Int J Syst Assur Eng Manag
https://doi.org/10.1007/s13198-024-02346-3

ORIGINAL ARTICLE

Imbalcbl: addressing deep learning challenges with small and imbalanced datasets

Saqib ul Sabha[1] · Assif Assad[1] · Sadaf Shafi[1] · Nusrat Mohi Ud Din[1] · Rayees Ahmad Dar[1] · Muzafar Rasool Bhat[2]

Received: 12 November 2023 / Revised: 12 February 2024 / Accepted: 7 April 2024

Abstract Deep learning, while transformative for computer vision, frequently falters when confronted with small and imbalanced datasets. Despite substantial progress in this domain, prevailing models often underachieve under these constraints. Addressing this, we introduce an innovative contrast-based learning strategy for small and imbalanced data that significantly bolsters the proficiency of deep learning architectures on these challenging datasets. By ingeniously concatenating training images, the effective training dataset expands from n to n^2, affording richer data for model training, even when n is very small. Remarkably, our solution remains indifferent to specific loss functions or network architectures, endorsing its adaptability for diverse classification scenarios. Rigorously benchmarked against four benchmark datasets, our approach was juxtaposed with state-of-the-art oversampling paradigms. The empirical evidence underscores our method's superior efficacy, outshining contemporaries across metrics like Balanced accuracy, F1 score, and Geometric mean. Noteworthy increments include 7–16% on the Covid-19 dataset, 4–20% for Honey bees, 1–6% on CIFAR-10, and 1–9% on FashionMNIST. In essence, our proposed method offers a potent remedy for the perennial issues stemming from scanty and skewed data in deep learning.

Keywords Small data · Imbalanced data · Deep learning · Imbalanced classification

1 Introduction

DEEP learning has garnered immense acclaim for its exceptional performance and versatility across a wide range of domains. such as Object Recognition (Hu et al. 2018), image classification (Jin et al. 2022; Krizhevsky et al. 2017), object detection (Ren et al. 2015), anomaly detection (Hossain et al. 2019) and numerous other fields. Despite the impressive achievements of deep learning models, they may encounter significant challenges when confronted with imbalanced data, which is a well-established and crucial problem in the field of machine learning. This challenge arises because most models are trained on balanced datasets and are not optimized to handle imbalanced data, which can lead to suboptimal performance, particularly when the training data among different classes is extensively skewed (Japkowicz and Stephen 2002; He and Garcia 2009).

Imbalanced data is a prevalent issue in many real-world applications, stemming from various applications, including fraud detection (Sanz et al. 2014), medical diagnosis (Bach et al. 2017), and activity recognition(Gao et al. 2016). The imbalanced nature of datasets can lead to reduced

✉ Saqib ul Sabha
saqib.sabha@islamicuniversity.edu.in

Assif Assad
assifassad@gmail.com

Sadaf Shafi
Daee.sadaf2011@gmail.com

Nusrat Mohi Ud Din
nusratmohiuddin92@gmail.com

Rayees Ahmad Dar
darrayes@gmail.com

Muzafar Rasool Bhat
muzafar.rasool@islamicuniversity.edu.in

[1] Department of Computer Science and Engineering, Islamic University of Science and Technology, Pulwama, Jammu and Kashmir 192122, India

[2] Department of Computer Science, Islamic University of Science and Technology, Pulwama, Jammu and Kashmir 192122, India

Published online: 01 May 2024