Adversarial OverSampling for imbalanced image data classification

Adam Wojciechowski, Mateusz Lango

Motivation

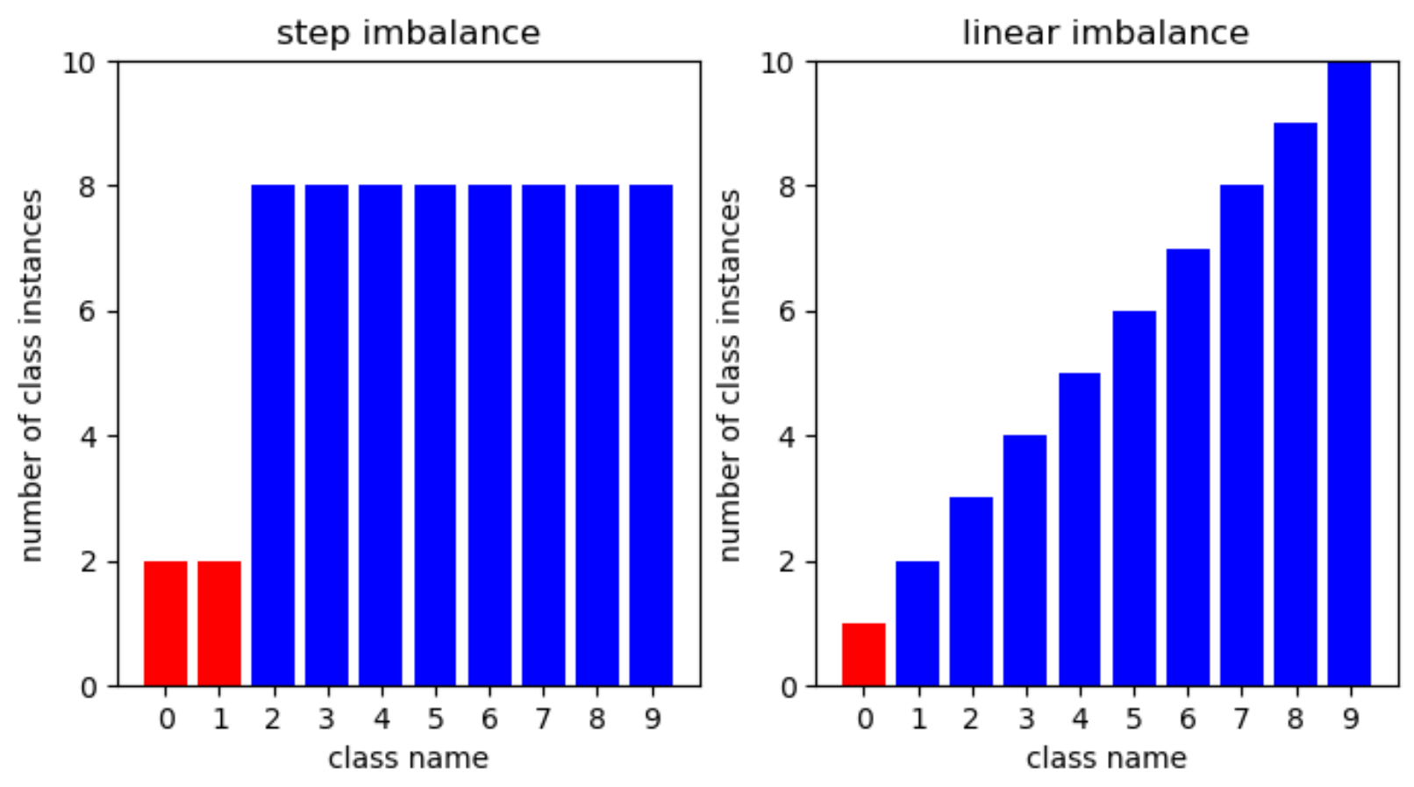
Multi-class imbalance is present in many real world machine learning classification problems. Unfortunately, popular methods for imabalanced data do not bring satysfying effects when applied to image datasets, therefore communnity concerning computer vision remains in need of efficient methods dealing with low model generalization levels and low multi-class accuracy metrics of underrepresented classes. We present the Adversarial OverSampling (AOS), an algorithm, which simultaneously upsamples and performs dataset and model specific data augmentation by generating adversarial examples from minority classes during training deep convolutional neural networks (CNNs) on multi-class imbalanced image datasets.

Multi-class imbalance

Multi-class imbalance is a common data characteristic for image classification problems, since it rarely happens, that the number of instances in every class is equal to each other. Researchers focus on two types of multi-class imbalance. *Step imbalance* is defined by parameters $\mu = \frac{|\{i \in \{1,...,N\}: C_i \text{ } is \text{ } minority\}|}{N}$, where:

- C_i is a set of examples in class i
- N is the total number of classes

and $\rho = \frac{max_i\{|C_i|\}}{min_i\{|C_i|\}}$. We call the second type of imbalance the *gradual imbalance*. It is defined by ρ and the number of instances in individual classes, that can be approximated by some function. For simplicity, below, we consider linear function, which creates *linear imbalance*.



Not addresing imbalance often has detrimental effects on training effectiveness resulting in low-quality recognition of minority classes. Also, the greater the μ or ρ parameter, the harder it is to train a model.

Related methods

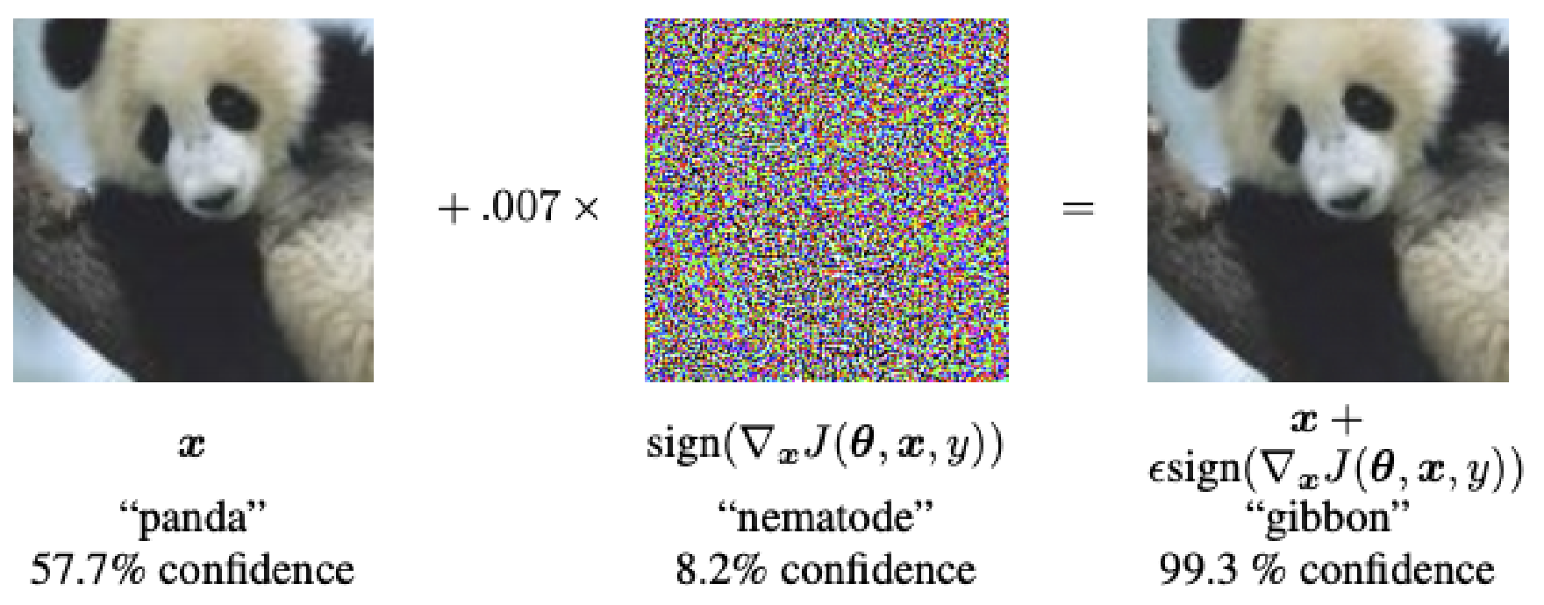
A Random OverSampling is a method, that can be performed by randomly choosing images from minority classes and copying them into the dataset. There exist more advanced oversampling methods like Synthetic Minority Oversampling TEchnique (SMOTE), which generates new images by per pixel linear interpolations of original images from the minority classes.

image a
SMOTE image
image b

All of the known methods are imperfect when apllied to image datasets. ROS is a valid and safe to use method, but it is also primitive, and when used alone, does not bring satisfying improvements. SMOTE can even lead to the degradation of classification performance, which can be partially explained by the fact, that the Euclidean distance used by it is not appropriate for image data.

Adversarial examples

Adversarial example is a general term for an transformed image, in a way, such that it is indistinguishable from original image to the human eye, yet attempt of classifying it results in misclassification. They are used to mislead CNNs often with malicious intentions.


+ .007 ×
=
$\boldsymbol{x}$
"panda"
57.7% confidence
$\text{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y))$
"nematode"
8.2% confidence
$\boldsymbol{x} + \epsilon\text{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y))$
"gibbon"
99.3 % confidence

There are different methods for creating adversarial examples. One of them, which is used in our implementation of AOS, is called Fast Sign Gradient Method (FSGM). It computes the gradient of the loss function with respect to the input image and uses it to create new image by adding it with certain intensity to the original image. The FSGM formula is as follows:

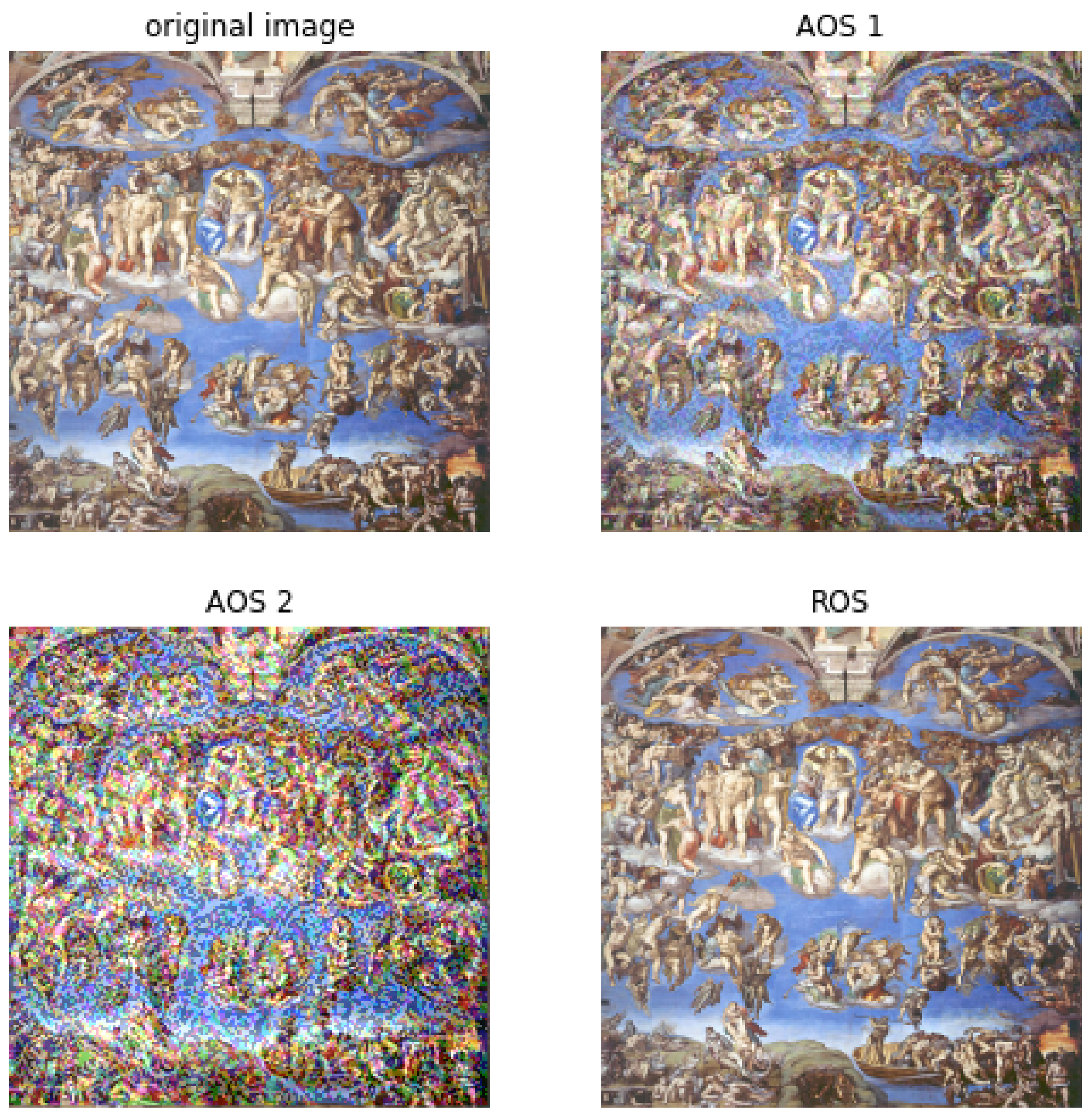
$$x' = x + \epsilon \cdot sign(\nabla_x J(\Theta, x, y))$$

where:

- x is original image
- ϵ is parameter controlling adversarial intensity
- Θ is model parameters
- y is label corresponding to original image
- x' is adversarial example
- J is Loss function

Adversarial OverSampling algorithm

Adversarial OverSampling is a hybrid method for addressing data imbalance, operating on data level as well as algorithmic level. The method needs two parameters: adversarial intenisty ϵ and augmentation intensity p. The method applies Random OverSampling during pre-training phase, creating base for further transformations. Subsequently, at the beggining of each batch, pick images from minority classes with probability $\mathcal{P}(\mathcal{X} \sim \mathcal{B}(1, p))$ and compute their signed gradients and add them to original images with intensity ϵ. Finalize by forwardpassing, backpropagating and updating Θ. Repeat.



In the figure above there is depicted an examplary set of instances of minority class within one batch. 'AOS 1' has $\epsilon = .05$ and 'AOS 2' has $\epsilon = .15$.

```
Algorithm 1: Adversarial OverSampling
$\epsilon$ ←input;
$p$ ←input;
batches←ROS(batches);
while batches_left is true do
    current_batch←next(batches);
    foreach image in current_batch do
        if image in minority_classes then
            decision ← randomize($\mathcal{P}$);
            if decision=1 then
                mask ←sign($\nabla_x J(\Theta, x, y)$);
                image ← $image$+$\epsilon$·mask;
            end
        end
    end
    forwardpass();
    backpropagate();
    update($\Theta$);
end
```

Results

In our research, we focused on comparing AOS to baseline data *BASE* and Randomly OverSampled data. Below we report the results of our experiments where the classification performance was measured by *F1-score*, which consolidates both *precision* and *recall* of the CNN, obtained from test sets. In Table 1 we present results on Intel's 'Image Scene Classification of Multiclass' dataset [3] and CIFAR-10, with imbalance parameters, which can be encountered in naturally imbalanced datasets. Let $\rho = 20$ and $\rho = 60$ respectively, $\epsilon = 0.007$ for both.

Table: 1

[%]	μ	**BASE**	**ROS**	**AOS**
INTEL	**0.4**	72.68	71.66	73.69
CIFAR-10	**0.4**	46.34	60.18	60.77
INTEL	**0.6**	61.32	62.48	66.33
CIFAR-10	**0.6**	35.77	52.67	54.10
INTEL	**0.8**	61.12	64.56	65.16
CIFAR-10	**0.8**	31.90	49.84	51.18

In Table 2 we present more extreme setup to emphasize the differences between performances of those approaches. Let $\rho = 500$ and $\epsilon = 0.1$.

Table: 2

[%]	μ	**BASE**	**ROS**	**AOS**
MNIST	**0.3**	83.39	86.60	92.92
CIFAR-10	**0.3**	44.66	43.83	45.41
MNIST	**0.9**	76.38	79.69	83.51
CIFAR-10	**0.9**	2.35	7.59	16.26

Conclusion

We propose a method for addressing multi-class imbalance applicable to image data and CNNs. Our method prooved to bring results superior to compared ones. We further conjecture, that the greater imbalance gets, the more efficiently our method handles the imbalance, for given problem.

Bibliography

[1] Mateusz Buda, Atsuto Maki, Maciej A. Mazurowski A systematic study of the class imbalance problem in convolutional neural networks 2017, Oct 15

[2] Ian J. Goodfellow, Jonathon Shlens, Christian Szegedy Explaining and Harnessing Adversarial Examples 2015, March 20

[3] https://www.kaggle.com/puneet6060/intel-image-classification

PUT
Poznan University of Technology

GHOST
Group of Horribly Optimistic STaticticians

PSNC
Poznan Supercomputing and Networking Center