Abstract

Title: *Strategies for dealing with small number of training samples in classification problem with unstructured data*

Insufficient amount of data or gaps in data are a common and frequently appearing problem in context of classification of unstructured data. Unadressed, those can lead to errors in results and eventually drive incorrect conclusions. Fortunately, advance in research offers various solutions to solve this problems.

The dissertation discusses various groups of methods to deal with a small number of training observations. The following three approaches have been proposed: *AttentionMix* (a data augmentation method, dedicated to the problem of text classification, that uses guidance when mixing observations), *StatMix* (a method of data augmentation using image statistics, dedicated to the problem of image classification in federated learning) and standard network training based on publicly available noisy data downloaded from the Internet.

The problem of classification of unstructured data often occurs when training models for various applications (e.g. classification of images or quantitative analysis of text). The goal in this case is to create a model that solves such a problem for unseen observations with the highest possible efficiency (i.e. has the ability to generalize).

Data augmentation is one of the regularization strategies for deep neural networks, which translates into better ability to generalize. Data augmentation is a strategy widely used in the literature to improve the results of the network and increase the ability to generalize. To a large extent, standard methods are used, such as rotation in the area of computer vision or substitution with synonyms in the area of natural language processing. Relatively recently, methods based on mixing observations have been proposed.

The methods proposed in this thesis show that there is still room for further research in this area. The *AttentionMix* method showed that mixing methods can be adapted to text data using engines dedicated to the text processing area. Thus showing flexibility in mixing methods for different modalities. The *StatMix* method, on the other hand, adapted the mixing process to the problem of federated learning where, in order to protect the

privacy of observations. In order to achieve that a significant reduction in the amount of data sent and used in the data augmentation process was proposed.

The results of the experiments showed that the methods of data augmentation by mixing observations, apart from their original application, are also flexible and suitable for other problems where their use translates into an increase in the effectiveness of the learned models. In addition, it is also shown that high-quality models can be trained using noisy data downloaded from the Internet.

Key words: *data augmentation, mixing observations, webly data, computer vision, natural language processing, federated learninig*