INFORMATICS INSTITUTE OF TECHNOLOGY

In Collaboration with

UNIVERSITY OF WESTMINSTER

A Novel Hybrid Deep Learning Technique to Address Class Imbalanced Datasets

A dissertation by

Mr. Qusai Saifuddin Sadikot

Supervised by

Mrs. Nipuni Perera

Submitted in partial fulfilment of the requirements for the BSc (Hons) in Computer Science degree at the University of Westminster.

April 2021

Abstract

Machine learning is one of the most prominent frontiers in artificial intelligence. Deep learning, which is a part of machine learning, has blown up in popularity in the scientific community, owing to the paper published by Geoffrey Hinton (Hinton, Osindero and Teh, 2006), due to its generalizing ability to complex and high dimensional datasets.

A big problem that any type of learning faces is the problem of class imbalanced data. This problem may significantly impact the performance and generalizing ability of ML / deep learning models. This is researched extensively with machine learning algorithms, but there is limited empirical and theoretical research with regards to deep learning, as this is still a relatively new field. An imbalanced learning technique that combats class imbalanced datasets with minimal fine-tuning of algorithm parameters or data relative to the level of imbalance will prove extremely useful in the field of deep learning.

The proposed solution is capable of addressing class imbalance with datasets that are severely skewed, as well as moderately skewed, and is experimented with different types of deep learning architectures and datasets. A novel imbalanced learning technique is proposed which includes a custom loss function to reduce the impact of easily classified instances on the total batch-wise loss. Along with that, a restricted dynamic sampling technique is introduced which dynamically alters the training set every epoch based on the class-wise F1 scores, calculated in a one-versus-all manner, obtained for the previous epoch. This technique provides performance improvements with deep learning models without any need to alter the training data distribution before training or alter the model components. It can also be extended to a variety of imbalance levels and not just significantly skewed datasets with high imbalance levels.

Keywords: Class imbalance, Data sampling, Imbalanced learning, Deep learning, Machine learning, Neural networks.