Evaluating data reduction techniques for supervised training

Yuwen Heng

Master of Science
Data Science, Technology, and Innovation
School of Informatics
University of Edinburgh
2020

Abstract

Training deep neural networks can be resources-consuming. The budget required is increasing with the size of the dataset. During the past few decades, many research is dedicated to developing training procedures to accelerate the convergence speed of deep learning. However, we still need the whole dataset to train the network and paying for a large dataset may not pay back well if we can use a smaller subset to achieve an acceptable performance. To solve this issue, we first adapted and evaluated three methods, Patterns by Ordered Projections (POP), Enhanced Global Density-based Instance Selection (EGDIS), and Curriculum Learning (CL), to reduce the size of two image datasets, CIFAR10 and CIFAR100, for the classification task. Based on the analysis, we present our two contributions: the Weighted Curriculum Learning (WCL) and a trade-off framework. The WCL outperforms POP and EGDIS in terms of both classification accuracy and time complexity. It achieves comparable performance compared with CL while keeping a portion of hard examples. The trade-off framework selects a subset of samples according to the acceptable relative accuracy and the dataset. In addition, the framework is also extended to predict the number of samples needed to achieve a particular accuracy with a given subset.

Acknowledgements

First and foremost, I would like to express my deepest gratitude to my supervisor, Dr Yang Cao, for offering me the opportunity to work with him on such an attracting and challenging project. His encouragement and valuable guidance helped me tackle the obstacles in my research path.

Furthermore, special thanks go to Professor Bob Fisher, Dr Pavlos Andreadis at the University of Edinburgh and Dr Jiacheng Ni at IBM for sharing me their knowledge about computer vision and deep learning. The programming skills and coursework experience that I learnt from them helped me to organise the experiments well.

Finally, I would like to send my love to my fiancee Danni Li for her accompany during the past three years . I wouldn't have the chance to study full-time without her full support.

Table of Contents

Chapter 1

Introduction

1.1 Motivation

1.2 Research Goal

1.3 Significance

1.4 Beneficiaries

Chapter 2

Background Research

In this chapter, we begin with presenting the necessary background to understand the deep convolutional neural network (deep CNN) and data reduction methods, as well as, other ideas required to understand our research method. In the next section we we start with the structure of CNN and the gradient descent training procedure. We then discuss the modern research outputs that can speed up the training procedure and outline their deficiencies. Next, we review the data reduction literature and present a CNN data reduction framework - use the network pre-trained on ImageNet to extract low-dimensional features and run the data reduction methods on extracted features. Furthermore, we cover the existing trade-off framework in the context of maximum-likelihood estimation machine learning algorithms and explain why it is not suitable for deep neural network. Finally, we present TAPAS, which is an accuracy predictor for deep neural network without training and has several properties that make it useful to build our trade-off framework.

2.1 Deep Convolutional Neural Network

2.2 Data Reduction Algorithms

2.3 Trade-off Framework

2.4 Accuracy Predictor

Chapter 3

Adapted Data Reduction Methods

3.1 Patterns by Ordered Projections

3.2 Enhanced Global Density-based Instance Selection

3.3 Curriculum Learning

3.4 Weighted Curriculum Learning

Chapter 4

Data Reduction Evaluations

4.1 Time Complexity

4.2 Classification Accuracy

Chapter 5

Trade-off Framework

5.1 Subset Selection Framework

Chapter 6

Trade-off Evaluation

6.1 Relative Accuracy Precision

Chapter 7

Conclusion and Future Work

Bibliography