Addressing Imbalanced Data in Machine Learning: Methods and Challenges

A Major Qualifying Project Report
Submitted to the faculty of the
WORCESTER POLYTECHNIC INSTITUTE
In partial fulfillment of the requirements for the
Degree of Bachelor of Science

Submitted By:

Joshua Jahnz
Robert Philpot

Date: April, 2024

Approved By:

Professor Ziming Zhang

Acknowledgements

We would like to express our sincere gratitude to our advisor, Professor Ziming Zhang, who contributed greatly to the successful completion of this paper. His expertise and insightful guidance were invaluable throughout the research process. We would also like to extend our heartfelt appreciation to our family and friends, who have provided us with the support and encouragement that we needed throughout the completion of this project in order to pursue our goals.

Abstract

Imbalanced datasets are prevalent in machine learning, posing significant challenges due to the underrepresentation of certain classes. This often leads to biased models with poor predictive performance on minority classes. The project we completed dives into various strategies to mitigate such biases, focusing on innovative methods that enhance model accuracy and fairness across different data distributions. We explore ten distinct techniques, including Nearest Neighbor Guidance (NNGuide), Parameter-Efficient Long-Tailed Recognition (PEL), and Ensemble Learning combined with data augmentation strategies like SMOTE. Each method was rigorously tested across popular datasets like CIFAR-10, CIFAR-100, and ImageNet-LT, utilizing metrics such as AUROC and F1 scores for a comprehensive evaluation. Our findings not only highlight the strengths and limitations of each approach but also guide the selection of appropriate techniques depending on the specific characteristics of the dataset. The insights from this research contribute to both theoretical and practical advancements in handling class imbalance, offering a pathway to more robust and equitable machine learning applications. This study underscores the necessity of tailored approaches to manage class disparities, paving the way for future innovations in the field.

Executive Summary

This paper explores various methods for addressing class imbalance in machine learning. Class imbalance is a persistent challenge that significantly impacts model performance across different applications. Our research systematically analyzes and compares the effectiveness of ten innovative methods designed to enhance Out-Of-Distribution (OOD) detection and long-tailed recognition using datasets such as CIFAR-10, CIFAR-100, and ImageNet-LT.

A few of the techniques studied include Nearest Neighbor Guidance (NNGuide), which improves OOD detection by integrating distance-based metrics with classifier confidence; Parameter-Efficient Long-Tailed Recognition (PEL), which optimizes the fine tuning of pre-trained models to enhance performance on under-represented classes; and Ensemble Learning combined with Data Augmentation strategies like SMOTE, which are crucial for mitigating the effects of class imbalance.

Our analysis consisted of thorough statistical evaluations and performance metrics, such as AUROC and F1 scores, to assess each method's effectiveness across various scenarios. Notably, methods like the Visual-Linguistic Transformer (VL-LTR) and Class-Balanced Distillation (CBD) demonstrated significant improvements in recognizing tail classes, showcasing their potential in real-world settings where data distribution is often skewed.

The insights gained from this comparative study not only contribute to the academic discourse on addressing class imbalance but also provide practical recommendations for deploying these methods in operational environments. Future work will focus on refining these approaches and exploring their applicability in other domains, potentially enhancing the robustness and fairness of machine learning models across diverse applications.

Contents

Introduction and Motivation

Machine learning has emerged as a powerful tool for extracting valuable insights and making predictions from vast amounts of data through a variety of different domains. However when faced with the challenge of imbalance datasets, where the distribution of classes is skewed with an uneven representation of classes, a conventional algorithm may struggle to learn from minority class examples and exhibit biases towards the majority class. Long-tailed class imbalance, a common problem in practical visual recognition tasks, often limits the practicality of deep network based recognition models in real-world applications, since they can be easily biased towards dominant classes and perform poorly on tail classes (Zhang et al., 2023). In many real-world scenarios, the minority class instances are scarce, resulting in limited availability of training for these classes which makes it challenging for algorithms to learn representative patterns and features from these minority classes. Imbalance datasets are prevalent in many real-world scenarios from fraud detection, medical diagnosis, anomaly detection, and text classification and more (Olawale, 2020). Misclassifying the minority class instances in these situations can come at a high cost, including financial loss, compromised security, or adverse health outcomes. The occurrence of rare events which are represented by the minority classes within these applications can be important when analyzing the datasets, making it essential to develop robust and effective techniques for handling class imbalance. There are many long-tailed learning methods that attempt to address these issues, however finding a state-of-art technique that is well generalized over multiple datasets while effectively handling imbalanced datasets can be challenging. In this context, the goal of this report is to explore different long-tailed learning methods and discuss the various strategies used to handle imbalance datasets and then determine which methods perform the best overall. We will examine the most recent state-of-the-art approaches aimed at improving the performance and fairness of machine learning models on imbalance datasets to highlight the recent advancements and best practices in the field of imbalanced learning. In this report we will explore the limitations of these long-tailed learning methods as well as discuss the strengths and weaknesses of the techniques the methods utilize when handling class imbalance at both the data and algorithm levels.

Related Works

Nearest Neighbor Guidance

The methods described in the paper "Nearest Neighbor Guidance for Out-of-Distribution Detection" focus on improving out-of-distribution (OOD) detection in machine learning models through a novel approach called Nearest Neighbor Guidance (NNGuide). Below, I've detailed the methods used in the paper, describing the fundamental approach, the specific techniques employed, and the evaluation framework used to test the effectiveness of these methods. The paper begins by outlining the importance of detecting OOD samples in machine learning systems, particularly those deployed in open-world environments where inputs may not always match the training distribution (Park et al., 2023). Traditional approaches, like classifier-based scores derived from a trained network, often suffer from overconfidence, incorrectly classifying far-OOD samples as in-distribution. The paper critiques these methods for their limitations and sets the stage for introducing their approach. The core of the paper's methodological contribution is the development of NNGuide, a new tool that combines the reliability of nearest neighbor methods with the detailed, precise scoring capabilities of classifier-based approaches. NNGuide works by leveraging both the distance-based metrics and classifier confidence scores to guide the detection of OOD samples. This is achieved by adjusting the classifier confidence of a test input based on its similarity to its nearest neighbors from a feature-embedded ID (in-distribution) data bank. The specific formula used modifies the base confidence score by incorporating the mean of the k-nearest neighbor distances, scaled by their respective confidences. This mechanism helps to mitigate the classifier's overconfidence by bounding the confidence scores based on neighbor proximity. The effectiveness of NNGuide is evaluated using comprehensive experiments on the ImageNet OOD detection benchmark, including scenarios involving natural distribution shifts in the data. They use standard OOD detection metrics like AUROC, FPR95, and AUPR to compare NNGuide against other established methods. Their results demonstrate that NNGuide consistently outperforms traditional approaches across various metrics and conditions. Extensive ablation studies are conducted to parse out the contributions of different components of NNGuide, demonstrating its robustness and adaptability to various architectures and configurations. The paper also delves into the theoretical aspects of NNGuide, providing propositions and proofs to support the method's efficacy. These theoretical insights underline how the guidance mechanism works under different conditions and contribute to a deeper understanding of why the method improves detection performance. The methods used in this paper represent a significant step forward in OOD detection, blending robust distance-based metrics with the nuanced capabilities of classifier-based systems through an innovative guidance approach. The detailed experiments and theoretical backing provided help establish NNGuide as a promising new tool for enhancing the reliability of machine learning models in varied and unpredictable environments.

Parameter-Efficient Long-Tailed Recognition

The paper titled "Parameter-Efficient Long-Tailed Recognition" introduces a fine-tuning method, PEL, designed to adapt pre-trained models for long-tailed recognition tasks efficiently. PEL addresses the challenge of fine-tuning pre-trained models on long-tailed datasets, where a few classes dominate the distribution. The method focuses on reducing overfitting and improving performance on tail classes without extensive additional training epochs or data. The core issue tackled by PEL is overfitting during fine-tuning, which particularly affects performance on less-represented (tail) classes (Shi et al., 2023). Two common approaches, full fine-tuning (adjusting all network parameters) and classifier fine-tuning (adjusting only the classifier), are found to exacerbate overfitting. PEL introduces parameter-efficient fine-tuning (PEFT), which

involves adjusting a small subset of model parameters, thus preserving the pre-trained model's discriminative power while efficiently adapting to new tasks. PEL employs various PEFT techniques, such as:
Bias-terms Fine-tuning (BitFit): Modifies only the bias terms in the model.

Visual Prompt Tuning (VPT): Introduces learnable prompts at different model layers.

Adapter and Low-Rank Adapter (LoRA): Adds small, trainable modules within the model's architecture without altering the majority of pre-trained weights.

AdaptFormer: A variation of the Adapter that integrates adjustments in parallel to existing model computations.

To expedite convergence and leverage semantic relationships inherent in class labels, PEL introduces a novel classifier initialization method. It utilizes textual features extracted from the CLIP model's textual encoder, providing a more effective starting point for the fine-tuning process. PEL's effectiveness is demonstrated across several long-tailed datasets, showing superior performance compared to existing state-of-the-art methods, particularly in handling tail classes. The method achieves significant improvements in accuracy with substantially fewer training epochs and without the need for additional training data. This methodological approach effectively leverages the strengths of pre-trained models while addressing the challenges specific to long-tailed recognition, providing a robust, efficient, and scalable solution for adapting deep learning models to imbalanced datasets.

Ensemble Learning

The methods discussed in the paper "A review of ensemble learning and data augmentation models for class imbalanced problems" explore innovative strategies to address the challenge of class imbalance in classification problems. The paper evaluates a variety of ensemble learning techniques, which are methods that combine multiple models to achieve better predictive performance than any single model could on its own. These techniques include bagging, a method that generates multiple versions of a predictor and uses them to get an aggregated predictor, boosting, an approach focused on training predictors sequentially, each trying to correct its predecessor, and stacking, which is training a new model to combine the predictions of several other models. Data augmentation plays a crucial role in addressing class imbalance by artificially enhancing the dataset with new, synthetic examples. The methods reviewed include:
SMOTE (Synthetic Minority Over-sampling Technique): This method generates synthetic samples from the minority class to balance the class distribution.

Random Oversampling and Undersampling: These methods involve replicating the minority class samples or reducing the majority class samples.

Advanced Techniques: The paper also discusses more sophisticated approaches like the use of Generative Adversarial Networks (GANs) to generate realistic synthetic data.

The paper presents a comprehensive framework for evaluating the performance of different combinations of data augmentation and ensemble learning methods. It uses several datasets to test these methods and utilizes metrics like the Area Under the ROC Curve (AUC), F1 score, and accuracy to assess model performance (Ahmad Khan et al., 2023). A significant contribution of this study is the integration of ensemble learning with data augmentation to tackle the unique challenges posed by imbalanced datasets effectively. The paper explores how different combinations of these methods can enhance model performance, particularly in scenarios where one class significantly outnumbers another. In summary, the paper provides a methodological review of how ensemble learning and data augmentation can be used synergistically to improve the outcomes in class-imbalanced situations. This approach not only mitigates the issue of overfitting but also ensures that the minority class is more accurately predicted, thereby enhancing overall model robustness.

µ2Net

The paper "A Continual Development Methodology for Large-Scale Multitask Dynamic ML Systems" presents a novel methodology for developing machine learning models, specifically focusing on dynamic multitask systems. The proposed methodology builds upon the existing µ2Net framework, introducing significant enhancements tailored for multitask learning (Gesmundo, 2022). The enhancements include a new scoring function that incorporates size and compute penalties. This function uses an exponential decay penalty to account for the complexity and efficiency of the model. Also, the paper extends the hyperparameter search space, allowing for greater flexibility in model training. This includes adjustments in learning rates, warm-up ratios, and optimization techniques. The methodology adds new types of mutation actions to modify the network architecture dynamically including layer cloning and removal, which allows selective cloning of layers to inherit parameters and optimizer states, as well as the removal of transformer layers to reduce model complexity and compute requirements, and also includes hyperparameter mutation which allows the adjustment of hyperparameters within a predefined search space to fine-tune model performance. A key aspect of the methodology is its focus on continual learning. Models are incrementally extended by learning new tasks and integrating new methodological improvements without starting from scratch. Also, the system undergoes multiple task iterations, where each iteration introduces methodological extensions or new tasks, progressively expanding the model's capabilities. The methodology is empirically evaluated to ensure its effectiveness in real-world scenarios. It compares the new method with existing baselines using a standard machine learning evaluation approach to assess performance trade-offs. The methodology also demonstrates the scalability of the method by applying it to a large number of image classification tasks, highlighting improvements in model quality and computational efficiency. In summary, the paper introduces a robust framework for developing scalable and efficient multitask machine learning systems that adapt and evolve through continual learning and dynamic architecture modifications. This approach not only enhances model performance but also optimizes computational resources, making it highly relevant for large-scale machine learning applications.

VL-LTR

The paper "VL-LTR: Learning Class-wise Visual-Linguistic Representation for Long-Tailed Visual Recognition" introduces innovative methods for enhancing visual recognition in long-tailed distributions by leveraging visual-linguistic data. The core component of the methodology is CVLP, which aims to align class-wise image and text data using contrastive learning. The approach takes advantage of both the visual features from images and linguistic information from text descriptions. This training strategy allows the model to understand and correlate high-level abstract information from texts with the low-level details from images, enhancing its recognition capabilities across variably represented classes. The LGR head is a novel addition that utilizes the visual-linguistic embeddings generated during CVLP (Tian et al., 2021). It applies these embeddings to guide the recognition process, particularly focusing on improving the model's performance in recognizing tail classes with fewer samples. The LGR head uses attention mechanisms to weigh the importance of different textual features relative to the visual input, dynamically adjusting the recognition process based on the content of both modalities. To refine the model's focus and reduce noise, the method includes an anchor sentence selection process. This process filters out less relevant or noisy text descriptions, selecting the most informative and discriminative textual data for training. This helps in maintaining high-quality linguistic data that enhances the model's learning and generalization capabilities. The paper validates these methods through extensive experiments on multiple benchmarks like ImageNet-LT, Places-LT, and iNaturalist 2018. The VL-LTR model demonstrates significant improvements over previous state-of-the-art methods, particularly in handling classes with fewer samples, showcasing the effectiveness of integrating visual and linguistic modalities in long-tailed recognition scenarios. In summary, the VL-LTR methodology effectively combines the strengths of visual and linguistic data to address the challenges posed by long-tailed distributions in visual recognition tasks, setting a new benchmark in the field.

Long-Tail Learning via Logit Adjustment

The paper titled "Long-Tail Learning via Logit Adjustment" proposes a sophisticated approach to tackle the challenges of long-tail distribution in machine learning classification tasks, focusing on enhancing model performance on rare or under-represented classes. The core methodological innovation in the paper is the use of logit adjustment to improve the performance on long-tailed datasets. The techniques include modifying the training loss to account for class frequencies by introducing an adjustment term directly into the softmax cross-entropy loss (Krishna Menon et al., 2021). This adjustment helps increase the relative margins between logits of rare versus dominant labels, promoting better generalization on rare classes. The adjustments also include adjusting the logits of a trained model based on class frequencies, aiming to recalibrate the output towards a more balanced prediction across classes. The paper emphasizes the statistical underpinning of the logit adjustment method. It is shown to be Fisher consistent for minimizing the balanced error, a metric that averages per-class errors. This statistical property ensures that the method not only improves empirical performance but is also grounded in robust theoretical principles. A reference implementation of the methods is made available, demonstrating the practical applicability of the techniques. The implementation details provide insights into integrating the logit adjustment into existing training pipelines. The methods are empirically validated on real-world datasets, where they demonstrate significant improvements in handling class imbalance. The results show better performance compared to existing techniques, especially in enhancing the detection and classification of under-represented classes. In summary, the paper provides a methodologically sound and empirically effective approach to addressing the challenges of long-tail learning in classification tasks. By adjusting logits based on class frequencies, the proposed techniques ensure more equitable learning across different classes, improving the model's performance on rare categories without compromising overall accuracy.

Relative Mahalanobis Distance

The paper "A Simple Fix to Mahalanobis Distance for Improving Near-OOD Detection" introduces the Relative Mahalanobis Distance (RMD) method to enhance Out-Of-Distribution (OOD) detection in neural networks, particularly focusing on near-OOD scenarios where traditional methods like the Mahalanobis Distance (MD) tend to fail. The traditional MD method is utilized for OOD detection by calculating the distance from a test input's feature map to class-conditional Gaussian distributions fitted to each class of the in-distribution data. The MD is computed as the minimum distance across all classes, which is then used as a confidence score for determining whether a sample is in-distribution or not. However, this method often fails for near-OOD detection because it does not effectively differentiate between subtly different but critical variations among classes. To address the shortcomings of MD in near-OOD scenarios, the paper proposes RMD, which adjusts the traditional MD by subtracting the distance from the test input to a "background" Gaussian distribution modeled on the entire training dataset without considering class labels (Ren et al. 2021). This background model acts as a baseline representation of the training data's overall feature space, providing a reference point that helps highlight deviations more indicative of OOD samples. The effectiveness of RMD is validated through experiments on challenging datasets, including CIFAR-100 vs. CIFAR-10, which are designed to test the model's ability to discern between closely related datasets (near-OOD). The results show significant improvements in AUROC scores for RMD over MD, demonstrating its superior capability in detecting near-OOD samples. The paper also compares RMD to other OOD detection techniques, showing that RMD provides a more stable and reliable measure for detecting near-OOD samples without requiring additional model training or complex hyperparameter tuning. In summary, this method enhances near-OOD detection accuracy by incorporating a relative measure against a comprehensive background model, making it more robust against the variations within near-OOD samples that traditional MD might overlook.

Class-Balanced Distillation

The paper titled "Class-Balanced Distillation for Long-Tailed Visual Recognition" by Iscen et al. introduces a novel framework called Class-Balanced Distillation (CBD), aimed at improving long-tailed visual recognition. This methodology is particularly effective in addressing the challenges posed by datasets where a few classes are vastly overrepresented compared to others. The proposed CBD employs a two-stage learning process. In the initial stage, multiple teacher models are trained using instance sampling. This traditional approach samples each instance with equal probability, irrespective of class frequency, which is effective for learning generalizable features but can underrepresented tail classes. The second stage introduces class-balanced sampling, where each class has an equal probability of being sampled, to focus on under-represented classes. This stage uses knowledge distillation from the teacher models to train a student model, incorporating the features learned by the teachers into the student model through a feature distillation process (Iscen et al., 2021). CBD enhances feature representations by using knowledge distillation, where the student model learns from the distilled knowledge of the teachers. This method allows the student to learn a more balanced representation across all classes, improving its ability to recognize tail classes without losing performance on head classes. An innovative aspect of CBD is the use of an ensemble of teacher models trained under different conditions (e.g., with different data augmentations) to provide a rich source of knowledge for the distillation process. This diversity helps in capturing a broader range of features and nuances, which is particularly beneficial for classes that are less represented. The effectiveness of CBD is demonstrated through extensive experiments on long-tailed recognition benchmarks such as ImageNet-LT and iNaturalist datasets. The results show that this method substantially outperforms traditional approaches, particularly in enhancing recognition accuracy for tail classes while maintaining or improving performance on head classes. In summary, the Class-Balanced Distillation method provides a robust and scalable approach to addressing the challenges of long-tailed distributions in visual recognition tasks. By intelligently leveraging the strengths of both instance and class-balanced sampling and enriching the training process with knowledge distillation, CBD achieves superior recognition performance across varied class frequencies.

Bilateral-Branch Network

This paper introduces a Bilateral-Branch Network (BNN) that incorporates two distinct learning branches: Conventional Learning Branch which utilizes a uniform sampler and focuses on learning universal patterns and representation learning from the original data distribution, and Re-Balancing Branch which has a reversed sampler tailored for the tail classes, focusing on the under-represented data to adjust and enhance classifier learning. Each of these branches performs a specific task to ensure comprehensive learning across classes that are both heavily populated as well as sparse. By having branches that encompass both learning methods, BNN utilizes an adaptive trade-off parameter which changes with the number of training stages. As the training progresses, the focus shifts towards balancing the branch, increasing the attention on the under-represented classes and improving the performance on these classes (Zhou et al., 2020). The trade-off between branches is controlled by an automatic adjusted parameter which allows the network to start with a focus on broad learning and shifts towards the imbalance in class representation. The outputs of both branches are combined using the automatic parameter to determine the final output of the network. This paper includes empirical results on several benchmark long-tailed datasets, such as the iNaturalist dataset, demonstrating that the BNN significantly outperforms existing methods. This performance is attributed to the effective handling of both the head and tail classes, with a balanced focus on pattern recognition and specific class adjustment. The BNN method effectively addresses the inherent problems of long-tailed distributions in visual recognition tasks by integrating two learning branches with an adaptive learning focus.

Momentum Contrast

The paper "Momentum Contrast for Unsupervised Visual Representation Learning" introduces a novel method known as Momentum Contrast (MoCo) for enhancing unsupervised learning in visual tasks (He et al., 2020). MoCo proposes a dynamic dictionary-based approach for contrastive learning, which addresses the challenges associated with maintaining consistency and size of the dictionary during training. This method involves two main components:
Dynamic Dictionary: MoCo utilizes a queue mechanism where the encoded representations of the current mini-batch are enqueued, and the oldest mini-batch is dequeued. This queue is decoupled from the mini-batch size, allowing the dictionary to grow significantly without being limited by the batch size.

Momentum-based Encoder Update: To maintain the consistency of the encoded representations in the dictionary, MoCo uses a momentum-based update for the key encoder. The key encoder is updated as a moving average of the query encoder, which ensures that the keys evolve smoothly and remain consistent over time, despite the changing input data.

MoCo conceptualizes the learning process as a dictionary look-up. An encoded query should match its corresponding key in the dictionary and be distinct from all other entries. This is facilitated by the contrastive loss function, which minimizes the distance between similar pairs and maximizes the distance between dissimilar pairs. The InfoNCE loss function is employed, which uses a softmax-based classifier to distinguish the correct key from a set of negative samples. MoCo is applied in the context of an instance discrimination pretext task, where a positive pair consists of different views (e.g., crops) of the same image, and all other pairs are considered negative. This task helps in learning robust visual features from unlabeled data. The effectiveness of MoCo is demonstrated through extensive experiments on datasets like ImageNet and Instagram-1B, showing significant improvements in tasks such as image classification, object detection, and segmentation. MoCo is shown to achieve competitive results with supervised pre-training methods, suggesting it as a viable alternative for various applications. In summary, MoCo provides an effective framework for unsupervised visual representation learning by innovatively managing the dictionary size and maintaining encoder consistency through momentum updates. This approach bridges the gap between unsupervised and supervised learning in many vision tasks.

Research Objectives

When addressing the challenges of dealing with imbalanced datasets, we came up with three research objectives:

1. Evaluate and compare the performance of the 10 different long-tailed learning methods across CIFAR-10, CIFAR-100, and ImageNet-LT datasets.
2. Identify the strengths and weaknesses of each long-tailed learning method in handling different degrees of class imbalance and dataset sizes.
3. Assess how well each long-tailed learning method generalizes across datasets with varying complexities, from balanced to highly imbalanced class distributions.

Performance Evaluation Across Datasets

In order to evaluate the performance of the ten long-tailed learning methods, we compare the performance of these methods across three distinct datasets: CIFAR-10, CIFAR-100, and ImageNet-LT. We utilize performance metrics including the accuracy, precision, and runtime when training each method using the three datasets. Our goal is to rank the ten long-tailed methods based upon which method is most effective based upon the metrics collected and by doing so, we can identify the most efficient long-tailed learning method across different datasets.

Strengths and Weaknesses Analysis

To identify the strengths and weaknesses of the ten long-tailed learning methods, we assess how the methods handle different degrees of class imbalance as well as dataset sizes. By identifying the patterns and trends, our goal is to better understand the strengths and limitations for each method. This will provide more insights into which long-tailed learning methods excel at certain imbalance ratios and dataset sizes. We hope to find common challenges and limitations these methods face when addressing imbalanced dataset scenarios.

Generalization Across Datasets

To assess how well the ten long-tailed learning methods are able to adapt to other datasets, we observe the behavior of each long-tailed learning method across diverse datasets. Our goal is to gain valuable insight into the transferability of knowledge learned from one dataset to another with varying complexities. Through this analysis, we hope to apply this information within real-world scenarios where long-tailed learning methods must adapt across diverse datasets.

Methodology and Analysis

Datasets:

There were three datasets that were mainly used in these papers, and we will use in our further evaluations, are CIFAR-10, CIFAR-100, and ImageNet-LT. CIFAR-10 and CIFAR-100 are popular datasets used for benchmarking image recognition algorithms. Both datasets contain roughly 60,000 images that are split into training and test sets. However, CIFAR-10 is split into 10 classes of 6,000 images per class, split up into broad categories such as dogs, cats, etc. This makes it easy for general image recognition as the images are so broad. This differs from CIFAR-100 as this dataset contains 100 classes with 600 images in every class. These classes are also grouped into 20 superclasses. This makes it easy for testing algorithms on precise classification as well as a larger group of class structures. Finally, ImageNet-LT differs from both of these greatly. This dataset is designed to be close to a real world example, as some classes are abundant but many are underrepresented. This is also a long-tail version of the ImageNet dataset, which contains over one million images over 1,000 classes. ImageNet-LT addresses the challenge of imbalanced datasets as it encompasses a skewed distribution of images across many classes, which makes it a great resource for evaluating algorithms.

Preprocessing

The preprocessing steps for each method varied slightly due to different requirements. A common step we encountered was resizing the images to a uniform size to ensure consistency in input dimensions across the dataset, changing the image size to 256x256 or 224x224 pixels. The pixel values of the images would also need to be normalized to standardize the input data and make the optimization more efficient to prevent issues like vanishing or exploding gradients during training. During testing, we applied class weights or resampling techniques to account for class imbalance when the model required training with such techniques to ensure that the evaluation reflects the model's performance under similar conditions as during training.

Performance Evaluation

In order to understand which methods achieved the best accuracy, we used a diverse range of methods, including Momentum Contrast (MoCo), Bilateral-Branch Network (BBN), Class-Balanced Distillation (CBD), Long-Tail Learning via Logit Adjustment, µ2Net+, VL-LTR (Visual-Linguistic Transformer for Long Tailed Recognition), Parameter-Efficient Fine-Tuning with Limited Data (PEL), Nearest Neighbor Guidance (NNGuide), Relative Mahalanobis Distance (RMD), and Learning Ensemble methods. Each method follows the same implementation as the pre-trained models to keep consistency within the evaluation. The datasets were divided into training, validation, and test sets to recreate scenarios with imbalanced datasets. The performance evaluation was based on measuring classification accuracy for each method followed by statistical analysis to compare the performance using the three datasets.

Strengths and Weaknesses Analysis

To assess the robustness of each method, the methods were simulated with different degrees of class imbalance within the datasets. We also took into factor the performance of the methods on datasets of different sizes, ranging from small to large. Secondly, we explored the sensitivity of each method to hyperparameters and training strategies, such as learning rate, batch size, and data augmentation techniques. The variations in hyperparameters gave insight to the performance and stability of the methods. We also take into account the generalization across datasets in order to identify which methods indicate their ability to generalize well to new data environments.

Results

Accuracy of Methods

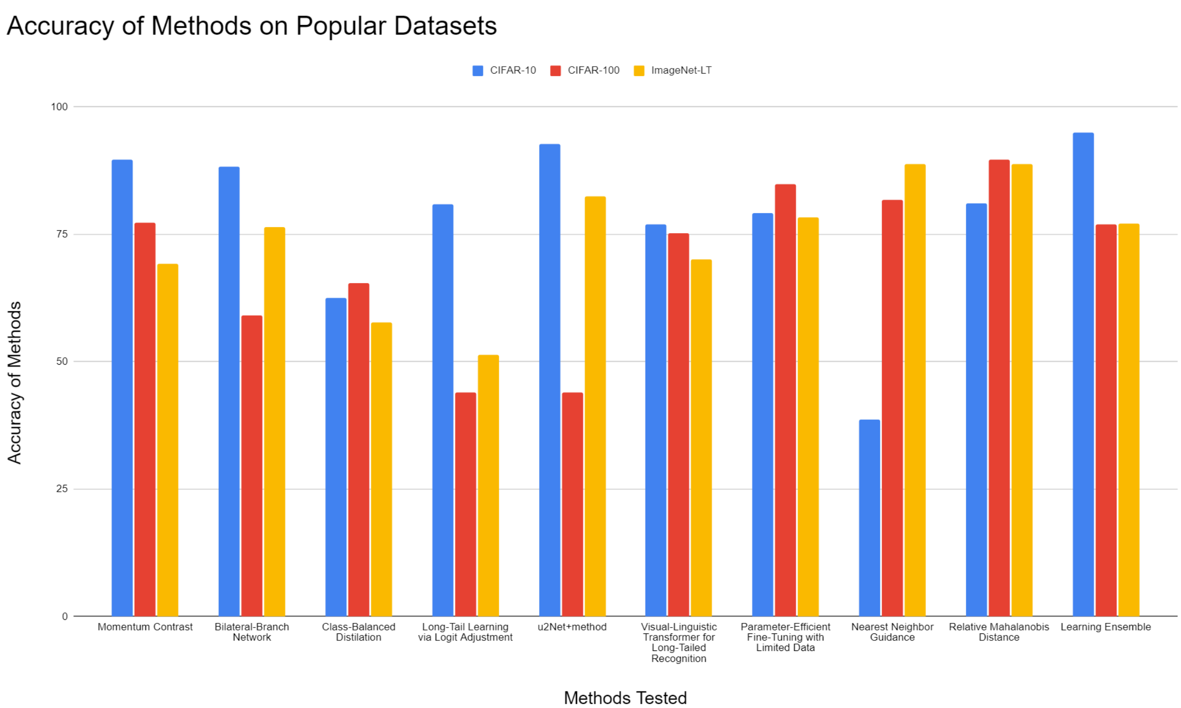

Figure 1: Barchart displaying the accuracy of methods across all three datasets

Method	CIFAR-10	CIFAR-100	ImageNet-LT
Momentum Contrast	89.7	77.3	69.3
Bilateral-Branch Network	88.3	59.12	76.5
Class-Balanced Distillation	62.5	65.4	57.7

Long-Tail Learning via Logit Adjustment	80.92	43.89	51.3
u2Net+method	92.67	43.89	82.5
Visual-Linguistic Transformer for Long-Tailed Recognition	77	75.2	70.1
Parameter-Efficient Fine-Tuning with Limited Data	79.1	84.9	78.3
Nearest Neighbor Guidance	38.6	81.74	88.82
Relative Mahalanobis Distance	81.01	89.71	88.82
Learning Ensemble	94.93	77.02	77.06

Table 1: Accuracy results of methods across three different datasets

From the results of our experiments, we were able to compile the highest accuracy achieved for each of the ten methods when being trained by CIFAR-10, CIFAR-100, and ImageNet-LT datasets. The Relative Mahalanobis Distance method was able to achieve an average accuracy score of 86.51%. Using the accuracy achieved from the experiments, we were able to determine the overall ranking of these methods based on the average accuracy score they achieved from all three datasets.

1. Relative Mahalanobis
2. Learning Ensemble
3. Parameter-Efficient Fine-Tuning with Limited Data
4. Momentum Contrast
5. Bilateral-Branch Network
6. Visual-Linguistic Transformer for Long-Tailed Recognition
7. u2Net+method
8. Nearest Neighbor Guidance
9. Class-Balanced Distillation
10. Long-Tail Learning via Logit Adjustment

Generalization Across Datasets

MoCo demonstrated strong generalization capabilities across datasets with varying complexities and class imbalances. The pretrained MoCo model performed well in classification tasks which lead to improved performance on datasets with different characteristics. BBN generalization capabilities depended on its ability to capture both local and global features effectively. Although BBN was able to perform well with datasets with similar characteristics it was trained on, its performance on datasets with significantly different characteristics resulted in significantly worse accuracy scores. Class-Balanced Distillation addressed class imbalance by utilizing a teacher model to distill knowledge to a student model. The generalization across datasets relied on the quality and diversity of the teacher model which led to a poor performance of accuracy when using the three datasets used in our experiment. Long-Tail Learning via Logit Adjustment relied on recalibrating class probabilities in order to alleviate the effects of class imbalance. However, it was unable to adapt to different data distributions and imbalance levels because of its limitations on existing techniques for handling imbalanced datasets. The u2Net+ method relied on its innovative design and method extensions aimed at enhancing the capabilities and performance of Machine Learning models, however it struggled with adapting to the CIFAR-100 dataset. The VL-LTR model leveraged transformer-based architectures for long-tailed recognition tasks (Tian et al., 2021). By integrating visual and linguistic information using Transformer-based architectures, VL-LTR consistently performed well with overall accuracy, particularly when given limited data for certain classes. PEL addressed generalization across datasets by focusing on fine-tuning pretrained models with limited data in order to improve generalization performance. PEL achieved high accuracy scores for all three datasets, displaying its effective and efficient ability to generalize across datasets. NNGuide utilized information from the "nearest neighbor" for the models decision-making process, enhancing its ability to detect samples that differ significantly from the training data distribution (Park et al., 2023). NNGuide was able to generalize well when adapting to large scale datasets. RMD successfully improved the robustness and reliability of OOD detection, leading to more accurate classification and anomaly detection from various datasets, displaying its ability to generalize well with different datasets. The Learning Ensemble method combined multiple models and techniques to improve generalization performance. By leveraging the diversity of individual models, ensembles reduce variance, improves generalization, and yields more reliable predictions, making the method a powerful approach at dealing with a diverse selection of datasets.

Scalability and Efficiency

MoCo demonstrated its ability to scale to different sized datasets as well as its efficiency in training and was able to outperform most of the other contrastive learning approaches. The computational efficiency of MoCo made it suitable for large-scale datasets with a limited amount of resources for training. While BBN offered a competitive performance compared to the other long-tailed learning methods, its scalability was limited by the computational resources required for its training, making it less efficient when compared to the other methods. Because the CBD method is impacted by the computational cost of distillation, this method is not as efficient at training multiple models when compared to the other methods. However, its innovative two-stage learning framework and knowledge distillation strategies achieves high performances for large scale imbalance datasets such as ImageNet-LT. The Long-Tail Learning via Logit Adjustment has a much lower computational cost when compared to the other methods while maintaining scalability for datasets of varying sizes. The u2Net+ method offered a competitive performance and scalability for both CIFAR-10 and ImageNet-LT datasets, however because of its dependency on its architectural design and computational requirements, it was unable to successfully adapt and perform well on the CIFAR-100 dataset. VL-LTR requires a high computational cost especially for large-scale datasets. While the transformer-based model offers state-of-the-art performance, its scalability is also limited by the computational resource requirements from the model. The PEL method depends on its fine-tuning techniques to achieve scalability and efficiency (Shi et al., 2023). Requiring only 10 epochs and fine-tuning far fewer model parameters when compared to other existing methods, PEL excels at being able to adapt to different sized data-sets while remaining quick and efficient. The NNGuide scalability and efficiency is dependent upon the computational cost of nearest neighbor search and guidance. The architecture of the model impacts the scalability as it is unable to perform well on smaller sized data sets like CIFAR-10. The RMD method achieves high scalability with low computational overhead as it depends on distance relative to class centroids (Ren et al. 2021). Because of its distance-based techniques, it is very efficient in handling class imbalance while maintaining scalability across datasets. The Learning Ensemble method combines multiple models which leads to higher computational requirements compared to methods that use single models. While this method can scale well for different sized data sets, the computational costs of the method can potentially hinder its scalability.

Conclusion

MoCo for Unsupervised Learning: MoCo is valuable for learning representations from large-scale unlabeled datasets, particularly in computer vision and machine learning tasks. Its applications extend to computer vision and other machine learning tasks where labeled data may be scarce or unavailable.
BBN for Multi-modal Tasks: BBN is designed for tasks requiring multi-modal input or feature fusion, improving performance in tasks like image captioning and multi-modal retrieval. Its versatility makes it particularly valuable in applications such as image captioning and multi-modal retrieval.
CBD for Long-Tailed Recognition: CBD addresses class frequency imbalances in datasets like ImageNet-LT, showing substantial improvements in accuracy, especially for tail classes. Its effectiveness in long-tailed recognition tasks highlights its importance in real-world applications where class imbalances are common.
VL-LTR for Visual-Linguistic Integration: VL-LTR integrates visual and linguistic

information using Transformer-based architectures, achieving high overall accuracy and significant improvements in few-shot learning. VL-LTR shows the most promise for tasks that require understanding both visual and textual content.

These models have shown effectiveness in their respective areas, and the choice for the best model depends on the specific requirements for the task at hand. For example, if dealing with unlabeled data, MoCo might be the best choice, while for long-tailed recognition tasks, CBD or VL-LTR would be more suitable. Overall, these models showcase the diverse capabilities of deep learning methods and their potential to address complex challenges across different domains.

Future Research

The rapid advancement of long-tailed learning techniques has significantly expanded the horizons of artificial intelligence, enabling groundbreaking progress across various domains. Future work can explore enhancements to the existing model architectures used in these methods, addressing challenges in domain adaptation and transfer learning can positively contribute to the advancement of machine learning and improve model performance and generalization across diverse datasets.

As deep learning models become increasingly complex, ensuring interpretability and explainability is paramount, especially in high-stakes applications such as healthcare or finance. Future work could focus on developing better techniques to interpret and explain the decision making behind these models, providing insight to the internal workings of these methods and increasing trust and transparency.

Addressing ethical considerations and mitigating biases in AI systems are crucial for responsible deployment. Future research should focus on developing methods and frameworks to detect and mitigate biases to ensure fairness and uphold ethical standards throughout the development and deployment lifecycle of the machine learning methods used. This could involve identifying biases in training data as well as understanding how biases propagate through the entire machine learning pipeline.

Bibliography

Ahmad Khan, Azal, et al. “A Review of Ensemble Learning and Data Augmentation Models for Class Imbalanced Problems: Combination, Implementation and Evaluation.” *Papers With Code*, 6 Apr. 2023, paperswithcode.com/paper/a-review-of-ensemble-learning-and-data.

Awe, Olawale. “Handling Imbalanced Dataset Classification in Machine Learning | IEEE Conference Publication | IEEE Xplore.” *Computational Strategies for Handling Imbalanced Data in Machine Learning*, 2020, ieeexplore.ieee.org/document/9362471.

Gesmundo, Andrea. “Papers with Code - A Continual Development Methodology for Large-Scale Multitask Dynamic ML Systems.” *A Continual Development Methodology for Large-Scale Multitask Dynamic ML Systems | Papers With Code*, 15 Sept. 2022, paperswithcode.com/paper/a-continual-development-methodology-for-large.

He, Kaiming, et al. “Papers with Code - Momentum Contrast for Unsupervised Visual Representation Learning.” *Momentum Contrast for Unsupervised Visual Representation Learning | Papers With Code*, 2020, paperswithcode.com/paper/momentum-contrast-for-unsupervised-visual.

Iscen, Ahmet, et al. “Papers with Code - Class-Balanced Distillation for Long-Tailed Visual Recognition.” *Class-Balanced Distillation for Long-Tailed Visual Recognition | Papers With Code*, 12 Apr. 2021, paperswithcode.com/paper/class-balanced-distillation-for-long-tailed.

Krishna Menon, Aditya, et al. “Papers with Code - Long-Tail Learning via Logit Adjustment.” *Long-Tail Learning via Logit Adjustment | Papers With Code*, 2021, paperswithcode.com/paper/long-tail-learning-via-logit-adjustment.

Park, Jaewoo, et al. “Papers with Code - Nearest Neighbor Guidance for out-of-Distribution Detection.” *Nearest Neighbor Guidance for Out-of-Distribution Detection | Papers With Code*, Sept. 2023, paperswithcode.com/paper/nearest-neighbor-guidance-for-out-of-1.

Ren, Jie, et al. “A Simple Fix to Mahalanobis Distance for Improving Near-Ood Detection.” *Papers With Code*, 16 June 2021, paperswithcode.com/paper/a-simple-fix-to-mahalanobis-distance-for.

Shi, Jiang-Xin, et al. “Papers with Code - Parameter-Efficient Long-Tailed Recognition.” *Parameter-Efficient Long-Tailed Recognition | Papers With Code*, 18 Sept. 2023, paperswithcode.com/paper/parameter-efficient-long-tailed-recognition.

Tian, Changyao, et al. “Papers with Code - VL-LTR: Learning Class-Wise Visual-Linguistic Representation for Long-Tailed Visual Recognition.” *VL-LTR: Learning Class-Wise*

Visual-Linguistic Representation for Long-Tailed Visual Recognition | *Papers With Code*, 26 Nov. 2021, paperswithcode.com/paper/vl-ltr-learning-class-wise-visual-linguistic.

Zhang, Yifan, et al. "Deep Long-Tailed Learning: A Survey." *arXiv*, 15 Apr. 2023, arxiv.org/pdf/2110.04596.pdf.

Zhou, Boyan, et al. "Papers with Code - BBN: Bilateral-Branch Network with Cumulative Learning for Long-Tailed Visual Recognition." *BBN: Bilateral-Branch Network with Cumulative Learning for Long-Tailed Visual Recognition* | *Papers With Code*, 2020, paperswithcode.com/paper/bbn-bilateral-branch-network-with-cumulative.