# Data Balancing Method for Training Segmentation Neural Networks *

Alexey Kochkarev[1 [0000−0003−4630−2832]], Alexander Khvostikov[1 [0000−0002−4217−7141]], Dmitry Korshunov[2 [0000−0002−8500−7193]], Andrey Krylov[1 [0000−0001−9910−4501]], and Mikhail Boguslavskiy[2 [0000−0002−1582−8033]]

[1]Faculty of Computational Mathematics and Cybernetics,
[2]Faculty of Geology, Lomonosov Moscow State University, Moscow, Russia
alexey.kochkarev@gmail.com khvostikov@cs.msu.ru
Dmit0korsh@gmail.com kryl@cs.msu.ru mikhail@geol.msu.ru

**Abstract.** Data imbalance is a common problem in machine learning and image processing. The lack of training data for the rarest classes can lead to worse learning ability and negatively affect the quality of segmentation. In this paper we focus on the problem of data balancing for the task of image segmentation. We review major trends in handling unbalanced data and propose a new method for data balancing, based on Distance Transform. This method is designed for using in segmentation convolutional neural networks (CNNs), but it is universal and can be used with any patch-based segmentation machine learning model. The evaluation of the proposed data balancing method is performed on two datasets. The first is medical dataset LiTS, containing CT images of liver with tumor abnormalities. Second one is a geological dataset, containing of photographs of polished sections of different ores. The proposed algorithm enhances the data balance between classes and improves the overall performance of CNN model.

**Keywords:** Image Segmentation, Data Balancing, Convolutional Neural Networks, Liver, Tumor, Geology.

## 1 Introduction

Data imbalance is a common issue in image segmentation [1]. If pixels corresponding to a particular “majority” class are far more numerous than pixels of one or more “minority” classes, the rarity of the “minority” class in the training data makes the training process less effective and worses the final results, as the learned model will tend to classify most pixels as members of the “majority” classes.

The problem of data imbalance is very common in medical problems and, in particular, detecting liver tumors. One of these problems is segmentation of CT images, since the volume and area of different organs and abnormalities differs a lot. In a typical CT

* The work was funded by RFBR, CNPq and MOST according to the research project 19-57-80014 (BRICS2019-394)

image of a liver tumor, the volume of healthy liver tissue is significantly greater than the volume of cancerous tissue [2].

Data imbalance problem also occurs in geological image segmentation. Some minerals are found in nature much less often than others. For example, photographs of polished sections, taken from lead-zinc fields contain a larger amount of Sphalerite ore (about 30 %) than Galena (about 7%).

The existent methods that are used to overcome class imbalance can be divided into two main categories [3].

The first category of methods is represented with algorithm-based methods. One of the approaches is cost-sensetive learning [4]. The idea is to assign different costs to classification mistakes for different classes. One common scheme involves assigning to each class a cost equal to the inverse of the proportion of this class in dataset. This leads to higher model penalization for rarest classes.

The second category of methods is represented with so-called data-based methods. They use sampling techniques to rebalance the distribution of classes during preprocessing. This involves either oversampling instances of the minority class or undersampling instances of the majority class, or even both [5]. One of the generalizations of these approaches is Synthetic Minority Over-Sampling Technique, or SMOTE [6]. This technique involves generating synthetic samples of the minority class to train on, thus reducing the class imbalance by artificially inflating the size of the minority class itself. It also should be noticed that SMOTE and other methods based on SMOTE [7][8] are designed for the machine learning-based classification problems, and either can not be used at all in the segmentation problem or demonstrate mediocre results.

In this paper we propose a data balancing method that focuses on modifying the class distribution in the dataset. The proposed method uses only data from the original set rather than replicating additional minority samples. The proposed method is specially created for segmentation problems and has a wide range of applications.

## 2 Proposed method

Segmentation is defined as finding a mapping of the source color image $I \in \mathbb{R}^{h \times w \times 3}$ of height $h$ and width $w$ to the annotation $S \in \mathbb{Z}^{h \times w}$, containing labels (or classes). Let us consider a set of classes $C = \{c_0, c_1, ..., c_N\}$.

For the convenience we internally store the annotation as a 2-D array of integers ($S = S(x, y)$), each value of which is just the index of the class value: $S(x, y) \in C$. The source images are stored as 3-D array of float numbers: $I = I(x, y),\ I(x, y) \in \mathbb{R}^3$.

The proposed method is created to be used with data that is fed into neural networks during training process. On every step of each training epoch the neural network gets a batch of square patches, taken from one of the dataset images $\{I_1(x, y), ..., I_m(x, y)\}$. Conventional random choice of patches can only increase the existing data imbalance, because there could occur objects that are smaller then a patch and for any patch covering case, the ratio between object and non-object pixels inside the patch can not be increased. The proposed method allows to choose patches, that contain most amount of pixels of the certain class. Thus, it is possible to keep balance of classes in images, that are fed into network. The proposed method consists of three main steps:

1. choose class, that was most rarely fed into the model;
2. choose image, which contains the greatest quantity of pixels of chosen class;
3. crop patch from the selected image, which contains the greatest quantity of pixels of chosen class;

All these stages will be considered in detail below.

### 2.1 Class choice

A square area in image $I_i(x, y)$ is called a patch and marked as $P(x, y)$. The appropriate area on corresponding annotation $S_i(x, y)$ is marked as $P_{Ann}(x, y)$. After patch is chosen, we sum the amount of pixels for each class:

$$s_j = |(x, y) : \ P_{Ann}(x, y) = c_j|, \ j = 0, 1, ..., N. \tag{1}$$

On this step we should choose class $C_k$, for which amount of pixels is the smallest:

$$s_k = min\{s_0, s_1, ..., s_N\}. \tag{2}$$

### 2.2 Image choice

For every image $I_i$ in dataset we compute weights for each class $\{w_0, w_1, ..., w_N\}$ as the amount of pixels of this class on the image:

$$w_j = |(x, y) : \ S_i(x, y) = c_j|, \ j = 0, 1, ..., N. \tag{3}$$

So, the image will be chosen from the dataset of $\{I_1, ..., I_m\}$ images with probability proportionally to its weight.

### 2.3 Patch choice

On this step we choose patch, which contains the greatest quantity of pixels of chosen class $C_k$. To perform this choice we first calculate the probability maps of choosing upper left pixel of the patch in current pixel. First, distance to class map is built. Let, annotation $S(x, y)$ to be consisted of two classes: Object ($c_1$) and Background ($c_0$):

$$S(x, y) \in \{c_0, c_1\}. \tag{4}$$

At first, we apply Distance Transform [9] and get a map $S_d(x, y)$, where each pixel is a distance to the nearest object pixel on the annotation $S(x, y)$:

$$S_d(x, y) = \begin{cases} 0, & S(x, y) = c_1 \\ \min\left(\|x - x_0, y - y_0\|_{L_2}, \forall S(x_0, y_0) = c_0\right), & S(x, y) = c_0 \end{cases}, \tag{5}$$

$$\text{where } \|x, y\|_{L_2} = \sqrt{x^2 + y^2}. \tag{6}$$

Consider a patch $P(x, y)$ of size $p \times p$ in annotation $S(x, y)$. The less the sum of pixels in a relevant area on $S_d(x, y)$, the more pixels of chosen class there are in this area. Let us define:

$$\tilde{S}_d(x, y) = 1 - \frac{S_d(x, y)}{\max\left(S_d(x, y)\right)}. \tag{7}$$

The more sum is inside a relevant area in $\tilde{S}_d(x, y)$, the more pixels of chosen class there are in chosen area. Summing up pixels on every rectangle area of $S(x, y)$, we get desired probability map $Pr(x, y)$, where each pixel is sum of distances from chosen class to background pixels. The higher the value in current pixel of $Pr(x, y)$, the higher the probability of getting most pixels of chosen class, if we choose upper left pixel of patch in current pixel.
The sum of pixels on the patch has to be calculated many times. Quick and efficient summation of pixels can be performed using Summed-Area table [10]. The value at any point $(x, y)$ in the summed-area table is the sum of all the pixels that are above and to the left of $(x, y)$, inclusive:

$$S_{int}(x, y) = \sum_{\substack{x' \leq x \\ y' \leq y}} S\left(x', y'\right). \tag{8}$$

The summed-area table can be computed efficiently in a single pass over the image, as the value in the summed-area table at $(x, y)$ is just:

$$S_{int}(x, y) = S(x, y) + S_{int}(x, y - 1) + S_{int}(x - 1, y) - S_{int}(x - 1, y - 1). \tag{9}$$

Once the summed-area table is computed, evaluating the sum of intensities over any rectangular area requires exactly four array references regardless of the area size:

$$\sum_{\substack{x_0 < x \leq x_1 \\ y_0 < y \leq y_1}} S(x, y) = S_{int}(D) + S_{int}(A) - S_{int}(B) - S_{int}(C), \tag{10}$$

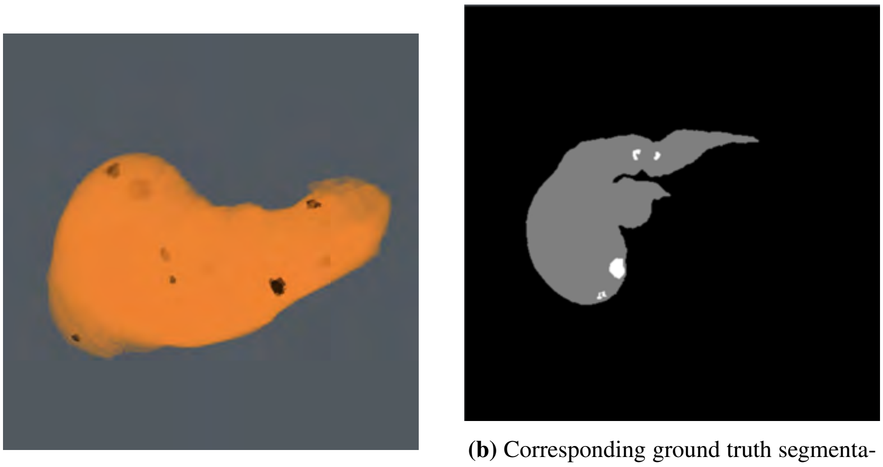
$$A = \left(x_0, y_0\right), B = \left(x_1, y_0\right), C = \left(x_0, y_1\right), D = \left(x_1, y_1\right). \tag{11}$$

## 3 Used datasets

In this paper we used two datasets to evaluate the proposed data balancing method. First one consists of 2-dimensional slices of liver Computer Tomography – Liver Tumor Segmentation Challenge [2] (LiTS). The second dataset consists of photographs of polished sections of ores, collected by geologists. Each dataset is fully-annotated.

### 3.1 LiTS dataset

LiTS dataset contains of 130 3-dimensional CT scans (Fig. 1a) of liver. The images' resolution is 512 × 512 × 75 pixels. We took 20 2-dimensional slices (Fig. 1b) from every scan, that contain liver and tumor pixels. Black pixels correspond to background, grey pixels correspond to liver, white pixels correspond to tumors (Fig. 1b).
The obtained dataset of chosen slices is highly imbalanced: total amount of pixels, corresponding to liver is 29.4%, while tumor abnormalities are present on 4.3% of pixels. Remaining 66.3% of pixels from LiTS dataset correspond to background.



**(a)** Visualization of 3D liver volume

**(b)** Corresponding ground truth segmentation of one of the axial slices (black – background, grey – liver, white – tumor)

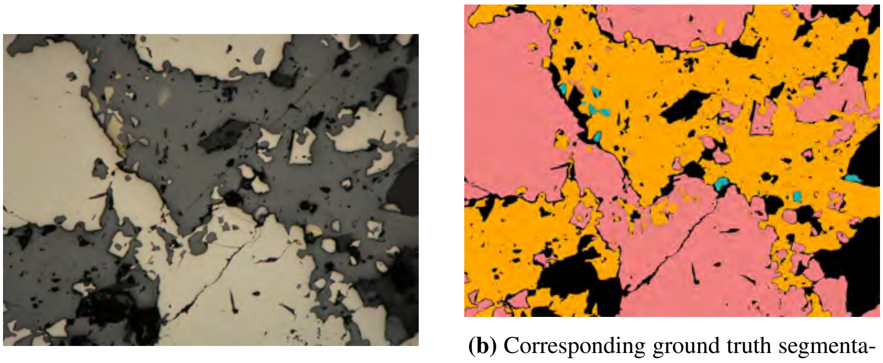
**Fig. 1.** A sample image and its annotation from LiTS dataset.

### 3.2 Polished sections of ores dataset

This dataset consists of 46 images with ground truth annotation made by expert geologist. The dimension of the images is 3396 × 2547. All the photos were taken by Canon G10 with ZEISS Axioscop 40 microscope.

Every image contains up to 4 classes (ores): Background (0 – Bg), Sphalerite (1 – Sh), Pyrite and Marcasite (2 – Py/Mrc) and Galena (3 – Gl).

The most common mineral in the images is Pyrite (pink color on GT) – 30.2%, Spha-



**(a)** Polished section photo

**(b)** Corresponding ground truth segmentation (black – background, pink – Pyrite, yellow – Sphalerite, green – Galena)

**Fig. 2.** A sample image and its annotation from ores dataset.

lerite (orange color on GT) is less common – 29.3%, Galena (green color on GT) is the rarest – only 6.7%, background (black on GT) is 33.7%.

## 4 Experiments and results

To evaluate the proposed data balancing method we chose 3000 patches from each dataset. First, the patches were chosen randomly, and after, the patches were chosen using the proposed method based on probability maps (Fig. 3, Fig. 4).

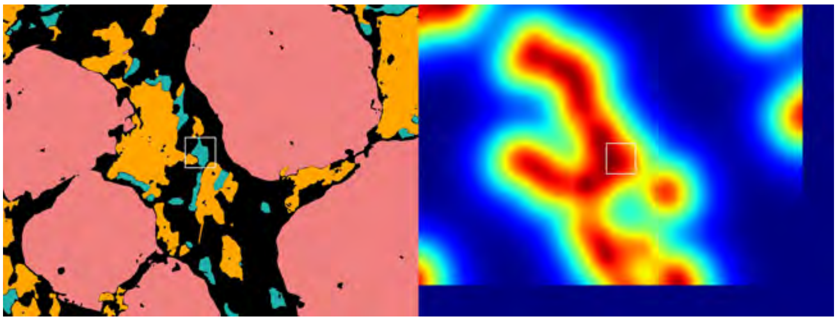


**Fig. 3.** Patch choosing with Galena ore on geological dataset. Left – original image, right – probability map of the proposed method responsible for patch selection. Probability map was built for Galena class.

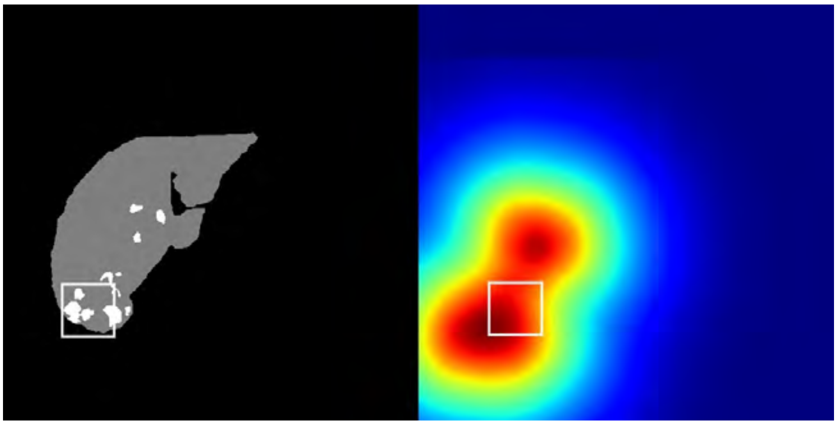


**Fig. 4.** Patch choosing with tumor on LiTS dataset. Left – original image, right – probability map of the proposed method responsible for patch selection. Probability map was built for tumor class.

The ratio of pixels for different classes improved in case of LiTS dataset significantly, but, as we see in Table 1, the quantity of pixels, corresponding to tumor is still less than for liver and background pixels. This imbalance can be explained by the fact that areas with tumor abnormalities are much smaller than selected patch. Even if we take patch according to probability map for selected class, we also get a big amount of pixels, corresponding to nearby classes. The decrease of patch size in this case is inappropriate from the point of view of training the neural network model.

**Table 1.** LiTS dataset experiment results, pixels ratio corresponding to different classes.

| LiTS Dataset | | | |
|---|---|---|---|
| Class labels | In dataset | Random (without balancing) | Proposed balancing |
| Background | 66.3% | 86.9% | 36.4% |
| Liver | 29.4% | 12.5% | 42.1% |
| Tumor | 4.3% | 0.6% | 21.5% |

On the opposite side, even the rarest ores present in areas which size is comparable with patch size even for small pathces. In this case the proposed method demonstrates its efficiency, resulting in well-balanced data, fed into network.

**Table 2.** Ores dataset experiment results, pixels ratio corresponding to different classes.

| Ores Dataset | | | |
|---|---|---|---|
| Class labels | In dataset | Random (without balancing) | Proposed balancing |
| Background | 33.7% | 26% | 25.2% |
| Sh | 29.3% | 55% | 25.1% |
| Py/Mrc | 30.2% | 16% | 24.8% |
| Gl | 6.7% | 3% | 24.9% |

The proposed balancing method was used while training convolutional neural network model based on U-Net [11] architecture. The model was trained in images from ores dataset. We used the $IoU$ metric to evaluate the performance of model::

$$IoU = \frac{|X \cap Y|}{|X| + |Y| - |X \cap Y|}, \tag{12}$$

where $X$ is ground truth annotation, $Y$ is prediction of network.

In multiclass cases $IoU$ is computed separately for each class, according to the ”one against all” principle.

The comparison of $IoU$ value for each class before and after balancing is presented on Fig. 5.

The balancing method improved training process and hence the final results.

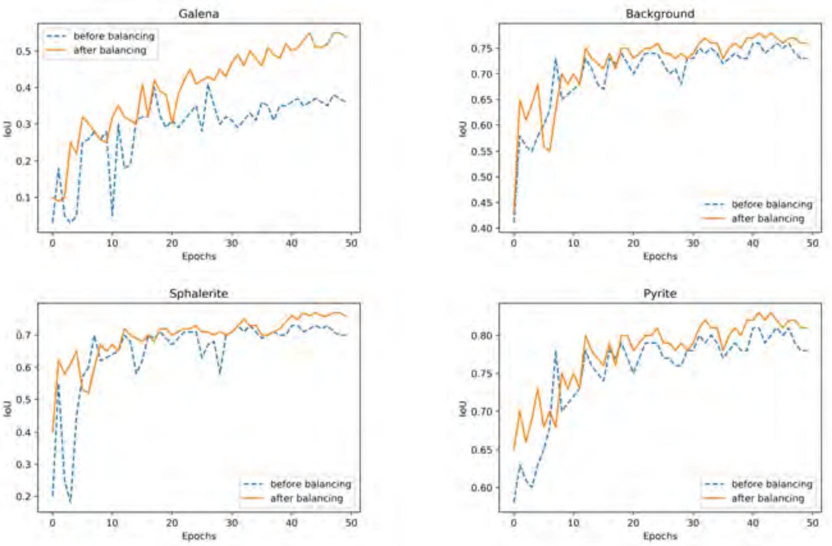


**Fig. 5.** $IoU$ over epoch for different classes while training UNet-based model for ores dataset.

## 5 Conclusion

The developed method of data balancing was tested on biomedical and geological datasets. It has shown its effectiveness in levelling data balance across classes for imbalanced datasets. The method can be used to handle the data imbalance problem in image segmentation task. It is universal and can be used with any patch-based segmentation machine learning model, including convolutional neural networks.

## References

1. He H., Garcia E. A.: Learning from imbalanced data. IEEE Transactions on knowledge and data engineering **21**(9), 1263 1284 (2009)
2. Christ P.: LiTS Liver Tumor Segmentation Challenge (LiTS17). URL https://competitions.codalab.org/competitions/17094 (2017)
3. Small H., Ventura J.: Handling unbalanced data in deep image segmentation. University of Colorado (2017)
4. Nashnush E., Vadera S.: Cost-sensitive Bayesian network learning using sampling. In: Recent Advances on Soft Computing and Data Mining. pp. 467 476. Springer (2014)
5. Buda M., Maki A., Mazurowski M. A.: A systematic study of the class imbalance problem in convolutional neural networks. Neural Networks **106**, 249-259 (2018)
6. Chawla N. V. et al.: SMOTE: synthetic minority over-sampling technique. Journal of artificial intelligence research **16**, 321-357 (2002)
7. Han H., Wang W. Y., Mao B. H.: Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning. In: International conference on intelligent computing. pp. 878-887. Springer (2005)

8. Maciejewski T., Stefanowski J.: Local neighbourhood extension of SMOTE for mining imbalanced data. In: 2011 IEEE symposium on computational intelligence and data mining (CIDM). pp. 104-111. IEEE (2011)
9. Fabbri R. et al.: 2D Euclidean distance transform algorithms: A comparative survey. ACM Computing Surveys (CSUR) **40**(1), 1-44 (2008)
10. Bradley D., Roth G.: Adaptive thresholding using the integral image. Journal of graphics tools **12**(2), 13 21 (2007)
11. Ronneberger O., Fischer P., Brox T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention. pp. 234-241. Springer (2015)