ABSTRACT

Semi-supervised learning (SSL) on class-imbalanced data poses a significant challenge. A common issue SSL methods face is that the model tends to focus on the class with the most data. The Class-Rebalancing Self-Training (CReST) method has been proposed. It aims to prevent the model from focusing on the class with the most data by updating it with a few examples using pseudo labels. However, the CReST method has not been widely studied. Because data is a scarce and costly resource, this undergraduate thesis will analyze the changes every five generations of SSL when CReST is applied.

In this undergraduate thesis, the author analyzes the workings of the CReST implementation in SSL to improve performance using a dataset with an imbalanced class distribution. The author employs the CIFAR 10 long-tailed dataset to test the performance of SSL using Python programming language on the Google Colab platform. CReST repeatedly retrains a baseline SSL model for every generation by extending the labeled set with pseudo-labeled samples from an unlabeled set, where minority class pseudo-labeled examples are picked more frequently according to predicted class distribution. To see what occurs during the implementation of CReST on SSL, the author plots diagrams of the changes every five generations by comparing the Recall, Precision, average accuracy recall per class, and data changes in each generation.

The results of the testing of the Class Rebalancing Self Training (CReST) Framework for imbalanced semi-supervised learning on the CIFAR-10 Long-Tailed dataset showed that the best generation was Generation 16 with an Average Accuracy Recall Per Class of 0.768. The study also revealed the reduction of pseudo-labels in the majority class and an increase in the minority class, as well as a decrease in precision in the majority class and an increase in recall in the minority class. The experiment also showed a decrease in Average Accuracy Recall, Per Class after Generation 16. Further research suggestions include addressing the over-sampling issue and exploring the application of the CReST framework in other areas of machine learning and AI.

Keywords: CReST, Semi-Supervised Learning, imbalance data, pseudo label,Semi-Supervised Learning generation