國立臺灣大學管理學院資訊管理學研究所

碩士論文

Department of Information Management

College of Management

National Taiwan University

Master’s Thesis

加權排序對比回歸學習用於資料不平衡之社群媒體熱門度預測

Weighted-Rank Contrastive Regression for Robust Learning on Imbalance Social Media Popularity Prediction

賴彥良

Yen-Liang Lai

指導教授: 莊裕澤 博士

Advisor: Yuh-Jzer Joung Ph.D.

中華民國 113 年 7 月

July, 2024

doi:10.6342/NTU202403142

Acknowledgements

讀碩班是我人生中一個重要的決定，其中經歷了大大小小的戰鬥，而寫論文只是其中一場免不了的戰役。回想當時無數個懷疑自己的日子，再回過頭審視自己兩年的變化，不由感嘆自己又成長了許多、又收成了一場難忘的勝利。比爾蓋茲說過:「大部分的人高估他們一年內能做的事，卻也低估了他們十年內能做到的事。」雖然我的人生階段還沒有一個十年期的里程碑，但做研究讓我對這句話有了深刻的體會。漫長的實驗讓人感到厭煩和不耐，在過程中點滴的累積經驗，卻也能慢慢堆砌成最終的成果。如果讓我再重來一次，我會告訴自己慢下來，多讀一些論文，多一些思考，多享受探索未知的過程。

我很感謝我的指導老師莊裕澤教授，總是給予我百分之百的自由去鑽研有興趣的領域，並且在過程中不斷要求細節，讓我能發揮最大的潛能。我也要感謝112A 實驗室的大家，陪我一起玩樂、一起投履歷、一起在實驗室度過每一個寫不出論文的日夜。感謝孫佚虹常常幫我丟垃圾跟泡乳清，感謝育緹每天的陪伴。最後我要感謝我的家人，在背後不斷的支持與鼓勵。

讀碩班是一個改變我人生的決定，讓我更認識自己的熱忱所在，讓我變得更加勇敢而謙遜。非常感謝這兩年所遭遇的人事點滴，也慶幸自己有好好享受這段痛苦且美好的過程。

doi:10.6342/NTU202403142

摘要

社群貼文的熱門程度往往反映受眾對於內容的喜愛程度，社群網紅或廠商可透過觀察貼文的按讚數變化，制定更有效的行銷策略，進而提升社群行銷的成效。因此，如何準確預測社群貼文的熱門度是一大關鍵。然而，現實世界的社群媒體資料具備不平衡的特性，極冷門與極熱門的貼文往往只具備很少的資料量，造成預測熱門度時的失準。有鑒於近年來對比學習在特徵學習的成功，以及將對比學習概念引進回歸任務的新興趨勢，本研究提出加權排序對比迴歸 (Weighted-Rank Contrastive Regression) 損失函數，以解決真實世界的回歸問題中數據不平衡的問題。我們在社群媒體熱門度預測資料集 (B. Wu et al., 2019) 上進行實驗，實驗結果顯示我們的方法優於傳統方法 (僅以 L1 損失函數進行擬合) 和當前的最先進的對比迴歸方法 Rank-N-Contrast (Zha et al., 2024)，尤其在處理高熱門貼文、缺乏負樣本數的異常值方面表現更佳。本研究所提出的加權排序對比迴歸損失函數不僅解決了社群媒體熱門度預測中的數據不平衡問題，更提供了一種可泛化的特徵學習方法，可推廣至其他任意的不平衡迴歸任務中。

關鍵字：社群媒體、熱門度預測、對比學習、不平衡回歸、網紅行銷

doi:10.6342/NTU202403142

Abstract

Social Media Popularity Prediction (SMPP) is the task of forecasting the level of engagement a social media post will receive. In SMPP, it is crucial for understanding audience engagement and enabling targeted marketing strategies. The popularity of social media posts often reflects the audience's preference for the content. Social media influencers or brands can design more effective marketing strategies by observing the changes in the number of likes on posts, thereby enhancing the effectiveness of social media marketing. However, the inherent imbalance in real-world social media data, where certain popularity levels are underrepresented, posed a significant challenge. In this study, we leveraged the recent success of contrastive learning and its growing integration into regression tasks, and introduced a weighted-rank contrastive regression loss to counteract the data imbalance challenges. Experiments on the Social Media Prediction Dataset (B. Wu et al., 2019) demonstrated that our method outperformed the vanilla approach (solely fit on L1 loss) and the current State of the art (SOTA) contrastive regression approach

doi:10.6342/NTU202403142

Rank-N-Contrast (Zha et al., 2024) , especially for challenging outliers with high popularity and few negative counterparts. The proposed weighted-rank contrastive regression loss not only addressed the inherent data imbalance in SMPP but also offered a robust representation learning solution that could be generalized to other real-world imbalanced regression tasks.

Keywords: Social Media, Popularity Prediction, Contrastive Learning, Imbalance Regression, Influencer Marketing

doi:10.6342/NTU202403142

Contents

Page

doi:10.6342/NTU202403142

doi:10.6342/NTU202403142

List of Figures

doi:10.6342/NTU202403142

doi:10.6342/NTU202403142

List of Tables

Chapter 1 Introduction

1.1 Background

Social media platforms have become deeply integrated into our daily lives, influencing how we communicate, access information, and consume content. For businesses and brands, social media represents a vast landscape of potential customers and a powerful tool for advertising and engagement. With the rise of social media, influencer marketing (Hayes, 2008) has emerged as an alternative to traditional marketing. This strategy involves brands collaborating with influencers (Freberg et al., 2010), who are social media users with established credibility and a loyal following within specific niches or industries. Influencers possess the ability to sway opinions and drive consumer behavior, making them valuable partners for brands aiming to reach targeted audiences. By partnering with influencers, brands could leverage this established trust and influence to promote their products or services in a more relatable and effective manner (X. Yang et al., 2019). Consumers are more likely to trust recommendations from individuals they admire and identify with, making influencer marketing a potent tool for enhancing brand awareness and driving sales.

A crucial aspect of influencer marketing is Social Media Popularity Prediction (SMPP), which is the task of forecasting the level of engagement a social media post will receive.

doi:10.6342/NTU202403142

This prediction is valuable for both content creators and businesses, as it informs content strategies and marketing decisions. For influencers, understanding which factors contribute to a post's success could help them refine their content strategies and maximize their reach. For brands, accurate popularity predictions could guide their influencers' selection process, ensuring they invest in partnerships that yield the highest return on investment. Social media platforms are dynamic environments where content popularity can fluctuate rapidly. Factors such as timing, visual appeal, content relevance, and user engagement all play a role in determining a post's success. Developing models that can accurately predict social media popularity is a complex task that requires a deep understanding of domain knowledge.

One of the main challenges in social media popularity prediction (SMPP) is the limited availability of real-world social media data. While some platforms offer APIs for data collection, access is often restricted or comes at a cost. This lack of publicly available data hampers research and development efforts in the field of social media popularity prediction. To address the scarcity of real-world data, the Social Media Prediction (SMP) Challenge (B. Wu et al., 2019) was established as an annual research event. The challenge provides participants with a well-structured dataset and a platform to develop and evaluate their models. Over the years, the SMP challenge has fostered significant advancements in social media popularity prediction, leveraging cutting-edge techniques in feature engineering and deep learning (Ding et al., 2019; Hsu et al., 2019; Lai et al., 2020; J. Wu et al., 2022). Another challenge in social media popularity prediction (SMPP) is the inherent imbalance in real-world social media data. The distribution of popularity metrics, such as number of likes, tends to be highly skewed, with a small number of posts achieving viral status while the majority receive mid-level engagement. This imbalance makes it difficult

doi:10.6342/NTU202403142

for traditional machine learning models to accurately predict popularity across the entire spectrum of values.

Although the SMP challenge has led to significant advancements in predicting social media popularity, the inherent data imbalance issue remains largely unaddressed, where the majority of social media posts fall into the mid-level popularity range, while posts at the extremes of low and high popularity suffer from data scarcity. This imbalance hinders the development of effective representations for these less frequent types of posts, and posed a significant obstacle to accurate prediction. Traditional approaches for handling imbalanced data primarily focus on categorical targets (Chawla et al., 2002; H. He & Ma, 2013; Yen & Lee, 2006), where the goal is to classify instances into distinct classes. However, numerous real-world applications involve continuous target variables, often with skewed distributions. For example, in computer vision, determining age from facial images involves a continuous target with inherent imbalances, similar challenges arise in medical applications where health metrics like heart rate and blood pressure are continuous and frequently exhibit skewed distributions across patients.

Y. Yang et al. (2021) first defined these challenges as Deep Imbalanced Regression (DIR) and proposed a smoothing approach to align feature and label space distributions for robust representation. Building on this, Gong et al. (2022) utilized a ranking loss as a regularizer to align feature and label similarities. Zha et al. (2024) further refined this approach by modeling the ranking loss as contrastive regression loss, thereby improving feature-label alignment which mitigating data imbalance issues. These recent advancements have highlighted the potential of contrastive regression as a promising approach for generating more robust representations in the face of imbalanced data.

doi:10.6342/NTU202403142

1.2 Research Motivation and Objectives

Inspired by Zha et al. (2024)'s success in modeling the feature-label alignment problem in a contrastive learning framework, we proposed a Weighted-Rank Contrastive Regression loss as a regularizer to address the data imbalance problem in social media popularity prediction. We integrated a weighted mechanism into the current state of the art (SOTA) Rank-N-Contrast (Zha et al., 2024). Our experiments on the Social Media Prediction Dataset (B. Wu et al., 2019) demonstrated the effectiveness of our approach in improving the accuracy and robustness of popularity predictions. The main contribution of this research are as follow:

- We proposed a Weighted-Rank Contrastive Regression loss that contrast the features by label distance.
- We demonstrated that by assigning weights to negative sample pairs enhanced the robustness of representation especially for those rare and extreme labels.
- We also proposed an simple end-to-end contrastive regression learning framework for multi-modal representation learning that can be extended to more complex architecture.

The remainder of this thesis is structured as follows. Chapter 2 provides a comprehensive literature review on Social Media Popularity Prediction and Imbalanced Regression. Chapter 3 details the proposed Weighted-Rank Contrastive Regression Loss, and the end-to-end multi-modal representation learning framework. In Chapter 4, we present an empirical evaluation of the proposed framework. Finally, Chapter 5 concludes the thesis by summarizing our research findings and discussing potential avenues for future work.

doi:10.6342/NTU202403142

Chapter 2 Literature Review

Social Media Popularity Prediction (SMPP) is a well-established research area with numerous studies proposing frameworks for predicting popularity of social media posts. Early studies relied on manually crafted features to enhance prediction accuracy (Gelli et al., 2015; Jin et al., 2010; McParlane et al., 2014). However, recent research has shifted towards leveraging pre-trained models for feature extraction (Ding et al., 2019; Kim et al., 2020; J. Wu et al., 2022; Xu et al., 2020), leading to significant advancements in performance over time. However, despite advancements in features and model architectures, the inherent imbalance distribution in social media data has often been overlooked, neglecting potential improvements from addressing this issue.

Y. Yang et al. (2021) proposed a feature distribution smoothing approach to tackle the imbalance regression problem. This approach enhanced the robustness of under-represented data and has effectively addressed the data imbalance. Subsequent studies (Gong et al., 2022; Keramati et al., 2023; Zha et al., 2024) have built upon this foundation, further demonstrating the potential of contrastive regression to revolutionize the field of imbalanced regression.

2.1 Social Media Popularity Prediction

Social Media Popularity Prediction (SMPP) refers to the task of forecasting the level of popularity a social media post will receive. Popularity, often measured by the number of likes, reflects the influencer's impact on viewers, making the prediction valuable for both content creators and businesses by informing content strategies and marketing decisions. Previous approaches to SMPP have employed two primary methods for feature extraction: manually preprocessed features and the utilization of pre-trained models.

2.1.1 Manually Preprocessed Features

Earlier studies in social media popularity prediction heavily relied on manually preprocessed features derived from textual, visual, or user profile data. For instance, Jin et al. (2010) employed upload frequency, upload time, and tags to predict image popularity on Flickr. McParlane et al. (2014) incorporated features from visual context (device type, size, orientation), visual content (scene type, number of faces, dominant color), user profile (gender, account type, number of uploads), and tags represented using TF-IDF vectors. Furthermore, Gelli et al. (2015) employed Name-Entity Recognition (NER) on image descriptions, identifying and counting entities like Location, Organization, and Person. These manually processed features have proven valuable and continue to be widely adopted in recent approaches. Notably, Ding et al. (2019) and Lai et al. (2020) also incorporated text features like caption length, and the number of tags. Table 2.1 summarizes the various manually processed features employed in these studies.

While offering valuable insights, manually preprocessed features require domain ex-

doi:10.6342/NTU202403142

Table 2.1: Manually processed Features

Feature Category	**Specific Features**
User Profile	upload frequency, upload time, # uploads
Visual Features	image size, orientation, scene type, # faces, dominant colors
Textual Features	TF-IDF (keywords), sentiment, NER, caption length, # tags

pertise and careful selection to avoid introducing bias. To address these limitations, some features have been replaced with more comprehensive ones extracted from pre-trained models. These pre-trained models have been exposed to vast amounts of data during their training process, learning to recognize subtle patterns and relationships that might not be readily apparent through manual preprocessing. These hidden insights provide a deeper understanding of the underlying data structure, resulting in more accurate predictions and improved performance in SMPP tasks.

2.1.2 Pretrained Models

In recent years, researchers have leveraged various pre-trained deep learning models to extract subtle features from multi-modal inputs like texts and images. For instance, Ding et al. (2019) and Xu et al. (2020) utilized a pre-trained ResNet (K. He et al., 2016) backbone for visual features and Word2Vec (Mikolov et al., 2013) for textual features. Meanwhile, J. Wu et al. (2022) employed a BERT (Bidirectional Encoder Representations from Transformers; (Devlin et al., 2019)) for text features and CLIP (Contrastive Language-Image Pre-training; (Radford et al., 2021)) for joint text-image features. These approaches have been effective in capturing subtle patterns that are difficult to achieve with manual feature engineering. Figure 2.1 illustrates a multimodal post encoder proposed by Kim et al. (2020), which effectively summarized the feature extraction process for social media posts.

doi:10.6342/NTU202403142

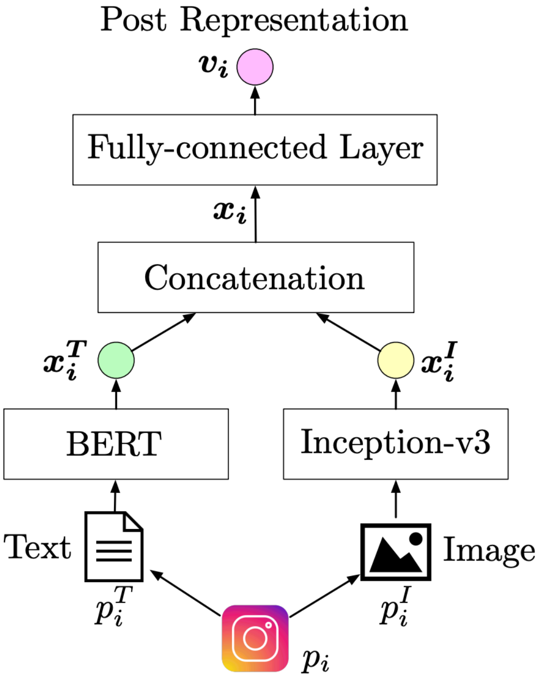


Figure 2.1: A typical feature extraction framework of social media posts (Kim et al., 2020).

The multimodal feature extraction framework is a more comprehensive approach that combines information from different modalities, such as text and images, to create a richer representation of social media posts. This approach recognizes that popularity is often influenced by a combination of factors, including the content's textual meaning, visual appeal, and the context in which it is shared. By integrating multimodal features, SMPP models can achieve better accuracy and robustness in their predictions, leading to more effective content strategies and targeted marketing campaigns.

2.2 Overcome the Imbalance Regression

Despite significant progress in Social Media Popularity Prediction (SMPP), a critical challenge remains largely unaddressed: the inherent imbalance within social media data. This is evident in datasets like Social Media Prediction Dataset (SMPD; (B. Wu et al., 2019)), where popularity distributions often exhibit a long-tail pattern, with the majority

doi:10.6342/NTU202403142

of posts having mid-level popularity and fewer samples as popularity increases (Figure 2.2). This imbalance poses a challenge for traditional models, hindering their ability to generalize across the entire popularity spectrum.

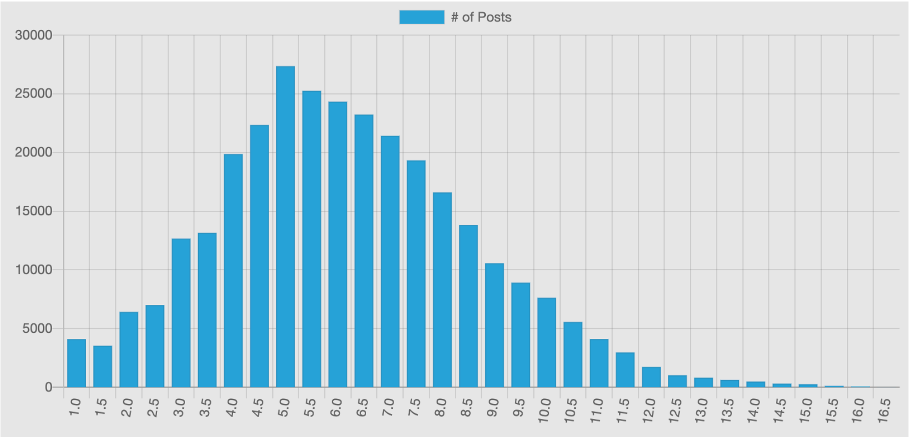


Figure 2.2: The data distribution of SMPD, where the x-axis denotes the log normalized popularity score (range from 1.0 to 16.5), and the y-axis denotes the number of posts. (B. Wu et al., 2019)

Previous attempts to address data imbalance have often relied on techniques like re-sampling or re-weighting, which primarily focus on categorical targets and are not directly applicable to regression tasks. To overcome this limitation, Y. Yang et al. (2021) introduced the concept of Deep Imbalanced Regression (DIR), and proposed a Feature Distribution Smoothing (FDS) approach that emphasized the significance of aligning feature and label spaces to enhance the generalization of under-represented samples. This innovative approach has ignited a surge of research in DIR, with recent methods incorporating contrastive learning to further enhance robustness for under-represented labels. This promising direction holds significant potential to address the challenges of imbalanced data in SMPP and various other real-world applications.

doi:10.6342/NTU202403142

2.2.1 Data Imbalance

Data imbalance is a common challenge in machine learning. It occurs when certain target values (labels) are significantly underrepresented in the training data. This leads to biased models in classification tasks, as they tend to favor the majority classes due to their over-representation. This bias negatively impacts overall performance metrics and can lead to inaccurate predictions for underrepresented classes. However, data imbalance also affects regression tasks, where the target variable is continuous, and specific ranges of values may be less frequent. This imbalance can cause features associated with minority label ranges to converge towards those of the majority due to insufficient training data, leading to increased prediction errors for these underrepresented ranges.

Traditional re-sampling methods primarily target classification tasks. However, some adaptations have been tailored for imbalanced regression. Random undersampling (Torgo et al., 2013, 2015) grouped lables in bins and randomly removes samples from majority bins to balance with minority bins. SMOTER (Torgo et al., 2013), a regression adaptation of SMOTE (Chawla et al., 2002), combines undersampling with synthetic minority sample generation to balance the data distribution. SMOGN (Branco et al., 2017) further improves SMOTER by adding gaussian noise to increase sample diversity. Despite the simplicity of re-sampling, these methods have limitations in imbalanced regression. Firstly, they do not fully account for the density of neighboring target values, which is crucial in determining a data point's representation. As illustrated in Figure 2.3, a low-frequency point within a dense neighborhood may be adequately represented, while one in a sparse neighborhood remains underrepresented. Secondly, linear interpolation techniques like SMOTE may be ineffective and can lead to a degradation in performance when

doi:10.6342/NTU202403142

generating synthetic samples for high-dimensional data, as is often the case with modern large pre-trained models.

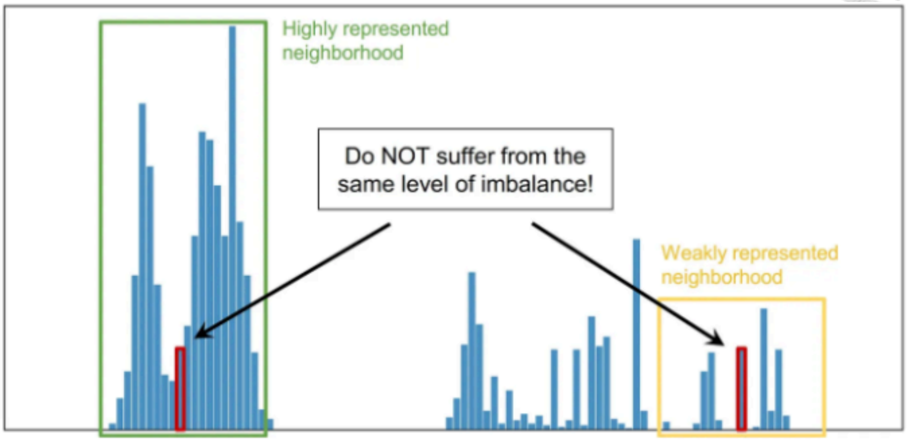


Figure 2.3: The importance of neighborhood density in regression (Y. Yang, 2021). Although the two red label bins contain the same amount of data, the left bin exhibits less imbalance due to the denser distribution of data points in its neighboring ranges, which provides a richer representation for the surrounding popularity levels.

Re-weighting techniques offer an alternative solution to address class imbalances by adjusting the contribution of each label to the overall loss. Dong et al. (2018) proposed a method that focuses on rectifying minority classes and mining hard samples from the majority. Cui et al. (2019) introduced a re-weighting scheme based on the effective number of samples per class to achieve a class-balanced loss. Cao et al. (2019) proposed a label-distribution-aware margin (LDAM) loss to minimize a margin-based generalization bound, which improved generalization on less frequent classes. However, the lack of distinct class boundaries in regression tasks makes the direct application of these re-weighting methods challenging and unsuitable for regression scenarios.

These limitations highlighted the need for innovative solutions to learn robust representations in imbalanced regression tasks, moving beyond traditional re-sampling or re-weighting techniques.

doi:10.6342/NTU202403142

2.2.2 Deep Imbalance Regression

Deep Imbalanced Regression (DIR), a concept introduced by Y. Yang et al. (2021), addresses the inherent imbalance often found in real-world regression tasks. Numerous applications, such as predicting age from facial images or predicting health metrics from patient populations, involve continuous target variables with skewed distributions. This imbalance is also prevalent in fields like economics, crisis management, and meteorology, where target variables often follow a normal distribution with rare occurrences at the extremes.

The challenge of imbalanced data is more intense in deep learning models due to their tendency to produce overconfident predictions, further amplifying the impact of skewed distributions. To tackle this issue, Y. Yang et al. (2021) defined DIR as the task of learning from such imbalanced data with regression tasks. The goal of DIR is to learn robust representations from imbalanced and skewed data, ensuring that these representations generalize effectively across the entire spectrum of target values.

Y. Yang et al. (2021) conducted an empirical experiment and found that the error distribution of categorical datasets differs from those with continuous label spaces. They utilized CIFAR100, a 100-class classification dataset, and the IMDB-WIKI dataset, a large-scale image dataset for age estimation with continuous labels for experiments. To simulate data imbalance, they sub-sampled both datasets and maintain the same label density distribution for the two sampled datasets. Figure 2.4 illustrates the distribution of the two sub-sampled datasets.

By plotting the test error distributions (Fig 2.5), Y. Yang et al. (2021) observed that the error distribution for CIFAR-100 correlates with the label density distribution. This is

doi:10.6342/NTU202403142

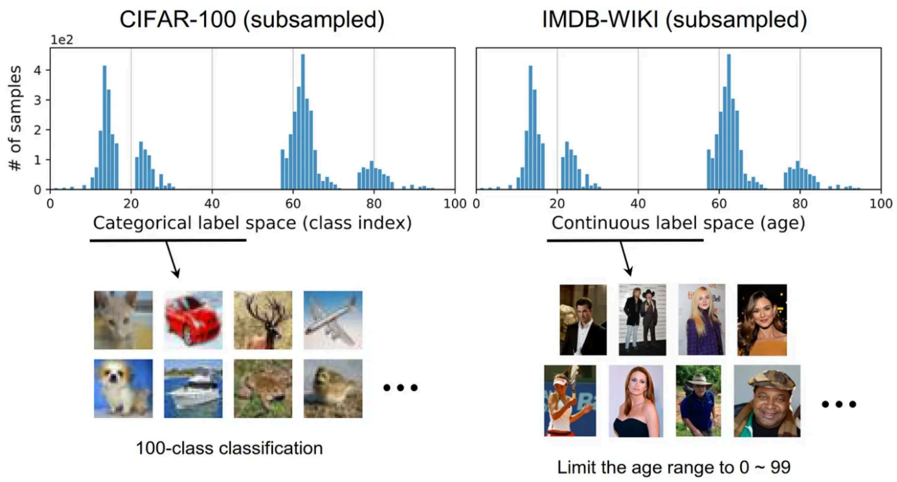


Figure 2.4: The sub-sampled datasets from both CIFAR-100 (classification) and IMDB-WIKI (regression) were curated to exhibit the same imbalance distribution. (Y. Yang, 2021)

expected since majority classes with more samples are learned better than minority classes. Interestingly, the error distribution for IMDB-WIKI differs significantly even when the label density distribution is the same as CIFAR-100. The IMDB-WIKI error distribution is much smoother and does not correlate well with the label density distribution (-0.47).

This phenomenon indicates that for continuous labels, the empirical label density fails to accurately capture the imbalance as perceived by the neural network model. In other words, the empirical label distribution does not represent the true underlying distribution of the regression dataset due to the inherent dependencies between data samples with similar label values.

To address this issue, They adopted the Label Distribution Smoothing (LDS) on the training data (Fig 2.6). Given a continuous empirical label density distribution, LDS convolves a symmetric kernel k with the empirical density distribution to extract a kernel-smoothed version that accounts for the overlap in information of data samples of nearby labels. The resulting effective label density distribution turns out to correlate well with the error distribution now, with a Pearson correlation of −0.83. This demonstrates that LDS

doi:10.6342/NTU202403142

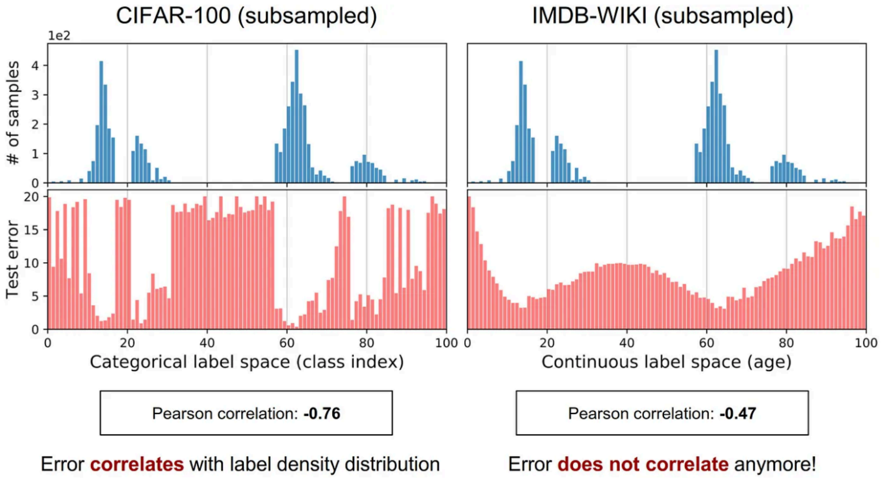


Figure 2.5: Despite both the classification and regression datasets being sub-sampled to exhibit the same level of imbalance, the distribution of test errors differs significantly between the two. (Y. Yang, 2021).

captures the real imbalance that affects regression problems.

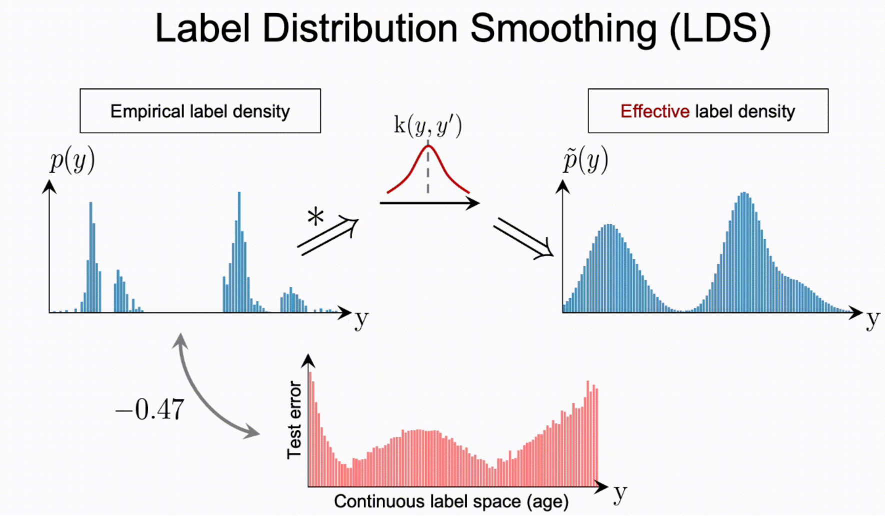


Figure 2.6: The test error correlation improved after LDS. (Y. Yang, 2021)

Motivated by the intuition that continuity in the target space should create a corresponding continuity in the feature space, Y. Yang et al. (2021) suggested that if the data is balanced, one expects the feature statistics corresponding to nearby targets to be close to each other. They first visualized the similarity between feature statistics (Fig 2.7) of an age prediction task. First, they selected age 30 as an anchor bin, denoted as b_0, and obtained the feature statistics (mean and variance) for this bin. Additionally, they calculate statistics for other bins and the cosine similarity between the feature statistics of b_0 and all other bins. The results are summarized in Figure 2.7. The figure also indicates regions with different data densities using the colors purple, yellow, and pink, where purple represents many-shot regions, yellow represents medium-shot regions and pink represents few-shot regions.

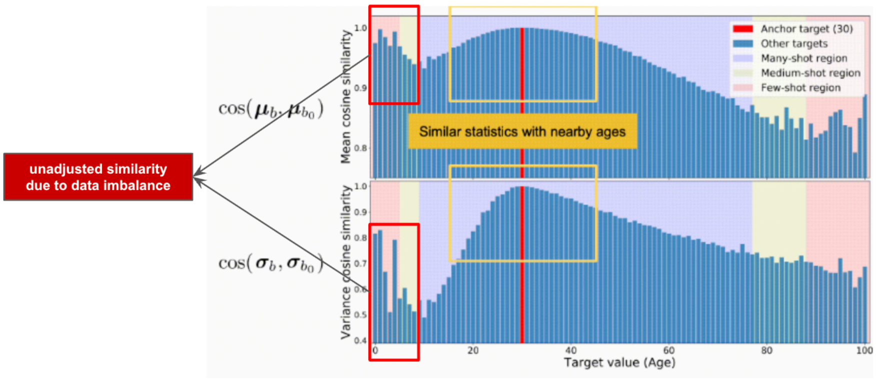


Figure 2.7: The cosine similarity of mean and variance w.r.t the anchor (age 30), where the x-axis denotes the target value (age) and the y-axis denotes the cosine similarity (Y. Yang, 2021).

Interestingly, Y. Yang et al. (2021) found that feature statistics around the anchor bin are highly similar to feature statistics at the anchor bin. Specifically, the cosine similarity of both the feature mean and feature variance for all bins between age 25 and 35 are within a few percent from their values at age 30 (the anchor age). They noticed that the anchor age at bin 30 falls in the many-shot region. Thus, the figure confirms the intuition that when there is enough data for continuous targets, the feature statistics of nearby bins are similar.

Inspired by these observations, Y. Yang et al. (2021) proposed feature distribution smoothing (FDS), which performs distribution smoothing on the feature space, essentially transferring feature statistics between nearby target bins. This procedure aims to calibrate potentially biased estimates of the feature distribution, particularly for underrepresented targets. They take a closer look at FDS and analyze its influence on network training.

doi:10.6342/NTU202403142

Similar to their previous setup, they plotted the feature statistics similarity for anchor age 0. As shown in Figure 2.8, due to very few samples in the target bin 0, the feature statistics can be largely biased, where age 0 shares high similarity with the region between ages 40-80. In contrast, with FDS, the statistics are better calibrated, resulting in high similarity only within its neighborhood and gradually decreasing similarity scores as the target value increases.

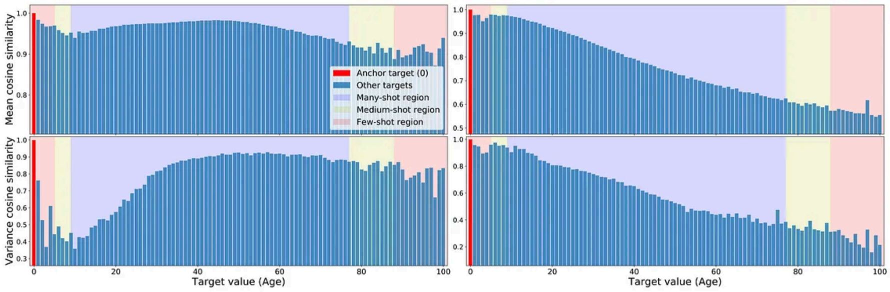


Figure 2.8: The cosine similarity of mean and variance before and after FDS. (Y. Yang, 2021)

To support the practical evaluation of imbalanced regression methods and facilitate future research,Y. Yang et al. (2021) curated five DIR benchmarks spanning computer vision, natural language processing, and healthcare. These benchmarks range from single-value prediction tasks like age, text similarity score, and health condition score to dense-value prediction tasks like depth.

- **IMDB-WIKI-DIR (vision, age)**: This dataset contains face images and corresponding ages for age estimation. The validation and test sets are balanced.

- **AgeDB-DIR (vision, age)**: Similar to IMDB-WIKI-DIR, AgeDB-DIR is also for age estimation from a single input image. However, the label distribution of AgeDB-DIR differs from IMDB-WIKI-DIR, despite having the same task.

- **NYUD2-DIR (vision, depth)**: Although it involves single target value prediction, NYUD2-DIR is based on the NYU2 dataset for depth estimation, a dense-value prediction task.

- **STS-B-DIR (NLP, text similarity score)**: This benchmark focuses on inferring the Semantic Textual Similarity score between two input sentences, a continuous value ranging from 0 to 5 with an imbalanced distribution.

- **SHHS-DIR (Healthcare, health condition score)**: This benchmark involves inferring a general health score (0-100) based on high-dimensional Polysomnography signals from patients' sleep data. The score distribution is also imbalanced.

The experiment results on the IMDB-WIKI-DIR dataset is illustrated in figure 2.9. Different methods are grouped into four sections based on their fundamental strategies. Within each group, Label Distribution Smoothing (LDS), Feature Distribution Smoothing (FDS), and their combination (LDS+FDS) are applied to the baseline method. The absolute improvements of LDS+FDS over the Vanilla model are reported. The table demonstrates that LDS and FDS achieve notable performance improvements regardless of the base training technique. Notably, for the few-shot region, they obtain over 40% relative improvements compared to the baseline model.

Y. Yang et al. (2021)'s work has significantly advanced the field of deep imbalanced regression (DIR). Their proposed feature distribution smoothing technique demonstrates promising potential in mitigating the negative effects of imbalanced data in regression tasks. Furthermore, their exploration of aligning label and feature spaces suggests a valuable avenue for creating more robust representations in regression models. Overall, their contributions have not only filled a critical gap in the existing literature but have also opened up new directions for future research in the area imbalance regression.

doi:10.6342/NTU202403142

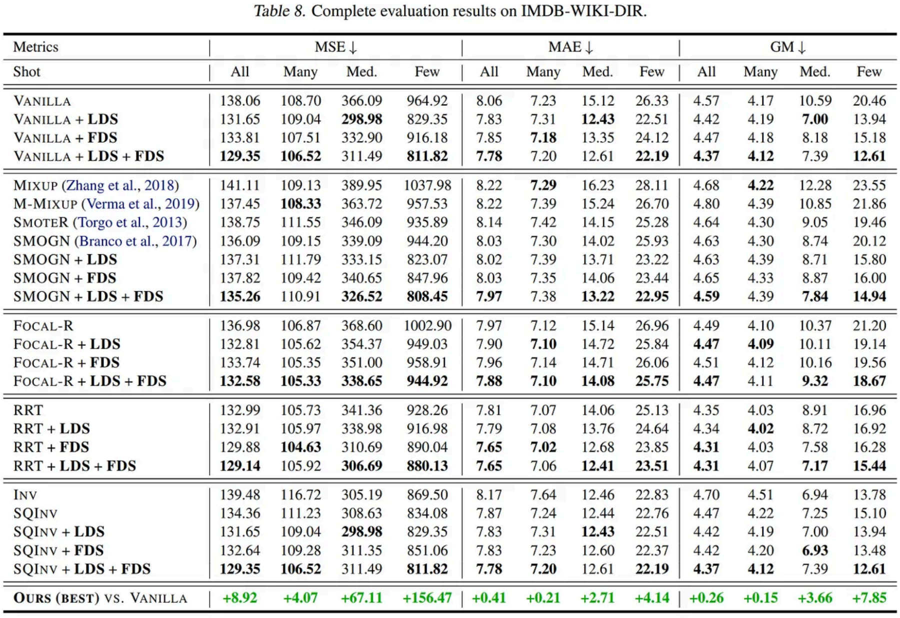


Table 8. Complete evaluation results on IMDB-WIKI-DIR.

Metrics	MSE ↓				MAE ↓				GM ↓			
Shot	All	Many	Med.	Few	All	Many	Med.	Few	All	Many	Med.	Few
VANILLA	138.06	108.70	366.09	964.92	8.06	7.23	15.12	26.33	4.57	4.17	10.59	20.46
VANILLA + **LDS**	131.65	109.04	**298.98**	829.35	7.83	7.31	**12.43**	22.51	4.42	4.19	**7.00**	13.94
VANILLA + **FDS**	133.81	107.51	332.90	916.18	7.85	**7.18**	13.35	24.12	4.47	4.18	8.18	15.18
VANILLA + **LDS** + **FDS**	**129.35**	**106.52**	311.49	**811.82**	**7.78**	7.20	12.61	**22.19**	**4.37**	**4.12**	7.39	**12.61**
MIXUP (Zhang et al., 2018)	141.11	109.13	389.95	1037.98	8.22	**7.29**	16.23	28.11	4.68	**4.22**	12.28	23.55
M-MIXUP (Verma et al., 2019)	137.45	**108.33**	363.72	957.53	8.22	7.39	15.24	26.70	4.80	4.39	10.85	21.86
SMOTER (Torgo et al., 2013)	138.75	111.55	346.09	935.89	8.14	7.42	14.15	25.28	4.64	4.30	9.05	19.46
SMOGN (Branco et al., 2017)	136.09	109.15	339.09	944.20	8.03	7.30	14.02	25.93	4.63	4.30	8.74	20.12
SMOGN + **LDS**	137.31	111.79	333.15	823.07	8.02	7.39	13.71	23.22	4.63	4.39	8.71	15.80
SMOGN + **FDS**	137.82	109.42	340.65	847.96	8.03	7.35	14.06	23.44	4.65	4.33	8.87	16.00
SMOGN + **LDS** + **FDS**	**135.26**	110.91	**326.52**	**808.45**	**7.97**	7.38	**13.22**	**22.95**	**4.59**	4.39	**7.84**	**14.94**
FOCAL-R	136.98	106.87	368.60	1002.90	7.97	7.12	15.14	26.96	4.49	4.10	10.37	21.20
FOCAL-R + **LDS**	132.81	105.62	354.37	949.03	7.90	**7.10**	14.72	25.84	**4.47**	**4.09**	10.11	19.14
FOCAL-R + **FDS**	133.74	105.35	351.00	958.91	7.96	7.14	14.71	26.06	4.51	4.12	10.16	19.56
FOCAL-R + **LDS** + **FDS**	**132.58**	**105.33**	**338.65**	**944.92**	**7.88**	**7.10**	**14.08**	**25.75**	**4.47**	4.11	**9.32**	**18.67**
RRT	132.99	105.73	341.36	928.26	7.81	7.07	14.06	25.13	4.35	4.03	8.91	16.96
RRT + **LDS**	132.91	105.97	338.98	916.98	7.79	7.08	13.76	24.64	4.34	**4.02**	8.72	16.92
RRT + **FDS**	129.88	**104.63**	310.69	890.04	**7.65**	**7.02**	12.68	23.85	**4.31**	4.03	7.58	16.28
RRT + **LDS** + **FDS**	**129.14**	105.92	**306.69**	**880.13**	**7.65**	7.06	**12.41**	**23.51**	**4.31**	4.07	**7.17**	**15.44**
INV	139.48	116.72	305.19	869.50	8.17	7.64	12.46	22.83	4.70	4.51	6.94	13.78
SQINV	134.36	111.23	308.63	834.08	7.87	7.24	12.44	22.76	4.47	4.22	7.25	15.10
SQINV + **LDS**	131.65	109.04	**298.98**	829.35	7.83	7.31	**12.43**	22.51	4.42	4.19	7.00	13.94
SQINV + **FDS**	132.64	109.28	311.35	851.06	7.83	7.23	12.60	22.37	4.42	4.20	**6.93**	13.48
SQINV + **LDS** + **FDS**	**129.35**	**106.52**	311.49	**811.82**	**7.78**	**7.20**	12.61	**22.19**	**4.37**	**4.12**	7.39	**12.61**
OURS (BEST) VS. VANILLA	**+8.92**	**+4.07**	**+67.11**	**+156.47**	**+0.41**	**+0.21**	**+2.71**	**+4.14**	**+0.26**	**+0.15**	**+3.66**	**+7.85**

Figure 2.9: The experiment results on the IMDB-WIKI-DIR dataset. (Y. Yang, 2021)

2.2.3 Contrastive Regression

Y. Yang et al. (2021) revealed that feature-label alignment is key in addressing Deep Imbalanced Regression (DIR) problems. Building upon this insight, recent studies have endeavored to achieve this alignment through tailored loss functions. Notably, Gong et al. (2022) introduced RankSim, incorporating a ranking loss as a regularizer to capture both local and distant relationships effectively. By aligning the sorted lists of neighbors in label and feature spaces, RankSim tackles the complexities of DIR. Experimental results on DIR benchmarks proposed by Y. Yang et al. (2021) showcase RankSim's superior performance than FDS.

Contrastive learning, a pairwise representation learning technique, excels at differentiating between semantically similar and dissimilar samples. It offers several distinct advantages:

- **Semantic Clustering**: Contrastive learning actively encourages the model to group together samples with similar labels while distancing those with dissimilar labels. This process results in semantically meaningful clusters within the learned representations. (Chen et al., 2020; Gao et al., 2021; Khosla et al., 2020)

- **Capture Invariant Features**: Contrastive learning facilitates the learning of key features shared by samples with the same label. This maximization of mutual information between latent representations ensures that the model captures the essential characteristics that define a particular class. (Bachman et al., 2019)

- **Robustness to Noise**: By prioritizing the relationships between data points over individual labels, contrastive learning enhances model resilience against noisy or erroneous data. This focus reduces the susceptibility to inconsistencies within the training data. (Tian et al., 2020)

Inspired by contrastive learning’ s success in robust representation learning, Zha et al. (2024) proposed Rank-N-Contrast (RNC), which modeled the ranking loss within a contrastive framework to address data imbalance: First, the samples are ranked according to their target distances, and then contrasted against each other based on their relative rankings. Second, given an anchor v_i, the likelihood of any other v_j is modeled to increase exponentially with respect to their similarity in the representation space. The set $S_{i,j} := v_k | k \neq i, d(y_i, y_k) \geq d(y_i, y_j)$ is introduced to denote the set of samples with higher ranks than v_j in terms of label distance w.r.t. v_i, where $d(\cdot,\cdot)$ is the distance measure between two labels (e.g., L_1 distance). The normalized likelihood of v_j given v_i and $S_{i,j}$ is then:

$$P(v_j \mid v_i, S_{i,j}) = \frac{\exp(\text{sim}(v_i, v_j)/\tau)}{\sum_{v_k \in S_{i,j}} \exp(\text{sim}(v_i, v_k)/\tau)},$$

Where $sim(\cdot,\cdot)$ is the similarity measure between two feature embeddings (e.g., negative L_2 norm) and τ denotes the temperature parameter. Maximizing $P(v_j|v_i, S_{i,j})$ effectively increases the probability that v_j outperforms other samples in the set and emerges at the top rank within $S_{i,j}$. As a result, they defined the per-sample RNC loss as the average negative log-likelihood over all other samples in a given batch:

$$\mathcal{L}^{(i)}_{\text{RNC}} = -\frac{1}{N-1}\sum_{j\neq i}\log\frac{\exp(\text{sim}(v_i,v_j)/\tau)}{\sum_{v_k\in S_{i,j}}\exp(\text{sim}(v_i,v_k)/\tau)}$$

The context of positive and negative pairs of RNC loss is illustrated in Fig 2.10. For a batch of data containing posts of popularity $\{1,3,4,8\}$, Two example positive pairs and corresponding negative pair(s) when the anchor is a post with popularity score 3 . When the anchor forms a positive pair with a post with popularity score 4, their label distance is 1, hence the corresponding negative samples are the post with popularity score 1 and the post with popularity score 8, whose label distances to the anchor are larger than 1. When the post with popularity score 1 creates a positive pair with the anchor, the post with popularity score 8 has a larger label distance to the anchor, thus serving as a negative sample.

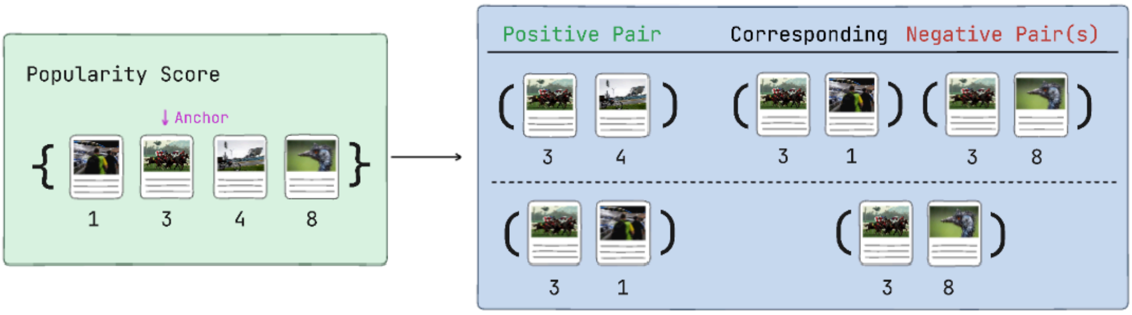


Figure 2.10: Two examples of positive and negative pairs of RNC loss for a batch of data containing posts of popularity $\{1,3,4,8\}$ when the anchor is the post with popularity score 3.

Intuitively, for an anchor sample i, Rank-N-Contrast contrasts it with every other sample j in the batch. This process enforces a higher feature similarity between i and j compared to i and any other sample k in the batch, provided the label distance between i and k is larger than that between i and j. Minimizing RNC loss aligns the orders of feature embeddings with their corresponding orders in the label space relative to anchor i. Extending this concept to all pairs in the batch, the method ensures that when the anchor is contrasted with the sample of closest label distance, their similarity surpasses that of all other samples in the batch. Similarly, when contrasting with the second closest sample, the similarity only needs to be greater than those with ranks of three or higher. By optimizing RNC loss, the model successfully aligns features with their corresponding label distances, maintaining high similarity among neighboring labels and less similarity with distant labels. Rank-N-Contrast has achieved state-of-the-art performance on the DIR benchmark proposed by Y. Yang et al. (2021), demonstrating superior ability in addressing the DIR problem.

Despite its successes, Rank-N-Contrast had a significant limitation: it did not consider varying label distances in negative samples. Its approach disregarded the impact of negative samples further from the anchor in the label space, which should ideally have provided a stronger contrastive signal than closer ones. Figure 2.11 illustrated this issue. The top image, replicating Figure 2.10, presented positive and negative pairs within a batch containing posts with popularity scores {1, 3, 4, 8}. The bottom image showed another batch with scores {1, 3, 4, 15}. In this scenario, for the positive pair {3, 4}, the negative samples became {3, 1} and {3, 15}. Similarly, for the positive pair {3, 1}, the negative sample became {3, 15}. Under Rank-N-Contrast, both negative samples $\{3, 8\}$ and $\{3, 15\}$ contributed equally to the overall Rank-N-Contrast loss, overlooking the impact of the more popular post with score 15.

doi:10.6342/NTU202403142

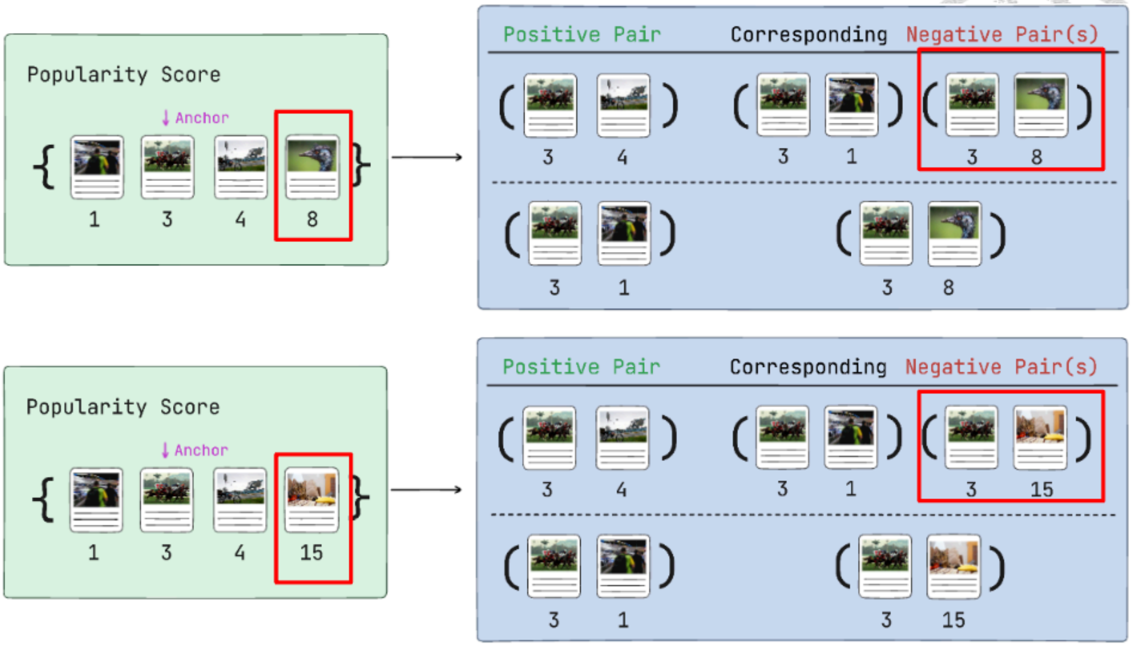


Figure 2.11: RNC loss treats negative pairs {3, 8} and {3, 15} equally in both batches, neglecting the impact of the larger label distance posed by the higher popularity score of 15.

This oversight regarding the varying distances from the respective anchors in the label space could negatively impact the feature representation of rare extreme labels, as their larger influence on the feature space was not adequately captured and reflected in the RNC loss. This challenge arises from the difficulty in sampling a sufficiently diverse batch of data. Theoretically, increasing the batch size could mitigate this issue by incorporating a broader range of samples. However, this solution is often impractical due to the following constraints:

- **Skewed Data**: Lack of extremely popular or extremely unpopular posts makes it hard to achieve true diversity by sampling a wide range of popularity levels.
- **Hardware Limitations**: Increasing batch size for better sample diversity can be computationally expensive and limited by hardware capabilities.

As a result, there is a clear need for practical approaches that can effectively address the limitation of Rank-N-Contrast, especially in situations where extreme skewed data distribution makes it infeasible to obtain a diverse range of samples.

2.3 Summary

We followed the multi-modal feature extraction framework proposed by Kim et al. (2020) as illustrated in Figure 2.1 for its simplicity. This framework included image and text features extracted from post content, supplemented by additional dense features such as user profile and time information available in the training data.

Instead of modifying the model architecture, we focused on refining Rank-N-Contrast (Zha et al., 2024) to overcome its limitation of not distinguishing between negative samples based on their label distances. We introduced a weighting mechanism that incorporated label distance information into the contrastive regression loss. Experimental results demonstrated that our approach fostered a more uniform feature space and significantly improved robustness on extremely rare and even unseen labels.

Chapter 3 Methodolodgy

3.1 Problem Definition

Given a new post v by user u, our objective is to predict its popularity s, defined as the expected number of attentions it would received if published at time t on social media. Popularity can be quantified using various dynamic indicators (e.g., views, likes, clicks) across different social media platforms. In our dataset, the ”view count” serves as a fundamental indicator of post popularity. To mitigate the wide variations in view counts among photos (ranging from zero to millions), a log-normalization function is applied:

$$s = \log_2 \frac{r}{d} + 1 \tag{3.1}$$

where s is the normalized popularity, r is the view count, and d is the number of days since posting.

3.2 Overview of Our Proposed Framework

We leveraged pre-trained visual and textual models as feature encoders to extract multi-modal features. These features were then concatenated with additional dense fea-

doi:10.6342/NTU202403142

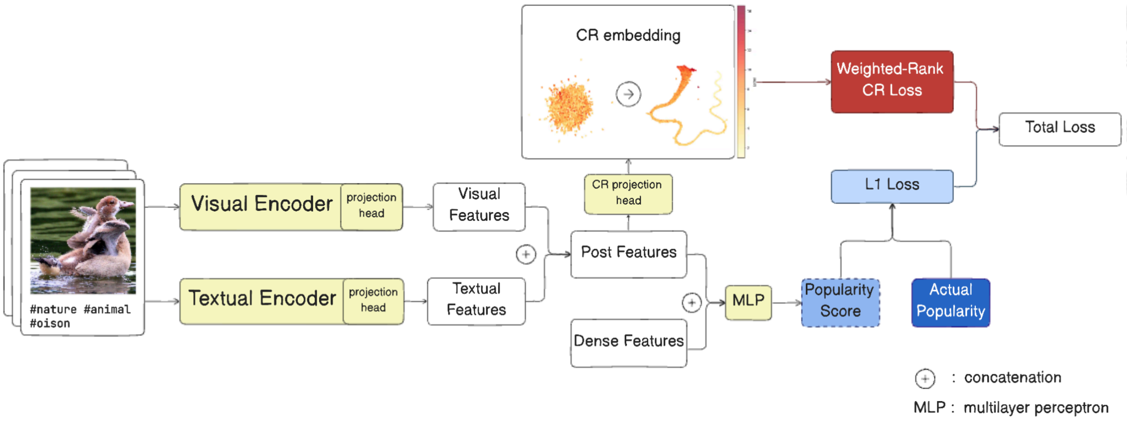


Figure 3.1: Overview of our framework

tures to create a comprehensive input for downstream prediction. The concatenated features were fed into a Multi-Layer Perceptron (MLP) to predict the popularity score. We also incorporated our proposed Weighted-Rank Contrastive Regression loss as a regularizer and calculated the contrastive regression loss alongside the L1 loss, these two losses were combined in a multi-task learning approach, with equal weighting assigned to each loss. This joint optimization process encouraged the feature encoders to learn more robust representations while simultaneously improving the prediction objective during training. Figure 3.1 illustrates an overview of our framework.

3.3 Post Representation Extraction

Following the approach of (Kim et al., 2020), we utilized pre-trained models to extract features from both the visual and textual components of the posts. For the visual features, the image preprocessing involved the following steps: (1) conversion to RGB color space, (2) resizing to a 224x224 pixel resolution, (3) subsequent normalization. After preprocessing, we employed the Vision Transformer (VIT; (Dosovitskiy et al., 2021)) to extract the visual features f_i. As for textual features, we utilized the hashtags within the social media posts, represented as a list of keywords. By concatenating these keywords, we then leveraged the Sentence Transformer (Reimers & Gurevych, 2019) to extract the textual features f_t. Finally, we concatenated f_i and f_t to obtain the comprehensive post features f_p.

3.4 Dense Features

Besides the visual and textual inputs, we also used the following dense features provided by the dataset: *userIsPro*: whether the user belong to pro member. *postCount*: The number of posted photo by the user. *photoFirstDateTaken*: The date of the first photo taken by the user. *postDate*: the publish timestamp of the post.

3.5 Weighted-Rank Contrastive Regression

We proposed Weighted-Rank Contrastive Regression loss that contrasts negative samples based on their relative label distance with respect to anchor, while following the ranking process of Rank-N-Contrast. The per sample Weighted-Rank Contrastive Regression loss can be denoted as:

$$\frac{1}{N-1}\sum_{j=1,j\neq i}^{N} -\log\left(\frac{e_\tau(v_i,v_j)}{\sum_{v_k\in S_{i,j}} w_{ik}\cdot e_\tau(v_i,v_k)}\right) \tag{3.2}$$

$$e_\tau = \text{exp}(\text{sim}(v_i,v_j)/\tau) \tag{3.3}$$

We simplified exp(sim$(v_i, v_j)/\tau$) to e_τ, where sim denotes the cosine similarity, and $/tau$ is a temperature hyperparameter that controls the sensitivity of the relationship between embedding similarity and the contrastive loss.

Following the approach in (Zha et al., 2024), for an anchor vector v_i and another sample v_j in the batch, we define $S_{i,j}$ as the set of samples that are further away in label distance from v_i compared to v_j. In our Weighted-Rank Contrastive Regression loss, we incorporate a weighting mechanism for negative sample pairs. The weight for a negative pair $\{v_i, v_k\}$ is represented as $w_{i,k}$.

To validate the effectiveness of our weighting mechanism, we conducted experiments on a curated dataset derived from the SMP dataset (see section 4.1) characterized by a skewed training data distribution and a balanced, uniform test set. We evaluated various weighting strategies including logarithmic, linear, quadratic, and exponential weighting on the uniform distributed test set. The results, presented in Table 3.1, support our hypothesis that a stronger emphasis on contrastive signals based on label distance leads to improved performance. Notably, the exponential weighting strategy, represented by $(1+\alpha)^d$, where d is the label distance, achieved the best performance. The quadratic weighting strategy, $d^2 + 1$, followed closely behind. However, the linear weighting $(d + 1)$ and logarithmic weighting log$(d + 1) + 1$ did not outperformed the baseline Rank-N-Contrast method. These findings reinforced our hypothesis that prioritizing distant negative samples in the contrastive loss can enhance the effectiveness of contrastive regression.

As a result, we required an exponential weighting on label distance to amplify the contrastive signal from more distant negative samples. Let $w_{i,k}$ denote the weight assigned to the negative sample pair $\{i, k\}$ and d denote the absolute label difference between sample i and k, Then, $w_{i,k}$ is calculated as in Equation 3.4, where α is a hyperparamter that controls the slope of $w_{i,k}$. In our experiment, we chose $\alpha = 0.4$ so that $w_{i,k}$ is bounded

doi:10.6342/NTU202403142

Table 3.1: Performance metrics of different weighting strategies.

	metrics	
Weighting Strategy	MAE	SRC
RNC (baseline)	2.198	0.838
$\log(d+1)+1$	2.715	0.510
$d+1$	2.642	0.579
d^2+1	2.175	0.838
$(1+\alpha)^d$	**2.142**	**0.841**

within the range of our label value.

$$w_{i,k} = (1+\alpha)^d \tag{3.4}$$

$$d = |y_i - y_k| \tag{3.5}$$

Reflecting on Figure 2.11, with Weighted-Rank Contrastive Regression loss, the negative pairs $\{3, 15\}$ and pair $\{3, 8\}$ will now be assigned weights of $(1+\alpha)^{|3-15|} = 1.4^{12}$ and $(1+\alpha)^{|3-8|} = 1.4^5$, respectively. Consequently, the post with the higher popularity score of 15 is encouraged to be further away from the anchor post 3 in the feature space under this weighted scenario. This weighting scheme ensures that negative samples with larger label distances from the anchor have a stronger influence on the contrastive loss, leading to more effective learning of feature representations, especially for rare and extreme labels.

We used a CR projection head to perform contrastive learning on the extracted post features f_p. After feature extraction, f_p were passed through the CR projection head. Here we denoted the output of CR projection head as f_p^{cr}. The Weighted-Rank Contrastive Regression loss was then computed on f_p^{cr}, enforcing the feature encoders to align the feature space with the corresponding label distances. In parallel, f_p was also fed into a Multi-Layer Perceptron (MLP) to generate a predicted popularity score. We calculated the L1 loss between this predicted score and the actual popularity score. Finally, we combined the Weighted-Rank Contrastive Regression loss and the L1 loss in a multi-task learning framework. Both losses were given equal weight, without emphasizing one over the other. This approach ensured that the model learns robust feature representations while simultaneously optimizing its predictive performance.

doi:10.6342/NTU202403142

Chapter 4 Empirical Experiments

4.1 Dataset

We utilized the Social Media Prediction Dataset (SMPD) proposed by (B. Wu et al., 2019), which was collected from Flickr, a major photo-sharing platform. SMPD comprises 486K social multimedia posts from 70K users, and incorporates diverse social media information such as anonymized photo-sharing records, user profiles, web images, text, timestamps, location data, categories and customize tags provided by users. Table 4.1 provides a detailed overview of the dataset statistics. For post feature extraction in our experiments, we utilized the post images and the associated customized tags.

Table 4.1: Dataset statistics for SMPD.

Dataset	#Post	#User	#Categories	Temporal Range (Months)	Avg. Title Length	#Customize Tags
SMPD	486k	70k	756	16	29	250k

4.2 Evaluation Metrics

We combined the Spearman Ranking Correlation (SRC) and Mean Absolute Error (MAE) to assess model performance. SRC quantifies the ordinal association between predicted and actual popularity rankings, while MAE measures the average prediction error.

doi:10.6342/NTU202403142

SRC is calculated as follows:

$$\mathrm{SRC} = \frac{1}{k-1}\sum_{i=1}^{k}\left(\frac{P_i-\bar{P}}{\sigma_P}\right)\left(\frac{\hat{P}_i-\tilde{P}}{\sigma_{\hat{P}}}\right) \tag{4.1}$$

where k is the number of samples, P_i is the actual popularity, $\hat{P}i$ is the predicted popularity,$\bar{P}$ and σ_P are the mean and standard deviation of actual popularity, and $\tilde{P}$ and $\sigma\hat{P}$ are the mean and standard deviation of predicted popularity.

MAE is calculated as:

$$\mathrm{MAE} = \frac{1}{k}\sum_{i=1}^{n}\left|\hat{P}_i - P_i\right| \tag{4.2}$$

The goal of SMPP is to enhance both ranking accuracy and prediction accuracy by minimizing the MAE and maximizing the SRC.

4.3 Hyperparameters and Experimental Settings

4.3.1 Hyperparameters

Table 4.2 outlines the training configuration including hardware specifications and hyperparameters. We fixed these settings in the main experiments discussed in Section 4.4. Specifically, we chose the largest batch size we can afford under the hardware limitation, and τ representing the temperature which controls the sensitivity of the feature similarity during contrastive learning.

Table 4.2: Training configuration.

hardware	RTX 4080
number of epochs	5
learning rate	3e-4
random seed	3407
batch size	128
τ	0.05

4.3.2 Experimental Settings

The detailed experimental settings of our framework is as below:

- **Backbone Model**: We used a vit checkpoint ”WinKawaks/vit-small-patch16-224” on huggingface for visual encoder, and a sentence-transformers checkpoint: “sentence-transformers/paraphrase-multilingual-MiniLM-L12-v2” for textual encoder.

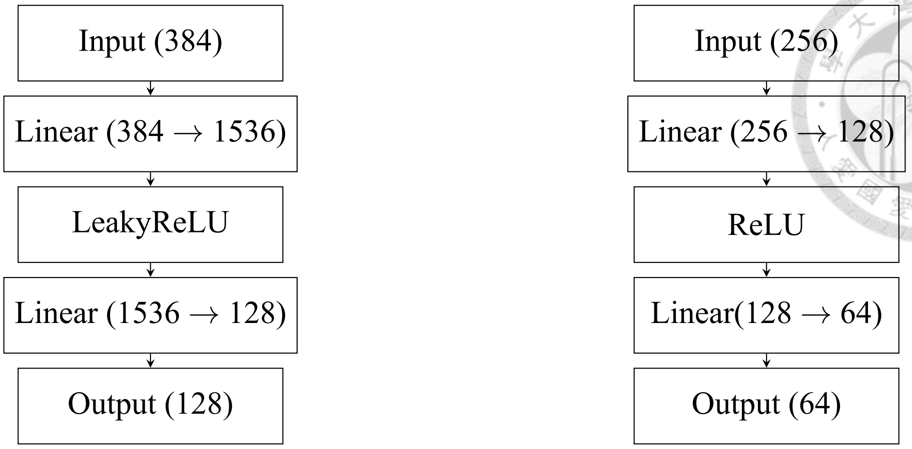
- **Visual and Textual Projection Head**: Both the visual and textual projection heads adhere to the architecture outlined in Figure 4.1a. The input feature tensors, initially of dimension 384, are first expanded to 1536 dimensions and then subjected to a non-linear transformation using LeakyReLU activation. Finally a linear layer project the output tensor to 128 dimensions.

- **CR Projection Head**: As illustrated in Figure 4.1b, the input size of 256 represents the concatenated visual and textual features. We then apply a ReLU non-linear transformation, and finally a linear layer reduce the output tensor to 64 dimensions.

4.4 Evaluation Results

To evaluate our proposed framework, we utilized the test API provided by the SMP Challenge (B. Wu et al., 2019). This API allows us to upload our prediction results and ob-

doi:10.6342/NTU202403142



(a) Textual and visual projection heads.

(b) The CR projection head.

tain the corresponding performance metrics through an online interface. Our experiments included three different modalities: text only, image only, and multi-modal inputs. The evaluation results, presented in Table 4.3, showcase the superior performance of Weighted-Rank Contrastive Regression compared to both the vanilla approach (direct fitting on L1 loss) and Rank-N-Contrast. Notably, Weighted-Rank Contrastive Regression (Weighted-Rank CR) outperformed the other methods in terms of both MAE and SRC across all three modalities.

Table 4.3: Performance metrics for different training objectives and input types.

Input	Vanilla (L1)		Rank-N-Contrast		Weighted-Rank CR	
	MAE	SRC	MAE↓	SRC↑	MAE↓	SRC↑
Tags	2.040	0.468	1.995 (+0.045)	0.483 (+0.015)	**1.925** (+0.115)	**0.499** (+0.031)
Image	2.262	0.301	2.214 (+0.048)	0.303 (+0.002)	**2.183** (+0.079)	**0.310** (+0.009)
Tags + Image	1.955	0.473	2.001 (-0.045)	0.501 (+0.028)	**1.901** (+0.054)	**0.504** (+0.031)

4.5 Additional Experiments

4.5.1 Curated Dataset

We curated two datasets with more imbalance distribution to test the robustness of Weighted-Rank CR. First, we sampled a subset of SMPD training dataset to create a more skewed distribution, with only few data points at both ends. Figure 4.2 illustrates the distribution of this sampled dataset.

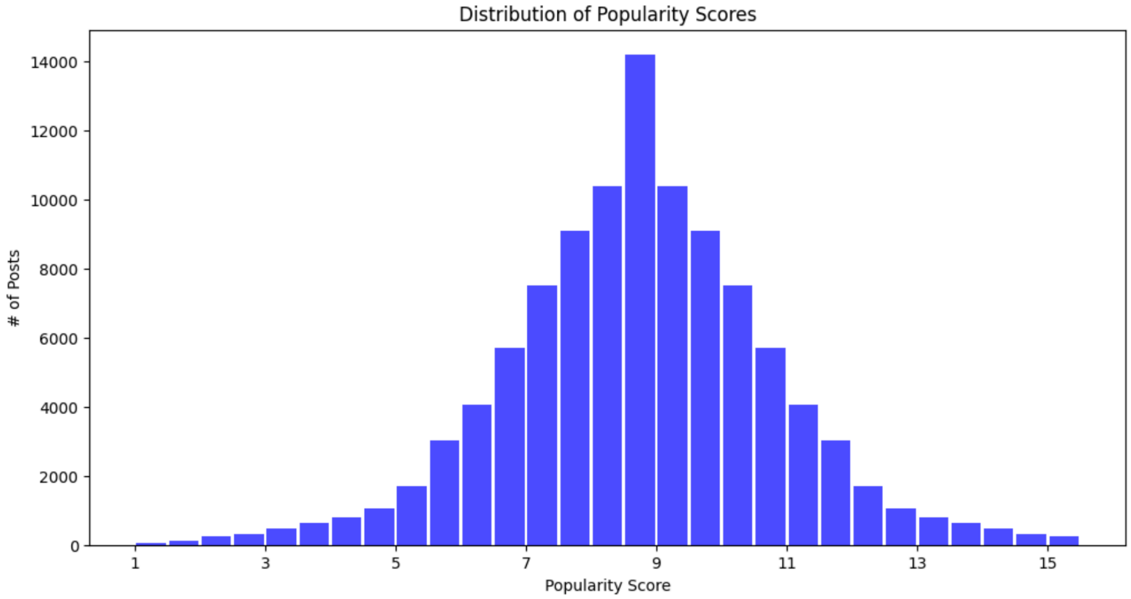
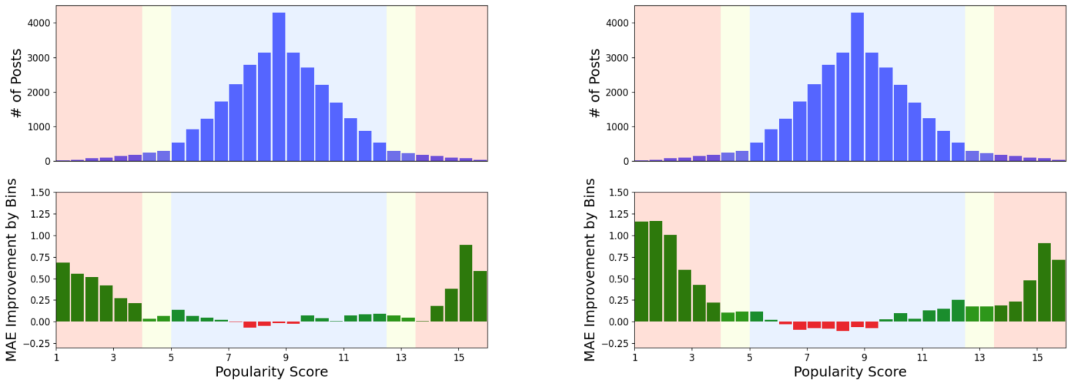


Figure 4.2: The distribution of the sampled dataset, with very few data points on both ends.

We visualized the MAE improvement across different label bins in Figure 4.3. The x-axis represents the label ranges, with the top portion of the figure depicting the data distribution (y-axis showing the number of posts), and the bottom portion displaying the MAE improvement (y-axis indicating the MAE difference). Positive values (in green) signify a lower MAE for that label bin, while negative values (in red) signify a higher MAE. The results demonstrate that contrastive regression substantially reduces the MAE for rarely seen data points, particularly at both extremes of the distribution. Furthermore, we vi-

doi:10.6342/NTU202403142

sualized the MAE improvement of Weighted-Rank CR over Rank-N-Contrast in Figure 4.4. The results demonstrated the superiority of Weighted-Rank Contrastive Regression compared to Rank-N-Contrast on the skewed-sampled dataset.



(a) The MAE improvement of Rank-N-Contrast over the Vanilla approach.

(b) The MAE improvement of Weighted-Rank-CR over the Vanilla approach.

Figure 4.3: The MAE improvement of both Rank-N-Contrast and Weighted-Rank-CR compared to the Vanilla approach. Positive values (in green) signify a lower MAE on the label bin, and negative values (in red) signify a higher MAE on the label bin.

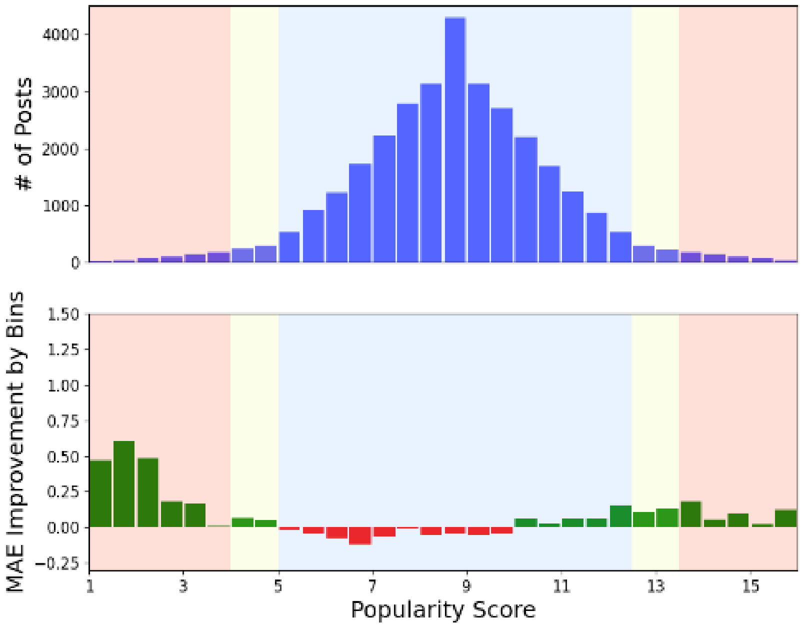
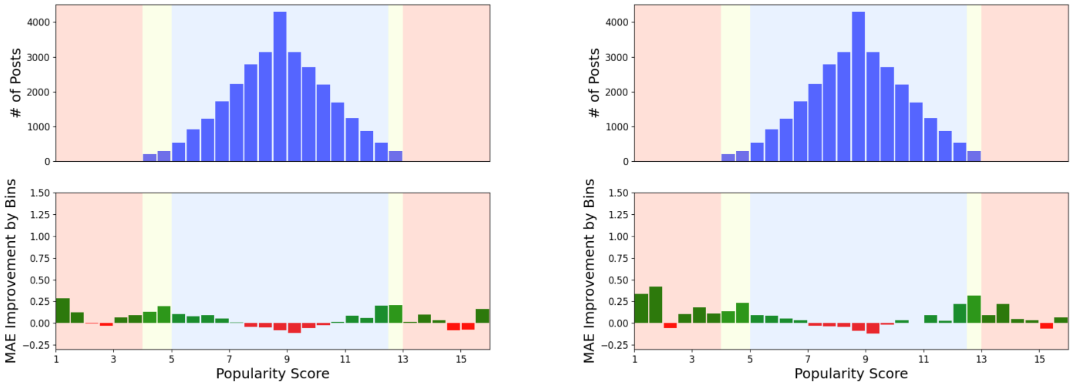


Figure 4.4: The MAE improvement of Weighted-Rank-CR over Rank-N-Contrast.

We also curated another more imbalanced dataset by removing data points with popularity scores below 4.0 and above 13.0. The MAE improvement across different label bins

doi:10.6342/NTU202403142

for this dataset is illustrated in Figure 4.5. Additionally, Figure 4.6 visualizes the MAE improvement of Weighted-Rank CR compared to Rank-N-Contrast on this more imbalanced dataset. The results consistently demonstrate the superiority of Weighted-Rank CR over Rank-N-Contrast even in this more challenging scenario.



(a) The MAE improvement of Rank-N-Contrast over the Vanilla approach.

(b) The MAE improvement of Weighted-Rank-CR over the Vanilla approach.

Figure 4.5: The MAE improvement of both Rank-N-Contrast and Weighted-Rank-CR compared to the Vanilla approach on a dataset that data points at both extremes are removed.

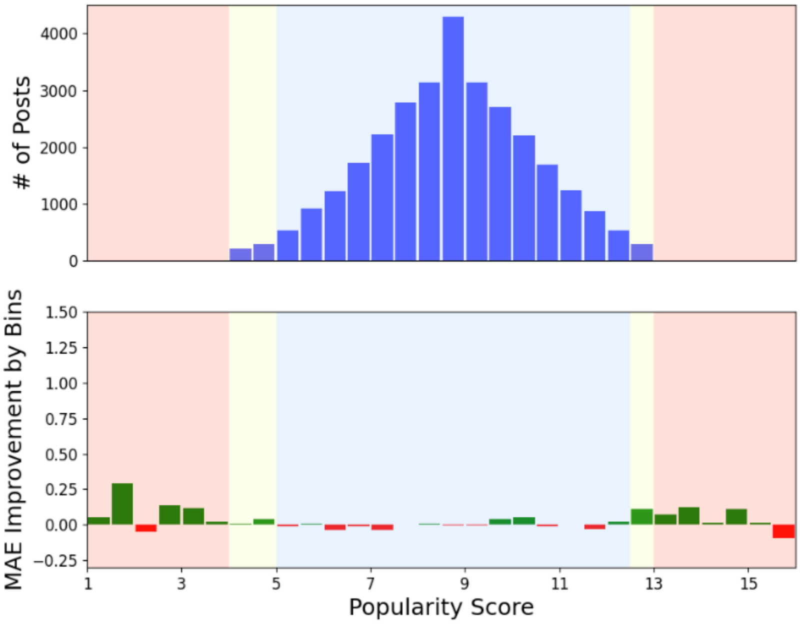


Figure 4.6: The MAE improvement of Weighted-Rank-CR over Rank-N-Contrast on a dataset that data points at both extremes are removed.

Chapter 5 Conclusion

5.1 Conclusion

In this thesis, we delved into the challenges of imbalanced regression in social media popularity prediction, highlighting the limitations of existing contrastive learning methods like Rank-N-Contrast. We proposed Weighted-Rank Contrastive Regression loss, a contrastive learning loss that incorporates label distance information into the Rank-N-contrast loss function, thereby enhancing the model's ability to learn discriminate representations for rare and extreme labels. We also proposed a simple end-to-end contrastive regression learning framework for multi-modal representation learning that can be extended to more complex architecture.

Our experiments on the Social Media Prediction Dataset (SMPD) demonstrated the effectiveness of Weighted-Rank CR, showcasing significant improvements in both ranking accuracy and prediction accuracy compared to the baseline methods. Notably, our approach proved particularly effective in handling imbalanced datasets, where rare labels are often underrepresented. Through extensive evaluation, we validated the superiority of Weighted-Rank CR across two curated datasets and various weighting strategies. Our findings highlight the importance of incorporating label distance information into contrastive learning for imbalanced regression tasks.

doi:10.6342/NTU202403142

This research contributes to the growing body of work addressing the challenges of imbalanced learning in Social Media Popularity Prediction (SMPP). The proposed Weighted-Rank CR method offers a promising avenue for future research, with potential applications in various domains where data imbalance poses a significant challenge.

5.2 Limitations and Future Works

Our current implementation employs a relatively simple weighting strategy that emphasizes contrastive signals based solely on label distance. However, more sophisticated weighting mechanisms could potentially further improve performance. For instance, incorporating additional factors like neighborhood density or feature similarity into the weighting function might lead to more nuanced and effective learning. There are several directions for future research emerged from this work:

- **Evaluation on Diverse SMPP Datasets**: Evaluating the performance of Weighted-Rank CR on a broader range of SMPP datasets with varying characteristics would provide a more comprehensive understanding of its generalization capabilities.

- **Application to Downstream Tasks**: Testing the learned representations on various downstream tasks, such as influencer classification and influencer recommendation. This could help determine the effectiveness of Weighted-Rank CR in real-world applications beyond popularity prediction.

- **Exploration of Advanced Weighting Strategies**: Investigating more sophisticated weighting mechanisms that incorporate additional information beyond label distance. This could involve incorporating feature similarity, sample density, or other

doi:10.6342/NTU202403142

relevant factors into the weighting function.

By addressing these limitations and pursuing these future directions, we believe that Weighted-Rank CR can be further enhanced and extended to a wider range of imbalanced regression problems.

doi:10.6342/NTU202403142

References

Bachman, P., Hjelm, R. D., & Buchwalter, W. (2019). Learning representations by maximizing mutual information across views. *Advances in neural information processing systems*, *32*.

Branco, P., Torgo, L., & Ribeiro, R. P. (2017). Smogn: A pre-processing approach for imbalanced regression. *First international workshop on learning with imbalanced domains: Theory and applications*, 36–50.

Cao, K., Wei, C., Gaidon, A., Arechiga, N., & Ma, T. (2019). Learning imbalanced datasets with label-distribution-aware margin loss. *Advances in neural information processing systems*, *32*.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). Smote: Synthetic minority over-sampling technique. *Journal of artificial intelligence research*, *16*, 321–357.

Chen, T., Kornblith, S., Norouzi, M., & Hinton, G. (2020). A simple framework for contrastive learning of visual representations. *International conference on machine learning*, 1597–1607.

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

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019, June). BERT: Pre-training of deep bidirectional transformers for language understanding. In J. Burstein, C. Doran, & T. Solorio (Eds.), *Proceedings of the 2019 conference of the north American chapter of the association for computational linguistics: Human language tech-*

nologies, volume 1 (long and short papers) (pp. 4171–4186). Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1423

Ding, K., Wang, R., & Wang, S. (2019). Social media popularity prediction: A multiple feature fusion approach with deep neural networks. *Proceedings of the 27th ACM International Conference on Multimedia*, 2682–2686.

Dong, Q., Gong, S., & Zhu, X. (2018). Imbalanced deep learning by minority class incremental rectification. *IEEE transactions on pattern analysis and machine intelligence*, *41*(6), 1367–1381.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2021). An image is worth 16x16 words: Transformers for image recognition at scale. *International Conference on Learning Representations*. https://openreview.net/forum?id=YicbFdNTTy

Freberg, K., Graham, K., McGaughey, K., & Freberg, L. A. (2010). Who are the social media influencers. *A study of puclic perceptions of personality. Puclic Relations Review, G model Pubrel-861*, *3*.

Gao, T., Yao, X., & Chen, D. (2021, November). SimCSE: Simple contrastive learning of sentence embeddings. In M.-F. Moens, X. Huang, L. Specia, & S. W.-t. Yih (Eds.), *Proceedings of the 2021 conference on empirical methods in natural language processing* (pp. 6894–6910). Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.emnlp-main.552

Gelli, F., Uricchio, T., Bertini, M., Del Bimbo, A., & Chang, S.-F. (2015). Image popularity prediction in social media using sentiment and context features. *Proceedings of the 23rd ACM international conference on Multimedia*, 907–910.

Gong, Y., Mori, G., & Tung, F. (2022). RankSim: Ranking similarity regularization for deep imbalanced regression. *International Conference on Machine Learning (ICML)*.

Haixiang, G., Yijing, L., Shang, J., Mingyun, G., Yuanyue, H., & Bing, G. (2017). Learning from class-imbalanced data: Review of methods and applications. *Expert systems with applications*, *73*, 220–239.

doi:10.6342/NTU202403142

Hayes, N. (2008). *Influencer marketing: Who really influences your customers?*. Taylor & francis.

He, H., & Ma, Y. (2013). Imbalanced learning: Foundations, algorithms, and applications.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Hsu, C.-C., Kang, L.-W., Lee, C.-Y., Lee, J.-Y., Zhang, Z.-X., & Wu, S.-M. (2019). Popularity prediction of social media based on multi-modal feature mining. *Proceedings of the 27th ACM International Conference on Multimedia*, 2687–2691.

Jin, X., Gallagher, A., Cao, L., Luo, J., & Han, J. (2010). The wisdom of social multimedia: Using flickr for prediction and forecast. *Proceedings of the 18th ACM international conference on Multimedia*, 1235–1244.

Keramati, M., Meng, L., & Evans, R. D. (2023). Conr: Contrastive regularizer for deep imbalanced regression. *CoRR*, *abs/2309.06651*. https://doi.org/10.48550/ARXIV.2309.06651

Khosla, P., Teterwak, P., Wang, C., Sarna, A., Tian, Y., Isola, P., Maschinot, A., Liu, C., & Krishnan, D. (2020). Supervised contrastive learning. *Advances in neural information processing systems*, *33*, 18661–18673.

Kim, S., Jiang, J.-Y., Nakada, M., Han, J., & Wang, W. (2020). Multimodal post attentive profiling for influencer marketing. *Proceedings of The Web Conference 2020*, 2878–2884.

Lai, X., Zhang, Y., & Zhang, W. (2020). Hyfea: Winning solution to social media popularity prediction for multimedia grand challenge 2020. *Proceedings of the 28th ACM International Conference on Multimedia*, 4565–4569.

McInnes, L., Healy, J., Saul, N., & Großberger, L. (2018). Umap: Uniform manifold approximation and projection. *Journal of Open Source Software*, *3*(29), 861. https://doi.org/10.21105/joss.00861

McParlane, P. J., Moshfeghi, Y., & Jose, J. M. (2014). " nobody comes here anymore, it's too crowded"; predicting image popularity on flickr. *Proceedings of international conference on multimedia retrieval*, 385–391.

doi:10.6342/NTU202403142

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. In Y. Bengio & Y. LeCun (Eds.), *1st international conference on learning representations, ICLR 2013, scottsdale, arizona, usa, may 2-4, 2013, workshop track proceedings*. http://arxiv.org/abs/1301.3781

Perozzi, B., Al-Rfou, R., & Skiena, S. (2014). Deepwalk: Online learning of social representations. *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, 701–710.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. (2021). Learning transferable visual models from natural language supervision. *International conference on machine learning*, 8748–8763.

Reimers, N., & Gurevych, I. (2019). Sentence-bert: Sentence embeddings using siamese bert-networks. In K. Inui, J. Jiang, V. Ng, & X. Wan (Eds.), *Emnlp/ijcnlp (1)* (pp. 3980–3990). Association for Computational Linguistics. http://dblp.uni-trier.de/db/conf/emnlp/emnlp2019-1.html#ReimersG19

Rolínek, M., Musil, V., Paulus, A., Vlastelica, M., Michaelis, C., & Martius, G. (2020). Optimizing rank-based metrics with blackbox differentiation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7620–7630.

Tian, Y., Sun, C., Poole, B., Krishnan, D., Schmid, C., & Isola, P. (2020). What makes for good views for contrastive learning? *Advances in neural information processing systems*, *33*, 6827–6839.

Torgo, L., Branco, P., Ribeiro, R. P., & Pfahringer, B. (2015). Resampling strategies for regression. *Expert systems*, *32*(3), 465–476.

Torgo, L., Ribeiro, R. P., Pfahringer, B., & Branco, P. (2013). Smote for regression. *Portuguese conference on artificial intelligence*, 378–389.

Wang, Y., Jiang, Y., Li, J., Ni, B., Dai, W., Li, C., Xiong, H., & Li, T. (2022). Contrastive regression for domain adaptation on gaze estimation. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 19376–19385.

doi:10.6342/NTU202403142

Wu, B., Cheng, W.-H., Liu, P., Liu, B., Zeng, Z., & Luo, J. (2019). Smp challenge: An overview of social media prediction challenge 2019. *Proceedings of the 27th ACM International Conference on Multimedia*.

Wu, B., Cheng, W.-H., Zhang, Y., Qiushi, H., Jintao, L., & Mei, T. (2017). Sequential prediction of social media popularity with deep temporal context networks. *International Joint Conference on Artificial Intelligence (IJCAI)*.

Wu, B., Mei, T., Cheng, W.-H., & Zhang, Y. (2016). Unfolding temporal dynamics: Predicting social media popularity using multi-scale temporal decomposition. *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence (AAAI)*.

Wu, J., Zhao, L., Li, D., Xie, C.-W., Sun, S., & Zheng, Y. (2022). Deeply exploit visual and language information for social media popularity prediction. *Proceedings of the 30th ACM International Conference on Multimedia*, 7045–7049.

Xu, K., Lin, Z., Zhao, J., Shi, P., Deng, W., & Wang, H. (2020). Multimodal deep learning for social media popularity prediction with attention mechanism. *Proceedings of the 28th ACM International Conference on Multimedia*, 4580–4584.

Yang, X., Kim, S., & Sun, Y. (2019). How do influencers mention brands in social media? sponsorship prediction of instagram posts. *Proceedings of the 2019 IEEE/ACM international conference on advances in social networks analysis and mining*, 101–104.

Yang, Y., Lv, H., & Chen, N. (2023). A survey on ensemble learning under the era of deep learning. *Artificial Intelligence Review*, *56*(6), 5545–5589.

Yang, Y. (2021). *Strategies and tactics for regression on imbalanced data* [accessed June 22, 2024]. https://towardsdatascience.com/strategies-and-tactics-for-regression-on-imbalanced-data-61eeb0921fca

Yang, Y., Zha, K., Chen, Y., Wang, H., & Katabi, D. (2021). Delving into deep imbalanced regression. *International conference on machine learning*, 11842–11851.

Yen, S., & Lee, Y. (2006). Under-sampling approaches for improving prediction of the minority class in an imbalanced dataset. *Lecture notes in control and information sciences*, *344*, 731.

Zha, K., Cao, P., Son, J., Yang, Y., & Katabi, D. (2024). Rank-n-contrast: Learning continuous representations for regression. *Advances in Neural Information Processing Systems*, *36*.

doi:10.6342/NTU202403142

Appendix A — Qualitative Visualization

A.1 Feature Similarity

We visualized the feature similarities across label bins relative to an anchor point for the vanilla approach (Figure A.1), Rank-N-Contrast (Figure A.2), and our proposed Weighted-Rank CR (Figure A.3). The visualization reveals greater uniformity (a larger utilized feature space) in our Weighted-Rank CR method compared to both the vanilla approach (Figure A.4) and Rank-N-Contrast (A.5).

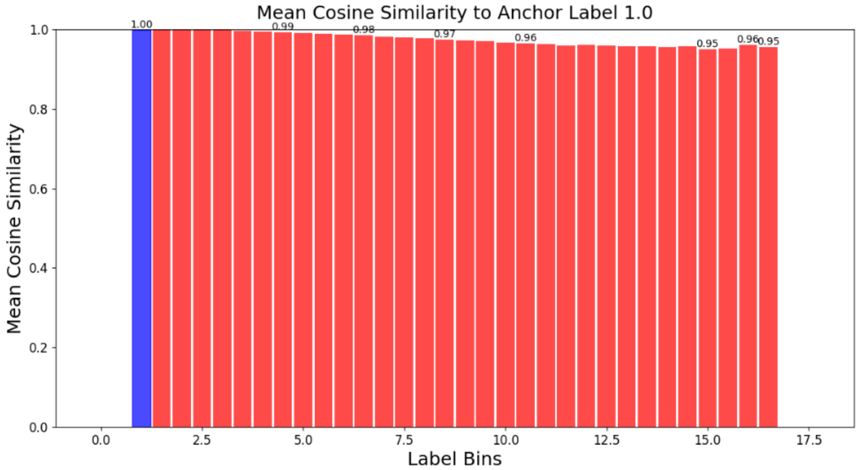


Figure A.1: The feature similarities relative to anchor 1.0 of the vanilla approach.

doi:10.6342/NTU202403142

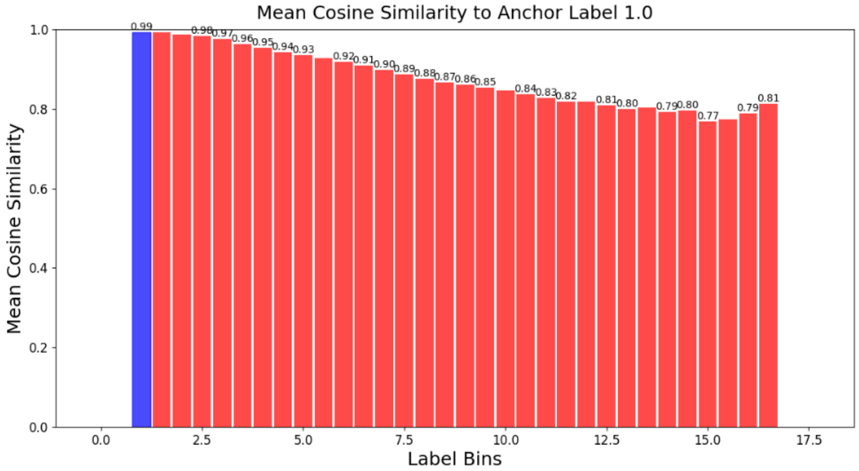


Figure A.2: The feature similarities relative to anchor 1.0 of Rank-N-Contrast.

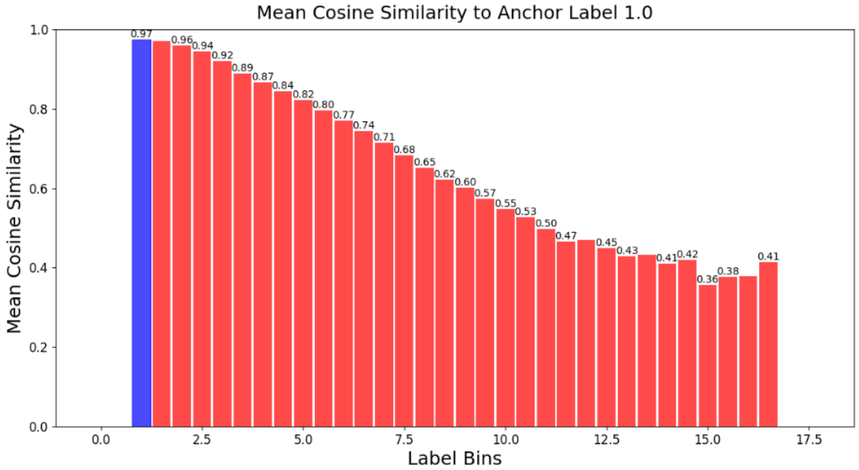


Figure A.3: The feature similarities relative to anchor 1.0 of Weighted-Rank-CR.

A.2 Feature Visualization

We employed UMAP (McInnes et al., 2018) to visualize the post features in a 2-dimensional space for both the vanilla approach and the Weighted-Rank CR method.

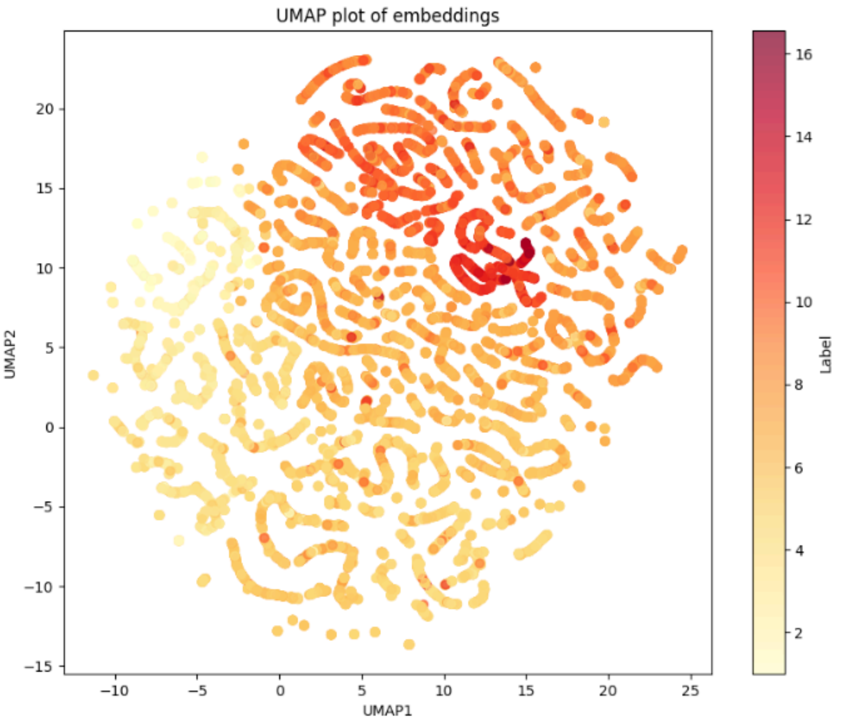


Figure A.4: The UMAP of the post features fit on the vanilla approach.

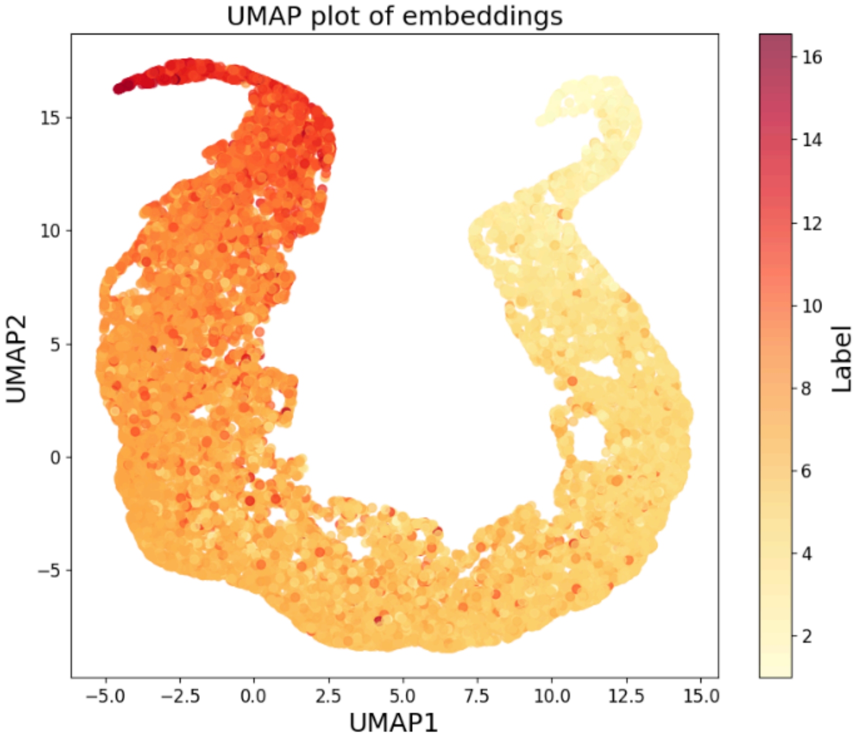


Figure A.5: The UMAP of the post features fit on Weighted-Rank CR.

doi:10.6342/NTU202403142