

QUANTIFYING AND MITIGATING CLASS IMBALANCE IN LONG-TAILED
VISUAL RECOGNITION

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

BY

ZEYNEP SONAT BALTACI

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

JULY 2022

Approval of the thesis:

**QUANTIFYING AND MITIGATING CLASS IMBALANCE IN
LONG-TAILED VISUAL RECOGNITION**

submitted by **ZEYNEP SONAT BALTACI** 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**

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

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

Examining Committee Members:

Assist. Prof. Dr. Ramazan Gökberk Cinbiş
Computer Engineering, METU

Assoc. Prof. Dr. Sinan Kalkan
Computer Engineering, METU

Prof. Dr. Pınar Duygulu Şahin
Computer Engineering, Hacettepe University

Date:

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: Zeynep Sonat Baltacı

Signature :

ABSTRACT

QUANTIFYING AND MITIGATING CLASS IMBALANCE IN LONG-TAILED VISUAL RECOGNITION

Baltacı, Zeynep Sonat

M.S., Department of Computer Engineering

Supervisor: Assoc. Prof. Dr. Sinan Kalkan

Co-Supervisor: Assist. Prof. Dr. Emre Akbaş

July 2022, 70 pages

Objects are distributed unevenly in real world, which manifests itself as a long-tailed distribution in realistic visual recognition datasets. Deep learning based approaches trained on such imbalanced datasets using conventional gradient-based training strategies exhibit unfair recognition performances towards classes that are under-represented in the dataset. This so-called class imbalance has been studied in the literature by measuring imbalance via either class frequency or class hardness, and using those measures to mitigate imbalance by sampling, loss weighting or calibration strategies. In this thesis, we argue and empirically show that sample frequency or hardness alone is not sufficient for capturing imbalance among classes. Then we propose a novel measure based on predictive uncertainty of a trained deep network and demonstrate that it can capture imbalance better than existing approaches. Finally, we incorporate our measure to existing imbalance mitigation methods: loss reweighting, resampling, margin-based methods, and two-stage training. We show that predictive uncertainty-based methods improve over or perform on par with existing baselines on long-tailed datasets CIFAR-10-LT, CIFAR-100-LT and ImageNet-LT.

Keywords: long-tailed visual recognition, class imbalance, predictive uncertainty, imbalance mitigation

ÖZ

UZUN KUYRUKLU GÖRSEL TANIMADA SINIF DENGESİZLİĞİNİN ÖLÇÜLMESİ VE AZALTILMASI

Baltacı, Zeynep Sonat

Yüksek Lisans, Bilgisayar Mühendisliği Bölümü

Tez Yöneticisi: Doç. Dr. Sinan Kalkan

Ortak Tez Yöneticisi: Dr. Öğr. Üyesi. Emre Akbaş

Temmuz 2022, 70 sayfa

Gerçek dünyada nesnelere, kategorilere eşit olmayan bir şekilde dağılır. Bu durum gerçekçi görsel tanıma veri kümelerinde uzun kuyruklu bir dağılım olarak kendini gösterir. Geleneksel gradyan tabanlı eğitim stratejileri kullanılarak bu tür dengesiz veri kümeleri üzerinde eğitilen derin öğrenme tabanlı yaklaşımlar, veri kümesinde yeterince temsil edilmeyen sınıflara karşı adil olmayan tanıma performansları sergiler. Bu sözde sınıf dengesizliği, literatürde, dengesizliği sınıf frekansı veya sınıf zorluğu yoluyla ölçerek ve bu ölçümleri örnekleme, kayıp fonksiyonu ağırlıklandırma veya kalibrasyon stratejileri yoluyla dengesizliği azaltmak için kullanılarak çalışılmıştır. Bu tezde, sınıflar arasındaki dengesizliği yakalamak için örnek frekansının veya zorluğunun tek başına yeterli olmadığını tartışıyor ve deneysel olarak gösteriyoruz. Ardından, eğitilmiş bir derin ağın tahmin belirsizliğine dayanan yeni bir ölçüt öneriyoruz ve bunun dengesizliği mevcut yaklaşımlardan daha iyi yakalayabildiğini gösteriyoruz. Son olarak, ölçümümüzü mevcut dengesizlik azaltma yöntemlerine dahil ediyoruz: kayıp fonksiyonunu yeniden ağırlıklandırma, yeniden örnekleme, marj tabanlı yöntemler

ve iki aşamalı eğitim. Tahmine dayalı belirsizliğe dayalı yöntemlerin, uzun kuyruklu CIFAR-10-LT, CIFAR-100-LT ve ImageNet-LT veri kümelerinde mevcut temellere göre iyileştirdiğini veya yakın performans gösterdiğini gösteriyoruz.

Anahtar Kelimeler: uzun kuyruklu görsel tanıma, sınıf dengesizliği, tahmine dayalı belirsizlik, dengesizliğin azaltılması

This thesis is dedicated to my father, who always did his best to help us achieve ours.

ACKNOWLEDGMENTS

First, I would like to express my gratitude to Assoc. Prof. Dr. Sinan Kalkan and Assist. Prof. Dr. Emre Akbař for their guidance and encouragement. They helped me believe in myself as a researcher. I feel fortunate to have them as my thesis supervisors.

I wish to extend my thanks to all the great people in our research group, especially Dr. Kemal Öksüz, for our fruitful discussions and their support throughout my study.

I am incredibly grateful for my mom, brother, and friends' accompany before and during my Master's degree. It feels reassuring to have them always by my side.

Lastly, I would like to thank my partner, Furkan. I couldn't imagine how this period of my life would be without his endless and unconditional support.

Throughout this study, we utilized the computational resources of Robotics and Artificial Intelligence Technologies Application and Research Center (ROMER) and METU ImageLab for the analysis and experiments conducted.

The experiments reported in this paper were partially performed at High Performance and Grid Computing Center (TRUBA resources) of Turkish Academic Network and Information Center (TÜBİTAK ULAKBİM).

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	xviii
CHAPTERS	
1 INTRODUCTION	1
1.1 Motivation	1
1.2 Problem Definition and Scope of the Thesis	2
1.3 Contributions	5
1.4 The Outline of the Thesis	5
2 BACKGROUND AND RELATED WORK	7
2.1 Primary Task: Long-tailed Visual Recognition	7
2.2 Addressing Class Imbalance	7
2.2.1 Resampling Methods	9
2.2.2 Reweighting Methods	10

2.2.3	Margin-based Methods	11
2.2.4	Two-stage Methods	11
2.2.5	Other Methods	12
2.3	Uncertainty Estimation in Deep Learning	13
2.3.1	Types of Uncertainty	14
2.3.2	Methods	14
2.4	Performance Evaluation	17
2.4.1	Top-1 Error Rate	17
2.5	Datasets	17
2.5.1	CIFAR-10-LT and CIFAR-100-LT	17
2.5.2	ImageNet-LT	18
2.6	Comparative Summary	18
3	A NOVEL MEASURE OF CLASS IMBALANCE: PREDICTIVE UN- CERTAINTY	21
3.1	Limitations of Existing Measures and Imbalance Mitigation Methods	21
3.2	Requirements for an Imbalance Measure	24
3.3	Measuring Imbalance with Predictive Uncertainty	24
3.3.1	Analyzing Class Imbalance with Predictive Uncertainty	25
4	MITIGATING CLASS IMBALANCE WITH PREDICTIVE UNCERTAINTY	39
4.1	Integrating Predictive Uncertainty into Existing Methods	39
4.1.1	Resampling Methods with Our Measure	41
4.1.2	Reweighting Methods with Our Measure	41
4.1.3	Margin-based Methods with Our Measure	42

4.1.4	Two-stage Methods with Our Measure	43
5	EXPERIMENTS	45
5.1	Datasets	45
5.2	Training Details	46
5.3	Experiments on Our Measure with Resampling Methods	48
5.4	Experiments on Our Measure with Reweighting Methods	49
5.5	Experiments on Our Measure with Margin-based Methods	50
5.6	Experiments on Our Measure with Two-stage Methods	50
5.6.1	Two-stage Resampling	50
5.6.2	Two-stage Reweighting	51
5.6.3	Bag of Tricks	51
5.7	Ablation Analysis on Uncertainty Estimation	53
5.7.1	Epistemic Uncertainty to Mitigate Class Imbalance	53
5.7.2	Calibrated Uncertainty to Mitigate Class Imbalance	53
5.7.3	Combining Class-Balanced Loss with Predictive Uncertainty	54
5.8	Discussion	54
6	CONCLUSION	57
6.1	Limitations and Future Work	57
	REFERENCES	61
	APPENDICES	
A	UNCERTAINTY ESTIMATION	69
A.1	Experiments on Estimating Uncertainty	69

A.1.1	Datasets	69
A.1.2	Model Selection	69
A.1.3	Training Details	70

LIST OF TABLES

TABLES

Table 3.1	Comparison of class imbalance measures.	25
Table 5.1	Top-1 error rates with standard deviations of one-stage resampling methods along with our measure on CIFAR-LT datasets.	49
Table 5.2	Top-1 error rates with standard deviations of one-stage reweighting methods along with our measure on CIFAR-LT datasets.	49
Table 5.3	Top-1 error rates with standard deviations of margin-based methods along with our measure.	50
Table 5.4	Top-1 error rates with standard deviations of the best performing two-stage resampling method [1] along with our measure on CIFAR-LT datasets.	51
Table 5.5	Top-1 error rates with standard deviations of the best performing two-stage reweighting methods [2, 3] along with our measure on CIFAR-LT datasets.	52
Table 5.6	Top-1 error rates with standard deviations of Bag of Tricks [1] along with our measure.	52
Table 5.7	Empirical effects of calibration on predictive and epistemic uncertainty as measures on CIFAR-10-LT.	54
Table 5.8	Top-1 error rates with standard deviations of one-stage reweighting with a linear combination of uncertainty and class-balanced weighted Focal loss.	55

LIST OF FIGURES

FIGURES

Figure 1.1	An illustration of a long-tailed dataset.	2
Figure 1.2	Subset of images from BRATS dataset [4].	3
Figure 2.1	Common groups of imbalance methods shown on the training pipeline	8
Figure 2.2	Uncertainty maps of frames in CamVid dataset [5].	13
Figure 2.3	Uncertainty types illustrated on a regression task.	15
Figure 2.4	Number of images in CIFAR-100-LT with imbalance ratio 100. .	19
Figure 3.1	Average top-1 error rates of classes in CIFAR-10-LT.	23
Figure 3.2	Average predictive uncertainty values of each class in CIFAR-10-LT.	27
Figure 3.3	Predictive uncertainty values of each class in CIFAR-10-LT unbalanced in reversed order.	28
Figure 3.4	Predictive uncertainty values of each class in CIFAR-100-LT. . .	28
Figure 3.5	Average accuracy of classes vs. soft-sampling ratio of CIFAR-10-LT with 50 imbalance ratio.	30
Figure 3.6	Average predictive uncertainty of classes vs. soft-sampling ratio of CIFAR-10-LT.	31

Figure 3.7	Average predictive uncertainty of classes vs. soft-sampling ratio of CIFAR-100-LT.	32
Figure 3.8	Training and validation values of balanced CIFAR-10.	33
Figure 3.9	Training and validation values of balanced CIFAR-100.	34
Figure 3.10	Training and validation values of CIFAR-10-LT with 50 imbalance ratio.	35
Figure 3.11	Training and validation values of CIFAR-10-LT with 100 imbalance ratio.	36
Figure 3.12	Training and validation values of CIFAR-100-LT with 50 imbalance ratio.	37
Figure 3.13	Training and validation values of CIFAR-10-LT with 100 imbalance ratio.	38
Figure 4.1	High-level scheme of our architecture.	40

LIST OF ABBREVIATIONS

CAM	Class Activation Maps
CB	Class-balanced
CE	Cross Entropy
CSCE	Cost-sensitive Cross Entropy
DE	Deep Ensembles
DRO-LT	Distributional Robustness Loss
DRW	Deferred Reweighting
IR	Imbalance Ratio
LDAM	Label-Distribution-Aware Margin Loss
LT	Long-tailed
MI	Mutual Information
PB	Progressively Balanced
SGD	Stochastic Gradient Descent

CHAPTER 1

INTRODUCTION

1.1 Motivation

Although there are solid and practical studies to solve visual recognition problems, it is not always possible to successfully transfer these solutions to daily applications. The data in the application areas where visual recognition models are used do not naturally have a balanced distribution. For example, in healthcare applications and the integration of autonomous systems into daily life, target classes have a very narrow representation within the dataset. Figure 1.2 shows examples from the BRATS, a brain tumor dataset. When we look at the image pixel distribution, we can see that the number of pixels including tumors is significantly lower than that of healthy tissues. Furthermore, the models that encounter samples outside the training data space may lead to performance drops. These performance drops affect the use of deep learning models in medical and safety applications and require expert interventions. In addition, collecting or creating a dataset is a human-involved process. The collected data can also be of direct human origin (e.g., social media data) and reflect personal opinions, trends, and social bias in datasets [6]. The decisions made by the models that encounter with such data they are trained in can also cause sociological and ethical problems. All these reasons make working on the class imbalance problem necessary and essential.

Class imbalance is defined in the literature as the unequal distribution of classes in the dataset. Figure 1.1 shows a *long-tailed* data distribution where there is a severe imbalance from overrepresented (*head*) to underrepresented (*tail*) classes. Continuing from this cardinality-based perspective, the aim of the methods that mitigate class

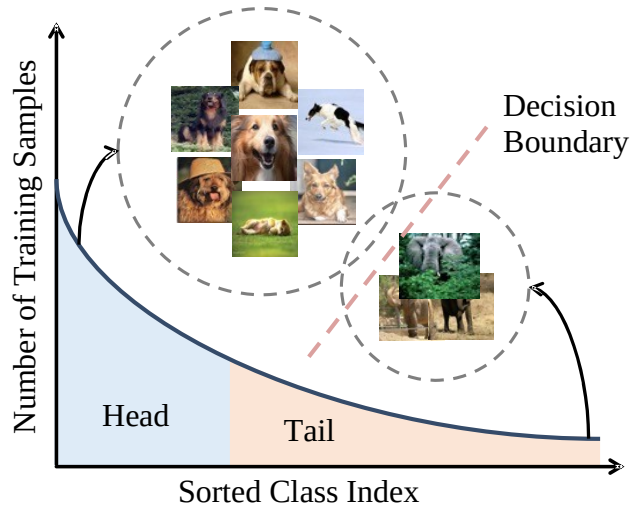


Figure 1.1: An illustration of a long-tailed dataset. Figure is retrieved from [7].

imbalance is to move the decision boundary of a standard classifier to a target decision boundary which separates the classes clearly. However, this is not the only factor that can cause imbalances and learning differences between classes. We should also account for the quality and representation span of the samples within data space. Although there are successful methods that try to solve the imbalance problem based on class cardinality or sample hardness, the lack of inclusiveness of these definitions alone also reveals the shortcomings of existing solutions and their potential for improvement.

The main motivations of this thesis are (i) whether we can extend the notion of class imbalance in the literature with a more general definition and (ii) if it is applicable as a solution to improve methods based on either the class cardinality or difficulty to mitigate imbalance. In the following sections, we define the problem we have identified in detail, explain the scope of this thesis and the proposed method.

1.2 Problem Definition and Scope of the Thesis

We assume that we have a supervised image classification problem, for a dataset \mathcal{D} of N examples and their classes $\{(\mathbf{x}_1, \mathbf{y}_1), \dots, (\mathbf{x}_N, \mathbf{y}_N)\}$, $\mathbf{x} \in \mathcal{X}$ and $\mathbf{y} \in \mathcal{Y}$. We are

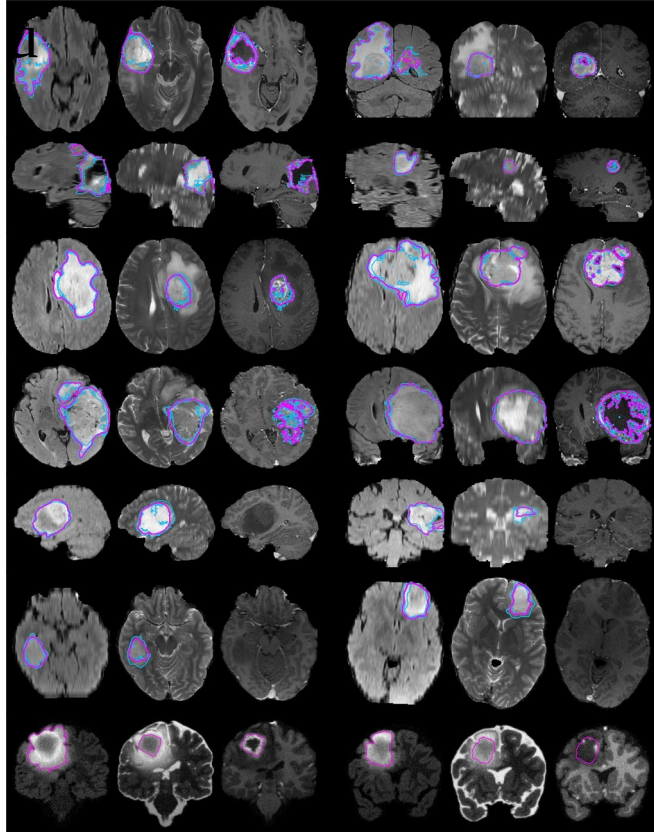


Figure 1.2: Subset of brain tumor images from BRATS dataset [4]. Purple and blue areas represent brain tumors. Image is retrieved from [4].

interested in finding the parameters (θ) of the mapping $f : \mathcal{X} \rightarrow \mathcal{Y}$ by solving the following optimization problem:

$$\theta^* \leftarrow \arg \min_{\theta} \mathcal{L}(f(\mathbf{x}; \theta), \mathbf{y}), \quad (1.1)$$

where $\mathcal{L}(\hat{\mathbf{y}}, \mathbf{y})$, e.g. Cross Entropy (CE), is a loss function measuring the dissimilarity of the prediction, $\hat{\mathbf{y}} = f(\mathbf{x}; \theta)$, with the correct output \mathbf{y} .

Such an approach is prone to an uneven distribution of samples across the classes in the dataset. Called *class imbalance* problem in literature [8], if class c_i has more samples than class c_j , i.e. $n_{c_i} > n_{c_j}$, class c_i will have more opportunities to update the parameters using vanilla optimization methods (e.g. Stochastic Gradient Descent [9]) for Eq. 1.1. This in turn will lead to sub-optimal performance for class c_j . The severity of the problem will increase if $n_{c_i} \gg n_{c_j}$, i.e. the imbalance between classes is higher. On the other hand, there may occur an imbalance among classes through its sample qualities and hardness. From this intuitive point of view, the definition of class imbalance relying on cardinalities is not inclusive, since it only reflects one side of imbalance, varying class frequencies.

We are interested in measuring the imbalance in a classification problem, with scalar values for each class denoted as $\mu_{c_0}, \mu_{c_1}, \dots, \mu_{c_N}$, such that $\mu_{c_j} > \mu_{c_i}$ iff. the mapping f favors c_i more than c_j . We would like to use the measure μ to analyze the presence of imbalance in possible imbalance sources: cardinality and hardness. We would like to investigate the comprehensiveness of our measure and develop better imbalance mitigation methods by using measure μ as a replacement for class cardinality and hardness in existing solutions to class imbalance.

We can group the existing set of approaches to address the class imbalance problem into four categories: resampling, loss reweighting, margin enforcing to class decision boundaries and two-stage training. These methods aim to balance different stages of training to mitigate the effects of class imbalance on model performance. Although these methods provide improvements over baselines, most of them depend on the frequencies of the classes. We suggest that the number of examples in a class does not broadly represent class imbalance. The distribution of the examples in data space and hardness should also be an essential consideration.

In this thesis, we limit ourselves to long-tailed visual recognition and perform our experiments and analyzes only on such datasets. However, our measure and the methods using our measure are generic and can be applied to other long-tailed classification problems.

1.3 Contributions

In this thesis, we make the following contributions:

- We argue and empirically show that existing imbalance mitigation strategies based on class frequency or hardness are not sufficient and comprehensive in different terms.
- We introduce a novel class imbalance measure based on predictive uncertainty, which reflects the uncertainty of a network on its prediction \hat{y} , analyze and compare our measure with class difficulty and frequency-based imbalance definitions. We show over various imbalance ratios on CIFAR-10-LT and CIFAR-100-LT datasets that imbalance can be represented by predictive uncertainty.
- We compute the predictive uncertainty for each class of an imbalanced dataset with an ensemble network and exploit our measure to mitigate class imbalance by loss reweighting and mini-batch resampling in one- and two-stage training and forcing a margin to have separable representations in feature space. Experiments show that our measure can be easily applied to a training pipeline as a mitigation method and improve the baseline.

1.4 The Outline of the Thesis

In Chapter 2, we identify our primary task as long-tailed (LT) recognition. We give a detailed background on existing imbalance mitigation methods and a comparative summary of recent studies. Following, we present an outline of uncertainty estimation in deep learning. We also provide brief information on performance evaluation metrics and datasets for ablation studies and experiments.

In Chapter 3, we suggest the requirements for an imbalance measure, discuss the limitations of methods summarized in Chapter 2 and propose predictive uncertainty as a measure of class imbalance.

In Chapter 4, we describe how we integrate our measure with existing mitigation methods within resampling, loss reweighting, margin-enforcement and two-stage training.

In Chapter 5, we present the results of our measure as a mitigation method with various architectures on long-tailed datasets.

Lastly, in Chapter 6, we discuss the limitations of our approach and provide future work.

CHAPTER 2

BACKGROUND AND RELATED WORK

2.1 Primary Task: Long-tailed Visual Recognition

A long-tailed dataset has a class distribution where the frequencies decrease severely among classes, leading some classes having remarkably more number of samples compared to others. The naming comes from the distribution where we plot number of samples in classes from the one with highest number of samples to the lowest (see Figure 1.1). The classes with more samples are called head and fewer samples are called tail respectively.

Long-tailed visual recognition is an image classification task on a long-tailed dataset, which we count as one of the factors giving rise to class imbalance. The main aim in long-tailed visual recognition is to mitigate the effects of class imbalance and provide a performance increase compared to an imbalanced training scheme. Although we argue that class cardinality is not enough to represent class imbalance, since long-tailed distribution guarantees varying class cardinality, we adopt long-tailed visual recognition as our primary task to analyze our measure and utilize as a mitigation method.

2.2 Addressing Class Imbalance

As we mathematically defined in Section 1.2, *imbalanced dataset* is defined as having two or more classes with a biased distribution of examples. Another interpretation of this definition is to have remarkably different class frequencies. In the real world, representatives of classes have varying occurrences, which makes the unseen test data

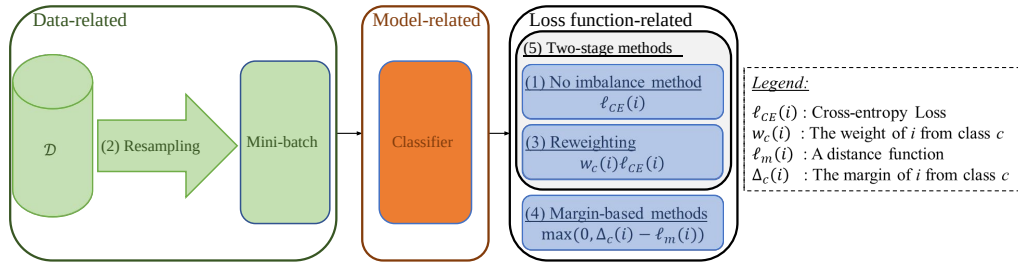


Figure 2.1: Different ways of handling imbalance. (1) The conventional Cross Entropy loss, (2) resampling methods aim to sample a balanced set of examples in the mini-batch, (3) reweighting methods assign a weight for each class for balanced training, (4) margin-based methods assign different margins for different classes to counter the imbalance and (5) two-stage training first supervises the classifier for feature learning without any imbalance method, which is followed by a resampling method to supervise only the classifier by freezing the features.

imbalanced. Thus, the neural networks trained on balanced datasets become vulnerable and over-confident during test time. This vulnerability can be dramatic in safety-critical applications such as medical diagnosis systems where the smaller class is the target and autonomous driving, where real-time decisions have severe consequences as we also mention in Section 1.1. Correspondingly, methodologies are proposed to mitigate class imbalance in datasets and make the models robust to underrepresented classes. In Figure 2.1 we can see a training pipeline for image classification. The first category is resampling which is a preprocessing method applied to the dataset before the training stage. An intervention to the resampling algorithm before giving the data to the model can prevent class imbalance. Other two mitigation categories are loss adaptation ways to mitigate class imbalance: loss reweighting and enforcing margin from representative samples to decision boundaries. Loss reweighting is basically weighting the loss of samples regarding its class or instance-based to change its contribution to the gradient update. Margin-based methods enforces a margin from the decision boundaries separating the class sample clusters during training to have better representations which will ease the classification process. Most of these works in the literature define class imbalance over the varying number of class frequencies in

data [2, 10, 3, 11]. From this point of view, most of the methods for long-tailed learning that lying under the aforementioned categories rely on the number of samples in classes. Additionally, novel loss functions, ensemble methods and disentangling feature learning and the main training pipeline as a complementary task are methods to improve classification performances in an imbalanced setting also some of which get their intuition from class cardinality. Various analyses, surveys and reviews have been conducted on the performance of neural networks on imbalanced datasets in visual tasks [7, 8, 12]. We briefly review some of these methodologies in this section within five categories.

2.2.1 Resampling Methods

This set of methods addresses class imbalance by sampling a balanced mini-batch from a dataset \mathcal{D} (green box in Figure 2.1):

$$\mathcal{L}(f(\mathbf{x}; \theta), \mathbf{y}) = \sum_{b=1}^B \mathcal{L}_b(f(\mathbf{x}_b; \theta), \mathbf{y}_b), \quad (2.1)$$

where samples in batch $b = (\mathbf{x}_b, \mathbf{y}_b)$ are samples with a probability p_i^c , where i is the sample index belonging to class c . More particularly, once the conventional random sampler is employed to sample a mini-batch, each example in the dataset will be sampled equally likely, i.e., the sampling probability of example i in the dataset is $p_i = \frac{1}{|\mathcal{D}|}$. On the other hand, a resampling method imposes a probability distribution over examples to promote tail classes, i.e., the classes with fewer examples in the dataset, during training. Accordingly, as the most naive methods equalize the sampling probabilities of examples of each class, denoted by p^c for class c : (i) Under-sampling achieves this by discarding some examples from overrepresented classes by setting $p_i^c = 0$, the sampling probability of example i from a class c , and (ii) over-sampling duplicates the examples from underrepresented class corresponding to increasing p_i^c for such classes [13]. Alike, Kang et al. [10] give uniform sampling probabilities to classes, referred as *class-balanced* (CB) sampling, which aim to equalize the number of samples from each class within mini-batches. Considering that the naive oversampling is prone to overfitting in favor of underrepresented classes, Peng et al. [11] combine oversampling with random sampling using an exponential weight

factor, yielding a more smooth distribution over p^c s, thereby smoothing the effect of oversampling. Similarly, Kang et al. [10] show that using random sampling in the early training; and as the training proceeds, gradually adjusting the sampling distribution over classes p^c to favor underrepresented classes, which they refer as *progressively balanced* (PB) sampling, is also useful. Overall, all of these methods rely on the number of examples to sample a mini-batch during training.

2.2.2 Reweighting Methods

Another common group of approaches [3, 14, 15, 16, 17, 2, 18] to handle the imbalanced data is to reweight the loss values of the examples such that the underrepresented classes are promoted, i.e., they have larger weights¹. Class-based reweighting assign w^c to losses of all examples belonging to class c :

$$\mathcal{L}_b(f(\mathbf{x}_b; \theta), \mathbf{y}_b) = \sum_{i=1}^{N_b} w_i^c \mathcal{L}_i, \quad (2.2)$$

where N_b is the batch size and w_i^c is the sample’s loss weight in class c . Similar to resampling methods, one naive approach is to set the weight of an example pertaining to class c as $w_i^c = \frac{1}{n_c}$ where n_c is the number of examples from class c instead of using equal weights for all classes, i.e., $w_i^c = 1$. Another approach following this objective, *cost-sensitive cross entropy* (CSCE) [2], sets $w_i^c = (\frac{n_{min}}{n_c})^\nu$ where n_{min} stands for the tail class frequency and ν is an exponential smoothing term. Additional to class cardinalities, there are methods that increase the contribution of samples to the objective function based on their class and sample hardness [19, 3]. A very common approach to balance out the loss accordingly by sample hardness is Focal loss [19], which regulates the negative probability of a sample with an exponential term γ to control its importance. Cui et al. propose a *class-balanced* (CB) loss [3] which also relies on class cardinalities along with sample effectiveness with a regulating hyperparameter β , where $w_i^c = \frac{1-\beta}{1-\beta^{n_c}}$. The common point of methods mainly depends on the class frequencies.

¹ We follow the same notation convention with resampling methods by replacing p by w in p_i, p_i^c and p^c .

2.2.3 Margin-based Methods

Another subcategory of loss adaptations is margin-based loss functions [20, 21, 22, 23, 24, 25, 26]. The main aim of these methods is to enforce a margin to the decision boundaries among classes in feature space or among logits to have separable representations, which leads to better classifier training. Maximizing the distance between the representative examples in a class to the decision boundary prevents a possible bias of the boundary from the overrepresented class to the underrepresented and prepares the ground for the classifier training. We took the works of Cao et al. [22] and Samuel et al. [26] as our baselines. Label-distribution-Aware Margin (LDAM) loss [22] suggests to minimize a margin-based generalization bound with an adaptation of CE. On the other hand, Distributional Robustness (DRO-LT) loss [26] learns margins among feature representations and improve the classification performance through representation learning.

2.2.4 Two-stage Methods

Although the reweighting and resampling methods mentioned above help long-tailed learning, these strategies may adversely affect the performance of head classes. Therefore, reweighting and resampling methods can be applied two-stage to avoid this drawback. Meta-learning and two-stage reweighting approaches [14, 15, 16, 17] are suggested to learn the head classes initially with a vanilla setting and gradually give importance to tail classes in later stages. Ren et al.’s method [14] meta-learns the loss weights from the validation set according to their gradient directions on the classifier network. It assigns the new loss weights at each iteration and updates the classifier weights simultaneously. Jamal et al. propose a method that pre-trains the model with a prior distribution overweights and defines them as two summed terms in contrast to Ren et al. [14]. Han et al.’s approach [15] learns an explicit mapping of the loss function to loss weights, assigning the learned weights to the sample losses in parallel with the classifier training. As a fine-tuning method, Park et al. [17] use the magnitude of the gradient vectors of samples as loss weights after imbalanced training, combined with a balancing term inversely proportional with the class frequencies and a hyper-parameter to down-weight the dominance of head classes. The main aim

of the two-stage approach is to address the lack of learning head classes. Zhang et al. [1] analyze numerous mitigation methods for class imbalance and come up with an optimal combination of these methods, *Bag of Tricks*. Although Bag of Tricks is a three-stage method, we handle it within this category, since it proposes two slightly different follow-up stages after imbalance training. It includes a data augmentation strategy where they create *class activation maps* (CAM) for images and generate data from the “more important” parts of the images according to these maps. Additionally, they combine the augmentation with class-balanced resampling.

2.2.5 Other Methods

We summarize various strategies that are worth mentioning but do not belong to any category mentioned above in this section. The main problem of imbalance is the distribution differences among source and target data. Various methods suggest solutions to class imbalance by modifying the models or objective functions to be robust to the distribution shift. Ren et al. [27] adapt the softmax function for label distribution shift with a scaling term over class frequencies and meta-learns the optimal resampling rate. Hong et al. [28] modify the logits to disentangle the source distribution and integrate the target distribution as the output probability in the inference phase. Following, they propose to disentangle the label distribution from the model in training phase by detaching the source prior and replacing it with uniform prior in model logits, which leads the model to accommodate any target label distributions. Zhang et al. [29] calibrates the classifier scores which aligns the distribution of model output with target distribution. On the other hand, Zhang et al. [10] analyzes the characteristics of feature and classifier learning distinctively with different strategies and suggest retraining the classifier with suitable sampling methods show significant improvements. Additionally, there are approaches utilizing text descriptions along with images for long-tailed image classification [30, 31] and ensemble models aim to reduce the variance and better span varying target distributions [32, 33].

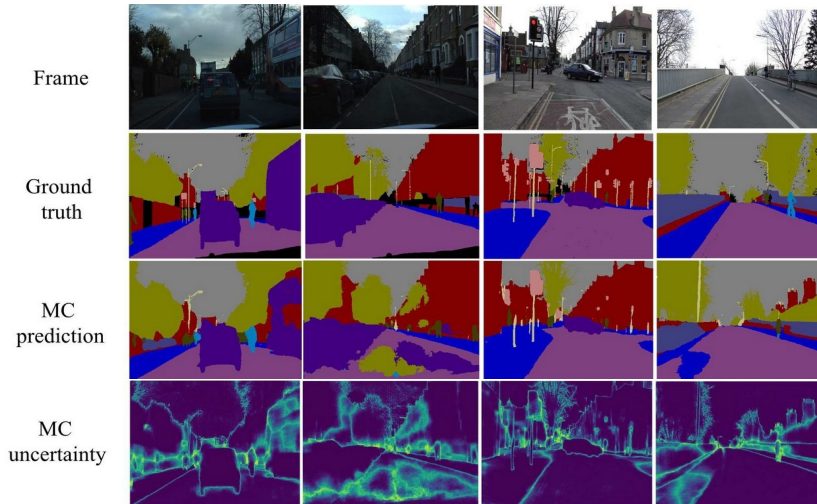


Figure 2.2: Uncertainty maps of frames in CamVid dataset [5]. MC stands for Monte Carlo Dropout [34]. Figure is from [35].

2.3 Uncertainty Estimation in Deep Learning

Uncertainty of a deep neural network can be defined as knowing the doubt in a model’s prediction. As deep learning models are integrated into our daily lives and used in applications that directly affect human life, instead of directly using the model output, measuring the uncertainty of the model on its output has gained importance. Illustrating the uncertainty estimation map of an image segmentation task on an autonomous vehicle’s cameras promote this importance. In Figure 2.2, we can see the ground truth image segmentation in row 2, and the output results in row 3. The estimated class of each pixel is the class which retrieves the highest probability score. However, these predictions do not give any information on the stochasticity of the model’s estimations or the other classes’ probabilities. Though, the estimated uncertainty map in row 4 show where the model is not that confident in terms of its estimations. Therefore, estimated uncertainty also gives an intuition on how much we should rely on the estimate of a model. Although some of the previous studies define uncertainty in various ways intuitively in their field of research, Gal [36] divide a model’s uncertainty into two commonly-held categories: *epistemic* and *aleatoric uncertainty* with two sub-categories, *homoscedastic* and *heteroscedastic*. We briefly

explain the types of uncertainty in the following section.

2.3.1 Types of Uncertainty

Figure 2.3 describes epistemic and heteroscedastic aleatoric uncertainty over a simple regression task. As illustrated, epistemic uncertainty is the uncertainty of the model when the training is done with not enough data. With more data to cover the data space, epistemic uncertainty can be decreased. On the other hand, aleatoric uncertainty represents the inherent noise in the data, which may be due to external errors or interference and the data itself. Homoscedastic aleatoric uncertainty assumes uniform observation noise for every data point. Another type of aleatoric uncertainty is heteroscedastic, representing different observation noises in different data points. Quantifying heteroscedastic aleatoric uncertainty helps when certain parts of the data have higher noise than the others.

2.3.2 Methods

With the rise of popularity of uncertainty estimation in deep learning, many methods specified over architectures and tasks or general frameworks have been proposed which are applicable to any neural network architecture. Gal et al. propose one of the highly-used methods to quantify predictive uncertainty in deep neural networks, *Monte Carlo (MC) Dropout* [34], which approximates a Bayesian model by performing MC sampling in test time while Dropout [38] layers are active. Gal et al. [34] derives predictive uncertainty as the entropy of the predictions retrieved from different MC samples. Gal also proves that predictive uncertainty inherently represent both epistemic and aleatoric uncertainty. Another baseline study on predictive uncertainty is *Deep Ensembles (DE)* [39], which consists of an ensemble of neural networks. The first moment of the predictions is the softmax probabilities of the ensemble network. For classification, we can adopt the entropy of predictions as in [34] as class-wise uncertainty estimates. Having M networks, a deep ensemble results in M outputs for an input, which represents a distribution over class probabilities in a classification task. Lakshminarayanan et al. [39] quantify the predictive uncertainty of a sample in

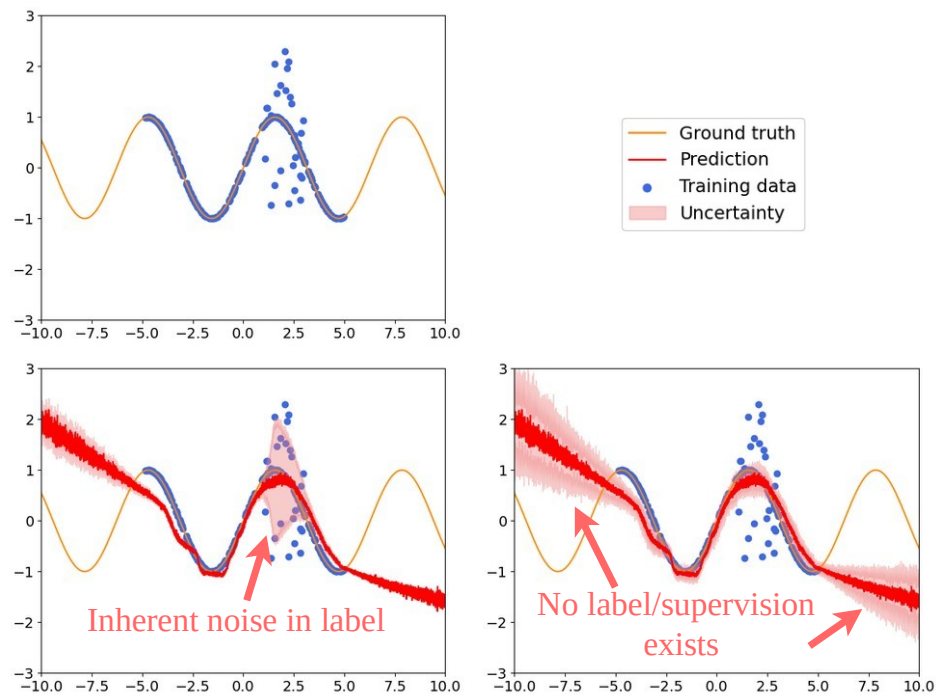


Figure 2.3: Uncertainty types illustrated on a regression task. The noisy data represents aleatoric uncertainty (bottom left). On the other hand, we can observe high epistemic uncertainty where the model does not have any supervision/data (bottom right). Figure is from [37].

validation set as:

$$\mu_i^{pu} = - \sum_{c=1}^C p_i^c \log p_i^c.$$

where i is the index of a sample from the validation set, C is the number of classes and p_i^c corresponds to:

$$p_i^c = \frac{1}{M} \sum_{m=1}^M p_i^{mc}, \quad (2.3)$$

where M is the number of networks in the ensemble and p_i^{mc} is the softmax probability given to class c for image i . A class' predictive uncertainty is the class-wise average of the samples in class c via the average entropy of the class probabilities of validation samples in a class:

$$\mu_c^{pu} = \frac{1}{N_{vc}} \sum_{i=1}^{N_{vc}} \mu_i^{pu}, \quad (2.4)$$

where N_{vc} is the total number of examples in class c in validation set. Based on the work of Gal et al. [36], a sample's epistemic uncertainty is quantified as the mutual information between the prediction and the posterior over the model parameters and approximate it by:

$$\mu_i^{eu} = \mu_i^{pu} - \frac{1}{M} \sum_{m=1}^M \left(\sum_{c=1}^C -p_i^{mc} \log(p_i^{mc}) \right). \quad (2.5)$$

An ensemble model consists of M neural networks with the same architecture and learning schedule, whose parameters are initialized randomly. This randomization is one of the factors that allows each network of the ensemble model to converge to different local minima [40]. Additionally, DE has a simpler training scheme appropriate for distributed computation, which makes it desirable for large-scale tasks. There exist single forward pass methods [41, 42, 43, 44, 45] which perform state-of-the-art compared to their former single forward pass uncertainty estimation techniques and obtain comparable results with DE while being computationally effective. We adopt DE in our experiments due to its simpler training pipeline and applicability to varying neural network architectures.

2.4 Performance Evaluation

2.4.1 Top-1 Error Rate

We evaluate the experiments on uncertainty estimation and imbalance mitigation methods with *top-1 error rate*, which represents the percentage of the number of samples classified incorrectly in a validation set:

$$\epsilon = 1 - \frac{N_t^v}{N^v},$$

where N^v is the number of samples in validation set and N_t^v is the samples correctly classified.

2.5 Datasets

We use artificially long-tailed datasets for our primary task. We inspect our measure on CIFAR-10-LT and CIFAR-100-LT [3] which are artificially made long-tail from original CIFAR datasets [46]. We also show the performance of our measure on an additional dataset, ImageNet-LT [47], with a mitigation method. ImageNet-LT is a long-tailed subset of the ImageNet dataset [48]. The detailed descriptions of the datasets and how they got imbalanced exist in following sections. For all datasets, we validate and test our models on the same set to be consistent with literature. Thus, test and validation sets for these datasets are the same and interchangeably used within this thesis.

2.5.1 CIFAR-10-LT and CIFAR-100-LT

CIFAR-10 and CIFAR-100 datasets, collected by Krizhevsky et al. [46], both contain 50000 training and 10000 test images within 10 and 100 number of classes, respectively. In CIFAR-10 there exist 5000 and 1000 samples per class in training and test data, respectively. These numbers are 500 and 100 for CIFAR-100. In CIFAR datasets all images are 32x32 pixels and colored. CIFAR-10 Long-tailed (LT) and

CIFAR-100-LT [3] are imbalanced versions of the original datasets with geometrically decreasing number of samples from head to tail classes in training set. Both datasets have a balanced test set with 10000 images as in their original version. The number of examples in training set depends on the *imbalance ratio* of the data defined as the number of samples in the class with lowest frequency over the class with highest frequency:

$$\text{IR} = \frac{n_{max}}{n_{min}},$$

and imbalance type which can be step-wise or exponential. step-wise unbalancing [49] is done by selecting a group of minority classes and decreasing their number of samples by multiplying it with the imbalance ratio. A subcategory of step-wise imbalance is linear imbalance, which linearly unbalances the classes from head to tail with a multiplicative factor. Lastly, exponential imbalance [3] which we generally see in naturally imbalanced dataset, is an unbalancing method where the number of samples in class c with index i is calculated as:

$$n_{c_i} = n_{max} * \text{IR}^{\left(\frac{i}{C-1}\right)}.$$

We adopt the exponential imbalanced CIFAR-10-LT and CIFAR-100-LT from [3] in our experiments and analysis. In Figure 2.4 we can observe an long-tailed version of CIFAR-100 with $\text{IR} = 100$.

2.5.2 ImageNet-LT

ImageNet dataset [48] is an open-source large-scale dataset with over 14 million images. ImageNet-LT [47] is a long-tailed version of ImageNet with 1000 categories and 115800 images in total. In ImageNet-LT the highest and lowest class frequencies are 1280 and 5, respectively, which makes $\text{IR} \approx 250$.

2.6 Comparative Summary

In literature, model and data uncertainty is leveraged to improve at many varying tasks such as object detection [50], bias mitigation [51], zero-shot image segmentation [52],

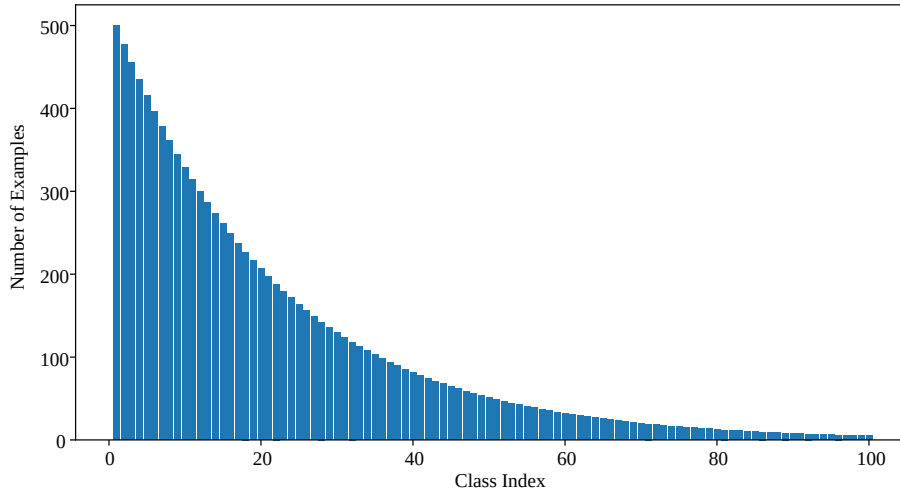


Figure 2.4: Number of images in CIFAR-100-LT with imbalance ratio 100 where the long-tailedness of the data is visible going through the higher class indices.

multi-task learning [53], image classification and face verification [21]. In this work, we rely on predictive uncertainty to introduce a measure of class imbalance and utilize it in long-tailed image recognition task. To our knowledge, the first approach that states a relation between predictive uncertainty and class imbalance is proposed by Khan et al. [21], where predictive uncertainty is derived from the training set with MC Dropout [34] and enforced as margins in feature space. However, this work does not analyze the relationship between class imbalance and predictive uncertainty in detail. In addition, we also interpret predictive uncertainty as a measure of class imbalance relying on our analysis and show that our measure can easily be adapted to any method and architecture in long-tailed visual recognition.

Accordingly, we first examine the relationship between the uncertainty of a model and the class imbalance problem in a dataset. Consequently, we utilize our measure to mitigate the negative effects of class imbalance. We take the study of Zhang et al. [1] as our baseline which sums up existing mitigation methods for imbalanced classification and conduct extensive experiments on varying methods including re-sampling, reweighting and two-stage training to obtain an ideal combination of these to mitigate imbalance. For margin-based methods, we adapt our work to the work of Cao et al. [22] and Samuel et al. [26]. We describe our approach and analysis

in Chapter 3, explain how we incorporate our measure in Chapter 4 and present the compared results in Chapter 5.

CHAPTER 3

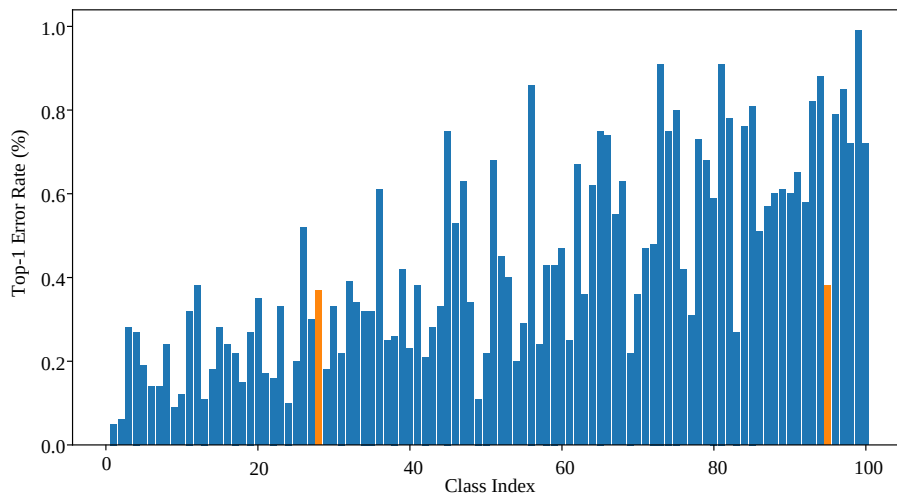
A NOVEL MEASURE OF CLASS IMBALANCE: PREDICTIVE UNCERTAINTY

Our main motivation in this thesis is to quantify class imbalance. In this chapter, we examine existing methods before presenting a quantitative measure of class imbalance. We discuss the shortcomings and limitations of these methods. Then, we specify some requirements for an imbalance measure to meet. Next, we introduce predictive uncertainty as a measure for class imbalance and analyze it over the suggested requirements in detail and compare it with existing methods.

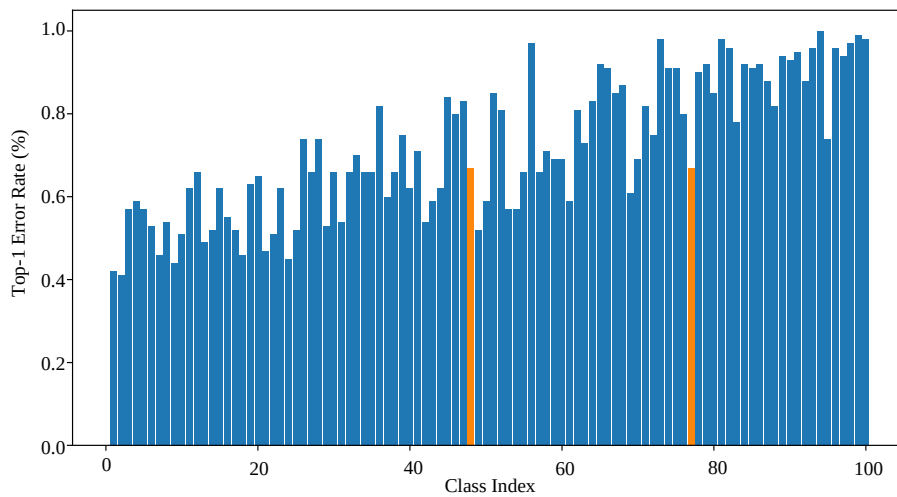
3.1 Limitations of Existing Measures and Imbalance Mitigation Methods

As we discussed in Section 2.2, class imbalance is quantified by the number of samples in each class. However, the number of samples across classes is not necessarily a measure of imbalance, as learning a class does not depend solely on class frequency. Additionally, hardness of the samples in classes can also lead to an imbalance among classes. As an example, see Figure 3.1 where we observe in varying imbalance ratios that, although some of the tail classes have significantly less number of samples than heads, they can be recognized with similar top-1 error rates. This simple analysis shows that class cardinality does not necessarily reflect the class characteristics or imbalance. There exist various methodologies to mitigate class imbalance based on the number of samples in a class, or a sample's effectiveness, which also relies on the number of samples in a class [3, 2, 22, 16, 10]. Another approach is to mitigate class imbalance through the hardness of samples [19] by increasing its contribution to the objective function which does not account for class frequency. We argue that these

two perspectives have potential for improvement with a comprehensive imbalance measure.



(a) Imbalance ratio 1/20



(b) Imbalance ratio 1/50

Figure 3.1: **(a)** Average top-1 error rates of classes in CIFAR-100-LT with imbalance ratio 1/20. **(b)** Average top-1 error rates of classes in CIFAR-100-LT with imbalance ratio 1/50. As in Figure 2.4, the classes in both datasets have exponentially decreasing frequencies. Although the top-1 error rates have an increasing trend from head to tail classes, there are many tail classes which have lower error rates despite having lower class frequencies compared to some head classes (which are both marked for the ease of comparison).

3.2 Requirements for an Imbalance Measure

Along with the number of samples in a class, another characteristic that varies between classes and may lead to imbalance is the representation quality of samples within a class. If the samples in a class do not include representative features of that class, no matter how many samples the class contains, it will be difficult for a model to learn the class. In this context, we argue that for a measure should meet two requirements to represent class imbalance while not depending only on class frequency:

Requirement 1: *The measure should have an increasing trend when the data gets more imbalanced.*

The primary requirement of our measurement is that the imbalance measure increases in parallel with the unbalance of the data, due to its definition. If our measure does not increase in line with an increasing imbalance, we can say that it is not representative.

Requirement 2: *The measure should not decrease with increasing number of same samples from the dataset.*

This requirement argues that our measurement should reflect the extent to which samples in a class can represent that class. If the information we obtain from a dataset is constant, we expect our measurement not to decrease.

3.3 Measuring Imbalance with Predictive Uncertainty

As we discuss in Section 2.3, predictive uncertainty tells us how reliable any predictions made by the model are and it encompasses both aleatoric and epistemic uncertainty. We argue that predictive uncertainty is suitable as a measure for class imbalance. We expect epistemic uncertainty to increase as classes become numerically uneven. Additionally, we suggest that in the case of an increase in the difficulty in class examples, aleatoric uncertainty will reflect this increase numerically. From this point of view, we argue that predictive uncertainty is more inclusive than current class imbalance quantification methods: class frequency and hardness (see Table 3.1). We analyze the predictive uncertainty measures we obtain from Deep Ensembles [39]

Table 3.1: Comparison of class imbalance measures. Our proposed measure considers both the amount of data and semantic complexity.

Criteria	Measures amount of data?	Measures semantic complexity?
Class cardinality ($\mu^\#$) [2, 3, 22, 16, 10]	Yes	No
Class difficulty (μ^L) [19, 3]	No	Yes
Predictive uncertainty (μ^{pu}) [Ours]	Yes	Yes

to see if the measures we get meet the requirements we present in Section 3.2 to represent class imbalance. We provide the details of uncertainty estimation in Appendix A.

3.3.1 Analyzing Class Imbalance with Predictive Uncertainty

We will focus on predictive uncertainty in this study, as it inherently comprises the uncertainty of a model due to the uneven distribution in training data and the uncertainty that comes from the data itself. We have identified the requirements of a class imbalance measure in Section 3.2. For each requirement, we design a setup and analyze predictive uncertainty with respect to these requirements. We also compare predictive uncertainty with class hardness and their behaviours on training and validation sets of CIFAR and CIFAR-LT datasets.

Requirement 1. *The measure should have an increasing trend when the data get more imbalanced.*

In order to see if our measurement meets this requirement, we need to choose a method to unbalance the data class-wise in a controlled manner. Then, in line with this imbalance, we expect our uncertainty measure to increase when the data gets more class-imbalanced.

Analysis Setup. In this thesis, we decide to unbalance the dataset by changing the class cardinality since the controlled increase of sample or class difficulty is a harder task. Following this, we apply a controlled unbalancing procedure to balanced datasets, CIFAR-10 and CIFAR-100 [46].

We utilized exponential unbalancing as in Equation 2.5.1 to obtain CIFAR-10 Long-

tailed (LT), and CIFAR-100 LT [3]. This method enables us to control the imbalance of the datasets and observe the effect of imbalance on predictive uncertainty measures. We analyze the effect of imbalance in class cardinalities on predictive uncertainty in 3 different datasets with 6 imbalance ratios: CIFAR-10-LT, CIFAR-10-LT unbalanced in reverse order of class index and CIFAR-100-LT with imbalance ratios $IR \in \{1, 2, 10, 20, 50\}$.

Observation 1. *Uncertainty values of classes increase when data gets more imbalanced for both CIFAR-10-LT, reversed CIFAR-10-LT and CIFAR-100-LT.*

To put it differently, the more a class loses samples compared to other classes in the dataset, the more its uncertainty increases. Figures 3.2, 3.3 and 3.4 show the relationship between the estimated average predictive uncertainty of a class and the imbalance ratio of CIFAR-10-LT, CIFAR-10-LT unbalanced with reversed order and CIFAR-100-LT datasets obtained with an ensemble network.

Observation 2. *Predictive uncertainties of classes within a dataset give information about other characteristics of classes along with class cardinality.*

We want to address the predictive uncertainty values of imbalance ratio 1, which stands for a balanced dataset. Although all classes have same number of examples, they have different uncertainty values. Moreover, the classes are not ordered in each imbalance ratio in parallel with their class cardinality. Although we interfere with the class frequencies to change the imbalance of the data, the disordered lines in the graphs of each dataset in Figures 3.2 and 3.4, from red to blue show that the uncertainty measure is not just related with the class frequency.

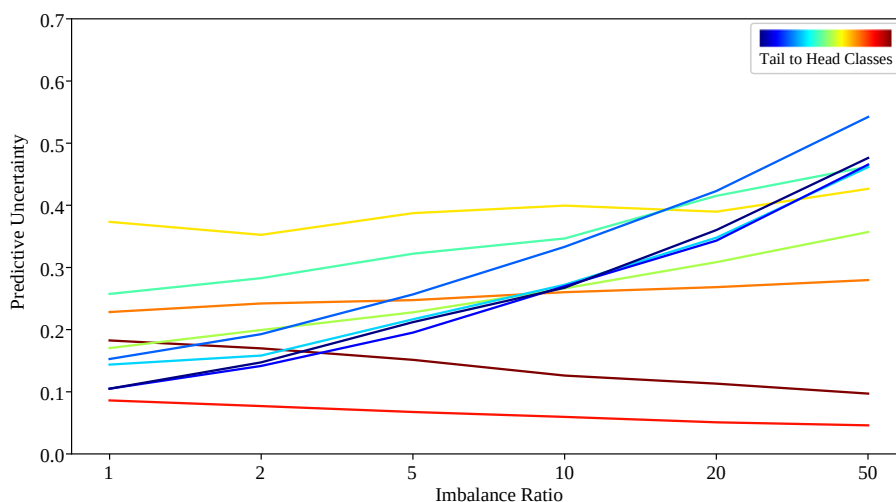


Figure 3.2: Average predictive uncertainty values of each class in CIFAR-10-LT. The color code goes from underrepresented (tail) to overrepresented (head) classes from blue to red, respectively. In relatively less imbalanced datasets we can observe disordered classes compared to their class cardinality. This shows that predictive uncertainty does not only reflect class cardinality but also class hardness.

Requirement 2. *The measure should not decrease with increasing number of same samples from the dataset.*

The imbalance measure should not decrease by reintroducing an information into the model. When we increase the number of samples by presenting the same samples, the class imbalance measure should not be affected. Therefore, predictive uncertainty should not decrease when we provide existing data to the model.

Analysis Setup. We utilized resampling as a preprocessing method to imitate this setup and we expect predictive uncertainty to be stable among increasing resampling since no new information is included in training with reiterated samples.

Peng et al. [11] define *soft-sampling* as the exponentially increasing resampling probability of classes from an imbalanced setting to class-aware sampling. In an imbal-

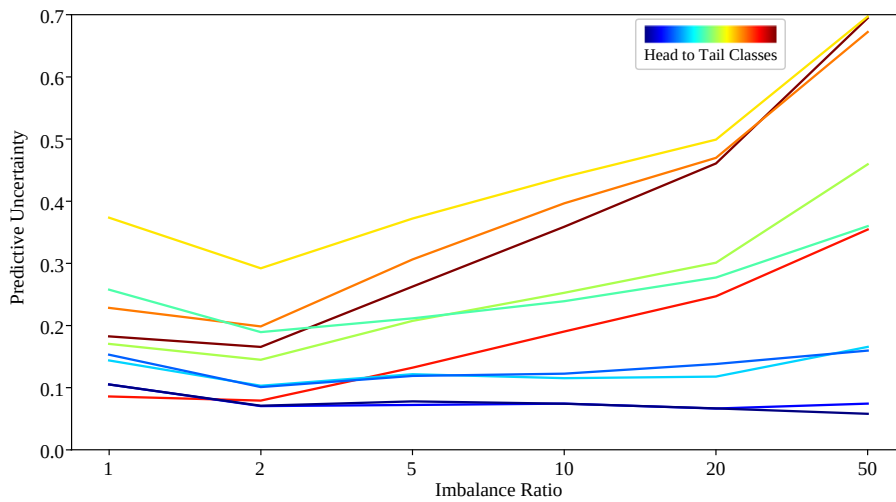


Figure 3.3: Predictive uncertainty values of each class in CIFAR-10-LT unbalanced in reverse order. The color code goes from overrepresented (head) to underrepresented (tail) classes from blue to red, respectively. We observe the same behaviour of increasing predictive uncertainty when we increase the imbalance independent from the class order.

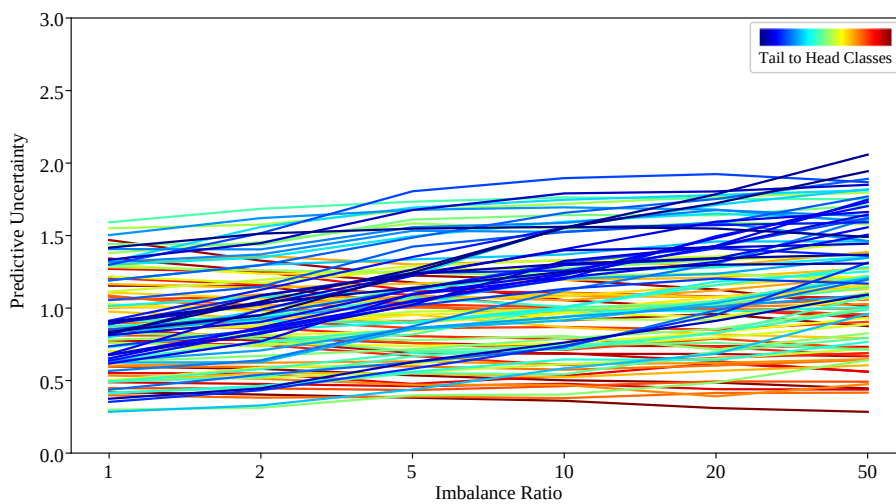


Figure 3.4: Predictive uncertainty values of each class in CIFAR-100-LT. The color code goes from underrepresented (tail) to overrepresented (head) classes from blue to red, respectively. Predictive uncertainty increases when we gradually unbalance the dataset.

anced setting, the sampling probability of a mini-batch sampler for a class c is p_n^c :

$$p_n^c = \frac{n_c}{N},$$

where n_c is the number of samples of class c in training set. Thus, each class' sampling probability is proportional to its cardinality. In a class-aware setting, each sample in the dataset has the same probability to be sampled regardless of its class:

$$p_a^c = \frac{1}{C}.$$

Soft-sampling exponentially weights between these two probabilities and to adjust a sampling probability favoring tail classes to a certain extend:

$$p^c = (p_a^c)^\lambda * (p_n^c)^{1 - \lambda},$$

where λ is the soft-sampling ratio which we will refer from now on. To observe the effect of increasingly resampling the data in favor of the underrepresented classes, we increase the soft-sampling ratio as in set $\{0, 0.1, 0.3, 0.5, 0.7\}$. We conduct this analysis on CIFAR-10-LT and CIFAR-100-LT datasets with imbalance ratios 50 and 100.

Observation. *Predictive uncertainty captures the information in the data space since continually introducing the samples in the dataset does not continuously decrease predictive uncertainty.*

We first examine the increasing accuracy in Figure 3.5 for CIFAR-10 with 50 imbalance ratio. Although there is a performance increase when we utilize soft-sampling, the model is still uncertain about its predictions. Moreover, we see a disordering in classes between average accuracy and predictive uncertainty, which supports that predictive uncertainties provide additional information of class characteristics. Although it may seem that the model performs with higher accuracy for certain classes, it does not always mean that the model is reliable on these classes. In addition, as we observe from Figures 3.6 and 3.7, increasingly oversampling the tail classes does not have a direct effect on the uncertainty measure in both CIFAR-10-LT and CIFAR-100-LT datasets with 50 and 100 imbalance ratios. This observation validates our uncertainty

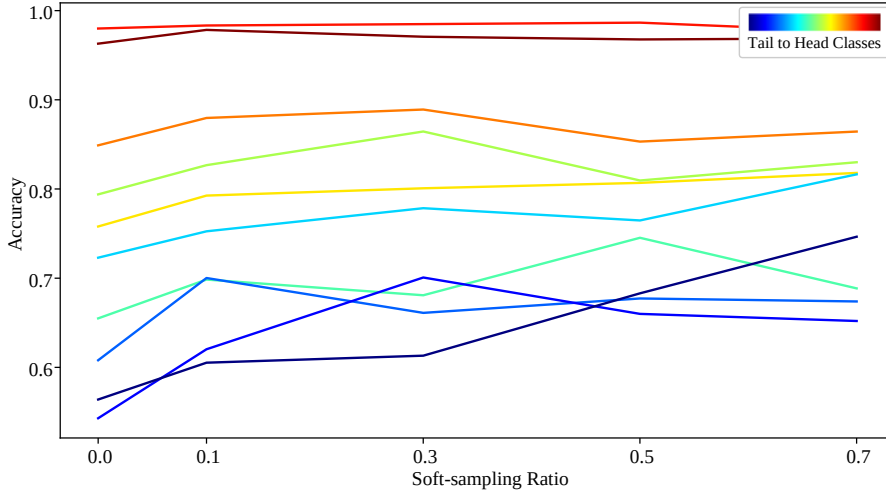
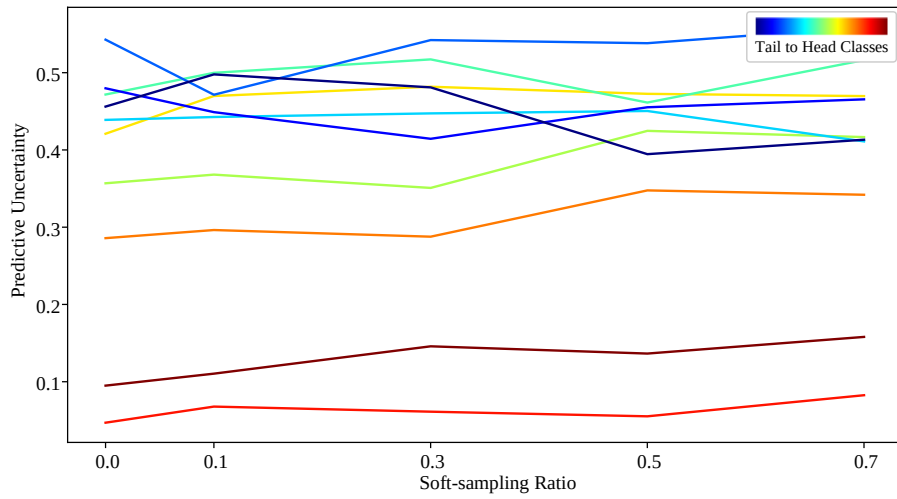


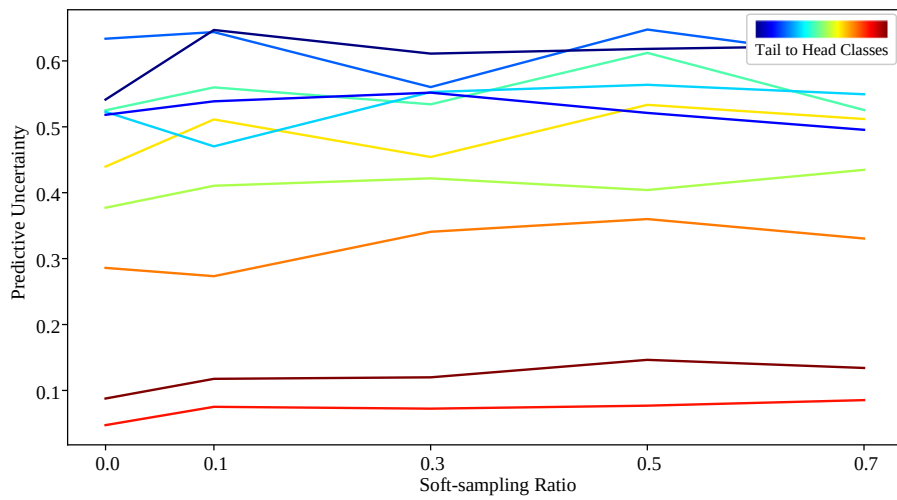
Figure 3.5: Average accuracy of classes vs. soft-sampling ratio of CIFAR-10-LT with 50 imbalance ratio. When we increase the number of samples in tail classes, we observe an increase in accuracy. However, predictive uncertainties of those classes remain non-decreased (see Figure 3.6).

representation meeting the requirement we specify.

Following the analysis of the requirements on predictive uncertainty, we examine the behaviour of class hardness together with predictive uncertainty and inverse class frequency. For CIFAR-LT datasets with 50 and 100 imbalance ratios, we analyze average predictive and epistemic uncertainty of classes along with inverse class frequency and CE loss values. We normalize the class averages to sum up to 1. From the Figures 3.10, 3.11, 3.12 and 3.13, we can observe a similar pattern for each dataset: For converged models, average validation and training losses on a balanced set show a similar trend with predictive uncertainty (see Figures 3.8 and 3.9 from validation and training sets of balanced CIFAR). However, methods based on hardness [19] relies on training losses assuming we do not have an access to test set labels. With an imbalanced setting, class-wise training loss aligns with inverse class cardinality when the model converges, showing a different behaviour from the balanced datasets. This analysis also support our approach where we utilize predictive uncertainties retrieved from the balanced validation set.

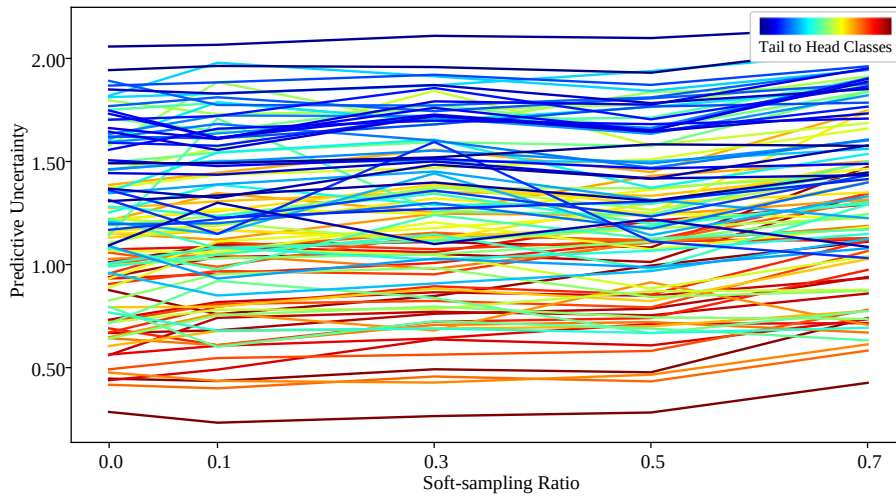


(a) Predictive Uncertainty of CIFAR-10-LT with 50 Imbalance Ratio

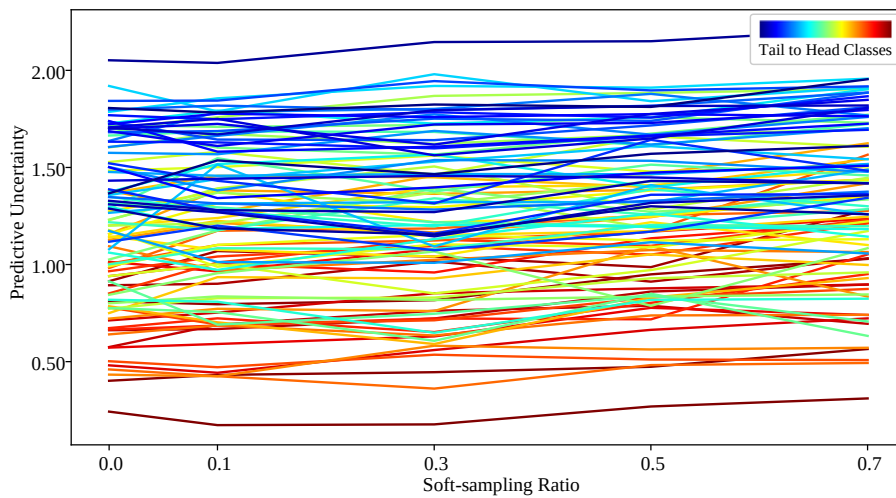


(b) Predictive Uncertainty of CIFAR-10-LT with 100 Imbalance Ratio

Figure 3.6: Average predictive uncertainty of classes vs. soft-sampling ratio of CIFAR-10-LT with 50 and 100 imbalance ratios. Resampling with increasing soft-sampling ratio does not decrease the average predictive uncertainty of classes for both imbalance ratios.

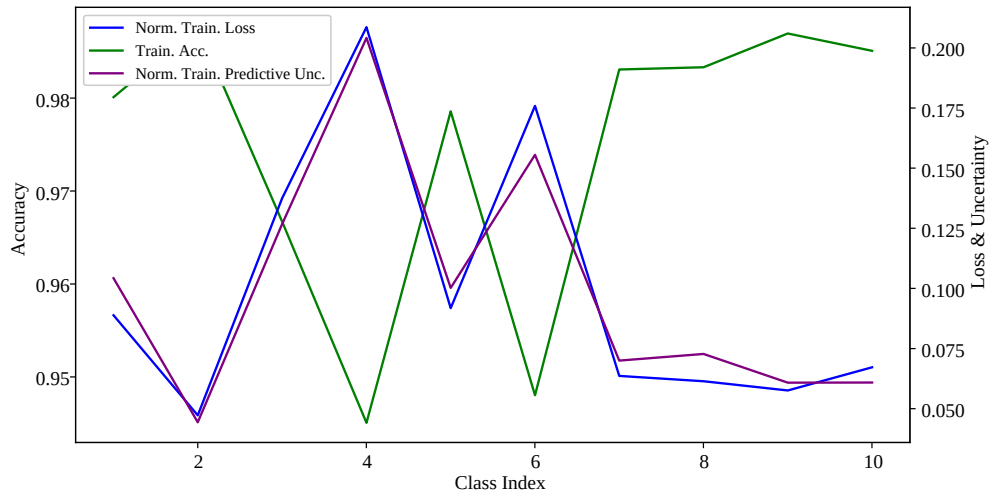


(a) Predictive Uncertainty of CIFAR-100-LT with 50 Imbalance Ratio

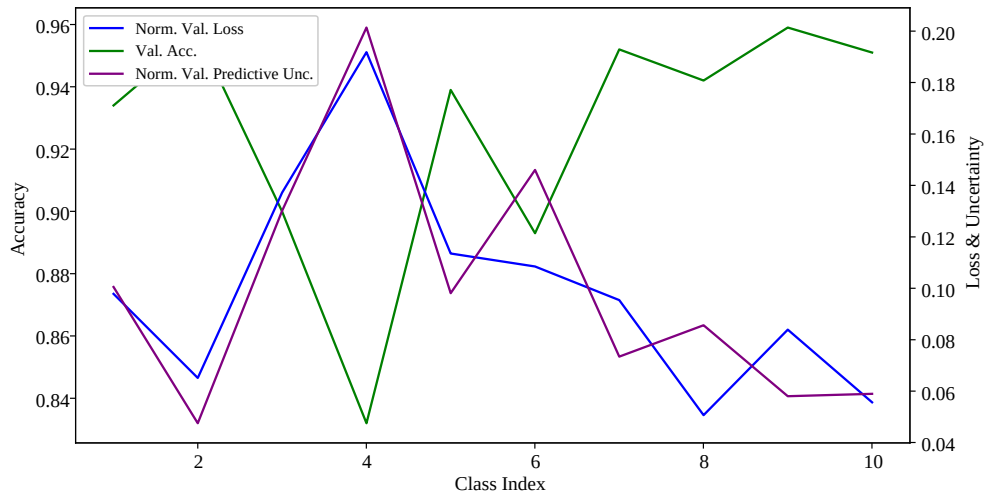


(b) Predictive Uncertainty of CIFAR-100-LT with 100 Imbalance Ratio

Figure 3.7: Average predictive uncertainty of classes vs. soft-sampling ratio of CIFAR-100-LT with 50 and 100 imbalance ratios. Resampling with increasing soft-sampling ratio does not decrease the average predictive uncertainty of classes for both imbalance ratios.

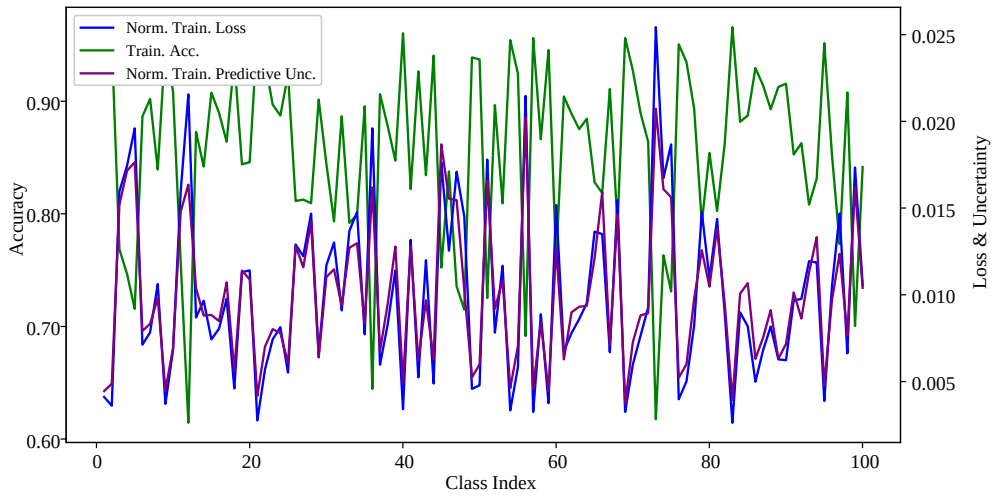


(a)

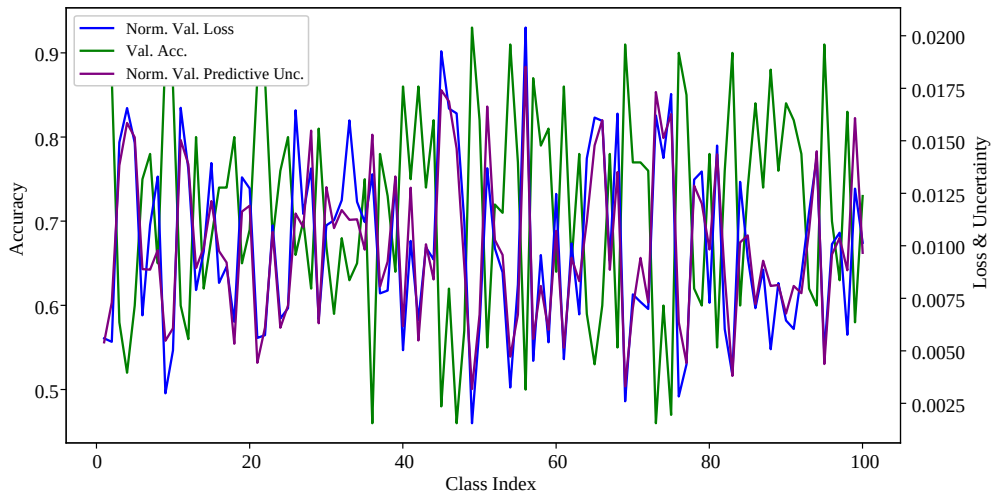


(b)

Figure 3.8: Training (a) and validation (b) average class-wise losses, uncertainty values and accuracies of balanced CIFAR-10 where both sets' classes have equal number of examples within.

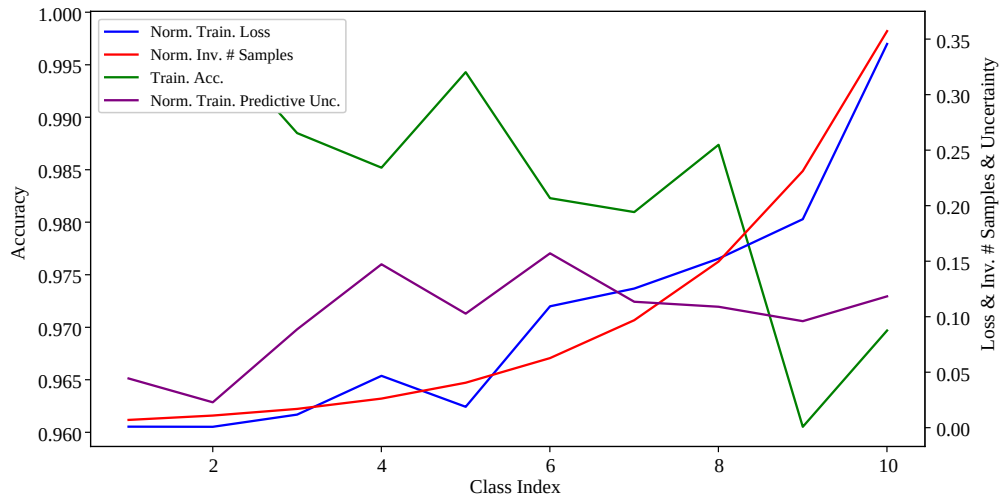


(a)

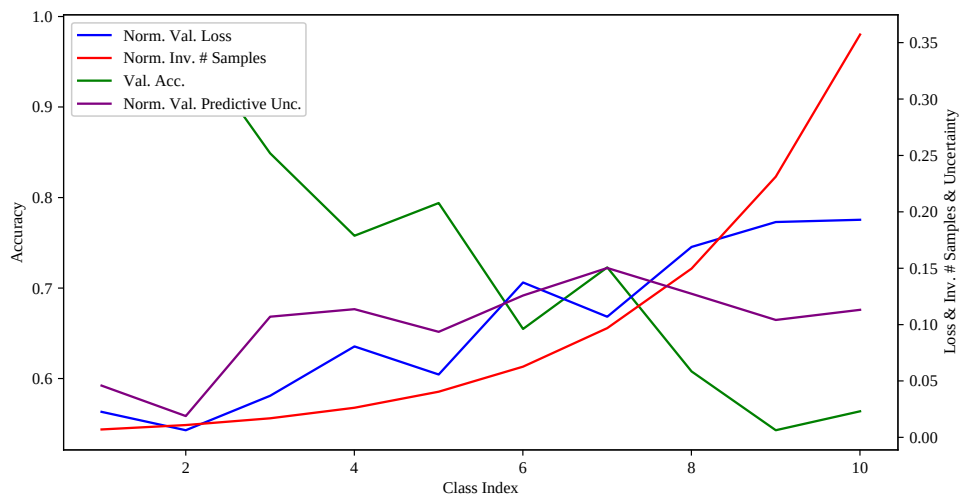


(b)

Figure 3.9: Training (a) and validation (b) average class-wise losses, uncertainty values and accuracies of balanced CIFAR-100 where both sets' classes have equal number of examples within.

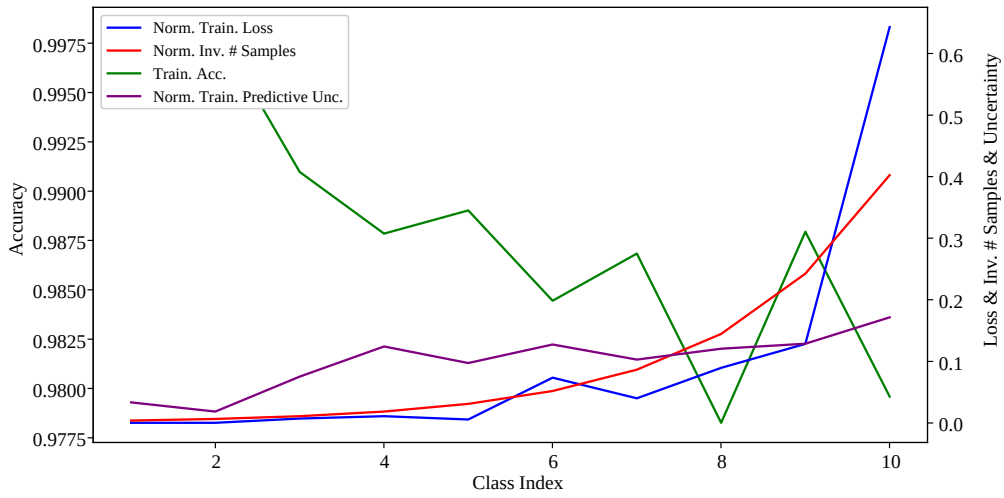


(a)

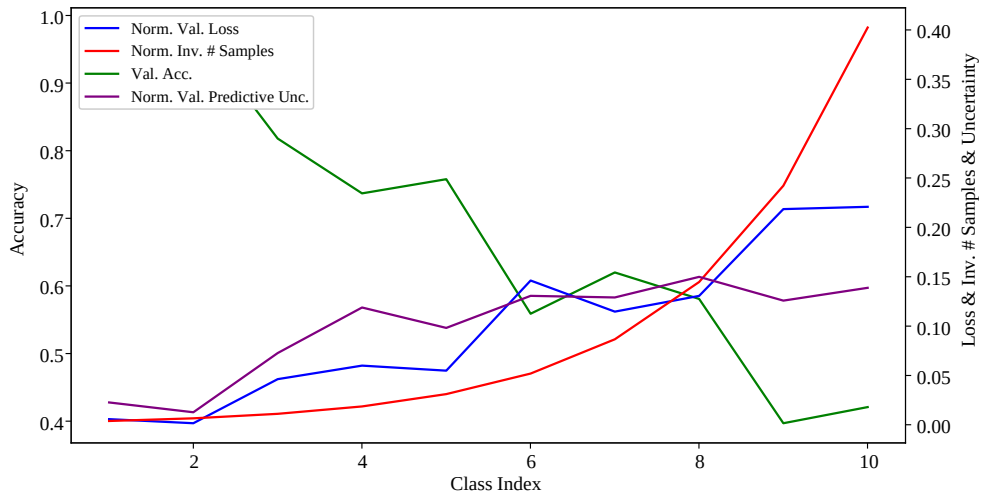


(b)

Figure 3.10: Training (a) and validation (b) average class-wise losses, uncertainty values, inverse number of samples and accuracies of CIFAR-10-LT with 50 imbalance ratio.

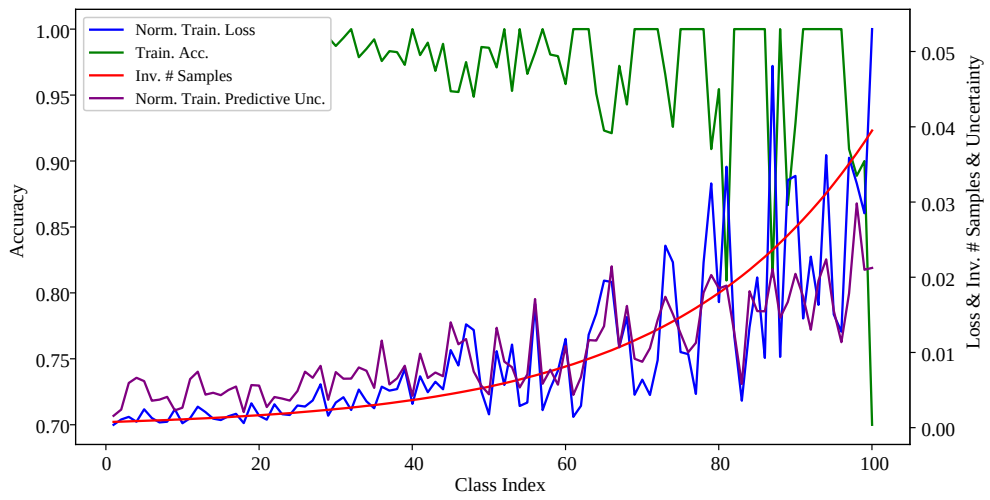


(a)

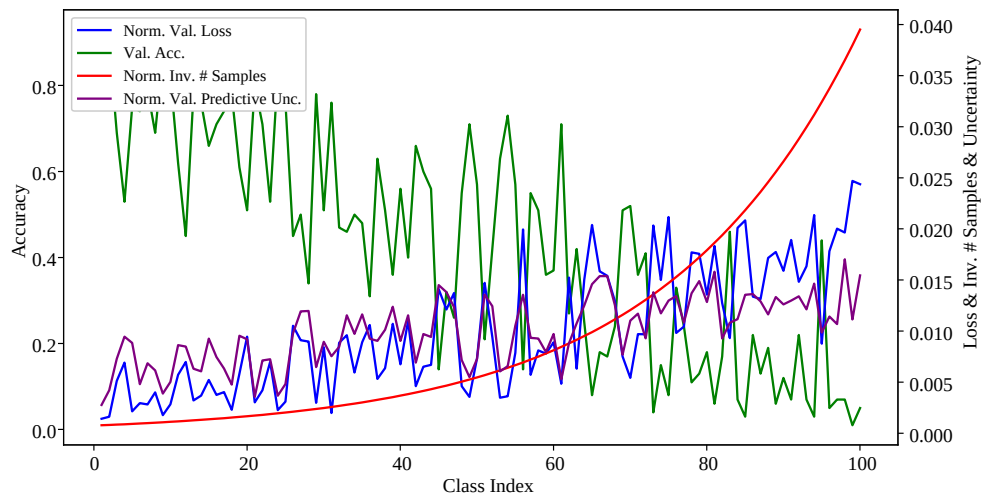


(b)

Figure 3.11: Training (a) and validation (b) average class-wise losses, uncertainty values, inverse number of samples and accuracies of CIFAR-10-LT with 100 imbalance ratio.

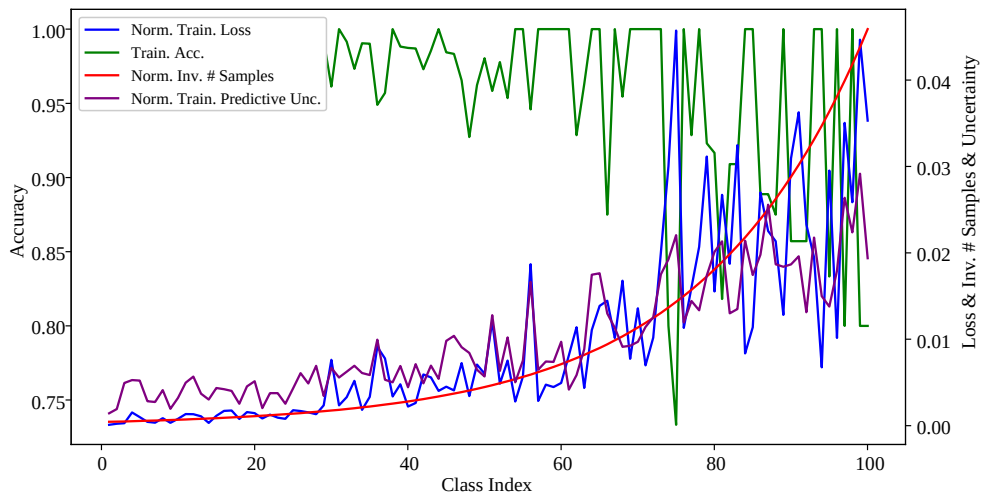


(a)

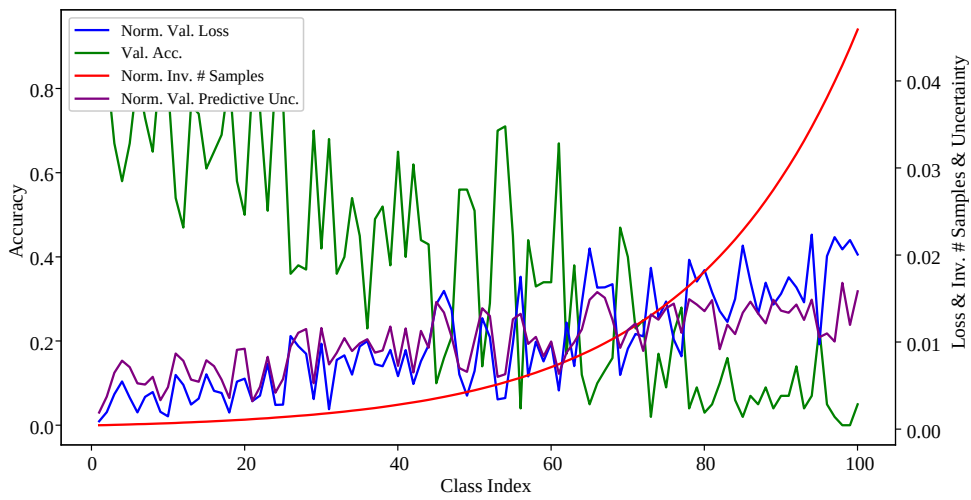


(b)

Figure 3.12: Training (a) and validation (b) average class-wise losses, uncertainty values, inverse number of samples and accuracies of CIFAR-100-LT with 50 imbalance ratio.



(a)



(b)

Figure 3.13: Training (a) and validation (b) average class-wise losses, uncertainty values, inverse number of samples and accuracies of CIFAR-10-LT with 100 imbalance ratio.

CHAPTER 4

MITIGATING CLASS IMBALANCE WITH PREDICTIVE UNCERTAINTY

Many methods have been presented to reduce and resolve class imbalance, a compelling feature of the long-tailed recognition problem (refer to Section 2.2). Although they touch on different aspects of the learning process, these methods are often built on class cardinality. We have shown, through certain requirements and analysis, that imbalance can be expressed in terms of predictive uncertainty. In this section we integrate our measure into several imbalance mitigation methods to investigate its effectiveness.

4.1 Integrating Predictive Uncertainty into Existing Methods

There exist various class imbalance mitigation methods, as we summarized in Section 2.2: resampling, reweighting, margin-based and two-stage methods. These methods can be combined with any network and included in a learning process. In this thesis, we chose the best performing examples of these methods (based on the work of Zhang et al. [1]) to test the usability of our measure in class imbalance reduction methods. In Figure 4.1, we visualize the high-level scheme of our architecture to utilize uncertainty to mitigate class imbalance. In this section, we go over these methods once again, and explain how we incorporate our imbalance measure based on class-wise predictive uncertainty into these methods in detail.

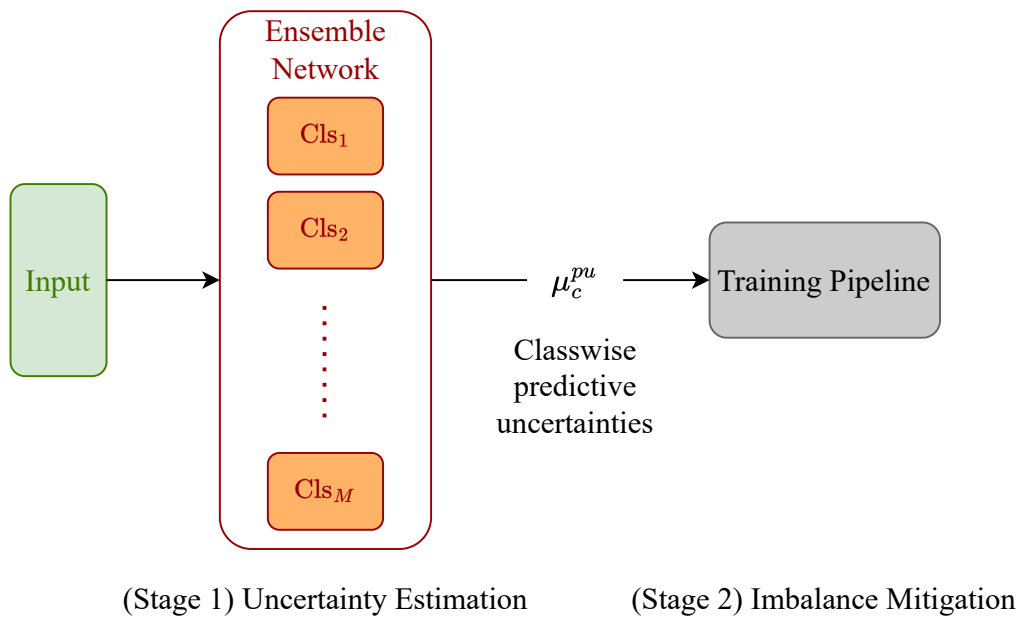


Figure 4.1: High-level scheme of our architecture to mitigate class imbalance. We first quantify class-wise predictive uncertainties of the validation set with an ensemble network and map it to a training pipeline (see a detailed version in Figure 2.1) to utilize them as mitigation methods.

4.1.1 Resampling Methods with Our Measure

To form a mini-batch during training, examples are sampled from each class with a certain probability. Resampling methods discussed in Section 2.2.1 assign varying probabilities p^c to each sample according to its class c where samples belonging to the same class have the equal sampling probabilities. The value assigned to p^c differs from method to method. Though, in the studies we present, it depends on class cardinality. We exploit our class imbalance measure predictive uncertainty, which we obtain from the validation set, as sampling probabilities p^c . Intuitively, if predictive uncertainty of a class is high, the model should favor that class during sampling to equalize that class with others in a mini-batch. We normalize raw uncertainty measures to sum up to 1 by:

$$p^c = \frac{\mu_c^{pu}}{\sum_{i=1}^C \mu_{c_i}^{pu}},$$

and assign it to p^c as class sample probability. In one-stage resampling we directly assign predictive uncertainties as sampling probabilities from the beginning of the training to the end.

4.1.2 Reweighting Methods with Our Measure

Existing reweighting methods change a sample i 's loss weight w_i according to its hardness or class cardinality (see Section 2.2.2). We suggest the uncertainty measures as weights w^c , where all samples in class c have the same weight, to reflect the uncertainty of the samples' contribution to the objective function. We normalize the class uncertainty measures and multiply it with the number of classes C :

$$w^c = \frac{\mu_c^{pu}}{\sum_{i=1}^C \mu_{c_i}^{pu}} * C,$$

to keep the objective scale same with compared methods. In one stage reweighting, we assign loss weights from the start and keep it until training finishes.

4.1.3 Margin-based Methods with Our Measure

Margin-based methods assign preset or learnable margins to samples in feature space to have separable representations for classifier learning [22, 26]. We take LDAM [22] and DRO-LT [26] methods as our baseline. Cao et al. formulate LDAM as:

$$\mathcal{L}_{\text{LDAM}} = -\log \frac{e^{z_y - \Delta_y}}{e^{z_y - \Delta_y} + \sum_{j \neq y} e^{z_j}},$$

where the margin is

$$\Delta_c = \frac{K}{n_c^{\frac{1}{4}}},$$

with a constant K . They optimize K for a fixed maximum margin. They provide a margin-based generalization bound based on class cardinalities which enforces a margin on the logits. We assign $\Delta_{j=c} = \mu_c^{pu}$ to utilize our measure, scaled accordingly to have the same fixed maximum margin 0.5:

$$\Delta_c = \mu_c^{pu} (0.5 / \max(\mu^{pu})),$$

and observe its empirical effects.

In DRO-LT loss [26], there is a three-stage training schedule with an imbalance learning first, following with a feature extractor learning and classifier tuning. We integrate our measure to feature learning part of DRO-LT loss and replaced it with the learnable margins of classes in the proposed robustness loss. The proposed robustness loss for a sample z is:

$$\mathcal{L}(z)_{\text{Robust}} = -\log \frac{e^{-d_z - \Delta_c}}{\sum_{z'} e^{-d_{z'} - \Delta_c}},$$

where d is the distance of sample z to its class centroid and Δ_c is the defined class-wise margin. In DRO-LT, Samuel et al. tested various scenarios of Δ_c being a constant, depending on class cardinality or a learnable parameter. We replace it with predictive uncertainty of each class in our experiments. We employ the normalized class uncertainties as margins between class distributions and scale it with the number of samples to keep the scale similar with the original paper:

$$\Delta_c = \frac{\mu_c^{pu}}{\sum_{i=1}^C \mu_{c_i}^{pu}} * C.$$

4.1.4 Two-stage Methods with Our Measure

Multi-stage imbalance training methods apply mitigation strategies with multiple stages throughout the training. The main purpose of training in more than one stage in the long-tailed recognition problem is to first learn the representatively strong classes and then tune the feature extraction or classification modules of the model on the representatively weak classes, thus not overfitting to any class. The first stage of two-stage models is imbalanced training, where we train the network with a standard loss function (in our baseline, CE loss) and learn especially head classes. After that, by integrating the resampling or reweighting methods we mentioned above, we enable the model to learn by favoring the tail classes. We took the two-stage approaches in resampling and reweighting stated by Zhang et al. [1] and compare them with the mitigation methods relying on our uncertainty measure. Additionally, we incorporate our measure into the Bag of Tricks [1] and replace the CAM-based class-balanced resampling with predictive uncertainty.

CHAPTER 5

EXPERIMENTS

In this section, we evaluate the use of our imbalance measure, predictive uncertainty, with different imbalance mitigation solutions.

5.1 Datasets

To have the exact data preprocessing, we apply the same procedure with our baselines to the data. We use the CIFAR-10-LT, CIFAR-100-LT, and ImageNet-LT datasets that we explained in Section 2.5 as follows:

- **CIFAR-LT**: For data augmentation, we adopt random horizontal flipping for CIFAR-LT experiments. We test the requirements over CIFAR-10-LT and CIFAR-100-LT datasets. We test our measure on CIFAR-LT datasets with 50 and 100 imbalance ratios.
- **ImageNet-LT**: For ImageNet-LT, we randomly crop the images to a size 224x224 and randomly horizontally flip. To align with the baseline in [1] we also apply varying color characteristic changes and for test set, we resize crop the images referencing the center as stated in [1].

Although we tested all the methods we integrated to reduce the imbalance with predictive uncertainty on CIFAR datasets, we were able to demonstrate ImageNet-LT on a single method we chose due to resource constraints.

5.2 Training Details

We utilize the official code base of Bag of Tricks from Zhang et al. [1] and report the reproduced results for each mitigation technique in one-stage and two-stage training. As the baseline, for CIFAR-10-LT and CIFAR-100-LT we train a residual network with 32 layers and for ImageNet-LT, 10 layers. We use CE loss as the objective function and SGD as the optimizer. We train the network on the CIFAR-LT datasets with a learning rate 0.1, momentum with coefficient 0.9 and weight decay with coefficient 0.0002, for 200 epochs with a batch size 128 on 1 GPU. For learning rate scheduling, we warm-up the learning rate in the first 5 epochs and multi-step scheduler with a multiplier of 0.01 at epochs 160 and 180. For ImageNet-LT baseline, the learning rate is set to 0.2 with momentum with coefficient 0.9 and weight decay with coefficient 0.0001. We train the network for 100 epochs with a batch size 512 on 4 GPUs with multi-step scheduling with a factor 0.1 at 60th and 80th epochs. The results we present are the average of 3 runs with different seeds for each experiment we conduct on the CIFAR-LT datasets except CAM-based experiments. We present the training details for each model categorized according to the imbalance mitigation method. We report our results on test set of CIFAR-LT and validation set of ImageNet-LT datasets following the code base [1]. Experiments for hyperparameter tuning are also conducted on the corresponding sets to be consistent with our reference.

Resampling methods. For one-stage resampling, we use CIFAR-10-LT and CIFAR-100-LT with 50 and 100 imbalance ratios as the datasets to demonstrate our results. For training, we adopt the same hyperparameters and schedule as the baseline we provide above and the other resampling methods. For the learning rate, we do a grid-search within the range [0.1, 0.7] with an additive increase by 0.1 for uncertainty sampling.

Reweighting methods. For one-stage reweighting, we also show our results on the CIFAR-LT datasets. We keep the training schedule same as the baseline except for the learning rate, where we do a grid-search within a range for our uncertainty-based resampling methods, as in one-stage uncertainty-based resampling. The additional hyperparameters β and γ for CB Focal loss [3] are {0.9999, 2}, {0.9999, 1}, {0.99, 1} and {0.9, 1} for CIFAR-10-LT and CIFAR-100-LT with imbalance ratios 50 and

100, respectively. For uncertainty-weighted focal loss, we adopt the same γ values.

Margin-based methods. We adapt our method to two margin-based methods: DRO-LT loss [26] and LDAM [22]. For DRO-LT Loss [26], we utilize the official implementation of the study for the experiments. The first stage of DRO-LT loss is the imbalance training with the hyperparameters, optimization and learning rate scheduling as our baseline. After training the network for 200 epochs, we train the feature extractor with DRO-LT loss replacing the margin of class centroids to decision boundaries with our uncertainty measure. DRO-LT loss adopts SGD optimizer with 0.005 learning rate and 0.0001 weight decay for 80-epoch feature extractor training. For classifier tuning, Samuel et al. freeze the feature extractor and tune the classifier with an additional 20 epoch with 0.01 learning rate and CB sampling. We do a grid-search for the learning rate of feature extractor training in the range [0.0025, 0.0075]. For LDAM [22], the hyperparameter settings provided in [1] for a scale multiplied with the margin-enforced logits and the maximum margin are 30 and 0.5 for each CIFAR-LT datasets. We did a hyperparameter search for the scale with the values {20, 30, 40}. We keep the exact configurations provided by Zhang et al. [1] for other hyperparameters.

Two-stage methods. For the first stage of the two-stage methods in Tables 5.4 and 5.5, we do imbalance training until the 160th epoch with the training details provided for our baseline. For deferred reweighting, CB Focal loss also adopts the same hyperparameters. In [1], the hyperparameter ν for CSCE is specified as 1, which corresponds to scaling the loss with the class cardinality. For deferred resampling, all resampling methods are applied after an imbalanced training at 160th epoch. For the second stages, we adopt the same training details and hyperparameter tuning schedule as in one-stage training. As a two-stage resampling method, we adapt our uncertainty measure to CAM-based resampling.

For the methods in the “Bag of Tricks” [1], the baseline for the CIFAR-LT datasets is adopted as the first stage training with 320 epochs and no learning rate scheduling. After the CAM-based data augmentation, the second stage is a 80-epoch training stage with the augmented data and mixup [54], 0.001 learning rate and CB resampling. In the second stage, a multi-step learning rate scheduling is applied at 40th

epoch. Lastly, 20-epoch third stage with multi-step learning rate scheduling at 8th and 16th epoch is applied with CB resampling with 0.001 learning rate. We adopt the same settings with “the Bag of Tricks” and replace the CB sampling with uncertainty-based sampling along with CAM-based data augmentation. We do a grid-search for learning rate in the range [0.001, 0.007] for second and third stage. For ImageNet-LT, the first stage of the baseline is a 160-epoch training with a multi-step scheduling at 120th and 160th epoch, with a factor of 0.1. Then, the model is trained additional 40 epochs with a learning rate 0.002, decreased by a multiplying factor 0.1 at 40th epoch. The last training runs 20 epochs with a learning rate of 0.002 and same multi-step scheduling is applied at 8th and 16th epochs. The inclusion of mix-up training and CAM-based resampling is the same as CIFAR-LT datasets. We do a grid-search for learning rate with a longer training of 80 epochs for second stage with a multi-step scheduling at 40th epoch and third stage with a training of 20 epochs.

5.3 Experiments on Our Measure with Resampling Methods

We adopt predictive uncertainty as sampling probabilities in varying resampling methods. We compare our measure with class-balanced and progressive resampling where the latter gradually adapts the probability from imbalance training to the former. First, we experiment directly sampling with uncertainty-based probabilities in one-stage training, which favors tail classes compared to head classes in mini-batches. In Table 5.1, we show our reproduced results of the compared methods and uncertainty-based sampling in row 3. Apart from CIFAR-10-LT with imbalance ratio 50, progressively balanced sampling improves the baseline consistently when applied in one-stage. Progressively increasing the number of samples from tail classes ensures the network learn the head classes first in initial epochs and promote tail classes in later iterations of training. Following that, we apply uncertainty-based sampling gradually during the training as in progressively balanced sampling, which surpasses the best performing method on CIFAR-10-LT with imbalance ratio 50 and CIFAR-100-LT.

Table 5.1: Top-1 error rates with standard deviations of one-stage resampling methods along with our measure on CIFAR-LT datasets. Our uncertainty-based resampling method improves other methods when it is applied both initially and progressively on different datasets.

Dataset Imbalance Ratio	CIFAR-10-LT		CIFAR-100-LT	
	50	100	50	100
Baseline (Vanilla ResNet)	24.26±0.35	30.18±0.52	57.14±0.08	61.48±0.40
PB sampling [10]	23.86±0.36	29.62±0.49	56.87±0.13	60.40±0.57
CB sampling [10]	23.33±0.14	30.33±0.30	61.13±0.52	66.34±0.30
Uncertainty sampling [Ours]	23.40±0.41	30.17±0.75	65.46±0.90	69.00±1.08
Progressive uncertainty sampling [Ours]	24.16±0.05	29.12±0.29	56.46±0.27	60.24±0.15

Table 5.2: Top-1 error rates with standard deviations of one-stage reweighting methods along with our measure on CIFAR-LT datasets. Results with (*) are retrieved from [1]. Uncertainty Focal loss improves CB Focal loss [3] on CIFAR-100-LT datasets.

Dataset Imbalance Ratio	CIFAR-10-LT		CIFAR-100-LT	
	50	100	50	100
Baseline (Vanilla ResNet)	24.26±0.35	30.18±0.52	57.14±0.08	61.48±0.40
Uncertainty CE [Ours]	22.78±0.66	28.27±0.68	54.86±0.49	59.17±0.53
Focal [19]	24.75*	29.62*	57.56*	61.90*
CB Focal [3]	21.29±0.39	26.64±0.59	57.60±0.48	61.41±0.35
Uncertainty Focal [Ours]	22.21±0.15	26.92±0.27	54.24±0.54	58.72±0.43

5.4 Experiments on Our Measure with Reweighting Methods

We apply normalized predictive uncertainty measures as fixed weights to class-wise losses along with CE and Focal losses. We present the results in Table 5.2 which shows class-balanced weighted Focal loss improving the baseline on the CIFAR-10-LT datasets. Using normalized uncertainty measures as class weights to Focal loss surpasses the baseline and the best performing method class-balanced Focal loss [1] within one-stage loss reweighting strategies with 2.9% less top-1 error rate on the CIFAR-100-LT with imbalance ratio 50 and 2.69% with imbalance ratio 100 respectively.

Table 5.3: Top-1 error rates with standard deviations of margin-based methods along with our measure. Results with (*) are retrieved from the original paper of DRO-LT loss [26]. Our measure improves or performs on par with these methods on different datasets.

Dataset	CIFAR-10-LT		CIFAR-100-LT	
	Imbalance Ratio		Imbalance Ratio	
	50	100	50	100
Baseline (Vanilla ResNet)	24.26±0.35	30.18±0.52	57.14±0.08	61.48±0.40
LDAM + DRW [22]	20.32±0.02	23.90±0.28	54.27±0.43	57.98±0.20
LDAM Unc. + DRW [Ours]	20.20±0.05	22.99±0.43	53.98±0.62	57.94±0.57
DRO-LT (shared ϵ) [26]	-	-	44.68*	54.34*
DRO-LT ($\epsilon\sqrt{n}$) [26]	-	-	42.80*	53.08*
DRO-LT (learned ϵ) [26]	14.35±0.19	17.22±0.12	46.40±0.10	51.75±0.15
DRO-LT ($\epsilon = \text{normalized } \mu_c^{pu}$) [Ours]	14.63±0.09	17.11±0.10	46.48±0.08	51.70±0.11

5.5 Experiments on Our Measure with Margin-based Methods

In margin-based methods we employ LDAM-DRW and DRO-LT as our baselines. As shown in Table 5.3, when we enforce our predictive uncertainty measures to LDAM as class margins to decision boundaries, we consistently improve the baseline on CIFAR-10-LT and CIFAR-100-LT with two imbalance ratios, 50 and 100. For DRO-LT loss, we adapt our measure as the margin from a class centroid to a decision boundary in feature space. We improve DRO-LT loss alternatives with the class cardinality based and shared ϵ values. Comparing with DRO-LT loss with learned ϵ , we perform slightly better on CIFAR-LT datasets with imbalance ratio 100 (see Table 5.3).

5.6 Experiments on Our Measure with Two-stage Methods

5.6.1 Two-stage Resampling

Applying uncertainty-based sampling in second stage improves the best performed CAM-based class-balanced sampling [1] on CIFAR-10-LT with imbalance ratios 50 and 100, 0.13% and 0.07% fewer top-1 error rate (see Table 5.4).

Table 5.4: Top-1 error rates with standard deviations of the best performing two-stage resampling method [1] along with our measure on CIFAR-LT datasets. Results with (*) are retrieved from [1]. CAM-based uncertainty sampling outperforms CAM-based CB sampling on CIFAR-10-LT datasets.

Dataset	CIFAR-10-LT		CIFAR-100-LT	
	50	100	50	100
Baseline (Vanilla ResNet)	24.26±0.35	30.18±0.52	57.14±0.08	61.48±0.40
CB sampling [10]	21.34*	29.04*	55.67*	59.56*
PB sampling [10]	24.58*	33.48*	56.93*	61.35*
CAM-based CB sampling [1]	18.66	22.62	53.56	57.70
CAM-based uncertainty sampling [Ours]	18.53	22.55	54.88	57.98
CAM-based progressive uncertainty sampling [Ours]	20.04	24.59	54.86	58.8

5.6.2 Two-stage Reweighting

In two-stage training, we adapt our measure to two-stage resampling and reweighting strategies. Uncertainty measure-weighted CE outperforms the best performing two-stage loss reweighting method on CIFAR-10-LT with 100 imbalance ratio, CB Focal loss, by 1.46%. Our method also outperforms cost-sensitive CE on CIFAR-100-LT with 50 and 100 imbalance ratios by 2.48% and 2.99% fewer top-1 error rate (see Table 5.5).

5.6.3 Bag of Tricks

We employ CAM-based uncertainty sampling with the optimal mitigation tricks’ combination in proposed in [1]. Our method remarkably improves Bag of Tricks baseline we reproduce on CIFAR-10-LT with 50 and 100 imbalance ratios with 1.2% and 2.26% top-1 error rate (see Table 5.6). Since Bag of Tricks is a combination of various mitigation methods, it is worth to analyze the integration of uncertainty further for CIFAR-100-LT and ImageNet-LT datasets.

Table 5.5: Top-1 error rates with standard deviations of the best performing two-stage reweighting methods [2, 3] along with our measure on CIFAR-LT datasets. Results with (*) are retrieved from [1]. Uncertainty-based reweighting of CE loss in second stage outperforms CB Focal [3] on CIFAR-100-LT and CIFAR-10-LT with 100 imbalance ratio.

Dataset	CIFAR-10-LT		CIFAR-100-LT	
	Imbalance Ratio		Imbalance Ratio	
	50	100	50	100
Baseline (Vanilla ResNet)	24.26±0.35	30.18±0.52	57.14±0.08	61.48±0.40
Focal [19]	23.77*	29.71*	57.32*	61.74*
CB Focal [3]	20.45±0.27	26.08±0.34	57.09±0.55	59.90±0.19
Cost-sensitive CE [2]	21.03±0.46	26.11±0.22	54.41±0.25	58.78±0.13
Uncertainty CE [Ours]	20.47±0.44	24.62±0.31	51.93±0.27	55.79±0.34

Table 5.6: Top-1 error rates with standard deviations of Bag of Tricks [1] along with our measure. Results with (*) are retrieved from [1]. CAM-based uncertainty sampling improves Bag of Tricks [1] on CIFAR-10-LT datasets.

Dataset	CIFAR-10-LT		CIFAR-100-LT		ImageNet-LT
	Imbalance Ratio		Imbalance Ratio		
	50	100	50	100	
Model	ResNet-32				ResNet-10
Baseline (Vanilla ResNet)	24.26±0.35	30.18±0.52	57.14±0.08	61.48±0.40	65.99*
Bag of Tricks [1]	17.04±0.31	22.00±0.07	47.73±0.39	51.12±1.52	56.87
Bag of Tricks with CAM-based uncertainty sampling [Ours]	15.84±0.09	19.74±0.40	48.49±0.41	52.02±0.42	57.54

5.7 Ablation Analysis on Uncertainty Estimation

Before we suggest predictive uncertainty as a class imbalance measure to mitigate the imbalance, we conduct ablation studies on epistemic uncertainty as a class imbalance measure. We examined the performance of predictive and epistemic uncertainty in improving class imbalance and the impact of calibrating our uncertainty estimates on the results. We empirically observe the effect of calibrated uncertainties on mitigating class imbalance when we utilize them as we propose in Chapter 4.

5.7.1 Epistemic Uncertainty to Mitigate Class Imbalance

Gal et al. [36] defines epistemic uncertainty of a neural network with the mutual information of the probability distribution retrieved from different Dropout activation as we state in Section 2.3. This does not include aleatoric uncertainty by definition since it does not reflect the variance of probabilities of classes within a distribution. Epistemic uncertainty reflects where we do not have enough data points in data space. Thus, it has an intuitive correlation with cardinality. Starting from this point of view, we compare predictive and epistemic uncertainty as a measure of class imbalance empirically. Table 5.7 shows the performance of uncertainty-based one-stage reweighting and resampling methods on CIFAR-10-LT. With slight differences, predictive uncertainty improves the baseline compared to epistemic uncertainty as a measure.

5.7.2 Calibrated Uncertainty to Mitigate Class Imbalance

Although modern neural networks perform well in more complex tasks, they may become no longer well-calibrated [55]. As we utilize ResNet architectures in our ensemble network, we also analyze the effects of calibrated ResNets on uncertainty estimation and mitigating class imbalance with uncertainty. We apply *temperature scaling*, where we soften the softmax by dividing the logits with a temperature value T [55]. Temperature scaling increases each networks output entropy and makes it more diffident on its predictions. We motivate this ablation study from the fact that,

Table 5.7: Empirical effects of calibration on predictive and epistemic uncertainty as measures on CIFAR-10-LT. Predictive uncertainty with performs best among these variations of estimated uncertainties.

Uncertainty Type	Predictive Unc.		Calibrated Predictive Unc.		Epistemic Unc.		Calibrated Epistemic Unc.	
	50	100	50	100	50	100	50	100
Uncertainty CE	23.67±0.14	28.30±0.76	23.36±0.12	29.03±0.93	23.65±0.54	29.15±0.74	23.71±0.10	29.31±0.43
Uncertainty Focal	22.92±0.28	28.22±0.32	23.68±0.31	29.95±0.49	22.82±0.48	28.40±0.57	23.67±0.47	29.19±0.11
Uncertainty Sampling	23.55±0.72	30.90±0.65	23.95±0.69	30.45±1.08	24.18±0.84	31.74±0.25	23.92±0.12	30.52±0.35

in the ensemble network, when all models are over-confident on a sample, this may lead to lower uncertainty, which can be misleading. We estimate the predictive and epistemic uncertainty from an ensemble with calibrated residual networks and conduct experiments with both calibrated entropy and mutual information as uncertainty measures and observe their effect on mitigating the imbalance. As we observe from Table 5.7, calibrated uncertainty does not perform well on mitigating the imbalance empirically¹.

5.7.3 Combining Class-Balanced Loss with Predictive Uncertainty

After the results we obtain from one-stage reweighting with our uncertainty-based measure, we do an ablation study on combining these two loss reweighting methods to analyze their effects further. We present our initial analysis in Table 5.8. Our initial results show that the linear combination of class-balanced and uncertainty-based weighting improves the baselines on both datasets. Although these results seem promising, we leave it as a future work since we believe that beyond a simple linear combination, a joint contribution of these methods should be analyzed in detail with different settings.

5.8 Discussion

We apply our measure for class imbalance, predictive uncertainty, to various class imbalance mitigation methods. We demonstrate the performance of our measure mainly

¹ In these experiments, we do not adopt advised weight decay and bias initialization settings for sigmoid-based losses in [3].

Table 5.8: Top-1 error rates with standard deviations of one-stage reweighting with a linear combination of uncertainty and class-balanced weighted Focal loss. The linear combination of these two methods improves their sole application with a slight hyperparameter tuning.

Dataset	CIFAR-10-LT		CIFAR-100-LT	
	50	100	50	100
Baseline (Vanilla ResNet)	24.26±0.35	30.18±0.52	57.14±0.08	61.48±0.40
Focal [19]	24.75*	29.62*	57.56*	61.90*
CB Focal [3]	21.29±0.39	26.64±0.59	57.60±0.48	61.41±0.35
0.7(CB Focal)+0.3(Unc. Focal)	21.64±0.17	26.07±0.44	53.67±0.29	58.68±0.34
0.5(CB Focal)+0.5(Unc. Focal)	21.61±0.31	25.61±0.14	53.95±0.51	58.27±0.42
0.3(CB Focal)+0.7(Unc. Focal)	21.75±0.37	26.10±0.55	53.75±0.11	59.27±0.19
Uncertainty Focal	22.21±0.15	26.92±0.27	54.24±0.54	58.72±0.43

on CIFAR-10-LT and CIFAR-100-LT datasets to be comparable with the literature. We only tune the learning rate and method-specific hyperparameters for the experiments. Yet, we observe performance improvements in each method for different datasets. These initial experiments we conduct on our uncertainty measure as a mitigation method shows that predictive uncertainty is a suitable measure to alleviate class imbalance in long-tailed recognition problems. Adapting our measure to different mitigation methods and gaining improvements on varying datasets shows the representative power of predictive uncertainty for class imbalance.

CHAPTER 6

CONCLUSION

In this thesis, we have studied the class imbalance problem. We have made two major contributions: (1) Introducing an uncertainty-based measure for quantifying class imbalance. (2) Using the proposed class imbalance measure in imbalance mitigation approaches.

In our first contribution, we have shown that predictive uncertainty can be a good source of information for the lack of data for a learning problem. We have introduced two requirements that can be expected from such a measure and demonstrated that predictive uncertainty satisfies these requirements very well.

In our second contribution, we have integrated our measure into existing imbalance mitigation solutions that originally rely on class cardinality or hardness. We selected many methods with different types of approaches, namely resampling, reweighting, margin-based and two-stage methods. On different datasets, we have shown that using our measure in those methods provides significant improvements over the baseline.

6.1 Limitations and Future Work

The main purpose of this thesis is to provide a novel measure for class imbalance and examine its contributions on long-tailed visual recognition. Counting this study as an initial step, there are certain limitations which we briefly list in this section. Following that, we suggest possible directions to follow to overcome these limitations.

First, obtaining the uncertainty measure and using it as a mitigation method forms

a (minimum) two-stage training pipeline. This increases the computation time if we have to train a model with our measure from scratch. Although we conduct our experiments with DE, using an ensemble for predictive uncertainty estimation has high computational complexity and is not resource friendly. To have an efficient training schedule along with uncertainty estimation, our main aim is to decrease the computational and time complexity of predictive uncertainty estimation and make the training process of the estimation network parallel with mitigating class imbalance. Nonetheless, many researchers utilize uncertainty estimation in different application areas and improve its comprehensiveness and reliability in the data space. Thus, the representative power of predictive uncertainty heavily relies on its quantification and the most recent works on uncertainty estimation determine the upper limit.

When we compare our measure with class hardness, we handle hardness over the contribution of samples to the objective function. Predictive uncertainty can be compared with different quantification methods for sample or class hardness. Along with this, joint contribution of hardness, cardinality and uncertainty-based methods to mitigate imbalance is a promising direction to follow.

We utilize class-wise predictive uncertainties of validation sets to mitigate class imbalance. Since we claim predictive uncertainty being comprehensive as a class imbalance measure, an ablation study of using the class-wise error rates or losses over the validation set by integrating them to existing methods for imbalance mitigation would be a valuable ablation study for comparison.

Although we experimentally observe the improvements of our measure as a mitigation method, particularly the uncertainty-based adaptations of margin enforcing methods can be examined theoretically in terms of its bounds and optimization.

If “ideal uncertainty” could be measured, it could be used in place of our measure to obtain the bounds of our uncertainty-based approach. In the literature, there is a study on the uncertainty of Bayes-optimal predictors where we theoretically have an infinite amount of data [56]. Although this work shows an insight for a definition of ideal uncertainty, there is much to explore on the notion of the “ideal” uncertainty. Starting from this point of view, obtaining the ideal uncertainty in data, what its correspondence and effect on the class imbalance would be an important research question

to answer.

Manipulating data is a common preprocessing method in deep learning. Although we show the effectiveness of predictive uncertainty in standard pipelines, the effects of data manipulation on predictive uncertainty worth an analysis.

Lastly, since we propose a measure for class imbalance, we have covered a subject that is open to evaluation on more datasets with varying characteristics to show its inclusiveness. Our next step in terms of the experiments is to broaden them on more realistic long-tailed datasets with additional mitigation methods. In larger and realistic datasets, the relationship with uncertainty estimation and imbalance, and its performance in mitigation methods can also be examined.

REFERENCES

- [1] 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.
- [2] N. Japkowicz and S. Stephen, “The class imbalance problem: A systematic study,” *Intell. Data Anal.*, vol. 6, no. 5, p. 429–449, 2002.
- [3] Y. Cui, M. Jia, T.-Y. Lin, Y. Song, and S. Belongie, “Class-balanced loss based on effective number of samples,” in *CVPR*, 2019.
- [4] B. H. Menze, A. Jakab, S. Bauer, J. Kalpathy-Cramer, K. Farahani, J. Kirby, Y. Burren, N. Porz, J. Slotboom, R. Wiest, L. Lanczi, E. Gerstner, M.-A. Weber, T. Arbel, B. B. Avants, N. Ayache, P. Buendia, D. L. Collins, N. Cordier, J. J. Corso, A. Criminisi, T. Das, H. Delingette, C. Demiralp, C. R. Durst, M. Dojat, S. Doyle, J. Festa, F. Forbes, E. Geremia, B. Glocker, P. Golland, X. Guo, A. Hamamci, K. M. Iftekharuddin, R. Jena, N. M. John, E. Konukoglu, D. Lashkari, J. A. Mariz, R. Meier, S. Pereira, D. Precup, S. J. Price, T. R. Raviv, S. M. S. Reza, M. Ryan, D. Sarikaya, L. Schwartz, H.-C. Shin, J. Shotton, C. A. Silva, N. Sousa, N. K. Subbanna, G. Szekely, T. J. Taylor, O. M. Thomas, N. J. Tustison, G. Unal, F. Vasseur, M. Wintermark, D. H. Ye, L. Zhao, B. Zhao, D. Zikic, M. Prastawa, M. Reyes, and K. Van Leemput, “The multimodal brain tumor image segmentation benchmark (brats),” *IEEE Transactions on Medical Imaging*, vol. 34, no. 10, pp. 1993–2024, 2015.
- [5] G. J. Brostow, J. Fauqueur, and R. Cipolla, “Semantic object classes in video: A high-definition ground truth database,” *Pattern Recognition Letters*, 2008.
- [6] A. Olteanu, C. Castillo, F. Diaz, and E. Kıcıman, “Social data: Biases, methodological pitfalls, and ethical boundaries,” *Frontiers in Big Data*, vol. 2, 2019.
- [7] Y. Zhang, B. Kang, B. Hooi, S. Yan, and J. Feng, “Deep long-tailed learning: A survey,” *ArXiv*, vol. abs/2110.04596, 2021.

- [8] S. Das, S. S. Mullick, and I. Zelinka, “On supervised class-imbalanced learning: An updated perspective and some key challenges,” *IEEE Transactions on Artificial Intelligence*, 2022.
- [9] S. Ruder, “An overview of gradient descent optimization algorithms,” *arXiv preprint arXiv:1609.04747*, 2016.
- [10] B. Kang, S. Xie, M. Rohrbach, Z. Yan, A. Gordo, J. Feng, and Y. Kalantidis, “Decoupling representation and classifier for long-tailed recognition,” in *Eighth International Conference on Learning Representations (ICLR)*, 2020.
- [11] J. Peng, X. Bu, M. Sun, Z. Zhang, T. Tan, and J. Yan, “Large-scale object detection in the wild from imbalanced multi-labels,” *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9706–9715, 2020.
- [12] K. Oksuz, B. Cam, S. Kalkan, and E. Akbas, “Imbalance problems in object detection: A review,” *IEEE Transactions on Pattern Analysis & Machine Intelligence*, vol. 43, no. 10, pp. 3388–3415, 2021.
- [13] L. Shen, Z. Lin, and Q. Huang, “Relay backpropagation for effective learning of deep convolutional neural networks,” in *ECCV*, 2016.
- [14] M. Ren, W. Zeng, B. Yang, and R. Urtasun, “Learning to reweight examples for robust deep learning,” in *Proceedings of the 35th International Conference on Machine Learning (J. Dy and A. Krause, eds.)*, vol. 80 of *Proceedings of Machine Learning Research*, pp. 4334–4343, PMLR, 2018.
- [15] J. Shu, Q. Xie, L. Yi, Q. Zhao, S. Zhou, Z. Xu, and D. Meng, “Meta-weight-net: Learning an explicit mapping for sample weighting,” in *NeurIPS*, 2019.
- [16] M. A. Jamal, M. Brown, M. Yang, L. Wang, and B. Gong, “Rethinking class-balanced methods for long-tailed visual recognition from a domain adaptation perspective,” *CoRR*, 2020.
- [17] 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 (ICCV)*, pp. 735–744, 2021.

- [18] K. R. M. Fernando and C. P. Tsokos, “Dynamically weighted balanced loss: Class imbalanced learning and confidence calibration of deep neural networks,” *IEEE Trans. Neural Netw. Learn. Syst.*, pp. 1–12.
- [19] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, “Focal loss for dense object detection,” in *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 2999–3007, 2017.
- [20] C. Huang, Y. Li, C. C. Loy, and X. Tang, “Learning deep representation for imbalanced classification,” in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5375–5384, 2016.
- [21] 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.
- [22] K. Cao, C. Wei, A. Gaidon, N. Arechiga, and T. Ma, “Learning imbalanced datasets with label-distribution-aware margin loss,” in *Advances in Neural Information Processing Systems*, 2019.
- [23] M. Hayat, S. H. Khan, S. W. Zamir, J. Shen, and L. Shao, “Gaussian affinity for max-margin class imbalanced learning,” *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 6468–6478, 2019.
- [24] A. K. Menon, S. Jayasumana, A. S. Rawat, H. Jain, A. Veit, and S. Kumar, “Long-tail learning via logit adjustment,” in *International Conference on Learning Representations*, 2021.
- [25] G. R. Kini, O. Paraskevas, S. Oymak, and C. Thrampoulidis, “Label-imbalanced and group-sensitive classification under overparameterization,” *Advances in Neural Information Processing Systems*, vol. 34, pp. 18970–18983, 2021.
- [26] D. Samuel and G. Chechik, “Distributional robustness loss for long-tail learning,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.
- [27] J. Ren, C. Yu, s. sheng, X. Ma, H. Zhao, S. Yi, and h. Li, “Balanced meta-softmax for long-tailed visual recognition,” in *Advances in Neural Information*

Processing Systems (H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, eds.), vol. 33, pp. 4175–4186, Curran Associates, Inc., 2020.

- [28] Y. Hong, S. Han, K. Choi, S. Seo, B. Kim, and B. Chang, “Disentangling label distribution for long-tailed visual recognition,” *arXiv preprint arXiv:2012.00321*, 2020.
- [29] S. Zhang, Z. Li, S. Yan, X. He, and J. Sun, “Distribution alignment: A unified framework for long-tail visual recognition.,” in *CVPR*, 2021.
- [30] C. Tian, W. Wang, X. Zhu, X. Wang, J. Dai, and Y. Qiao, “VI-ltr: Learning class-wise visual-linguistic representation for long-tailed visual recognition,” *arXiv preprint arXiv:2111.13579*, 2021.
- [31] T. Ma, S. Geng, M. Wang, J. Shao, J. Lu, H. Li, P. Gao, and Y. Qiao, “A simple long-tailed recognition baseline via vision-language model,” *ArXiv*, vol. abs/2111.14745, 2021.
- [32] X. Wang, L. Lian, Z. Miao, Z. Liu, and S. Yu, “Long-tailed recognition by routing diverse distribution-aware experts,” in *International Conference on Learning Representations*, 2021.
- [33] Y. Zhang, B. Hooi, L. Hong, and J. Feng, “Test-agnostic long-tailed recognition by test-time aggregating diverse experts with self-supervision,” *arXiv preprint arXiv:2107.09249*, 2021.
- [34] Y. Gal and Z. Ghahramani, “Dropout as a bayesian approximation: Representing model uncertainty in deep learning,” in *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48*, ICML’16, p. 1050–1059, JMLR.org, 2016.
- [35] P.-Y. Huang, W.-T. Hsu, C.-Y. Chiu, T.-F. Wu, and M. Sun, “Efficient uncertainty estimation for semantic segmentation in videos,” in *Computer Vision – ECCV 2018* (V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, eds.), (Cham), pp. 536–552, Springer International Publishing, 2018.
- [36] Y. Gal, “Uncertainty in deep learning,” 2016.

- [37] H. Xu, Z. Zhou, Y. Wang, W. Kang, B. Sun, H. Li, and Y. Qiao, “Digging into uncertainty in self-supervised multi-view stereo,” in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 6058–6067, IEEE Computer Society, 2021.
- [38] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, “Dropout: A simple way to prevent neural networks from overfitting,” *Journal of Machine Learning Research*, vol. 15, no. 56, pp. 1929–1958, 2014.
- [39] B. Lakshminarayanan, A. Pritzel, and C. Blundell, “Simple and scalable predictive uncertainty estimation using deep ensembles,” in *Advances in Neural Information Processing Systems (I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, eds.)*, vol. 30, Curran Associates, Inc., 2017.
- [40] S. Fort, H. Hu, and B. Lakshminarayanan, “Deep ensembles: A loss landscape perspective,” *ArXiv*, vol. abs/1912.02757, 2019.
- [41] J. Antorán, J. U. Allingham, and J. M. Hernández-Lobato, “Depth uncertainty in neural networks,” 2020.
- [42] J. van Amersfoort, L. Smith, Y. W. Teh, and Y. Gal, “Uncertainty estimation using a single deep deterministic neural network,” 2020.
- [43] J. van Amersfoort, L. Smith, A. Jesson, O. Key, and Y. Gal, “Improving deterministic uncertainty estimation in deep learning for classification and regression,” *CoRR*, vol. abs/2102.11409, 2021.
- [44] J. Mukhoti, A. Kirsch, J. R. van Amersfoort, P. H. S. Torr, and Y. Gal, “Deterministic neural networks with appropriate inductive biases capture epistemic and aleatoric uncertainty,” *ArXiv*, vol. abs/2102.11582, 2021.
- [45] B. Lakshminarayanan, D. Tran, J. Liu, S. Padhy, T. Bedrax-Weiss, and Z. Lin, “Simple and principled uncertainty estimation with deterministic deep learning via distance awareness,” in *Advances in Neural Information Processing Systems 33*, 2020.
- [46] A. Krizhevsky, “Learning multiple layers of features from tiny images,” tech. rep., 2009.

- [47] Z. Liu, Z. Miao, X. Zhan, J. Wang, B. Gong, and S. X. Yu, “Large-scale long-tailed recognition in an open world,” in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.
- [48] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, “Imagenet: A large-scale hierarchical image database,” in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009.
- [49] M. Buda, A. Maki, and M. A. Mazurowski, “A systematic study of the class imbalance problem in convolutional neural networks,” *Neural networks : the official journal of the International Neural Network Society*, vol. 106, pp. 249–259, 2018.
- [50] Y. Li, X. Chen, L. Quan, and N. Zhang, “Loss rescaling by uncertainty inference for single-stage object detection,” in *2020 IEEE International Conference on Image Processing (ICIP)*, pp. 698–702, 2020.
- [51] S. R. S, N. Ravikumar, A. J. Bulpitt, and D. C. Hogg, “Epistemic uncertainty-weighted loss for visual bias mitigation,” in *Workshop on Fair, Data Efficient and Trusted Computer Vision*, 2022.
- [52] P. Hu, S. Sclaroff, and K. Saenko, “Uncertainty-aware learning for zero-shot semantic segmentation,” in *Advances in Neural Information Processing Systems* (H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, eds.), vol. 33, pp. 21713–21724, Curran Associates, Inc., 2020.
- [53] A. Kendall, Y. Gal, and R. Cipolla, “Multi-task learning using uncertainty to weigh losses for scene geometry and semantics,” *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7482–7491, 2018.
- [54] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, “mixup: Beyond empirical risk minimization,” *International Conference on Learning Representations*, 2018.
- [55] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, “On calibration of modern neural networks,” in *Proceedings of the 34th International Conference on Machine Learning* (D. Precup and Y. W. Teh, eds.), vol. 70 of *Proceedings of Machine Learning Research*, pp. 1321–1330, PMLR, 2017.

- [56] M. Jain, S. Lahlou, H. Nekoei, V. Butoi, P. Bertin, J. Rector-Brooks, M. Korablyov, and Y. Bengio, “Deup: Direct epistemic uncertainty prediction,” *ArXiv*, vol. abs/2102.08501, 2021.
- [57] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016.

APPENDIX A

UNCERTAINTY ESTIMATION

A.1 Experiments on Estimating Uncertainty

As our measure depends on it, as the first step of the thesis, we need to make an uncertainty estimation on the datasets. In this section, the datasets and models we used for uncertainty estimation, training details and our results are included.

A.1.1 Datasets

On the model we choose to measure the imbalance, we use CIFAR-10-LT and CIFAR-100-LT with 50 and 100 imbalance ratios and ImageNet-LT, on which the baseline methods we will compare are applied. For data augmentation, we adopt the same preprocessing in Section 5.1.

A.1.2 Model Selection

We showed in Section 3.3 that, with an increasing sampling ratio, Deep Ensembles [39] estimates stable predictive uncertainties. Following this ablation study, we leverage Deep Ensembles [39] to estimate predictive uncertainty due to its reliable performance and ease of use. We keep each network’s complexity the same with our baselines and use ResNet variations (10 and 32 depending on our baseline) as the networks in an ensemble for each dataset. We experiment on an ensemble of $M \in \{3, 5, 7\}$ networks and set $M = 5$ throughout the experiments relying on the performances for all datasets. After obtaining the pipeline for our measure which ensures the requirements

we specify in Section 3.2, we use the uncertainty estimation architecture we set for ImageNet-LT.

A.1.3 Training Details

For the experiments we conduct, the ensembles consist of $M = 5$ networks adopted from our baseline [1], e.g. for CIFAR-LT we have ResNet architectures [57] with 32 layers and for ImageNet-LT, ResNet-10. We train the ensemble with the same setting of our baselines with vanilla ResNets for each dataset. For CIFAR-10-LT and CIFAR-100-LT, we employ SGD optimizer with learning rate 0.1, momentum 0.9 and weight decay 0.0002. We train 200 epochs with batch size 128 on 1 GPU. For learning rate scheduling, we use initial warm-up until the 5th epoch and multi-step scheduling at 160th and 180th epochs with a multiplying factor 0.01. For ImageNet-LT, we optimize using SGD with learning rate 0.2, momentum 0.9 and weight decay 0.0001. We employ the same warm-up strategy with CIFAR-LT experiments and use multi-step learning rate scheduling training with 100 epochs, decreasing the learning rate with factor 0.1 at 60th and 80th epochs. We trained the ensemble with batch size 512 on 4 GPUs. We obtain the predictive uncertainties for each class by averaging each samples' uncertainties within the class after the ensemble converges.