Image classification methods for the imbalanced data

The aim of this project is to carry out research in the area of the imbalanced data classification in the image recognition task. The data can be described as imbalanced if the number of instances belonging to one of the considered classes (the majority class) is larger than the number of instances of one of the other classes (the minority class). Standard machine learning algorithms are poorly equipped for dealing with the imbalance in data, usually producing a bias towards the the majority class. At the same time, data imbalance is prevalent in the practical setting, where most of the datasets display some level of imbalance between the classes. Because of that, a significant amount of research was dedicated to the problem of dealing with the imbalance in data. Various approaches to dealing with the imbalance have been proposed in the literature. Between them, two main methodologies can be distinguished. Firstly, the data-level methods, in which the training data of the algorithm is manipulated in some manner, either by increasing the number of the minority examples (oversampling) or decreasing the number of the majority examples (undersampling), with the goal of balancing the class distributions. Secondly, the classifier-level methods, in which the standard classification algorithms are adjusted on incorporate the information about the data imbalance. However, despite the abundance of a various methods of dealing with the imbalance, many of them are ill-suited for the image data.

The issue of the image data imbalance affects many practical domains, such as: histopathology, autonomous vehicles, face recognition, remote sensing and urban rescue. Due to the prevalence of the issue, developing effective methods of dealing with the imbalance in the image recognition task can have great theoretical and practical impact on the whole field of the image recognition.

In this project we form a research hypothesis that **it is possible to design effective strategies for dealing with the data imbalance in the image recognition task**. To evaluate the validity of this hypothesis we will propose both the data-level and classifier-level methods for the image data. The designed methods will be focused on, but not limited to, being usable with the convolutional neural networks, which became the dominant paradigm in the image recognition task. In the project we will explore the following main research directions:

1. Research into adapting the training process of the convolutional neural networks to accommodate for the data imbalance.
2. Research into adjusting the existing data resampling strategies to the limitations of image data.
3. Research into designing a novel data resampling strategies based on the neural structures and the specific properties of the images.