MakiFlow.
[18] Kingma, D.P. A method for stochastic optimization / D.P. Kingma, J. Ba // arXiv preprint arXiv:1412.6980, 2014.
[19] Chollet, F. Xception: Deep learning with depthwise separable convolutions / F. Chollet // Proceedings of the IEEE conference on computer vision and pattern recognition, 2017. – P. 1251-1258.
[20] 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.
[21] Hawkins, D.M. The problem of overfitting / D.M. Hawkins // Journal of chemical information and computer sciences. – 2004. – Vol. 44(1). – P. 1-12. DOI:10.1021/ci0342472.

Application of the gradient descent for data balancing in diagnostic image analysis problems

A.V. Mukhin[1], I.A. Kilbas[1], R.A. Paringer[1,2], N.Yu. Ilyasova[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. The article proposes an algorithm for data balancing based on gradient descent. The proposed algorithm is able to partially mitigate the influence of the data imbalance problem which is commonly seen in the tasks of diagnostic image analysis. The authors have investigated the influence of the proposed algorithm on the accuracy of a fully convolutional neural network. The neural network was trained on unbalanced data as well as on the balanced by the algorithm. Recommendations on how to use the proposed algorithm are also formulated.