MITIGATING CLASS IMBALANCE IN LONG-TAILED VISUAL RECOGNITION THROUGH THE USE OF INTRINSIC DIMENSIONALITY

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES
OF
MIDDLE EAST TECHNICAL UNIVERSITY

BY

ÇAĞRI ESER

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR
THE DEGREE OF MASTER OF SCIENCE
IN
COMPUTER ENGINEERING

JANUARY 2024

Approval of the thesis:

MITIGATING CLASS IMBALANCE IN LONG-TAILED VISUAL RECOGNITION THROUGH THE USE OF INTRINSIC DIMENSIONALITY

submitted by **ÇAĞRI ESER** in partial fulfillment of the requirements for the degree of **Master of Science in Computer Engineering Department, Middle East Technical University** by,

Prof. Dr. Halil Kalıpçılar
Dean, Graduate School of **Natural and Applied Sciences** _____

Prof. Dr. Mehmet Halit S. Oğuztüzün
Head of Department, **Computer Engineering** _____

Prof. Dr. Sinan Kalkan
Supervisor, **Computer Engineering, METU** _____

Assoc. Prof. Dr. Emre Akbaş
Co-supervisor, **Computer Engineering, METU** _____

Examining Committee Members:

Assoc. Prof. Dr. Ramazan Gökberk Cinbiş
Computer Engineering, METU _____

Prof. Dr. Sinan Kalkan
Computer Engineering, METU _____

Prof. Dr. Erkut Erdem
Computer Engineering, Hacettepe University _____

Date:11.01.2024

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Name, Surname: Çağrı Eser

Signature :

ABSTRACT

MITIGATING CLASS IMBALANCE IN LONG-TAILED VISUAL RECOGNITION THROUGH THE USE OF INTRINSIC DIMENSIONALITY

Eser, Çağrı
M.S., Department of Computer Engineering
Supervisor: Prof. Dr. Sinan Kalkan
Co-Supervisor: Assoc. Prof. Dr. Emre Akbaş

January 2024, 67 pages

Natural image datasets used in the field of visual recognition are often imbalanced in terms of the number of samples between class categories in the dataset. This problem, defined commonly as *class imbalance*, results in sub-optimal performance on these under-represented classes for deep learning models which are trained with such datasets. Attempts to remedy this problem include re-sampling, loss re-weighting and other calibration methods which generally use the number of samples as the primary factor in their mitigation strategy, ignoring other factors. In this thesis, we argue that model performance in a dataset depends on the difficulty of individual class categories as well as the number of samples present in the dataset. We use the concept of intrinsic dimensionality (ID) to express this idea of difficulty and explore the different definitions and estimation strategies for calculating ID inside a dataset. We further investigate the relationship between ID and class imbalance. Lastly, we report our results on using class ID estimation for class imbalance mitigation on long-tailed variations of natural image datasets – MNIST-LT, CIFAR-10-LT and CIFAR-100-LT.

Keywords: class imbalance, intrinsic dimension, long-tailed visual recognition, class imbalance mitigation

ÖZ

UZUN KUYRUKLU GÖRSEL TANIMADA SINIF DENGESİZLİĞİNİN ÖZ BOYUT KULLANIMI İLE AZALTILMASI

Eser, Çağrı
Yüksek Lisans, Bilgisayar Mühendisliği Bölümü
Tez Yöneticisi: Prof. Dr. Sinan Kalkan
Ortak Tez Yöneticisi: Doç. Dr. Emre Akbaş

Ocak 2024, 67 sayfa

Görsel tanıma alanında kullanılan doğal görüntü veri kümeleri sıklıkla veri kümesi içerisindeki sınıf kategorileri arasındaki örnek sayısı bakımından dengesiz durumdadır. *Sınıf dengesizliği* olarak tanımlanan bu problem, bu veri kümeleri ile eğitilen derin öğrenme modellerinin az temsil edilen sınıflarda idealin altında performans göstermesine sebep olur. Bu problemin çözülmesi için yeniden örnekleme, yeniden kayıp fonksiyonu ağırlıklandırma ve çeşitli kalibrasyon yöntemleri gibi genellikle temel etken olarak örnek sayısını kullanılıp diğer etkenler ihmal edilmektedir. Bu tezde bir veri kümesi içerisindeki model performasının veri kümesi içerisindeki örnek sayısı ile birlikte sınıf kategorilerinin bireysel zorluğuna da bağlı olduğu savunuyoruz. Bu zorluk düşüncesini iç boyut (İB) kavramını kullanarak ifade ediyoruz ve bir veri kümesi içerisinde İB hesaplanmasında kullanılan farklı tanımları ve stratejilerini araştırıyoruz. İç boyut ve sınıf dengesizliği arasındaki ilişkiyi araştırıyoruz. Son olarak, sınıf İB tahminlerini kullanarak MNIST-LT, CIFAR-10-LT ve CIFAR-100-LT uzun-kuyruklu doğal görüntü kümelerinde sınıf dengesizliği azaltma deneylerimizde elde ettiğimiz

sonuçları sunuyoruz.

Anahtar Kelimeler: sınıf dengesizliği, öz boyut, uzun kuyruklu görsel tanıma, sınıf dengesizliği azaltılması

This thesis is dedicated to my loved ones, who have always supported me in any challenge that I had to overcome.

ACKNOWLEDGMENTS

Firstly, I would like to acknowledge Prof. Dr. Sinan Kalkan, Assoc. Prof. Dr. Emre Akbaş and Dr. Kemal Öksüz for their support, feedback and encouragement throughout this master's program. It is with their continual guidance that I have grown as a researcher and as a person.

Secondly, I would like to thank all of the members of our research group, as their opinions and discussions have aided my work to blossom into the thesis that it is now.

I would also like to extend my thanks to Kuartis for their understanding during this process and for offering hardware support whenever there is any need.

Finally, I would like to thank my mother, my father and my brothers for their continual support throughout this process, as they have always done before.

This work was partially supported by the Scientific and Technological Research Council of Turkey (TÜBİTAK) through the project titled "Addressing Class Imbalance in Visual Recognition Problems by Measuring Class Imbalance and Using Epistemic Uncertainty (DENGE)" (project no. 120E494).

TABLE OF CONTENTS

ABSTRACT v

ÖZ vii

ACKNOWLEDGMENTS x

TABLE OF CONTENTS xi

LIST OF TABLES xv

LIST OF FIGURES xvi

LIST OF ABBREVIATIONS xix

CHAPTERS

1 INTRODUCTION 1

1.1 Motivation 1

1.2 Problem Definition and Scope of the Thesis 3

1.3 Contributions 3

1.4 The Outline of the Thesis 4

2 BACKGROUND AND RELATED WORK 5

2.1 Imbalance Problems in Computer Vision 5

2.2 Long-tailed Visual Recognition 5

2.3 Class Imbalance Mitigation 6

2.3.1 Re-sampling 7

2.3.2 Re-weighting 7
2.3.3 Margin-based Methods 8
2.3.4 Other Mitigation Methods 8
2.4 Intrinsic Dimensionality 9
2.4.1 Model ID 10
2.4.2 Data ID 11
2.4.2.1 Projective Methods 11
2.4.2.2 Geometric Methods 12
2.5 Datasets 13
2.5.1 Long-Tailed Datasets 13
2.5.2 CIFAR 13
2.5.3 MNIST 14
2.5.4 Other Datasets 14
2.6 Performance Evaluation 14
2.6.1 Top-1 Accuracy 14
2.6.2 F1 Score 15
3 UNDERSTANDING COMPLEXITY THROUGH MODEL-BASED INTRINSIC DIMENSIONALITY 17
3.1 Model-based Intrinsic Dimensionality 17
3.1.1 Calculating Model-based ID for Individual Class Categories . 17
3.1.2 Model-based ID and Class Complexity 18
3.1.3 Effects of Model Architecture and Class Imbalance 19
3.1.4 Drawbacks of Model-based ID 20

4 A DETAILED LOOK INTO DATA-BASED INTRINSIC DIMENSIONALITY 27

4.1 Data-based Intrinsic Dimensionality 27

4.1.1 Definition of Data-based ID 27

4.2 ID Estimators 27

4.2.1 Maximum Likelihood Estimation 28

4.2.2 TwoNN 29

4.2.3 FisherS 30

4.3 Extending Data ID to Class-Based ID 32

4.4 Comparison of ID Estimators 32

4.4.1 Synthetic Datasets 32

4.4.2 Natural Image Datasets 35

4.4.3 Comparing Data-based ID to Model-based ID 37

4.5 Data-based ID as an Imbalance Measure 38

4.5.1 Requirements for an ID-based Imbalance Measure 38

4.5.2 Requirement Analysis 39

4.5.2.1 Requirement 1 39

4.5.2.2 Requirement 2 40

5 MITIGATING CLASS IMBALANCE THROUGH INTRINSIC DIMENSIONALITY 47

5.1 Experiment Setup 47

5.1.1 Loss Re-weighting with Data-based ID 48

5.1.2 Re-sampling with Data-based ID 48

5.1.3 Re-margining with Data-based ID 48

5.2 Class Imbalance Mitigation Experiments 49
5.2.1 Class Imbalance Mitigation though Re-weighting 50
5.3 Class Imbalance Mitigation though Re-sampling 51
5.4 Class Imbalance Mitigation though Re-margining 51
5.5 Summary 52
6 CONCLUSION 55
6.1 Limitations and Future Work 56
REFERENCES 57
APPENDICES
A EXPERIMENT DETAILS 65
A.1 Experiment setup 65
A.2 Additional Results 66

LIST OF TABLES

TABLES

Table 5.1 Top-1 Accuracy results for re-weighting methods on CIFAR-10-LT . 50

Table 5.2 Top-1 Accuracy results for re-weighting methods on CIFAR-100-LT 51

Table 5.3 Top-1 Accuracy results for re-sampling methods on CIFAR-10-LT . 51

Table 5.4 Top-1 Accuracy results for re-sampling methods on CIFAR-100-LT . 52

Table 5.5 Top-1 Accuracy (x100) results for margin-based methods on long-tailed CIFAR datasets. 52

Table 5.6 Top-1 Accuracy (x100) results for margin-based methods on long-tailed CIFAR datasets. 53

Table 5.7 Top-1 Accuracy results using logit adjustment on long-tailed CIFAR datasets. 53

Table A.1 F1 results for re-weighting methods on CIFAR-10-LT 66

Table A.2 F1 results for re-sampling methods on CIFAR-10-LT 66

Table A.3 F1 results for re-weighting methods on CIFAR-100-LT 66

Table A.4 F1 results for re-sampling methods on CIFAR-100-LT 67

LIST OF FIGURES

FIGURES

Figure 1.1 An illustration of a dataset with a long-tailed class distribution. Retrieved from [1]. 2

Figure 2.1 Class sample distribution in LVIS [2], a long-tailed instance segmentation dataset. Retrieved from [1]. 6

Figure 2.2 An illustration of ID estimation of a synthetic spiral dataset at different locations and resolutions. Retrieved from [3]. 11

Figure 3.1 Comparison of performance metrics for 10 one-vs-all sub-tasks (one for each class category, fully-connected model) trained on CIFAR-10-LT with an imbalance factor of 0.1. 18

Figure 3.2 Comparison of recall scores for one-vs-all digit classification sub-tasks with a LeNet architecture on MNIST-LT under various imbalance factors. 19

Figure 3.3 An example of the model-based ID estimation process for a LeNet model on the MNIST dataset. 20

Figure 3.4 Comparison of d_{int90} estimates for one-vs-all experiments on MNIST-LT with varying imbalance factors. The baseline model is fully-connected model with a depth of 1 and a layer width of 50. The imbalance ratios of the datasets are 1 (red), 10 (blue) and 50 (purple), respectively. 21

Figure 3.5 Averaged F1 scores for LeNet models trained on MNIST-LT under various imbalance factors, with randomized long-tail ordering. 5 random orderings were chosen for each subtask. 21

Figure 3.6 Comparison of ID estimations and F1 scores for one-vs-all LeNet models on MNIST-LT with IF = 0.05. (a) Default ordering long-tailed version. (b) Randomized ordering averaged ($N = 5$) long-tailed results. 23

Figure 3.7 Effects of changing model width for a fully-connected model on balanced MNIST 24

Figure 3.8 Effects of changing model depth for a fully-connected model on balanced MNIST 24

Figure 3.9 Effects of increasing imbalance ratio for a fully-connected model on MNIST 25

Figure 3.10 An example of a failed model-based ID estimation process for a LeNet model on the CIFAR-10 dataset. 25

Figure 4.1 Point-wise ID estimations of MLE [4] on mixtures of Gaussian distributions. (a) Equal mixture of two Gaussian distributions. (b) Mixture for two Gaussian distributions with changing ratio of sample contribution. Here, a dataset with a ratio of 0.0 only includes samples from the first component, 1.0 only includes samples from the second component, and 0.5 means equal contribution of samples from both components. 33

Figure 4.2 Estimating the ID of CIFAR-10 class categories with MLE [4] (a) Using the isolated manifold of each class category. (b) Using the manifold created by the entire dataset. 34

Figure 4.3 ID estimates of MLE [4] on a simple Gaussian distribution with an ID of 5, in a space with extrinsic dimensionality of 10. We use the unbiased MLE estimator of [5], highlighted in blue. 35

Figure 4.4 ID estimates of MLE [4] for synthetic Gaussian data and their expected values. 36

Figure 4.5 The trend of ID estimates of MLE [4] for synthetic Gaussian data of increasing dimensionality. 37

Figure 4.6 ID estimates of several methods for balanced CIFAR-10. 38

Figure 4.7 ID estimates of LIDL [6] CIFAR-10-LT with different imbalance ratios. 39

Figure 4.8 ID estimates of FisherS on CIFAR-10-LT with different imbalance ratios. 40

Figure 4.9 ID estimates of several methods for balanced CIFAR-10. 41

Figure 4.10 Normalized ID estimates of several methods for balanced CIFAR-10. 41

Figure 4.11 ID estimates of several methods for CIFAR-10-LT with an imbalance ratio of 100. 42

Figure 4.12 Comparison of Model-based and Data-based ID estimation on MNIST. 42

Figure 4.13 Comparison of model-based and data-based ID estimation on CIFAR-10-LT under increasing class imbalance. 43

Figure 4.14 Average ID estimates of various estimators for CIFAR-10-LT under increasing class imbalance. 44

Figure 4.15 Average ID estimates of MLE for individual class categories of MNIST-LT under increasing class imbalance. 44

Figure 4.16 Average ID estimates of FisherS for different datasets under increasing class imbalance. 45

Figure 4.17 ID estimates of FisherS for individual class categories of datasets under with increasing percentage of identical samples. 45

LIST OF ABBREVIATIONS

2D	2 Dimensional
3D	3 Dimensional
CE	Cross Entropy
LT	Long-tail
ID	Intrinsic Dimension
IR	Imbalance Ratio
IF	Imbalance Factor
MLE	Maximum likelihood estimator

CHAPTER 1

INTRODUCTION

1.1 Motivation

The field of visual recognition has witnessed unprecedented progress with the emergence of deep learning [7, 8, 9]. Aside from visual recognition, deep learning based models have been used in several other fields with great success (including autonomous vehicles [10], healthcare [11], biology [12], and game solving [13], to name a few). In general, such models are trained with data of different modalities and characteristics, in which the distribution of classes within the datasets are artificially balanced [1].

Datasets which have substantial imbalance between class categories are described as being *long-tailed*, where classes with a relatively higher number of samples are called *head* classes, and those with lower number of samples are called *tail* classes. Figure 1.1 illustrates an example of a long-tailed natural image dataset, where more common animals (e.g., dog, bird) are head classes and less frequent animal categories (e.g., eagle) are in the long tail.

Aside from the aforementioned class imbalance problem, other types of imbalance are also present in widely-used datasets. Such examples include spatial imbalance, scale imbalance and foreground-background class imbalances, which are common problems in visual recognition tasks [14]. Mitigating such forms of imbalance is an active research area in the long-tailed visual recognition community.

Despite the plethora of approaches proposed in the literature for mitigating class imbalance, there is only a handful of studies measuring class imbalance. The literature often relies on cardinality as a measure of class imbalance, which, however, is limited

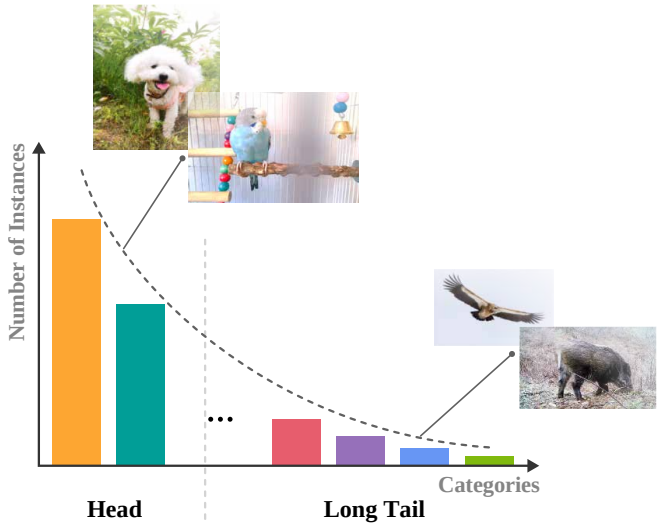


Figure 1.1: An illustration of a dataset with a long-tailed class distribution. Retrieved from [1].

since a class with less samples but distinctive features can easily be recognized with more accuracy than a class with more samples but ambiguous features. A promising approach for quantifying imbalance is *uncertainty estimation* [15, 16, 17], where the goal is to quantify the amount of uncertainty of the data distribution in order to increase model robustness around uncertain regions. Uncertainty is often split into two categories [18, 19]; aleatoric uncertainty which represents uncertainty caused by the inherent randomness of the environment, and epistemic uncertainty which represents uncertainty caused by the lack of available information.

In this thesis, we focus on the class imbalance problem present on the multi-class classification task. Therefore, our main motivation is to find a novel approach to the class imbalance problem and improve model robustness against class imbalance using this new approach. We accomplish this goal by using intrinsic dimensionality, a concept that attempts to quantify the inherent structure of a system. Our work includes an in-depth exploration of various definitions of ID and their interactions with tasks and datasets, and also includes several experiments that feature using ID alongside available mitigation methods that show that ID can be effective at increasing robustness against class imbalance.

We define the class imbalance problem more formally in the following section.

1.2 Problem Definition and Scope of the Thesis

We assume the standard supervised classification setting, in which we have a dataset consisting of N examples with pairs $\{(x_1, y_1), ..., (x_N, y_N)\}$, where $x_i \in X$ is a sample from the dataset and $y_i \in Y$ is the corresponding class category. The goal in gradient-based learning is to learn a mapping $f : X \rightarrow Y$ with parameters θ such that the following term is minimized:

$$\theta^* \leftarrow \arg\min_{\theta} \sum_{i=1}^{N} \mathcal{L}(f(x_i; \theta), y_i), \tag{1.1}$$

where $\mathcal{L}(\hat{y_i}, y_i)$ is a loss function (e.g., cross entropy) which is minimized when the predicted class category $\hat{y_i} = f(x_i; \theta)$ is the same as the ground truth category y_i.

This simple setup of training a model treats each sample and each class in the dataset equally, and does not factor in the different characteristics of individual class categories. In a long-tailed dataset, this leads to the model adjusting itself more to classes with more examples and thus under-performing for classes with a lower number of examples. This forms the basis of the *class imbalance* problem.

In terms of solving the class imbalance problem, there are multiple strategies found in common literature. These strategies are commonly grouped into categories such as weighted sampling, loss weighting and margin-based methods [1, 20, 21]. These categories are not comprehensive, and each method offers its own advantages and disadvantages. Moreover, the default parameters of weight-based strategies are generally calculated from the proportion of samples of each class in the dataset, which does not take into account the specific difficulties of each class category.

In this thesis, we focus on the argument that these methods should not solely rely on the number of samples, and propose that class-specific details of a dataset should be extracted and taken into consideration for potentially improved results.

1.3 Contributions

Our contributions are as follows:

- We analyze different types of intrinsic dimensionality measures of models and datasets in detail.
- We investigate the relationship between class imbalance and intrinsic dimensionality, and evaluate using ID as a measure for class imbalance.
- We further investigate the use of ID as a way to improve model robustness against class imbalance, and report our results on long-tailed variations of CIFAR-10 and CIFAR-100.

1.4 The Outline of the Thesis

This thesis is structured as follows:

- In Chapter 2, we provide background information for the class imbalance problem, a brief literature review of the common solutions against class imbalance, and a brief overview of intrinsic dimensionality.
- Chapter 3 includes our analysis of model-based ID, in which we explore the relation between the model-dataset-imbalance trio.
- In Chapter 4 we present a detailed overview of data-based intrinsic dimensionality methods and their performance in synthetic and natural image datasets, and evaluate data-based ID as a potential measure of class imbalance.
- In Chapter 5, we investigate possible uses of data ID to increase model robustness against class imbalance on long-tailed versions of CIFAR-10 and CIFAR-100 and report the results of our experiments on these datasets under different imbalance factors.
- Chapter 6 summarizes the contributions of this thesis, and briefly discusses limitations and future work.
- Appendix A includes further details of our experiment setup and additional results on using ID estimates in mitigating class imbalance.

CHAPTER 2

BACKGROUND AND RELATED WORK

In this chapter, we provide background and review relevant studies.

2.1 Imbalance Problems in Computer Vision

Visual recognition is a collection of a variety of different tasks which lends the opportunity for several forms of imbalance to occur. Furthermore, each sub-field of visual recognition has their own share of unique imbalance problems. As an example in object detection, Oksuz et al. [14] classify types of imbalances into groups such as class imbalance, spatial imbalance and scale imbalance, which is then categorized into several smaller subgroups of potential imbalance.

In the long-tailed visual classification setup, the primary type of imbalance is *class imbalance*, where there is a sizable discrepancy in the number of samples between different class categories. This situation, if not mitigated properly, causes models to focus more on classes with more samples and ignore classes with less examples to optimize their objective more effectively, leading to sub-optimal performance on tail classes.

2.2 Long-tailed Visual Recognition

Long-tailed visual recognition is the task of performing visual recognition (such as image classification, instance segmentation, scene recognition, etc.) on a dataset that follows a long-tail distribution. Zhang et al. [20] define the long-tail distribution as

a small portion of classes having a much larger number of sample points (i.e., head classes) but other classes being associated with only a few samples (i.e., tail classes). This notion of head classes and tail classes can be observed in Figure 2.1. Long-tailed learning is a challenging topic in visual recognition as models experience difficulty in learning under-represented classes with classical training environments which treat each class category equally.

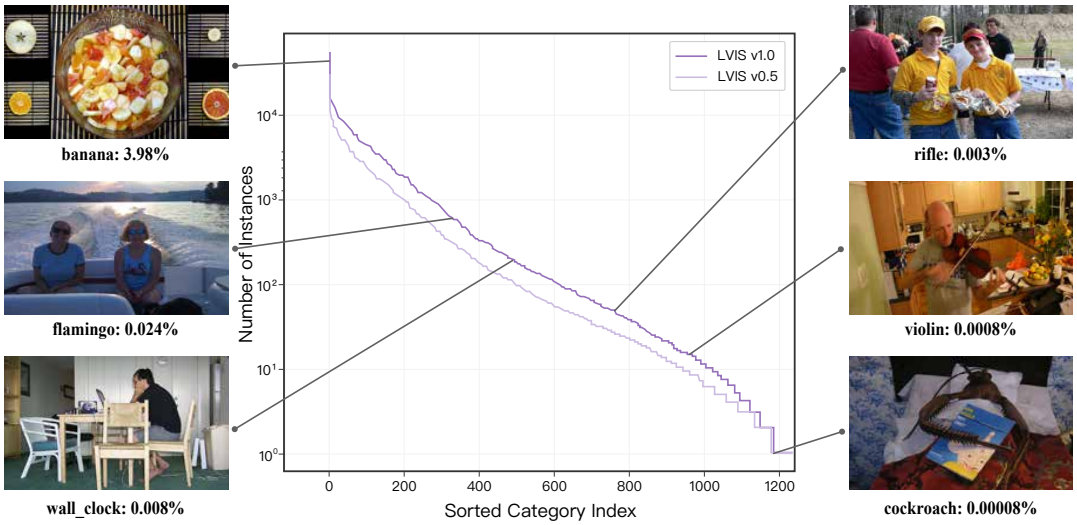


Figure 2.1: Class sample distribution in LVIS [2], a long-tailed instance segmentation dataset. Retrieved from [1].

2.3 Class Imbalance Mitigation

As class imbalance has been recognized to be a prevalent problem in long-tailed visual recognition, many studies have investigated possible solutions for mitigating this form of imbalance. Similarly to the categorization of imbalance types, mitigation strategies for class imbalance have been grouped into common categories such as re-sampling, loss re-weighting, and margin-based methods [20, 21]. We go into some classical methods of imbalance mitigation below.

2.3.1 Re-sampling

Deep neural networks for visual recognition tasks are conventionally trained with some form of stochastic gradient descent (SGD), where the model processes mini-batches of training data which consist of uniformly-sampled random examples from the training data. In this setup, the simplest idea to make a model focus more on tail classes is to increase the probability of sampling from tail classes (i.e., oversampling, possibly by copying samples from under-represented data [22]) and reducing the probability of sampling from the more represented classes (i.e., undersampling, possibly by removing samples [23]). Re-sampling strategies focus on the sampling probability of class categories to increase robustness against class imbalance. Class-balanced sampling gives each class category an equal probability of being sampled, removing the effect of number of samples per class [24]. Square-root sampling is an intermediate step between random and class-balanced sampling, using the square root of the number of samples for a class to determine sampling probability. Progressively-balanced sampling, as the name suggests, progressively changes the sampling probability of samples from uniform to class-based sampling [25, 26]. Nevertheless, most of these methods primarily focus on the number of samples of each class in the dataset.

2.3.2 Re-weighting

While re-sampling methods attempt to solve the imbalance problem in the sampling stage, re-weighting methods try to mitigate it by modifying the loss function. We assume the conventional setup in the multi-class classification setting from Section 1.2, in which case the total loss in the fixed re-weighting scheme is equal to

$$\sum_{i=1}^{N} \mathcal{L}(f(x_i;\theta), y_i) = \sum_{c}^{C} w_c \mathcal{L}_c, \tag{2.1}$$

where $c \in C$ is a class category in the dataset, $\mathcal{L}_c$ is the loss accumulated by class c in the non-weighted case and w_c corresponds to the weight term being applied for class c by the re-weighting strategy. Trivially, it is possible to use the inverse of number of samples for each class ($w_c = \frac{N}{n_c}$, where $N = max_c\{n_c\}$ and n_c is the number of samples for class c) as the fixed weight, which can be described as weighted softmax

in the conventional multi-class classification setting. Cui et al. [27] propose class-balanced loss, where this weighting is determined inversely by the *effective number* of samples of each class ($w_c = \frac{1-\gamma}{1-\gamma^{n_c}}$, where γ is a scaling parameter). In addition to class-based re-weighting, instance-based loss re-weighting is also possible. Focal loss [28] attempts to mitigate class imbalance by conditioning the trained model to more difficult examples using prediction hardness, while IB loss [29] modifies the weight of each sample based on its influence on the decision boundary.

2.3.3 Margin-based Methods

Margin-based methods [26, 30, 31] focus on modifying the margins of the class decision boundaries for increasing model robustness against class imbalance. Cao et al. [26] propose label-distribution-aware margin (LDAM) loss, in which the training labels are used directly to compute class margins; tail classes are assigned larger margins than head classes, leading to easier separation during model training. Similarly, DRO-LT [31] introduces distributional robustness loss, which punishes classes for not having compact representations in feature space, therefore also increasing the margin for tail classes. Khan et. al propose uncertainty-based margin learning [16], which combines re-margining with uncertainty quantification and computes the class margins from the uncertainty estimate of each class.

2.3.4 Other Mitigation Methods

Aside from re-sampling, re-weighting and margin-based methods, there exist other kinds of methods to handle class imbalance. These methods are flexible, and can be combined in a variety of ways to offer additional robustness against class imbalance.

Data augmentation methods attempt to handle the imbalance problem by modifying the data such that the effect of class imbalance is reduced. Naive image augmentation methods such as flipping, cropping and rotation are not very effective for improving performance in tail classes [1]. However, other augmentation strategies such as mixup training [32, 33, 34] have been shown to be effective against class imbalance. Similarly, SMOTE [35] and its variants over the years have been used for re-sampling

using data augmentation effectively and easily.

Two-stage methods (also known as decoupled training) [24, 36, 37] are another family of mitigation methods that focus on separating the representation learning (imbalanced training) stage from the classifier learning (balanced fine-tuning) stage. The first stage trains a feature extractor model that learns a good representation for the imbalanced data; this model is then fixed in-place and the second stage aims to improve tail class performance using the methods described in this section (such as *deferred* re-sampling and re-weighting).

Logit adjustment [38, 39, 40], is another form of class imbalance mitigation where the objective is to adjust the prediction logits of a trained model to improve tail class performance. Logit adjustment is flexible, and can be done during training or applied post-hoc for a trained model. LADE [40] is an example of a logit adjustment method that uses the test set labels to apply post-hoc adjustment to calibrate a trained model for the test set.

Among other mitigation methods are ensemble models [41, 42], which are both used frequently in imbalance mitigation as well as uncertainty quantification. Knowledge distillation [43, 44] is another family of methods where the aim is to transfer the knowledge from several teacher models into a student model. Metric learning [45, 46], meta-learning [47, 48] methods also exist, but are not as popular in the general literature.

2.4 Intrinsic Dimensionality

The field of pattern recognition and machine learning involves solving challenging problems in datasets of non-trivial number of dimensions. In this context, intrinsic dimensionality is associated with the number of dimensions being required to solve a particular task or for representing a set of data. Naturally, with such a general description, there have been many different definitions of intrinsic dimensions in the literature [49, 50]. Fukunaga [51] defines ID as the minimum number of parameters needed to note the observed properties of some set of data. Li et al. [52] define the ID of the objective landscape (which we refer to as *model*-based ID from now on) in a

deep learning problem as the number of free parameters required for a restricted deep learning model to reach a fixed margin (e.g., 90% or 100%) of the performance of a well-taught unrestricted model of the same architecture.

A common assumption in many studies is that data in high dimensions is not truly high-dimensional; that a dataset of dimension D is embedded in this high-dimensional space through a mapping from a manifold of (intrinsic) dimension d, with $d \ll D$ [4]. With this assumption, it is clear that intrinsic dimensionality is closely related to the manifold hypothesis; specifically, ID represents the topological dimensions of the data manifold [3]. We define this description of ID as *data ID* in further sections.

Depending on the estimator, ID can be estimated locally, globally or in a point-wise manner [50]. An example can be seen in Figure 2.2, where the estimated ID of the data can change depending on the location and scale of the estimator.

Intrinsic dimensionality has been associated with different areas such as dimensionality reduction [53], density estimation [54], manifold learning [55], and data analysis [56]. One important detail is that these works are generally about the ID of a dataset, and class-based ID analysis has not been explored in the literature. To the best of our knowledge, calculating the ID of individual class categories and using ID in an attempt to mitigate the effect of class imbalance in long-tailed visual recognition tasks is a novelty introduced by this work.

We briefly go over definitions of ID in the following subsections.

2.4.1 Model ID

In the gradient-based deep learning model context, Li et al. [52] define and calculate the intrinsic dimension of the objective landscape. Specifically, for an optimization problem of fixed dimension D, the ID of a solution is defined to be the co-dimension of the solution set inside $\mathbb{R}^D$. For estimating the ID in complicated problem spaces the authors initially train a model with D parameters. They then suggest random subspace training, in which they use the same model, but with only d free parameters ($d < D$), resulting in d degrees of freedom for model optimization. By incrementally increasing the number of free parameters d and comparing the restricted model's per-

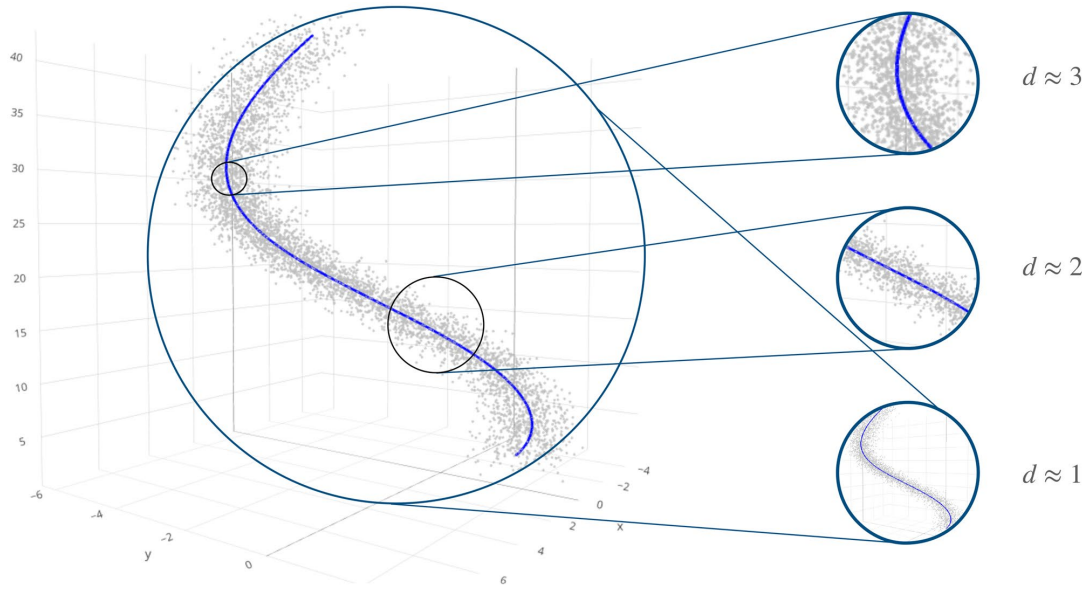


Figure 2.2: An illustration of ID estimation of a synthetic spiral dataset at different locations and resolutions. Retrieved from [3].

formance against the unrestricted original model, they find a restricted model with d_{int} free parameters which shows similar performance as the free model, and thus d_{int} is defined to be the estimated ID of this problem.

This definition of ID is unique in the fact that it is not just dependent on not just the dataset, but also on the model architecture and the selected loss function (and the selected learning task, as well).

2.4.2 Data ID

Data ID estimation methods are commonly categorized into two families: projective and geometric methods.

2.4.2.1 Projective Methods

Projective ID estimation methods estimate the number of dimensions of the manifold by projective manipulation of the data. Linear projection-based methods frequently involve Principal Component Analysis (PCA) [57], and its various variants, such as

probabilistic [58], bayesian[59], sparse, spherical, and local PCA [54, 60]. On the other hand, nonlinear projective methods also exist, such as nonlinear PCA, Isomap [56] and LLE [61]. In general, these methods are trained by minimizing the projection error, thereby finding the best subspace for the manifold.

2.4.2.2 Geometric Methods

Geometric ID estimation methods consist of strategies for estimating ID through geometric analysis of the data. These methods can be further categorized as fractal, graph-based and nearest-neighbor methods.

Fractal methods are based on the fractal geometry and the fractal dimension, the primary examples being the correlation dimension (CD) method [62, 63], which assumes the number of points in a hypersphere of radius r scales exponentially with the dimension of the underlying n-manifold, along with box-counting and Hausdorff dimensions. One detail of fractal-based methods is that they are able to estimate non-integer intrinsic dimensions for datasets.

Nearest-neighbor methods assume close points in the data can be considered as uniformly distributed inside a d-dimensional hypersphere. MLE [4] is a method that estimates ID using this assumption, selecting the dimension that maximizes this likelihood. TwoNN [64] estimates ID by only using the distances of the first two nearest neighbors of each point.

Besides the methods mentioned above, other methods also exist. One example is DANCo [65], which estimates ID using the normalized nearest-neighbor distances combined with computed angles of pair-wise neighbors. Tempczyk et al. [6] propose LIDL as a method of ID estimation using approximate likelihood, which uses a generative model for density estimation. Albergante et al. propose FisherS, a method for estimating ID using Fisher separability [66].

2.5 Datasets

In this thesis, we perform our experiments on simple and explainable multi-class natural image datasets that can easily be transformed into long-tailed datasets.

2.5.1 Long-Tailed Datasets

As our task is long-tailed visual recognition, we use long-tailed versions of well-studied image classification datasets MNIST, CIFAR-10 and CIFAR-100 as our natural image datasets. The imbalance levels of these datasets can be modified artificially, by re-sampling the original dataset such that the class categories show an exponential decrease when ordered from most examples to less examples. In such an imbalanced distribution, we use the term imbalance factors (IF) and imbalance ratio (IR) to specify the imbalance relationship between classes:

$$\text{IR} = \frac{N_{max}}{N_{min}}, \tag{2.2}$$

$$\begin{aligned} \text{IF} &= \frac{N_{min}}{N_{max}} \\ &= \frac{1}{\text{IR}}, \end{aligned} \tag{2.3}$$

where N_{max} and N_{min} refers to the number of samples of the most frequent and least frequent class, respectively.

2.5.2 CIFAR

Krizhevsky et al. [67] introduce CIFAR-10 and CIFAR-100 for multi-class classification, which are composed of 50000 training images and 10000 test images of size 32x32x3 (three color channels). CIFAR-10 includes 10 classes, therefore there are exactly 5000 training examples and 1000 test examples per class. CIFAR-100 on the other hand has 100 classes, meaning that there are 500 training and 100 test images per class.

These datasets have been modified and used extensively for long-tailed visual recognition purposes. In this thesis, we adopt the method by Cao et al. [26] for long-tail modification of these datasets to obtain CIFAR-10-LT and CIFAR-100-LT.

2.5.3 MNIST

MNIST [68] is a database of 60000 training images of hand-written digits (digits 0 through 9) of size 28x28, with 10000 test examples. The distribution of training class labels is not overly imbalanced. Using the same long-tail method as Cao et. al [26] on CIFAR datasets, it is possible to obtain long-tailed MNIST variants which we use in our analysis of model ID.

2.5.4 Other Datasets

Aside from the mentioned datasets, other datasets have been used in long-tailed visual recognition. Places-LT [69], LVIS[2], iNaturalist-2018[70], ImageNet-LT [69] are such examples of long-tailed datasets that have been used frequently in the literature. Once again, we consider more simple and explainable datasets such as MNIST and CIFAR in our experiments involving intrinsic dimensionality.

2.6 Performance Evaluation

In our experiment evaluations, we consider top-1 accuracy and the F1 score as our primary metrics.

2.6.1 Top-1 Accuracy

For our experiments on class imbalance mitigation methods we report the top-1 accuracy, which conventionally is defined as the percentage of examples in a validation set that were classified correctly using only the top prediction of a trained model:

$$\text{Top-1 Accuracy} = \frac{N_t^v}{N^v}, \tag{2.4}$$

where N^v is the size of the validation set, and N_t^v is the number of examples in the validation set which were correctly classified.

2.6.2 F1 Score

For our experiments on class imbalance where accuracy is not an appropriate measure for the task, we use the F_1 score:

$$\begin{aligned} F_1 &= \frac{2}{\text{recall}^{-1} \times \text{precision}^{-1}} \\ &= \frac{2 \times \text{recall} \times \text{precision}}{\text{recall} + \text{precision}}. \end{aligned} \tag{2.5}$$

Precision and recall are defined as:

$$\text{Precision} = \frac{TP}{TP + FP}, \tag{2.6}$$

$$\text{Recall} = \frac{TP}{TP + FN}, \tag{2.7}$$

where TP, FP and FN are the number of true positives, false positives and false negatives respectively.

CHAPTER 3

UNDERSTANDING COMPLEXITY THROUGH MODEL-BASED INTRINSIC DIMENSIONALITY

3.1 Model-based Intrinsic Dimensionality

Following the work by Li et al. [52], we calculate the intrinsic dimension of the objective landscape using fully connected and LeNet [68] networks on MNIST, CIFAR-10, CIFAR-100 and their long-tailed variations using random subspace optimization (as mentioned in Section 2.4.1). Furthermore, we extend their work and calculate ID for each class category. This is done by formulating the initial multi-class classification task into N sub-tasks of one-vs-all classification for each class category in the datasets ($N = 10$ for MNIST and CIFAR-10, and $N = 100$ for CIFAR-100), and estimating ID for each subtask individually.

3.1.1 Calculating Model-based ID for Individual Class Categories

We calculate model-based ID for each class category by formulating the N-class categorization task into N sub-tasks of one-vs-all classification, and estimate the ID of each of these sub-tasks. Importantly, we note that using accuracy as our performance metric for model comparison (as used by Li et al.) is not ideal in this long-tailed one-vs-all classification setup due to the compounding imbalance effects. As a result, accuracy scores do not represent the effect of the long-tail clearly, as shown in Figure 3.1. In comparison, a metric that is more robust against imbalance (e.g., recall, F1-score) is a better fit for the one-vs-all classification task (see Figure 3.2).

For the ID estimate, we follow the authors [52] in using d_{int90}, i.e, the number of

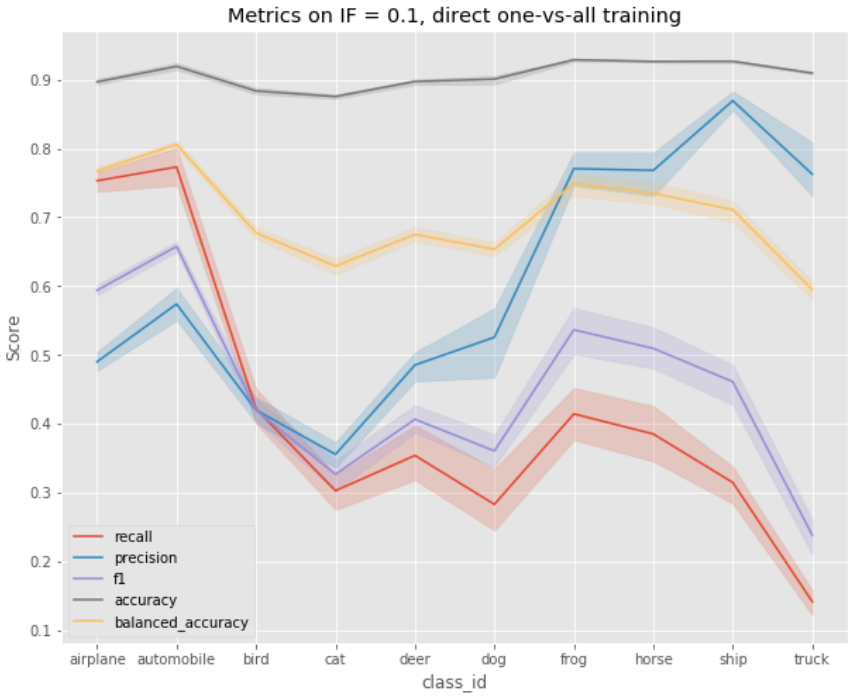


Figure 3.1: Comparison of performance metrics for 10 one-vs-all sub-tasks (one for each class category, fully-connected model) trained on CIFAR-10-LT with an imbalance factor of 0.1.

parameters of the trained restricted model which reaches 90% of the evaluation performance of the free model, and we use F1 score as the evaluation metric instead of accuracy due to the factors mentioned previously. An example of this ID estimation process can be seen in Figure 3.3, where the ID of a class category ("3", in this case) is successfully estimated by the number of free parameters required for the trained model to surpass the necessary one-vs-all classification performance (which is calculated to be 120 for this class).

3.1.2 Model-based ID and Class Complexity

Figure 3.4 shows our results for long-tailed MNIST variants. This figure illustrates a correlation between estimated model-based ID and increasing class imbalance: tail classes are more difficult to learn, and the model-based ID estimate increases accordingly. Interestingly, we see a unique spike around the category “4”, which suggests that the long-tail (i.e., the number of examples) is not the sole factor in determin-

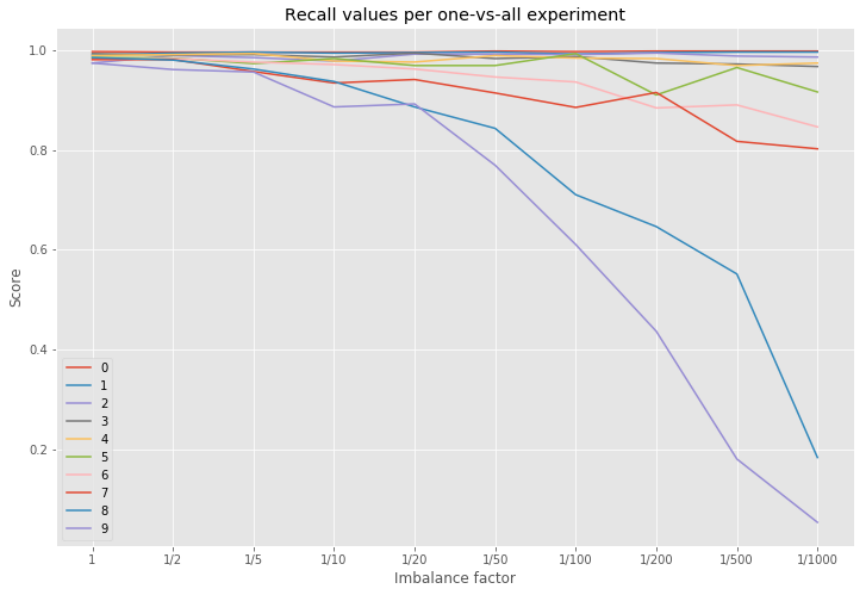


Figure 3.2: Comparison of recall scores for one-vs-all digit classification sub-tasks with a LeNet architecture on MNIST-LT under various imbalance factors.

ing the complexity of a class, which is what we argued as part of our motivation in Section 1.1.

To observe the effect of the long-tail distribution more clearly, we perform multiple experiments with the long-tail ordering of the classes changing each time, and report the average "metaclass" ID estimates in Figure 3.5 and Figure 3.6. Here, 0 corresponds to the average randomized class with the most samples, and 9 corresponds to the class with the least number of samples. Again, we see the trend of increasing difficulty in tail classes under increasing class imbalance, although this is observed more smoothly through randomized classes.

3.1.3 Effects of Model Architecture and Class Imbalance

In order to understand the effects of model architecture and the chosen dataset on the ID estimate more clearly, we conduct simple experiments and briefly go over their results. The setup is relatively straightforward: we use simple neural networks with fully-connected layers as our baseline models, and observe the change in estimated ID by changing only one of the selected hyperparameters. Figure 3.9 highlights that

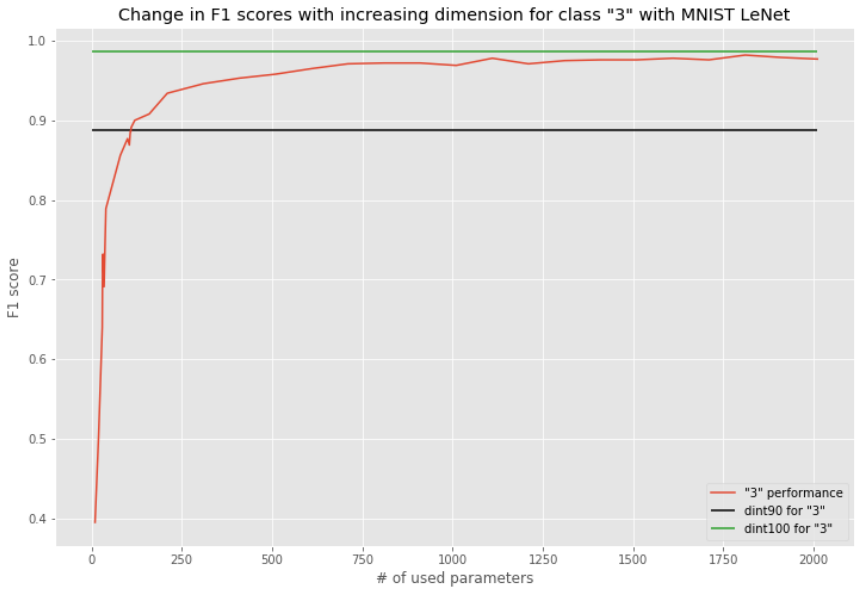


Figure 3.3: An example of the model-based ID estimation process for a LeNet model on the MNIST dataset.

increasing the width of the neural network's parameters increases the estimated ID of the problem. While this can be attributed to increasing the model's capacity, Figure 3.8 shows that increasing a fully-connected model's number of layers does not cause such an observable increase in the model-based ID, despite also increasing the model's capacity. Figure 3.9 furthermore illustrates that increasing the imbalance ratio, and therefore the class imbalance, results in the increase in estimated model ID, specifically in the tail classes.

3.1.4 Drawbacks of Model-based ID

While the analysis of model-based ID and its properties is appealing, this method of estimating ID has some shortcomings that makes us unable to use these estimates for class imbalance mitigation purposes. One of these shortcomings is that there is no guarantee that a model with specific hyperparameters will reach the d_{int90} threshold, which means that there is no way to calculate the ID of that specific configuration. An example is provided in Figure 3.10, which illustrated the failed process of calculating the ID of the CIFAR-10 "cat" class with a LeNet architecture: increasing the number of trainable parameters is not enough to pass the d_{int90} threshold to successfully

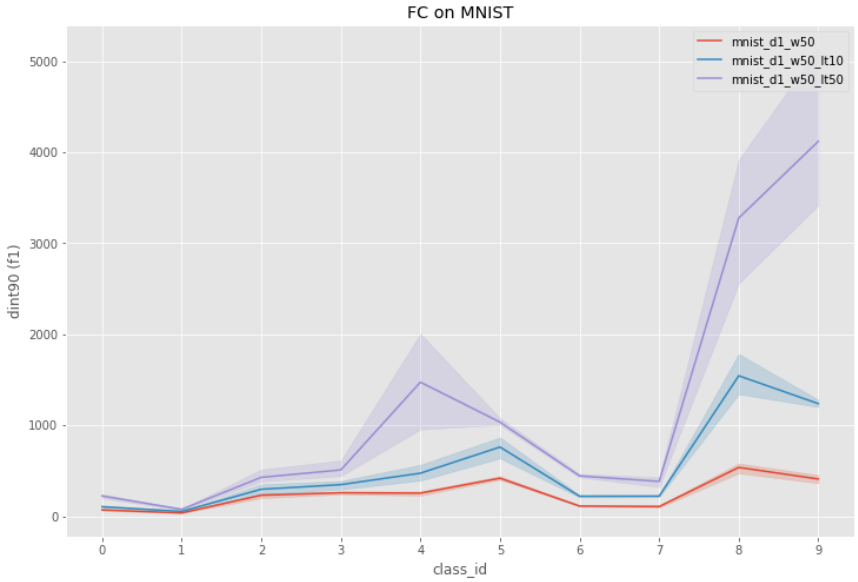


Figure 3.4: Comparison of d_{int90} estimates for one-vs-all experiments on MNIST-LT with varying imbalance factors. The baseline model is fully-connected model with a depth of 1 and a layer width of 50. The imbalance ratios of the datasets are 1 (red), 10 (blue) and 50 (purple), respectively.

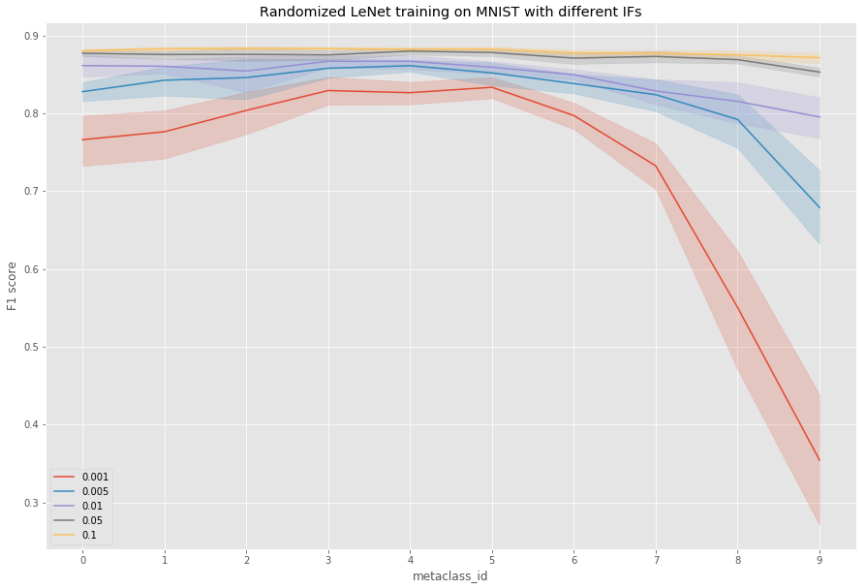
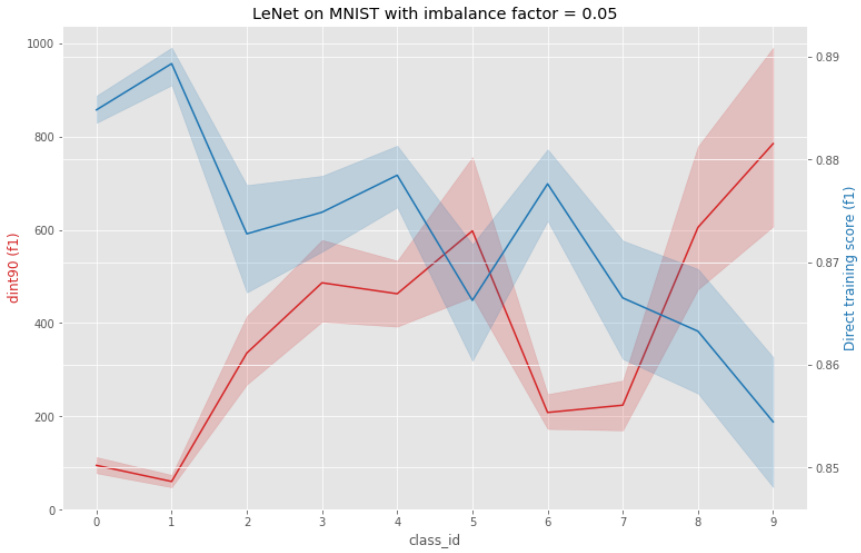


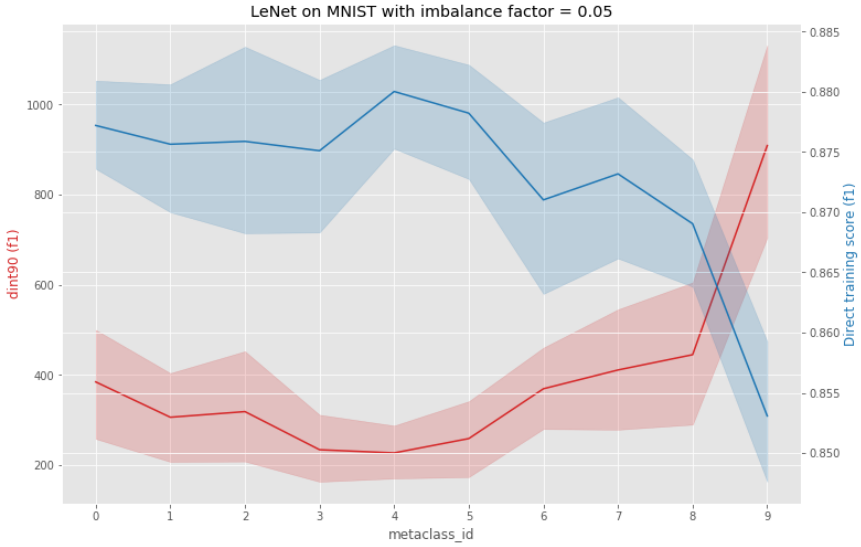
Figure 3.5: Averaged F1 scores for LeNet models trained on MNIST-LT under various imbalance factors, with randomized long-tail ordering. 5 random orderings were chosen for each subtask.

estimate the intrinsic dimension of this class category.

Furthermore, this method of calculating requires training an initial model that learns the dataset "sufficiently well", and afterwards training many restricted models for multiple iterations for each subtask until the restricted model "converges", does not scale well as the dataset and task specification gets increasingly more complex.



(a) Default ordering LT



(b) Randomized ordering LT

Figure 3.6: Comparison of ID estimations and F1 scores for one-vs-all LeNet models on MNIST-LT with IF = 0.05. (a) Default ordering long-tailed version. (b) Randomized ordering averaged ($N = 5$) long-tailed results.

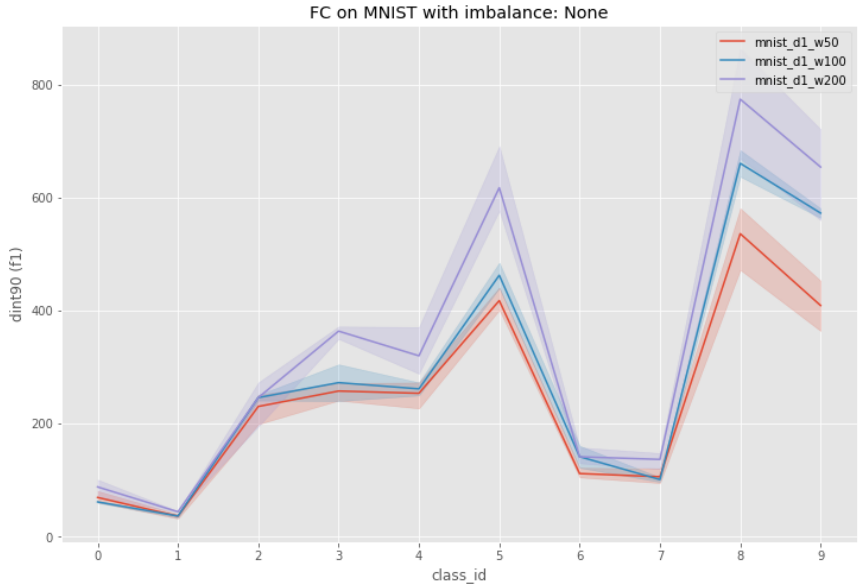


Figure 3.7: Effects of changing model width for a fully-connected model on balanced MNIST

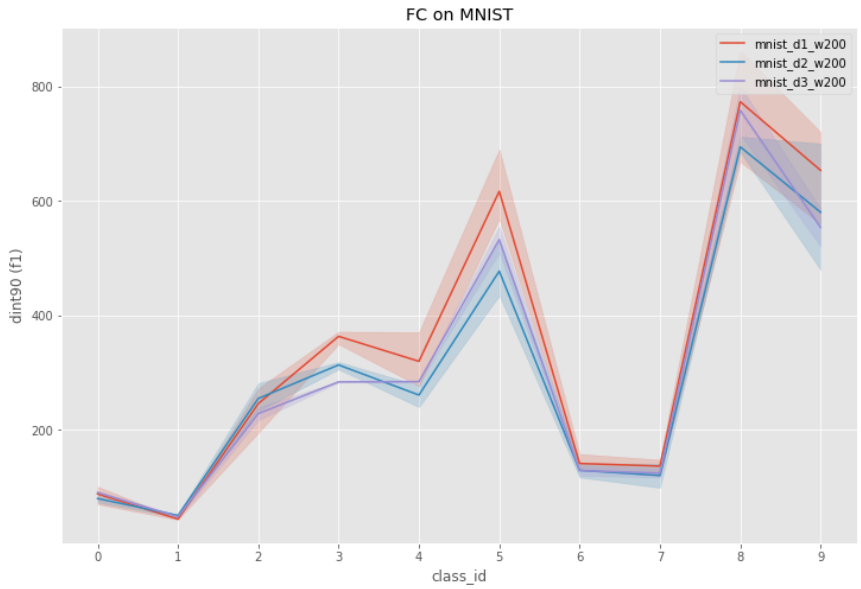


Figure 3.8: Effects of changing model depth for a fully-connected model on balanced MNIST

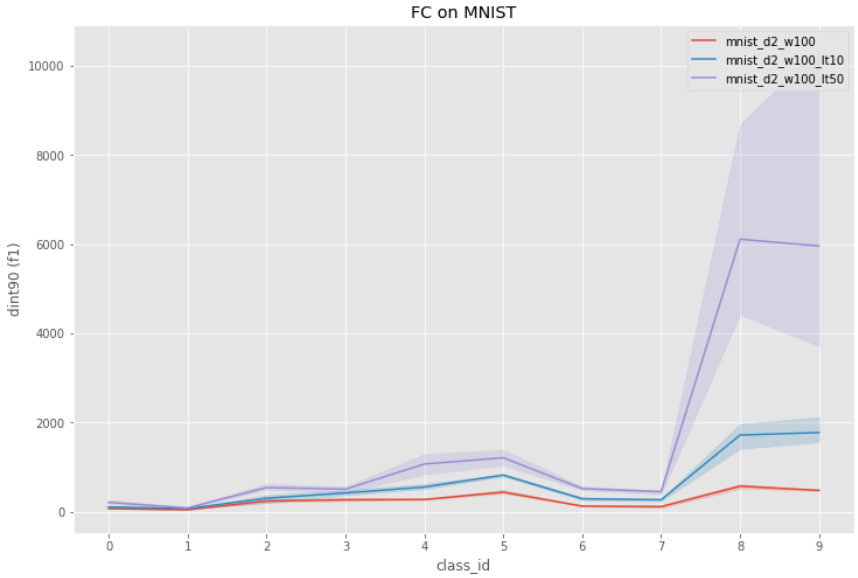


Figure 3.9: Effects of increasing imbalance ratio for a fully-connected model on MNIST

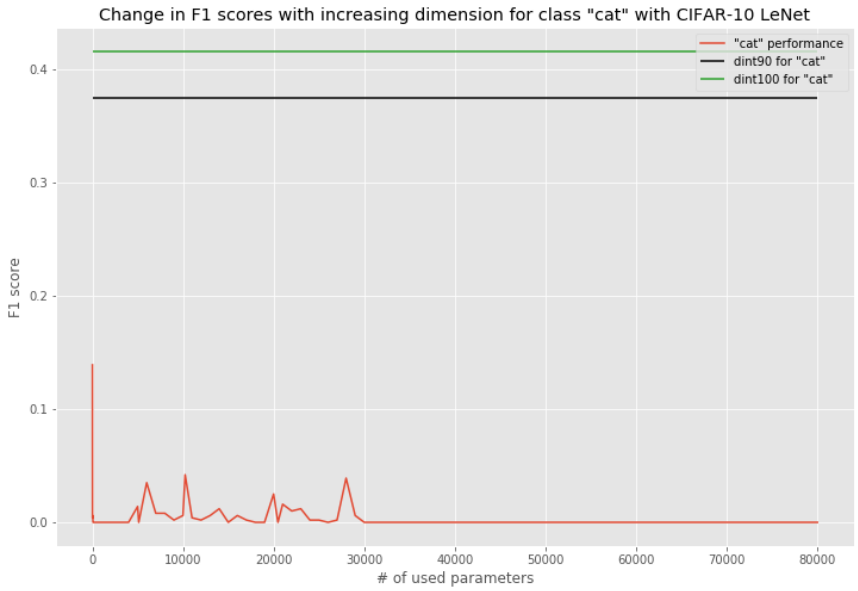


Figure 3.10: An example of a failed model-based ID estimation process for a LeNet model on the CIFAR-10 dataset.

CHAPTER 4

A DETAILED LOOK INTO DATA-BASED INTRINSIC DIMENSIONALITY

4.1 Data-based Intrinsic Dimensionality

As mentioned in Chapter 2, a significant amount of work on intrinsic dimensionality focuses on calculating the ID of a set of data points. We start by introducing some definitions of ID used in mathematics, computer science and statistics.

4.1.1 Definition of Data-based ID

We adopt the manifold-based definition of Ceruti et al. [65]. Let $X_N = \{x_i\}_{i=1}^N \subset \mathbb{R}^D$ be a dataset of N points residing in D-dimensional space. We assume that the data points lie in a lower-dimensional manifold $\mathcal{M} \subseteq \mathbb{R}^d$ with $d < D$ that is embedded in the higher dimensional space. In this case, d is said to be the intrinsic dimension of this set of data.

Alternatively, the ID is the minimum number of parameters to represent the data without information loss. Some definitions define it as the number of variables that best approximates the data instead (e.g., [66]), discarding the requirement of no information loss. Nevertheless, all of these definitions are valid ways to describe the concept of ID, and the estimators presented in the following section all attempt to estimate the same quantity.

4.2 ID Estimators

We try a wide range of ID estimation methods in order to calculate ID.

4.2.1 Maximum Likelihood Estimation

Levina and Bickel [4] apply the maximum likelihood principle to the distances of points to their nearest neighbors to obtain an estimate of the dimension of the underlying manifold. Specifically, MLE treats i.i.d. observations of data points around the neighborhood of each point as a homogeneous Poisson process, and derives the MLE by maximizing the log-likelihood of this observed process.

MLE assumes that we have i.i.d. observations $X_1, ..., X_n$ in $\mathbb{R}^p$ which represent an embedding of a lower-dimensional sample, i.e., $X_i = g(Y_i)$, where Y_i are sampled from an unknown smooth density f on $\mathbb{R}^m$, with unknown $m \leq p$, and g is a continuous and sufficiently smooth mapping, such that close neighbors in $\mathbb{R}_m$ are mapped to close neighbors in the embedding.

The basic idea of MLE is to fix a point x, assume $f(x)$ is constant in a small sphere $S_x(R)$ of radius R around x, and treat the observations as a homogeneous Poisson process in $S_x(R)$. They define the following inhomogeneous process $\{N(t,x), 0 \leq t \leq R\}$,

$$N(t,x) = \sum_{i=1}^{n} 1\{X_i \in S_x(t)\}, \tag{4.1}$$

which counts observations within distance t from fixed point x. MLE approximates this binomial process by a Poisson process and suppresses the dependence on x, where the rate $\lambda(t)$ of the process $N(t)$ is then

$$\lambda(t) = f(x)V(m)mt^{t-1}. \tag{4.2}$$

Letting $\theta = \log f(x)$, they obtain the log-likelihood of the observed process $N(t)$:

$$L(m,\theta) = \int_0^R \log \lambda(t) dN(t) - \int_0^R \lambda(t)dt. \tag{4.3}$$

In particular, the MLEs must satisfy the following likelihood equations below:

$$\frac{\partial L}{\partial \theta} = \int_0^R dN(t) - \int_0^R \lambda(t)dt = N(R) - e^{\theta}V(m)R^m = 0, \tag{4.4}$$

$$\begin{aligned} \frac{\partial L}{\partial m} &= \left(\frac{1}{m} + \frac{V'(m)}{V(m)}\right) N(R) + \int_0^R \log(t) dN(t) \\ &\quad - e^{\theta}V(m)R^m \left(\log(R) + \frac{V'(m)}{V(m)}\right) \\ &= 0. \end{aligned} \tag{4.5}$$

Substituting Equation 4.4 into Equation 4.5 gives the MLE for m:

$$\hat{m}_r^x = \left[\frac{1}{N(R,x)} \sum_{j=1}^{N(R,x)} \log \frac{T_k(x)}{T_j(x)} \right]^{-1}. \tag{4.6}$$

In our case, it is more convenient to fix the number of neighbors k, rather than the radius of the sphere R. In this case, we use the alternative ID presented by the authors in which the estimate in Equation 4.6 becomes:

$$\hat{m}_k^x = \left[\frac{1}{k-1} \sum_{j=1}^{k-1} \log \frac{T_k(x)}{T_j(x)} \right]^{-1}. \tag{4.7}$$

One important detail of MLE is that as a nearest-neighbor based method, MLE is able to provide different estimations of ID depending on the neighborhood (number of neighbors) being considered for each point, effectively providing distinct ID estimates for different scales.

In our experiments with MLE, we use the suggested unbiased estimator proposed in [5], similar to works such as [71].

4.2.2 TwoNN

Facco et al. [64] propose to estimate the ID of a dataset by using only the first two nearest neighbors of each point in the dataset.

Similarly to MLE, let i be a point in the dataset, and consider the list of its first k nearest neighbors in the dataset; let r_1, r_2, ..., r_k be a sorted list of their distances from i, and let $r_0 = 0$. With this definition, the volume of the hyperspherical shell enclosed between two successive neighbors $l-1$ and l is given by:

$$\Delta v_l = \omega_d (r_l^d - r_{l-1}^d), \tag{4.8}$$

where d is the dimensionality of the space in which the points are embedded and ω_d is the volume of the d-sphere with unitary radius. Moreover, the authors show that all the Δv_l are independently drawn from an exponential distribution with rate equal to the density ρ, where

$$P(\Delta v_l \in [v, v+dv]) = \rho e^{-\rho v} dv, \tag{4.9}$$

provided that the density is assumed to be constant around the point i. Considering two shells Δv_1 and Δv_2, with R being the quantity $\frac{\Delta v_i}{\Delta v_j}$ and assuming constant density, Facco et al. are able to compute exactly the probability distribution function of R:

$$\begin{aligned} P(R \in [\bar{R}, \bar{R} + d\bar{R}]) &= \int_0^\infty dv_i \int_0^\infty dv_j \rho^2 e^{-\rho(v_i+v_j)} \mathbf{1}\left\{ \frac{v_j}{v_i} \in [\bar{R}, \bar{R} + d\bar{R}] \right\} \\ &= d\bar{R} \frac{1}{(1+\bar{R})^2}, \end{aligned} \tag{4.10}$$

where $\mathbf{1}$ represents the indicator function. To obtain the probability density function (pdf) for R, they divide by $d\bar{R}$:

$$g(R) = \frac{1}{(1+R)^2}. \tag{4.11}$$

The pdf in Equation 4.11 does not depend explicitly on the dimensionality d, which appears only in the definition of R. In order to work with a cumulative distribution function (cdf) depending explicitly on d, they define the quantity $\mu \doteq [1, +\infty)$, where R and μ are related by equality

$$R = \mu^d - 1. \tag{4.12}$$

Equation 4.12 allows to find an explicit formula for the distribution of μ:

$$f(\mu) = d\mu^{-d-1} \mathbf{1}_{[1,+\infty]}(\mu), \tag{4.13}$$

and the cdf is obtained by integration:

$$F(\mu) = (1 - \mu^{-d} \mathbf{1}_{[1,+\infty]}(\mu). \tag{4.14}$$

Functions f and F are independent of the local density, but depend explicitly on the intrinsic dimension d. Finally, the intrinsic dimension d can be estimated with

$$\frac{\log(1 - F(\mu))}{\log(\mu)} = d. \tag{4.15}$$

4.2.3 FisherS

Albergante et al. [66] propose FisherS, another method to estimate ID by using Fisher separativity analysis on the given dataset.

For this ID estimation method, the authors assume that the dataset X is normalized by applying the following steps:

1. centering

2. projecting onto the linear subspace spanned by first k principal components, where k may be relatively large

3. whitening

4. normalising each vector to unit length.

In particular, normalizing each vector is necessary for comparing the data distribution with a unit sphere. Similarly, choosing the number of principal components to retain with PCA aims to avoid having excessively small eigenvalues of the covariance matrix. An effective way to estimate k, as reported by the authors, is by selecting the largest k (in their natural ranking) such that the corresponding eigenvalue λ_k is not smaller that λ_1/C, where C is a predefined threshold.

After such normalization of X, it is said that a point $x \in X$ is Fisher-linearly separable from the cloud of points Y with parameter α, if

$$(\mathbf{x}, \mathbf{y}) \leq \alpha(\mathbf{x}, \mathbf{x}), \tag{4.16}$$

for all $y \in Y$, where $\alpha \in [0, 1)$. If Equation 4.16 is valid for each point $x \in X$ such that Y is the set of points $y \neq x$, then the dataset X is said to be Fisher-separable with parameter α.

In order to quantify deviation from perfect separability, Albergante et al. introduce $p_\alpha(\mathbf{y})$, the probability that a point y is separable from all other points. They further define $\bar{p}_\alpha(\mathbf{y})$ as a mean value of the distribution of $p_\alpha(y)$ over all data points.

In this formulation, p_α is equivalent to:

$$p_\alpha = \bar{p_\alpha} = \frac{(1-\alpha^2)^{\frac{n-1}{2}}}{\alpha\sqrt{2\pi n}}. \tag{4.17}$$

With Equation 4.17, the distribution of p_α for a uniform sampling from an n-sphere is then a delta function centered in $\bar{p_\alpha}$, and the effective dimensions of this data set can be evaluated by comparing the quantity $p_{\{}\bar{alpha}$ for this data set against $\bar{p_\alpha}$ for the equidistributions on a sphere.

With this method, if $\bar{p_\alpha}$ can be empirically estimated for a given α, then the effective dimension can be estimated by solving Equation 4.17 with respect to n:

$$n_\alpha = \frac{1}{-\ln(1-\alpha^2)} W\left(\frac{-\ln(1-\alpha^2)}{2\pi \bar{p}_\alpha^2 \alpha^2 (1-\alpha^2)}\right), \tag{4.18}$$

where $W(x)$ is the Lambert function and n_α is consequently the intrinsic dimension for this data set.

4.3 Extending Data ID to Class-Based ID

Depending on the estimator being used, ID can be computed locally, globally, or in a point-wise basis. Figure 4.1a shows an example of MLE being calculated in a point-wise manner for a synthetic dataset of a mixture of two Gaussian mixtures of different intrinsic dimensionality, and Figure 4.1b shows the ID estimates of a Gaussian mixture with varying degrees of sample contribution between the two Gaussian distributions.

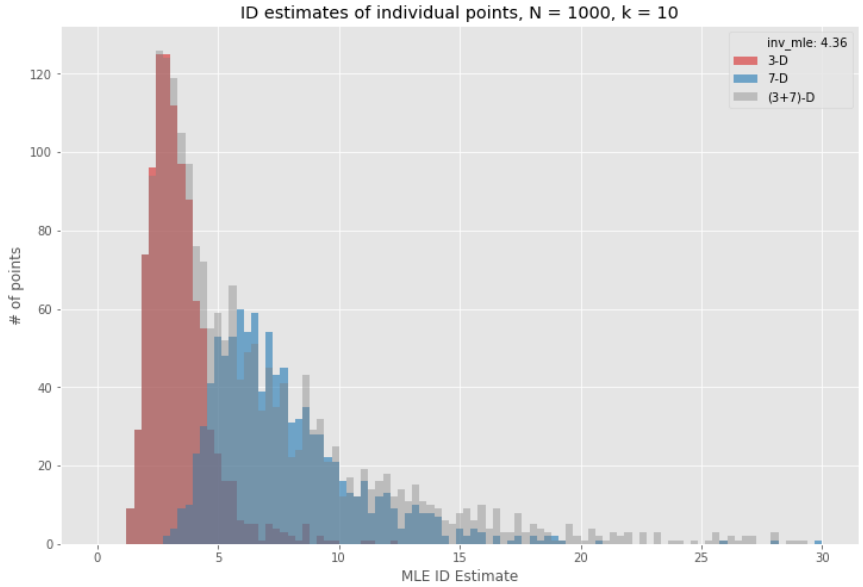
Depending on the ID estimation method used, it is possible in a multi-class setting to calculate class IDs using neighbors from the isolated class manifold (which in Figure 4.2a we call the isolated manifold), and also from the manifold of the entire dataset (which in Figure 4.2b we call the entire manifold). In our experiments, we use the isolated class manifold as the results are more explainable than those obtained with using the entire dataset manifold.

4.4 Comparison of ID Estimators

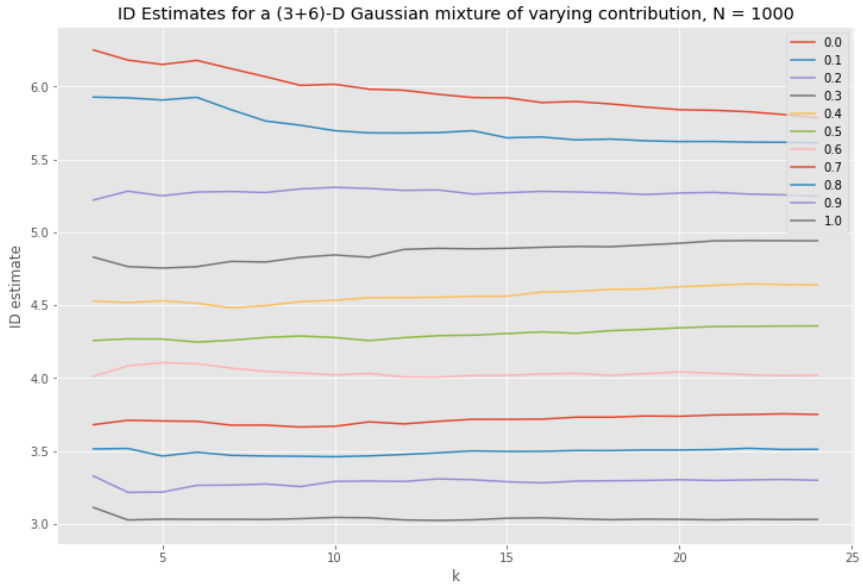
We consider various ID estimators in the literature and compare their class-wise ID estimation results on synthetic and natural image datasets.

4.4.1 Synthetic Datasets

It is useful to analyze the properties of data-based ID on synthetic datasets. For example, it is trivial to generate datasets from known distributions (such as Gaussians)

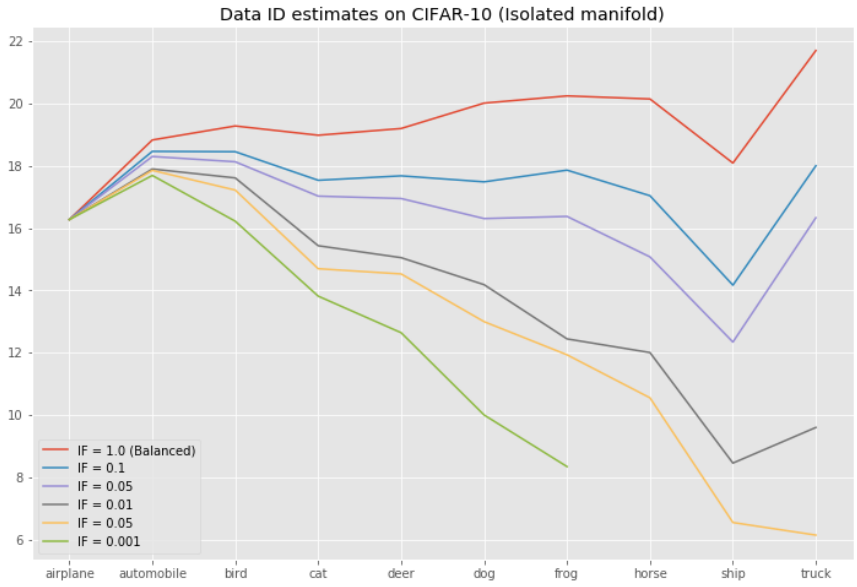


(a) Equal mixture

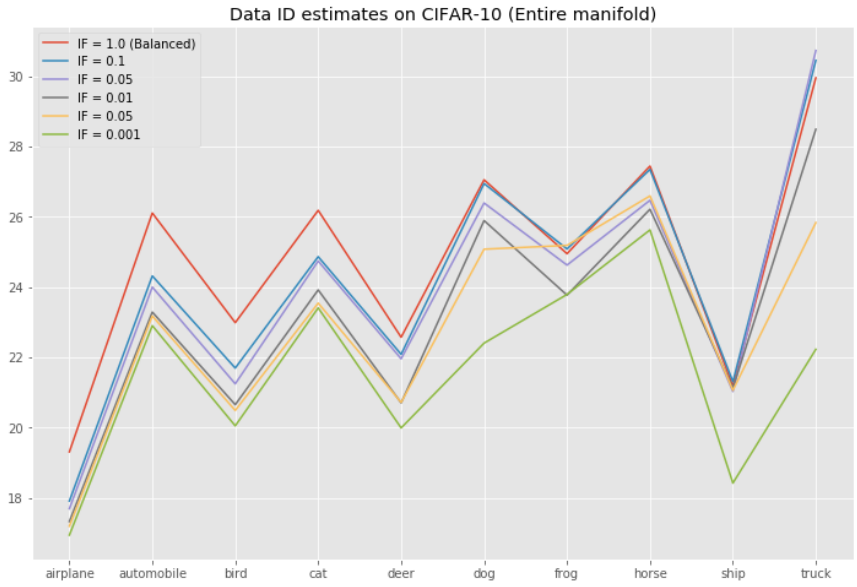


(b) Mixture of two Gaussian distributions of changing relative contribution.

Figure 4.1: Point-wise ID estimations of MLE [4] on mixtures of Gaussian distributions. (a) Equal mixture of two Gaussian distributions. (b) Mixture for two Gaussian distributions with changing ratio of sample contribution. Here, a dataset with a ratio of 0.0 only includes samples from the first component, 1.0 only includes samples from the second component, and 0.5 means equal contribution of samples from both components.



(a) Isolated manifold



(b) Entire manifold

Figure 4.2: Estimating the ID of CIFAR-10 class categories with MLE [4] (a) Using the isolated manifold of each class category. (b) Using the manifold created by the entire dataset.

which we can use to verify the effectiveness of ID estimators; in fact, given that this is the only reliable way to acquire ground truth data, many ID estimators are

evaluated using synthetic datasets. Figure 4.3 shows the ID estimate of MLE for a simple Gaussian distribution and justifies the use of the unbiased estimator proposed by MacKay and Ghahramani [5].

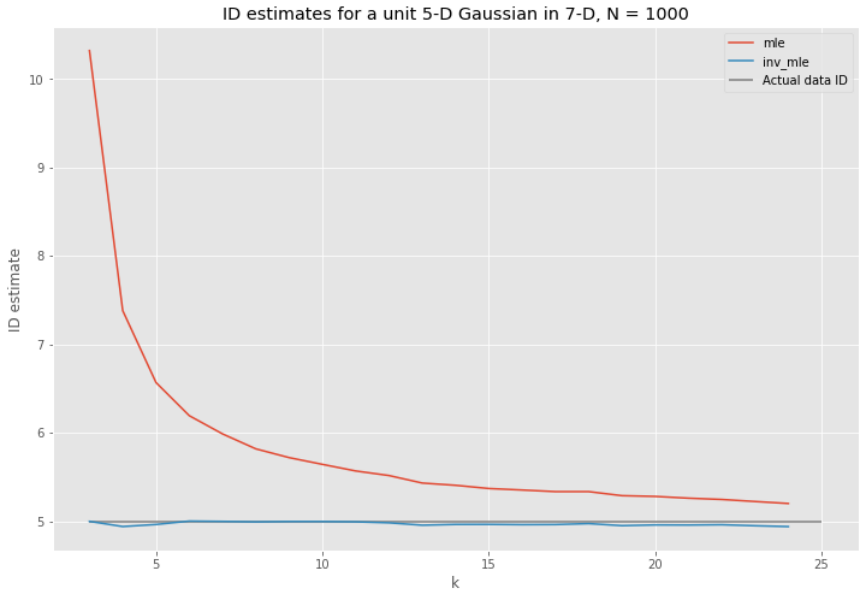


Figure 4.3: ID estimates of MLE [4] on a simple Gaussian distribution with an ID of 5, in a space with extrinsic dimensionality of 10. We use the unbiased MLE estimator of [5], highlighted in blue.

One problem that is apparent for some estimators (such as MLE) is that they do not scale well to increasing dimension above a specified level. Figure 4.4 shows the ID estimates of ID trailing below the actual dimensionalities of synthetically generated Gaussian datasets. Figure 4.5 further indicates that this trend of ID underestimation is quadratic or perhaps exponential with regards to the number of samples.

4.4.2 Natural Image Datasets

Differently from synthetic datasets, natural image datasets (such as MNIST and CIFAR) are a collection of natural images; they do not have associated ground truth ID values. It is possible to estimate the ID of these datasets like any other dataset using these estimators and make comparisons [71], although we do not have a useful metric to determine the estimator that provides "the best fit" to these datasets. Figure 4.6 shows the estimations of several ID estimators for the balanced CIFAR-10 dataset,

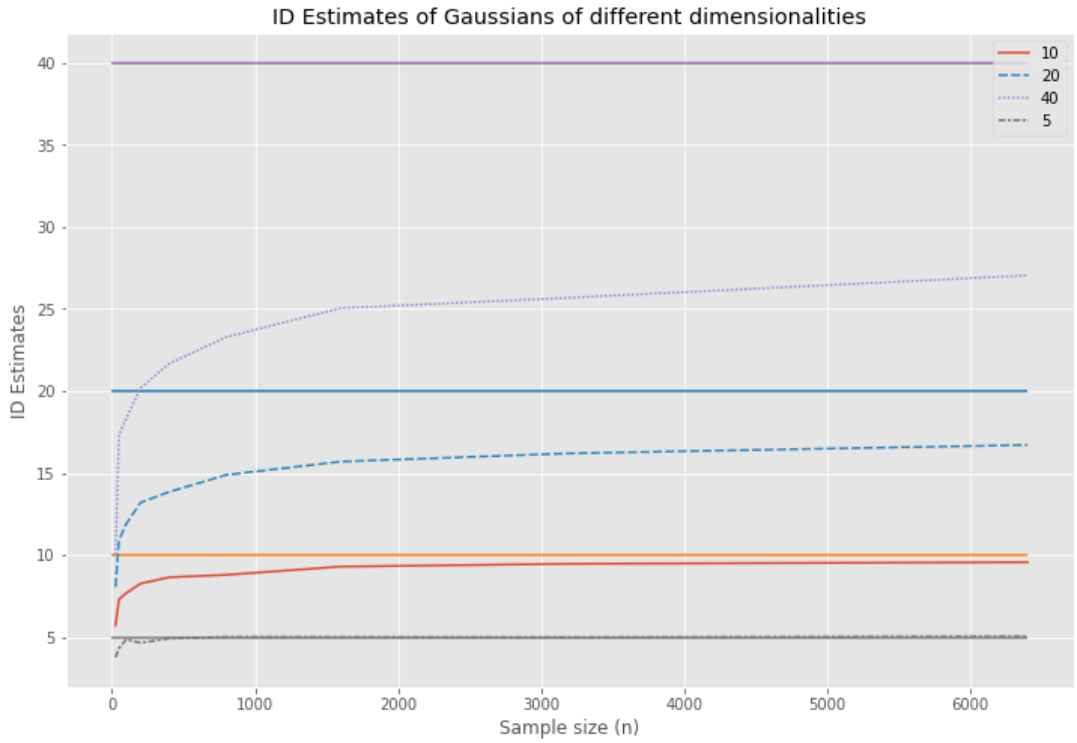


Figure 4.4: ID estimates of MLE [4] for synthetic Gaussian data and their expected values.

which all estimate ID in their own "magnitudes".

However, as our goal to use the estimated ID between class categories to increase model robustness, the relative ID ordering between categories is more important to our discussion rather than the exact ID values themselves. We note that as the intrinsic dimension represents the complexity of a class category, estimators whose ID estimates change drastically with changes in sample size are too volatile to be used as reliable ID estimators for natural image datasets. As an example of this phenomenon, we present the case of estimating the class ID values of CIFAR-10-LT categories under increasing class imbalance: Figures 4.2 and 4.7 show the estimations of MLE [4] and LIDL [6], another recent ID estimator, in which the ID estimates are shown to behave erratically with increasing class imbalance and are also quite sensitive to sample size. On the other hand, Figure 4.8 shows that the ID estimates of FisherS [66] are relatively stable with respect to change in the number of samples.

Therefore, we use the ID estimations of the TwoNN [64] and FisherS [66] methods

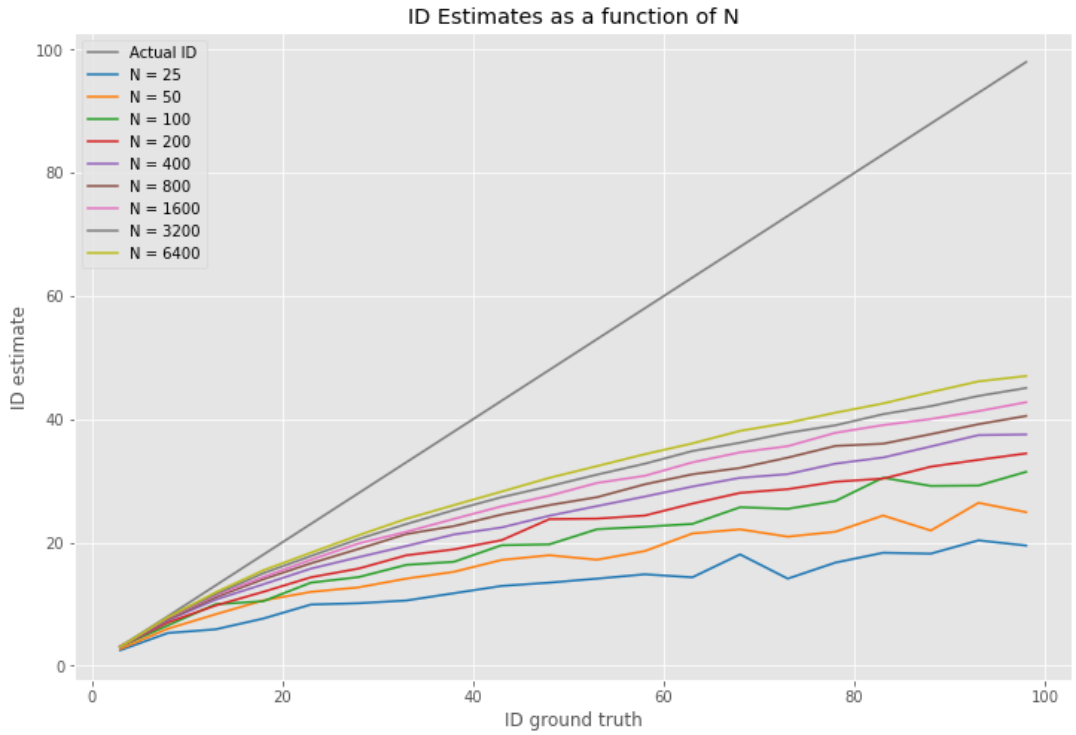


Figure 4.5: The trend of ID estimates of MLE [4] for synthetic Gaussian data of increasing dimensionality.

in our experiments involving the use of ID in improving model robustness against imbalance in Chapter 5. Furthermore, in preparation for weighted re-sampling and loss re-weighting, we normalize each estimator's ID estimate and scale it according to the number of categories, as shown by Figure 4.10 and Figure 4.11.

4.4.3 Comparing Data-based ID to Model-based ID

In Figure 4.12 and Figure 4.13 we show the model-based and data-based ID estimations of class categories in MNIST and CIFAR-10, respectively. It is intriguing to observe that ignoring the difference in magnitudes, model-based and data-based ID estimates are similar in terms of "peaks" and "valleys". This is more clearly observed in Figure 4.12 with MNIST, whereas for CIFAR-10 it is made clearer under a small amount of class imbalance (Figure 4.13). Nevertheless, these figures indicate that despite the difference in definition and estimation strategy, these two applications of ID are similar.

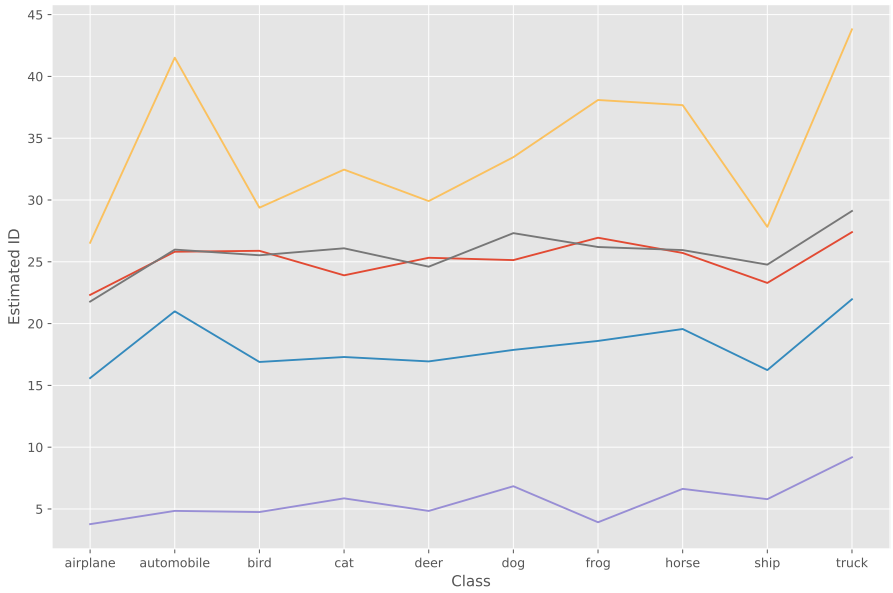


Figure 4.6: ID estimates of several methods for balanced CIFAR-10.

4.5 Data-based ID as an Imbalance Measure

In addition to our discussion of data-based ID, we evaluate whether this definition of ID can be used as a measure of class imbalance.

4.5.1 Requirements for an ID-based Imbalance Measure

For an imbalance measure, the primary requirements that have to be satisfied are as follows:

Requirement 1. The imbalance measure should increase when data imbalance increases.

Requirement 2. The measure should not decrease when identical samples are added into the dataset.

We look at these requirements below.

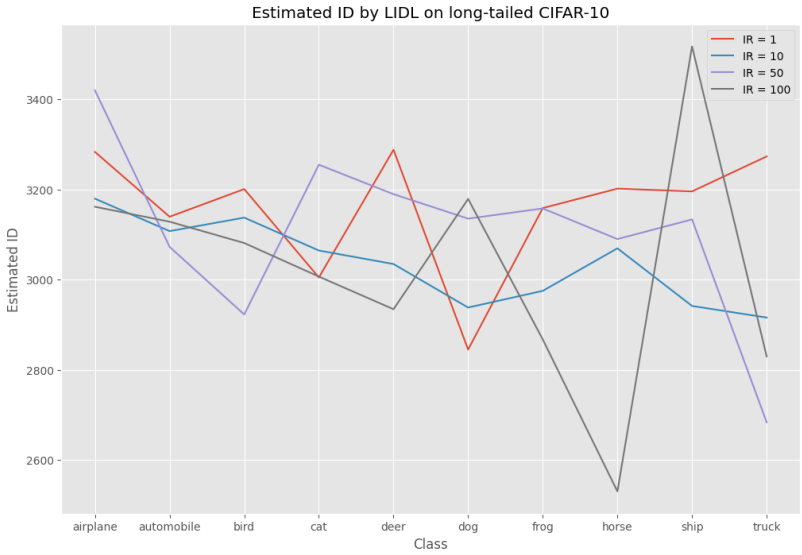


Figure 4.7: ID estimates of LIDL [6] CIFAR-10-LT with different imbalance ratios.

4.5.2 Requirement Analysis

4.5.2.1 Requirement 1

Measures satisfying this requirement should increase in value when class imbalance increases. In our observations, we see in Figure 4.14 that most data-based ID estimators actually decrease on average value when data is further imbalanced, thus failing the requirement of being a good imbalance measure. An example is shown by MLE [4] on MNIST in Figure 4.15, where the ID estimate is shown to be decreasing in the tail classes. However, FisherS [66] appears to be an exception to this trend; moreover, as shown by Figure 4.16, it is shown to be non-decreasing in all of our used datasets (MNIST, CIFAR-10, CIFAR-100) through increasing class imbalance.

It should be briefly stated that although model-based ID also fulfills the first requirement (recalling back to Figure 3.4), it is difficult to use it meaningfully as an imbalance measure due to the problems mentioned at the end of Chapter 3.

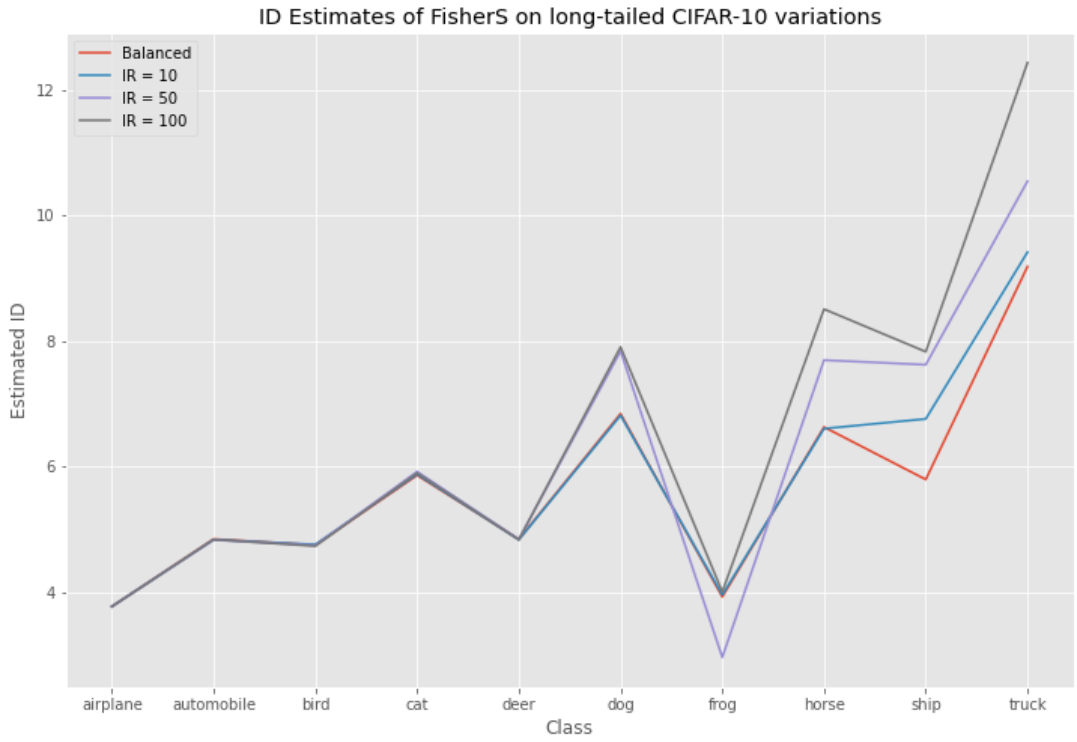


Figure 4.8: ID estimates of FisherS on CIFAR-10-LT with different imbalance ratios.

4.5.2.2 Requirement 2

For a measure to successfully qualify as an imbalance measure, it is also required to be non-decreasing in the case of replication of identical samples. In order to test this requirement, for various ratios, we randomly sample images from the dataset and add these identical samples back into the dataset, thus increasing their sample size while keeping the number of unique samples in the dataset constant. With this configuration, we observe in Figure 4.17 that FisherS is also shown to be decreasing in CIFAR-100, therefore concluding that these data ID estimators are not equipped to be good imbalance measures.

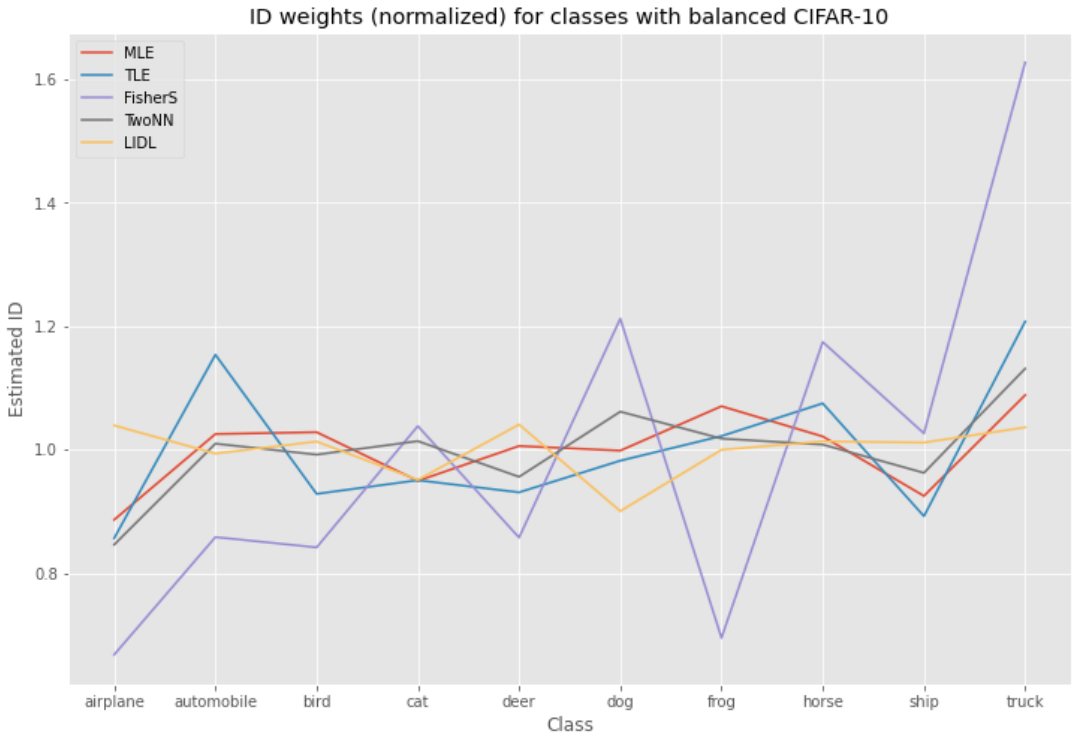


Figure 4.9: ID estimates of several methods for balanced CIFAR-10.

Figure 4.10: Normalized ID estimates of several methods for balanced CIFAR-10.

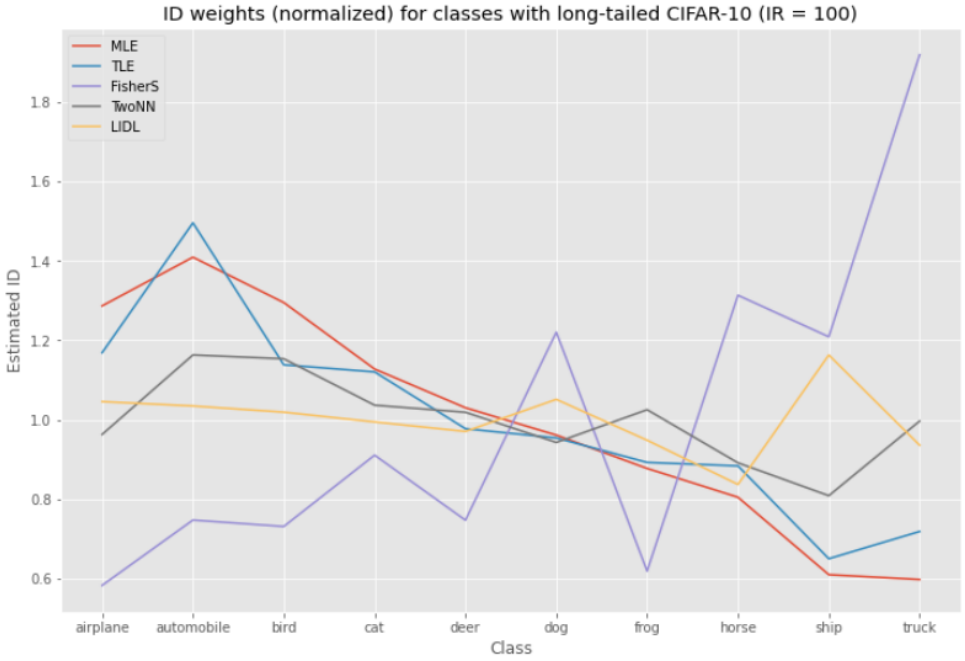


Figure 4.11: ID estimates of several methods for CIFAR-10-LT with an imbalance ratio of 100.

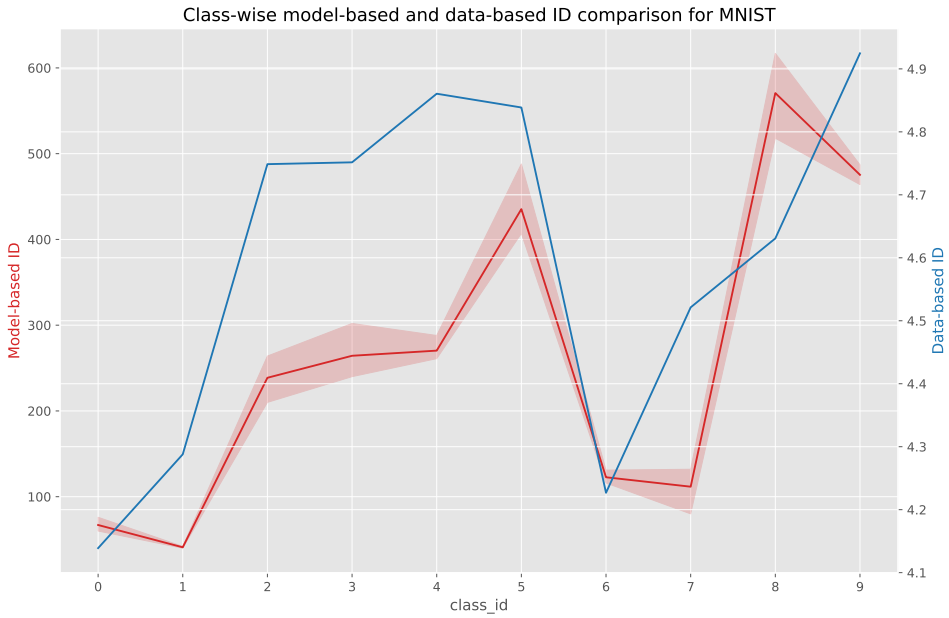
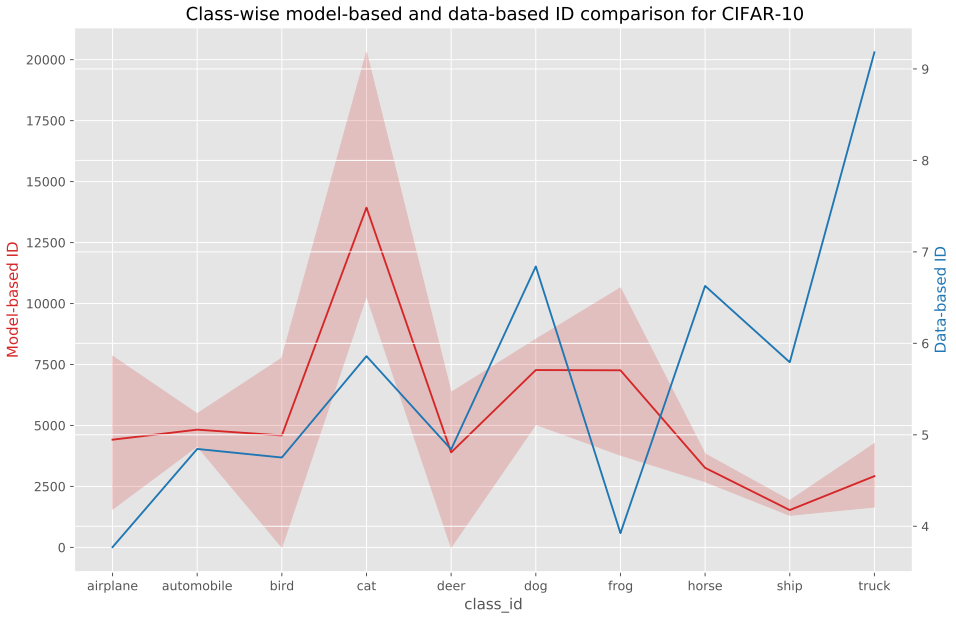
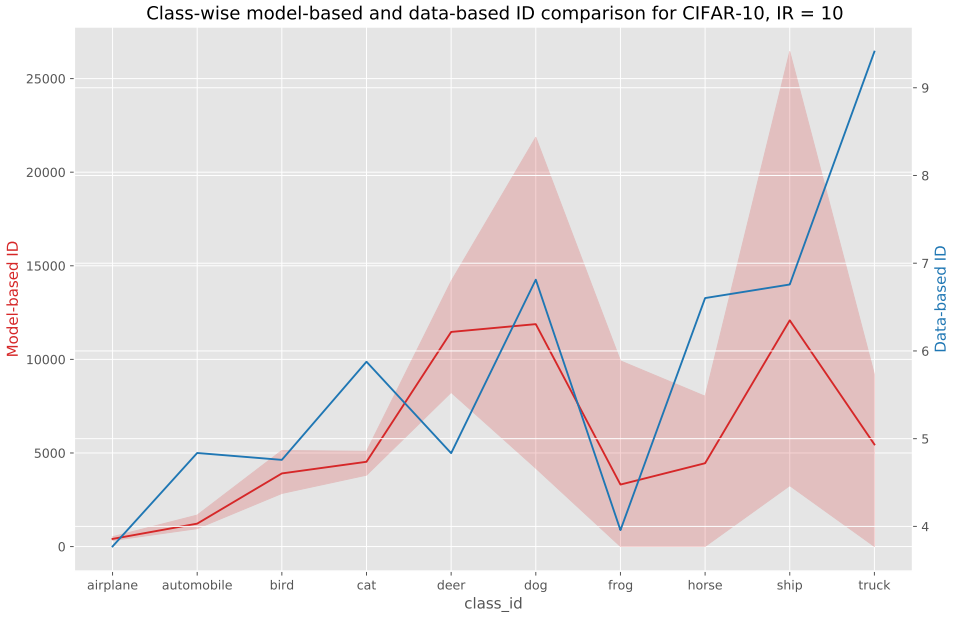


Figure 4.12: Comparison of Model-based and Data-based ID estimation on MNIST.



(a) Balanced



(b) Long-tailed with IR = 10

Figure 4.13: Comparison of model-based and data-based ID estimation on CIFAR-10-LT under increasing class imbalance.

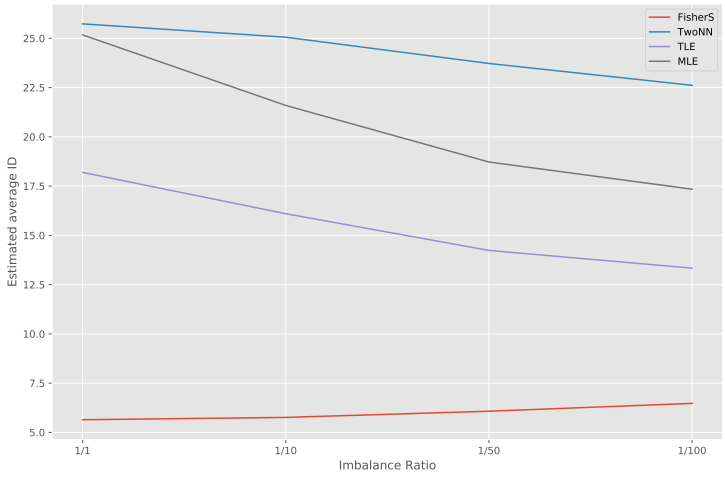


Figure 4.14: Average ID estimates of various estimators for CIFAR-10-LT under increasing class imbalance.

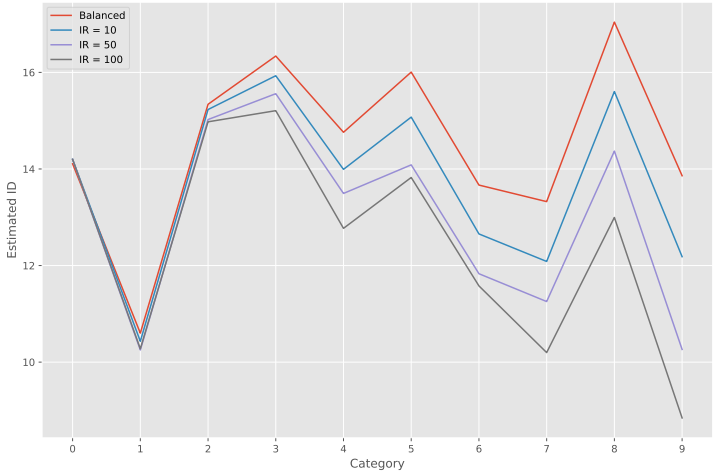


Figure 4.15: Average ID estimates of MLE for individual class categories of MNIST-LT under increasing class imbalance.

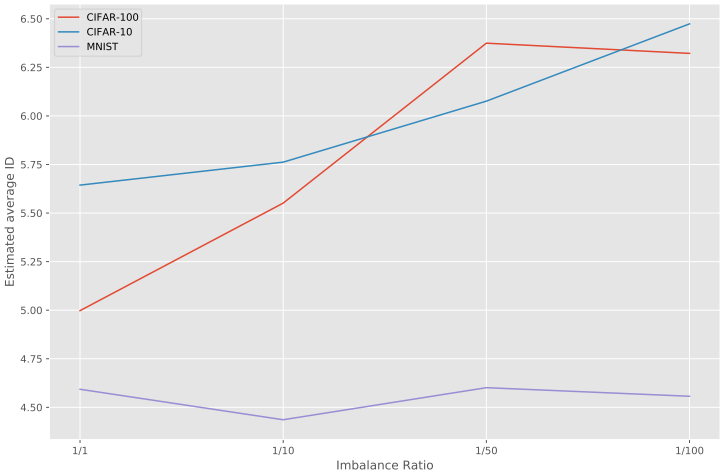


Figure 4.16: Average ID estimates of FisherS for different datasets under increasing class imbalance.

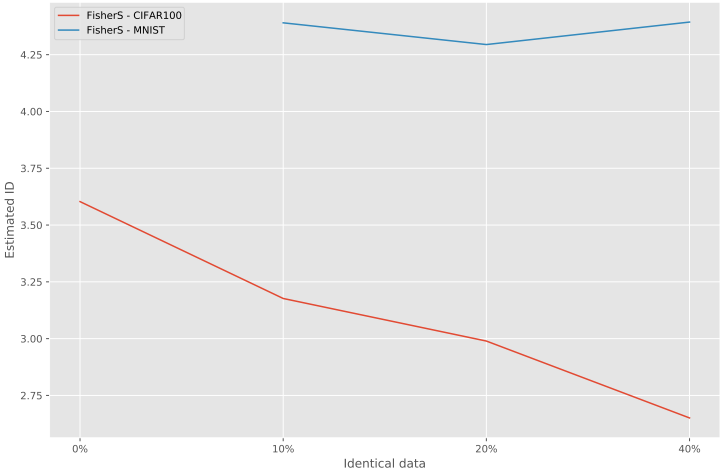


Figure 4.17: ID estimates of FisherS for individual class categories of datasets under with increasing percentage of identical samples.

CHAPTER 5

MITIGATING CLASS IMBALANCE THROUGH INTRINSIC DIMENSIONALITY

In this chapter, we use the estimated data-based ID values of individual class categories in several strategies of mitigating class imbalance, and report our findings below.

5.1 Experiment Setup

We use the framework proposed in the Bag of Tricks study [21] for our experiments in re-sampling, re-weighting and margin-based methods with ID. A detailed explanation of our model training, hyperparameter selection and hardware is provided in Appendix A.

In our ID experiments, we use the TwoNN [64] and FisherS [66] estimators for our experiments involving the use of data-based ID. As discussed in the end of Chapter 4, we choose these methods because of their relative robustness against changing imbalance factors, and also because they represent two distinct families of estimating intrinsic dimensionality.

We use top-1 accuracy as the evaluation metric for our re-balancing comparisons, following similar work. Additionally, our results using the F1 score as the evaluation metric can also be found in Appendix A.

We briefly go into detail as to how we incorporate ID into these families of class imbalance mitigation methods.

5.1.1 Loss Re-weighting with Data-based ID

Recalling back to Section 2.3.2, we apply ID-based class weights by replacing the variable ω_c in Equation 2.1 with the data-based ID estimate of that class category. In other words, the total loss with ID-informed loss re-weighting is equal to

$$\sum_{c}^{C} \hat{d}_c \mathcal{L}_c, \tag{5.1}$$

where $c \in C$ is a class category in the dataset, $\mathcal{L}_c$ is again loss accumulated by class c and $\hat{d}_c$ corresponds to the normalized estimated data-based ID of that class category.

5.1.2 Re-sampling with Data-based ID

Similarly to re-weighting schemes, weighted re-sampling is done by increasing the probability of samples to be included in mini-batches for mini-batch stochastic gradient descent, either by over-sampling or under-sampling specific class categories in the dataset. Following similar work, we apply ID-based class weights to re-sampling by normalizing the ID estimates for all classes such that their sum is 1, assigning this weight as the class sampling probability, and afterwards selecting samples from this class by performing uniform sampling across data points of this class category. In other words, the sampling probability for a sample n of class category c is equal to

$$\begin{aligned} p_n &= p_c \times \frac{1}{N_c}, \\ p_c &= \frac{\hat{d}_c}{\sum_i^C \hat{d}_c}, \end{aligned} \tag{5.2}$$

where N_c is the number of samples for class category c, and $\hat{d}_c$ is the data-based ID estimate of class category c.

5.1.3 Re-margining with Data-based ID

In our margin-based class mitigation analysis, we apply class ID estimates with the LDAM [26] and DRO-LT [31] frameworks.

In the case of LDAM, we apply our class ID estimates to the LDAM loss defined in [26] as

$$\begin{aligned} \mathcal{L}_{\mathrm{LDAM}}((x,y);f) &= -\log \frac{e^{z_y-\Delta_y}}{e^{z_y-\Delta_y+\sum_{j\neq y} e^{z_j}}}, \\ \Delta_y &= \frac{C}{n_j^{\frac{1}{4}}}, \end{aligned} \tag{5.3}$$

where f is a model, (x, y) is an example and $z_j = f(x)_j$ denotes the j-th output of the model for the j-th class, with Δ_j being the margin for class category j with constant C.

Similar to [15], we apply the ID estimate for each class by setting the margin as

$$\Delta_c = 0.5 \times \frac{\hat{d}_c}{\max_c \hat{d}_c}, \tag{5.4}$$

such that the maximum margin is 0.5.

The DRO-LT [31] framework is a multi-stage method that includes using distributional robustness loss, representation learning (i.e., feature extraction) and model fine-tuning. In particular, the robustness loss in [31] for a sample z is equal to

$$\begin{aligned} \mathcal{L}(z)_{Robust} &= -\sum_{c\in C} w(c) \sum_{z\in S_c} \log \frac{e^{-d(\hat{\mu_c},z)-\Delta_c}}{\sum_{z'} e^{-d(\hat{\mu_c},z')-\Delta'_c}}, \\ \Delta_c &= 2\varepsilon_c, \end{aligned} \tag{5.5}$$

where $d(\hat{\mu_c}, z)$ measures the distance between the sample z and the estimated centroid of its class $\hat{\mu_c}$, and Δ_c is the class margin for class c. Following [15], we utilize class ID estimates by assigning the margin as

$$\Delta_c = \frac{\hat{d}_c}{\sum_c \hat{d}_c} \times C. \tag{5.6}$$

In the case of learnable ID margins, we initialize the margin values ε_c with the normalized ID estimates $\hat{d}_c$ and allow these epsilon values to update during the training process.

5.2 Class Imbalance Mitigation Experiments

With the modifications taken in the previous section, we are ready to use intrinsic dimensionality for class imbalance mitigation purposes.

5.2.1 Class Imbalance Mitigation though Re-weighting

We apply our class ID estimations to loss re-weighting in the long-tailed CIFAR datasets using the strategy described in Section 5.1.1, and present them in Table 5.1 and Table 5.2 alongside other re-weighting strategies. These methods include standard cross-entropy with equal weights, focal loss as defined by Li et al. [28], and weighted cross-entropy with weights derived from sample sizes of classes. Inverse CE is defined using Equation 2.1 with fixed class weights of inverse sample size, $w_c = \frac{N}{N_c}$, where N is the size of the dataset, and N_c is the number of samples for class c. We define log-inverse weighted CE as a smoother variant of inverse CE for exponentially long-tailed datasets with $w_c = \frac{N}{\log(N_c)}$ if $N_c > 2$ and $w_c = \frac{N}{N_c}$ otherwise.

In the tables below, we additionally define three meta-classes (frequent, common and rare) and report the average score for class categories in these meta-classes. Specifically, for long-tailed datasets of $N = 10$ class categories we choose a 3-4-3 split, while for $N = 100$ we choose a 33-34-33 split; meaning that sorted for number of samples, the first 33 head classes are considered "frequent", the next 34 classes are considered "common" and the last 33 tail classes are considered to be in the "rare" meta-class.

With loss re-weighting, we see that inverse CE performs well on long-tailed CIFAR-10 data, while ID-based re-weighting shows its effect most clearly on Table 5.2, especially as class imbalance increases.

Table 5.1: Top-1 Accuracy results for re-weighting methods on CIFAR-10-LT

Imbalance Ratio	10				50				100			
Meta-class	F	C	R	Avg	F	C	R	Avg	F	C	R	Avg
CE	0.936	0.840	0.844	0.870	**0.928**	0.753	0.611	0.763	**0.928**	0.724	0.492	0.716
Focal Loss[28]	0.932	0.841	0.836	0.867	0.925	0.750	0.601	0.758	0.835	0.710	0.582	0.709
CE (inverse)	0.922	**0.843**	**0.857**	**0.871**	0.89	**0.771**	**0.700**	**0.786**	0.848	0.709	**0.598**	0.717
CE (log-inverse)	0.933	0.840	0.837	0.867	0.918	0.756	0.633	0.768	0.920	**0.726**	0.504	0.718
ID - TwoNN	**0.940**	0.838	0.835	0.868	0.925	0.755	0.625	0.767	0.882	0.690	0.555	0.707
ID - FisherS	0.938	0.833	0.846	0.868	0.920	0.747	0.650	0.770	0.923	0.724	0.520	**0.722**

5.3 Class Imbalance Mitigation though Re-sampling

Similarly to the previous section, we apply our class ID estimations to loss re-weighting in the long-tailed CIFAR datasets using the strategy described in Section 5.1.2, and present them in Table 5.3 and Table 5.4 alongside other methods. Here, Inverse CE and log-inverse CE are defined similarly as in Section 5.2.1, and the calculated class weights are using class sampling probabilities, identically to Section 5.1.2. Interestingly, ID-based re-sampling does very well in CIFAR-10-LT (Table 5.3), while the equally weighted cross-entropy dominates other methods in CIFAR-100-LT instead (Table 5.4).

Table 5.2: Top-1 Accuracy results for re-weighting methods on CIFAR-100-LT

Imbalance Ratio	10				50				100			
Meta-class	F	C	R	Avg	F	C	R	Avg	F	C	R	Avg
CE	0.701	0.564	0.417	0.560	0.677	0.427	0.183	0.429	0.659	**0.391**	0.117	**0.389**
Focal Loss[28]	**0.709**	0.566	0.407	0.561	**0.684**	0.425	0.151	0.420	**0.669**	0.374	0.094	0.379
CE (inverse)	0.655	**0.582**	**0.447**	0.562	0.561	0.380	0.165	0.369	0.525	0.325	0.105	0.318
CE (log-inverse)	0.694	0.578	0.435	**0.569**	0.679	0.426	**0.187**	0.431	0.658	0.365	0.107	0.377
ID - TwoNN	0.708	0.568	0.423	0.566	0.679	**0.459**	0.171	**0.437**	**0.669**	0.380	0.097	0.382
ID - FisherS	0.687	0.564	0.427	0.559	0.672	0.432	**0.187**	0.430	0.663	0.382	0.121	**0.389**

Table 5.3: Top-1 Accuracy results for re-sampling methods on CIFAR-10-LT

Imbalance Ratio	10				50				100			
Meta-class	F	C	R	Avg	F	C	R	Avg	F	C	R	Avg
CE	**0.936**	0.840	0.844	**0.870**	**0.928**	0.753	0.611	0.763	**0.928**	**0.724**	0.492	0.716
CE (inverse)	0.889	0.843	**0.869**	0.864	0.797	**0.774**	**0.720**	0.765	0.735	0.693	0.581	0.672
CE (log-inverse)	0.921	0.840	0.844	0.865	0.910	0.756	0.665	0.775	0.887	0.718	0.546	0.717
ID - TwoNN	0.934	0.824	0.848	0.864	0.907	0.762	0.696	**0.785**	0.904	0.695	0.540	0.711
ID - FisherS	0.909	**0.857**	0.831	0.865	0.902	0.761	0.684	0.780	0.824	0.696	**0.643**	**0.719**

5.4 Class Imbalance Mitigation though Re-margining

Finally, our results on using ID for class imbalance mitigation using margin-based methods (as described by Section 5.1.3) for the long-tailed CIFAR datasets are provided in Table 5.6, alongside other methods. We observe that margin-based methods

Table 5.4: Top-1 Accuracy results for re-sampling methods on CIFAR-100-LT

Imbalance Ratio	10				50				100			
Meta-class	F	C	R	Avg	F	C	R	Avg	F	C	R	Avg
CE	**0.701**	0.564	0.417	**0.560**	**0.677**	**0.427**	0.183	**0.429**	**0.659**	**0.391**	0.117	**0.389**
CE (inverse)	0.462	0.566	**0.450**	0.493	0.161	0.361	0.170	0.232	0.145	0.286	**0.128**	0.187
CE (log-inverse)	0.618	**0.576**	0.425	0.540	0.525	0.390	0.185	0.367	0.455	0.303	0.095	0.284
ID - TwoNN	0.648	0.565	0.431	0.548	0.607	0.388	**0.188**	0.394	0.594	0.357	0.104	0.352
ID - FisherS	0.616	0.557	0.448	0.541	0.525	0.397	0.183	0.369	0.508	0.334	0.102	0.315

comfortably pass the cross-entropy based baseline model. ID-based re-margining is not successful in boosting the performance of LDAM [26], however, DRO-LT [31] together with class ID performs very well on long-tailed versions of CIFAR-10, obtaining the best accuracy score from all mitigation methods explored in this thesis for imbalance ratios of 10 and 50.

Table 5.5: Top-1 Accuracy (x100) results for margin-based methods on long-tailed CIFAR datasets.

Dataset	CIFAR-10			CIFAR-100		
Imbalance Ratio	10	50	100	10	50	100
CE	87.00	76.29	71.58	56.04	42.89	38.90
LDAM + DRW [26]	87.76	**81.64**	**77.19**	**57.57**	**46.44**	**41.83**
LDAM + DRW + ID	**87.81**	81.12	76.39	57.23	45.38	41.46
DRO-LT (learned ϵ) [31]	91.02	85.88	**82.39**	**64.02**	**53.95**	**48.57**
DRO-LT + ID (ε = normalized ID)	90.56	85.39	81.73	63.40	53.12	48.52
DRO-LT + ID (ε = learned + normalized ID)	**91.10**	**85.91**	82.34	63.69	53.63	48.52

Similar to margin-based mitigation, we can use ID with logit adjustment [38], as shown in Table 5.7.

5.5 Summary

This chapter includes the integration of data-based class-wise ID into the most common strategies of class imbalance mitigation. Our results show that while ID does not provide guaranteed increase of performance for all methods, it can be used for

Table 5.6: Top-1 Accuracy (x100) results for margin-based methods on long-tailed CIFAR datasets.

Dataset	CIFAR-10			CIFAR-100		
Imbalance Ratio	10	50	100	10	50	100
CE	87.00	76.29	71.58	56.04	42.89	38.90
LDAM + DRW [26]	87.76	**81.64**	**77.19**	**57.57**	**46.44**	**41.83**
LDAM + DRW + ID	**87.81**	81.12	76.39	57.23	45.38	41.46
DRO-LT (learned ϵ) [31]	91.02	85.88	**82.39**	**64.02**	**53.95**	**48.57**
DRO-LT + ID (ε = normalized ID)	90.56	85.39	81.73	63.40	53.12	48.52
DRO-LT + ID (ε = learned + normalized ID)	**91.10**	**85.91**	82.34	63.69	53.63	48.52

Table 5.7: Top-1 Accuracy results using logit adjustment on long-tailed CIFAR datasets.

Dataset	CIFAR-10	CIFAR-100
Imbalance Ratio	100	100
CE (baseline)	0.716	0.389
Logit adjustment + ID ($\tau = 1$)	0.734	0.409
Logit adjustment + ID ($\tau = \tau^*$)	0.741	0.410
Logit adjustment [38]	**0.773**	**0.441**

improving the performance of many of the available mitigation methods, including loss re-weighting (in Table 5.2), re-sampling (in Table 5.3), as well as margin-based methods (with DRO-LT in Table 5.6).

CHAPTER 6

CONCLUSION

In this thesis, we focused on understanding and mitigating the class imbalance problem using the concept of intrinsic dimensionality. In light of this, we achieved the following:

1. Using and extending model-based ID for understanding and analyzing ID in relation to class imbalance and class complexity.

2. Using data-based ID for analysis of class manifolds and class imbalance, and as a potential imbalance measure.

3. Integrating class-wise estimated ID into several methods for class imbalance mitigation purposes.

Our first contribution was the analysis of model-based ID and its relation to class imbalance. We extended the definition of model-based ID for individual classes and demonstrated that model-based ID increases accordingly with increasing class imbalance, and non-uniform changes in ID can be attributed to the inherent complexity of each class. We observed the effects that choosing a model architecture has on the ID of the task.

Our second contribution was the analysis on data-based ID. Similarly to model-based ID, we provided detailed analysis of data-based ID estimation methods and their behavior with both synthetic and natural image datasets. As an aside, we also evaluated several data-based ID estimators on their possibilities of being used as imbalance measures.

Lastly, our final contribution was integrating class-wise estimated ID into several class imbalance mitigation methods in the current literature and our results from our experiments with ID-enriched class imbalance mitigation strategies. We showed that by using ID with the available methods in current literature, it is possible to further increase their robustness against class imbalance without explicitly training a model.

6.1 Limitations and Future Work

The main limitation of using ID estimates in natural image datasets that ID cannot be calculated and evaluated qualitatively on these datasets as there is no ground truth ID for such data. This has the effect of adding some uncertainty into the results, and varying class imbalance mitigation performance improvement depending on the quality of the ID estimate. Unfortunately, this is a fundamental problem that ID estimation models cannot solve, and ultimately forces them to calibrate themselves on synthetic data which can be generated with a ground truth intrinsic dimension.

In terms of future work, there are multiple opportunities that can be identified as potential sources of interest. One such idea is coming up with a more efficient way of computing model-based ID. The main problem of this definition of ID is the requirement of iteratively training many intermediate restricted models that have no guarantee to "converge" to an unrestricted model. Another such opportunity is to incorporate ID into other class imbalance mitigation strategies in a more end-to-end approach. Injecting dynamically computed class-wise ID iteratively into the training process of deep learning models is one such approach; many other strategies can be considered upon further investigation.

REFERENCES

[1] L. Yang, H. Jiang, Q. Song, and J. Guo, "A survey on long-tailed visual recognition," *International Journal of Computer Vision*, vol. 130, pp. 1837–1872, may 2022.

[2] A. Gupta, P. Dollar, and R. Girshick, "Lvis: A dataset for large vocabulary instance segmentation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5356–5364, 2019.

[3] F. Denti, D. Doimo, A. Laio, and A. Mira, "The generalized ratios intrinsic dimension estimator," *Scientific Reports*, vol. 12, no. 1, p. 20005, 2022.

[4] E. Levina and P. Bickel, "Maximum likelihood estimation of intrinsic dimension," *Advances in neural information processing systems*, vol. 17, 2004.

[5] D. MacKay and Z. Ghahramani, "Comments on 'maximum likelihood estimation of intrinsic dimension'by e," *Levina and P. Bickel*, 2005.

[6] P. Tempczyk, R. Michaluk, L. Garncarek, P. Spurek, J. Tabor, and A. Golinski, "Lidl: Local intrinsic dimension estimation using approximate likelihood," in *International Conference on Machine Learning*, pp. 21205–21231, PMLR, 2022.

[7] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems*, vol. 25, 2012.

[8] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

[9] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[10] P. Sun, H. Kretzschmar, X. Dotiwalla, A. Chouard, V. Patnaik, P. Tsui, J. Guo, Y. Zhou, Y. Chai, B. Caine, *et al.*, "Scalability in perception for autonomous driving: Waymo open dataset," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2446–2454, 2020.

[11] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pp. 234–241, Springer, 2015.

[12] J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, *et al.*, "Highly accurate protein structure prediction with alphafold," *Nature*, vol. 596, no. 7873, pp. 583–589, 2021.

[13] D. Silver, T. Hubert, J. Schrittwieser, I. Antonoglou, M. Lai, A. Guez, M. Lanctot, L. Sifre, D. Kumaran, T. Graepel, *et al.*, "Mastering chess and shogi by self-play with a general reinforcement learning algorithm," *arXiv preprint arXiv:1712.01815*, 2017.

[14] K. Oksuz, B. C. Cam, S. Kalkan, and E. Akbas, "Imbalance problems in object detection: A review," *IEEE transactions on pattern analysis and machine intelligence*, vol. 43, no. 10, pp. 3388–3415, 2020.

[15] Z. S. Baltaci, K. Oksuz, S. Kuzucu, K. Tezoren, B. K. Konar, A. Ozkan, E. Akbas, and S. Kalkan, "Class uncertainty: A measure to mitigate class imbalance," *arXiv preprint arXiv:2311.14090*, 2023.

[16] S. Khan, M. Hayat, S. W. Zamir, J. Shen, and L. Shao, "Striking the right balance with uncertainty," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 103–112, 2019.

[17] A. Ashukha, A. Lyzhov, D. Molchanov, and D. Vetrov, "Pitfalls of in-domain uncertainty estimation and ensembling in deep learning," *arXiv preprint arXiv:2002.06470*, 2020.

[18] E. Hüllermeier and W. Waegeman, "Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods," *Machine Learning*, vol. 110, pp. 457–506, 2021.

[19] M. Valdenegro-Toro and D. S. Mori, "A deeper look into aleatoric and epistemic uncertainty disentanglement," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 1508–1516, IEEE, 2022.

[20] Y. Zhang, B. Kang, B. Hooi, S. Yan, and J. Feng, "Deep long-tailed learning: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

[21] Y. Zhang, X. Wei, B. Zhou, and J. Wu, "Bag of tricks for long-tailed visual recognition with deep convolutional neural networks," in *AAAI*, pp. 3447–3455, 2021.

[22] H. Han, W.-Y. Wang, and B.-H. Mao, "Borderline-smote: a new over-sampling method in imbalanced data sets learning," in *International conference on intelligent computing*, pp. 878–887, Springer, 2005.

[23] C. Drummond, R. C. Holte, *et al.*, "C4. 5, class imbalance, and cost sensitivity: why under-sampling beats over-sampling," in *Workshop on learning from imbalanced datasets II*, vol. 11, pp. 1–8, 2003.

[24] B. Kang, S. Xie, M. Rohrbach, Z. Yan, A. Gordo, J. Feng, and Y. Kalantidis, "Decoupling representation and classifier for long-tailed recognition," *arXiv preprint arXiv:1910.09217*, 2019.

[25] Y. Cui, Y. Song, C. Sun, A. Howard, and S. Belongie, "Large scale fine-grained categorization and domain-specific transfer learning," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4109–4118, 2018.

[26] K. Cao, C. Wei, A. Gaidon, N. Arechiga, and T. Ma, "Learning imbalanced datasets with label-distribution-aware margin loss," *Advances in neural information processing systems*, vol. 32, 2019.

[27] Y. Cui, M. Jia, T.-Y. Lin, Y. Song, and S. Belongie, "Class-balanced loss based on effective number of samples," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9268–9277, 2019.

[28] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proceedings of the IEEE international conference on computer vision*, pp. 2980–2988, 2017.

[29] S. Park, J. Lim, Y. Jeon, and J. Y. Choi, "Influence-balanced loss for imbalanced visual classification," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 735–744, 2021.

[30] F. Wang, J. Cheng, W. Liu, and H. Liu, "Additive margin softmax for face verification," *IEEE Signal Processing Letters*, vol. 25, pp. 926–930, jul 2018.

[31] D. Samuel and G. Chechik, "Distributional robustness loss for long-tail learning," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9495–9504, 2021.

[32] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond empirical risk minimization," *arXiv preprint arXiv:1710.09412*, 2017.

[33] V. Verma, A. Lamb, C. Beckham, A. Najafi, I. Mitliagkas, D. Lopez-Paz, and Y. Bengio, "Manifold mixup: Better representations by interpolating hidden states," in *International conference on machine learning*, pp. 6438–6447, PMLR, 2019.

[34] H.-P. Chou, S.-C. Chang, J.-Y. Pan, W. Wei, and D.-C. Juan, "Remix: rebalanced mixup," in *Computer Vision–ECCV 2020 Workshops: Glasgow, UK, August 23–28, 2020, Proceedings, Part VI 16*, pp. 95–110, Springer, 2020.

[35] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: synthetic minority over-sampling technique," *Journal of artificial intelligence research*, vol. 16, pp. 321–357, 2002.

[36] Y. Li, T. Wang, B. Kang, S. Tang, C. Wang, J. Li, and J. Feng, "Overcoming classifier imbalance for long-tail object detection with balanced group softmax," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10991–11000, 2020.

[37] B. Zhou, Q. Cui, X.-S. Wei, and Z.-M. Chen, "Bbn: Bilateral-branch network with cumulative learning for long-tailed visual recognition," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9719–9728, 2020.

[38] A. K. Menon, S. Jayasumana, A. S. Rawat, H. Jain, A. Veit, and S. Kumar, "Long-tail learning via logit adjustment," *arXiv preprint arXiv:2007.07314*, 2020.

[39] S. Zhang, Z. Li, S. Yan, X. He, and J. Sun, "Distribution alignment: A unified framework for long-tail visual recognition," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2361–2370, 2021.

[40] Y. Hong, S. Han, K. Choi, S. Seo, B. Kim, and B. Chang, "Disentangling label distribution for long-tailed visual recognition," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 6626–6636, 2021.

[41] R. Rahaman *et al.*, "Uncertainty quantification and deep ensembles," *Advances in Neural Information Processing Systems*, vol. 34, pp. 20063–20075, 2021.

[42] B. Lakshminarayanan, A. Pritzel, and C. Blundell, "Simple and scalable predictive uncertainty estimation using deep ensembles," *Advances in neural information processing systems*, vol. 30, 2017.

[43] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, 2015.

[44] J. Gou, B. Yu, S. J. Maybank, and D. Tao, "Knowledge distillation: A survey," *International Journal of Computer Vision*, vol. 129, pp. 1789–1819, 2021.

[45] C. Huang, Y. Li, C. C. Loy, and X. Tang, "Learning deep representation for imbalanced classification," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5375–5384, 2016.

[46] Q. Dong, S. Gong, and X. Zhu, "Class rectification hard mining for imbalanced deep learning," in *Proceedings of the IEEE international conference on computer vision*, pp. 1851–1860, 2017.

[47] T. Hospedales, A. Antoniou, P. Micaelli, and A. Storkey, "Meta-learning in neural networks: A survey," *IEEE transactions on pattern analysis and machine intelligence*, vol. 44, no. 9, pp. 5149–5169, 2021.

[48] S. Li, K. Gong, C. H. Liu, Y. Wang, F. Qiao, and X. Cheng, "Metasaug: Meta semantic augmentation for long-tailed visual recognition," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5212–5221, 2021.

[49] P. Campadelli, E. Casiraghi, C. Ceruti, and A. Rozza, "Intrinsic dimension estimation: Relevant techniques and a benchmark framework," *Mathematical Problems in Engineering*, vol. 2015, pp. 1–21, 2015.

[50] F. Camastra and A. Staiano, "Intrinsic dimension estimation: Advances and open problems," *Information Sciences*, vol. 328, pp. 26–41, 2016.

[51] K. Fukunaga, *Introduction to statistical pattern recognition*. Elsevier, 2013.

[52] C. Li, H. Farkhoor, R. Liu, and J. Yosinski, "Measuring the intrinsic dimension of objective landscapes," in *International Conference on Learning Representations*, 2018.

[53] D. Mo and S. H. Huang, "Fractal-based intrinsic dimension estimation and its application in dimensionality reduction," *IEEE Transactions on knowledge and data engineering*, vol. 24, no. 1, pp. 59–71, 2010.

[54] M. Fan, N. Gu, H. Qiao, and B. Zhang, "Intrinsic dimension estimation of data by principal component analysis," *arXiv preprint arXiv:1002.2050*, 2010.

[55] A. M. Farahmand, C. Szepesvári, and J.-Y. Audibert, "Manifold-adaptive dimension estimation," in *Proceedings of the 24th international conference on Machine learning*, pp. 265–272, 2007.

[56] J. B. Tenenbaum, V. d. Silva, and J. C. Langford, "A global geometric framework for nonlinear dimensionality reduction," *science*, vol. 290, no. 5500, pp. 2319–2323, 2000.

[57] H. Hotelling, "Analysis of a complex of statistical variables into principal components.," *Journal of educational psychology*, vol. 24, no. 6, p. 417, 1933.

[58] M. E. Tipping and C. M. Bishop, "Probabilistic principal component analysis," *Journal of the Royal Statistical Society Series B: Statistical Methodology*, vol. 61, no. 3, pp. 611–622, 1999.

[59] C. Bishop, "Bayesian pca," *Advances in neural information processing systems*, vol. 11, 1998.

[60] K. Fukunaga and D. R. Olsen, "An algorithm for finding intrinsic dimensionality of data," *IEEE Transactions on computers*, vol. 100, no. 2, pp. 176–183, 1971.

[61] S. T. Roweis and L. K. Saul, "Nonlinear dimensionality reduction by locally linear embedding," *science*, vol. 290, no. 5500, pp. 2323–2326, 2000.

[62] P. Grassberger and I. Procaccia, "Measuring the strangeness of strange attractors," *Physica D: nonlinear phenomena*, vol. 9, no. 1-2, pp. 189–208, 1983.

[63] F. Camastra and A. Vinciarelli, "Estimating the intrinsic dimension of data with a fractal-based method," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 24, no. 10, pp. 1404–1407, 2002.

[64] E. Facco, M. d'Errico, A. Rodriguez, and A. Laio, "Estimating the intrinsic dimension of datasets by a minimal neighborhood information," *Scientific reports*, vol. 7, no. 1, p. 12140, 2017.

[65] C. Ceruti, S. Bassis, A. Rozza, G. Lombardi, E. Casiraghi, and P. Campadelli, "Danco: An intrinsic dimensionality estimator exploiting angle and norm concentration," *Pattern recognition*, vol. 47, no. 8, pp. 2569–2581, 2014.

[66] L. Albergante, J. Bac, and A. Zinovyev, "Estimating the effective dimension of large biological datasets using fisher separability analysis," in *2019 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8, IEEE, 2019.

[67] A. Krizhevsky, G. Hinton, *et al.*, "Learning multiple layers of features from tiny images," 2009.

[68] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[69] Z. Liu, Z. Miao, X. Zhan, J. Wang, B. Gong, and S. X. Yu, "Large-scale long-tailed recognition in an open world," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2537–2546, 2019.

[70] G. Van Horn, O. Mac Aodha, Y. Song, Y. Cui, C. Sun, A. Shepard, H. Adam, P. Perona, and S. Belongie, "The inaturalist species classification and detection dataset," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 8769–8778, 2018.

[71] P. Pope, C. Zhu, A. Abdelkader, M. Goldblum, and T. Goldstein, "The intrinsic dimension of images and its impact on learning," *arXiv preprint arXiv:2104.08894*, 2021.

APPENDIX A

EXPERIMENT DETAILS

A.1 Experiment setup

For our experiments with class imbalance mitigation using class-wise ID, we adopt the following setup.

- **Data:** We use CIFAR-10-LT and CIFAR-100-LT as described by Section 2.5. During model training, we use random horizontal flipping alongside channel normalization as our preprocessing step, with long-tail imbalanced ratios of 10, 50 and 100.

- **Training:** For our experiments using loss re-weighting, re-sampling and LDAM [26] in re-margining, we use code from the Bag of Tricks [21] framework. Some hyperparameters for commonly-trained cross-entropy based models include the SGD optimizer, momentum with a value of 0.9, and a weight decay of 2×10^{-4}. The models for re-weighting and re-sampling are trained for 200 epochs with a batch size of 128, on a RTX2060 Super. Other hyperparameters, such as learning rate scheduling and deferred re-weighting (with LDAM), are kept as-is from [21]. Our experiments with DRO-LT [31] use the codebase provided by the authors. The first stage of training (i.e., representation learning with cross-entropy) is taken as-is, and the fine-tuning stage using ID alongside learnable class margins is described in Section 5.1.3, with the same hardware as the other experiments.

A.2 Additional Results

This section includes additional results that we have obtained on long-tailed CIFAR-10 and CIFAR-100 datasets using F1 as the evaluation metric instead of top-1 accuracy. F1 is a better metric than top-1 accuracy for imbalanced multi-class classification; however, as the CIFAR test sets contain the same number of samples per class category, in practice F1 does not offer any meaningful advantage over top-1 accuracy. This can also be observed by comparing the results here to the ones in Chapter 5.

Table A.1: F1 results for re-weighting methods on CIFAR-10-LT

Imbalance Ratio	10				50				100			
Meta-class	F	C	R	Avg	F	C	R	Avg	F	C	R	Avg
CE	**0.880**	0.843	0.899	0.870	0.781	0.760	0.747	0.762	0.748	0.730	0.651	0.711
Focal Loss [28]	0.871	**0.845**	0.894	0.867	0.778	0.757	0.738	0.757	0.749	0.689	0.685	0.705
CE (inverse)	0.878	0.843	**0.904**	**0.871**	**0.803**	0.764	**0.805**	**0.788**	0.746	0.694	**0.721**	0.718
CE (log-inverse)	0.872	0.843	0.896	0.867	0.788	0.760	0.763	0.769	0.743	0.734	0.658	0.713
ID - TwoNN	0.874	0.843	0.897	0.868	0.779	**0.767**	0.756	0.767	**0.757**	0.668	0.690	0.701
ID - FisherS	0.878	0.840	0.898	0.868	0.788	0.756	0.777	0.771	0.745	**0.736**	0.675	**0.720**

Table A.2: F1 results for re-sampling methods on CIFAR-10-LT

Imbalance Ratio	10				50				100			
Meta-class	F	C	R	Avg	F	C	R	Avg	F	C	R	Avg
CE	**0.880**	**0.843**	0.899	**0.870**	0.781	0.760	0.747	0.762	0.748	**0.730**	0.651	0.711
CE (inverse)	0.867	0.832	**0.907**	0.864	0.759	0.745	**0.801**	0.766	0.689	0.649	0.694	0.674
CE (log-inverse)	0.870	0.839	0.901	0.866	0.785	0.764	0.781	0.775	**0.749**	0.706	0.693	0.714
ID - TwoNN	0.868	0.836	0.901	0.865	**0.796**	**0.771**	**0.801**	**0.787**	0.737	0.707	0.686	0.709
ID - FisherS	0.872	0.839	0.895	0.865	0.787	0.766	0.796	0.781	0.736	0.688	**0.748**	**0.720**

Table A.3: F1 results for re-weighting methods on CIFAR-100-LT

Imbalance Ratio	10				50				100			
Meta-class	F	C	R	Avg	F	C	R	Avg	F	C	R	Avg
CE	0.592	0.573	0.495	0.553	0.493	0.444	0.256	0.398	0.455	**0.415**	0.171	**0.347**
Focal Loss [28]	0.588	0.570	0.493	0.550	0.49	0.444	0.218	0.384	0.449	0.406	0.138	0.331
CE (inverse)	0.567	**0.578**	**0.529**	0.558	0.403	0.397	0.235	0.345	0.359	0.345	0.145	0.283
CE (log-inverse)	**0.597**	0.577	0.517	**0.563**	0.491	0.448	**0.262**	**0.400**	0.450	0.393	0.158	0.334
ID - TwoNN	**0.597**	0.574	0.509	0.560	**0.498**	**0.466**	0.234	**0.400**	**0.457**	0.396	0.144	0.333
ID - FisherS	0.584	0.568	0.505	0.552	0.485	0.454	0.260	**0.400**	0.455	0.409	**0.174**	0.346

Table A.4: F1 results for re-sampling methods on CIFAR-100-LT

Imbalance Ratio	10				50				100			
Meta-class	F	C	R	Avg	F	C	R	Avg	F	C	R	Avg
CE	**0.592**	**0.573**	0.495	**0.553**	**0.493**	**0.444**	0.256	**0.398**	**0.455**	**0.415**	**0.171**	**0.347**
CE (inverse)	0.437	0.511	0.502	0.483	0.117	0.281	0.200	0.200	0.075	0.218	0.147	0.147
CE (log-inverse)	0.543	0.563	0.501	0.536	0.390	0.392	**0.258**	0.346	0.309	0.316	0.136	0.254
ID - TwoNN	0.563	0.566	0.505	0.544	0.440	0.408	0.253	0.367	0.404	0.390	0.153	0.316
ID - FisherS	0.541	0.551	**0.522**	0.538	0.389	0.401	0.253	0.348	0.354	0.343	0.145	0.281