Utilizing Contrastive Learning To Address Long Tail Issue in Product Categorization

Lei Chen
Rakuten Institute of Technology
Boston, MA, USA
lei.a.chen@rakuten.com

Tianqi Wang
University of Buffalo
Buffalo, NY, USA
twang47@buffalo.edu

ABSTRACT

Neural network models trained in a supervised learning way have become dominant. Although high performances can be achieved when training data is ample, the performance on labels with sparse training instances can be poor. This performance drift caused by imbalanced data is named as long tail issue and impacts many NN models used in reality. In this talk, we will firstly review machine learning approaches addressing the long-tail issue. Next, we will report on our effort on applying one recent LT-addressing method on the item categorization (IC) task that aims to classify product description texts into leaf nodes in a category taxonomy tree. In particular, we adopted a new method, which consists of decoupling the entire classification task into (a) learning representations using the K-positive contrastive loss (KCL) and (b) training a classifier on balanced data set, into IC tasks. Using SimCSE to be our self-learning backbone, we demonstrated that the proposed method works on the IC text classification task. In addition, we spotted a shortcoming in the KCL: false negative (FN) instances may harm the representation learning step. After eliminating FN instances, IC performance (measured by macro-F1) has been further improved.

KEYWORDS

imbalanced data, long tail, item categorization, contrastive learning

ACM Reference Format:
Lei Chen and Tianqi Wang. 2022. Utilizing Contrastive Learning To Address Long Tail Issue in Product Categorization. In *Proceedings of the 31st ACM International Conference on Information and Knowledge Management (CIKM '22), October 17–21, 2022, Atlanta, GA, USA.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3511808.3557522

1 INTRODUCTION

Nowadays with the leaps of using neural networks to train powerful models in a supervised learning way, many models show impressive performances and some are even close to human's performances. However, on corner cases with only skew and sporadic supervision signals, such as users without frequent online behavior footsteps, news on rare topics, and so on, existing models face a great challenge, which is frequently called as long-tail (LT) issue and has gotten attention from the CIKM community [11, 17].

CIKM '22, October 17–21, 2022, Atlanta, GA, USA

ACM ISBN 978-1-4503-9236-5/22/10.
https://doi.org/10.1145/3511808.3557522

LT issue impacts many models used in practice. For example, item categorization (IC) is an e-commerce NLP task to classify a product into a node in a category taxonomy tree. Only a few head labels have sufficient samples, while a large number of tail labels contain few samples. Consequently, sporadic supervision on these tail labels tends to cause unsatisfactory IC performance. For online e-commerce sites, given massive users and their wide-spreading shopping interests/intentions, daily query/visiting volumes on these tail labels are still considerable. Therefore, poor performance on these tail labels tend to bring very negative user experiences and hurt sales too. Therefore, algorithms that can handle the LT-issue have become important.

In this talk, we will firstly introduce LT-addressing methods that mostly are proposed from the computer vision (CV) research area. After briefly summarizing several major approach groups, we will focus on two (but related) groups that fundamentally change the design on LT-addressing approaches. Since 2019, separation of the learning of feature representations and the learning of a classifier head and only applying the balanced data set on the latter has become a leading approach. Later, in the feature representation learning stage, contrastive learning (CL) has been suggested since its training process is not negatively impacted by labels' skewed distributions. Then, we will report on our work of applying these latest LT-addressing methods in the IC task.

2 RELATED WORK

Many methods have been proposed to address the LT issue. One group of those methods resamples the data to balance the label distribution, e.g., SMOTE [2]. Another group of methods assigns different weights to samples based on their label frequencies, e.g., Focal loss [12], class-balanced (CB) loss [5], Label-Distribution-Aware Margin loss (LDAM) [1] and so on. In addition, few shot learning [13] and transfer learning [14] methods are also proposed for mitigating the LT issue.

Recently, a *two-stage* training strategy (exampled in [9, 16]), which decouples the learning a feature encoder and the learning of a classifier, has become influential in computer vision and shows its superior performance on addressing the LT issue. In [9], an image encoder is first trained on the original dataset in a supervised learning way and then a classifier is trained on resampled dataset (balanced) on the frozen encoding representations obtained in the first stage. In [16], a Bilateral Branch Network (BBN) is used to dynamically interpolate these two training stages.

Contrastive learning (CL) has been found to be effective in providing high-quality encoders in a simple self-learning fashion. For

example, in computer vision, SimCLR [3] uses the consistence between an anchor image and its transformed version and the inconsistence between the anchor and other instances in a batch (in-batch negative instances) to guide encoder training. Inspired by the success of SimCLR in computer vision, CL-based text representation learning has also become a hot research topic in natural language processing (NLP). SimCSE [6] uses dropout operations to be an effective text augmentation and can learn effective text representations.

Regarding the two-stage training strategy, self-learning which discards the influence of label distribution has been used in its representation learning stage, e.g., [8, 15]. Besides simply using self-supervision, including the supervision signal from existing labels can improve the representation learning [10]. However, introducing semantics information may suffer from the LT issue and hurt the performance of tail classes. To address this issue, K-positive contrastive (KCL) loss [8] is proposed to learn balanced feature representations. An instance is called *false negative* (FN), if any in-batch negative instance shares the label carried by the anchor image. FN samples are found to be harmful to CL methods and corresponding mitigation methods are proposed [4, 7].

3 APPLICATION ON ITEM CATEGORIZATION

In this talk, we will demonstrate our work on utilizing contrastive learning to address the LT issue in the IC task. Our framework uses unsupervised *SimCSE* [6] for data augmentation and *K-positive contrastive loss (KCL)* [8] to learn balanced feature embeddings. Moreover, we recognize false negative (FN) instances exist in KCL and apply two different strategies: *FN attraction* and *FN elimination* to cancel them. The experimental results on three Amazon product category datasets show that the contrastive learning methods help on improving the model performance on tail classes and the FN mitigating can further improve CL-based LT-addressing method.

Table 1 reports our concrete results on the three Amazon product category datasets, i.e., Automotive, Electronics, and Beauty, with labels of 953, 500, and 229. The Baseline model is simply fine-tuning a BERT-base model using a cross-entropy (CE) loss. cRT method described in [9] is used to represent a conventional LT-addressing method. In addition, using KCL to train text encoder in the first stage is included. Note that the CL on texts is based on applying SimCSE method here. On top of this, methods mitigating the FN issue are tried. We can find that the KCL method works better than cRT and always can improve IC performance measured in macro-F1 ($F1_m$). Both FN mitigating methods are helpful to further improve performance.

My talk intends to advocate an attention to the value of addressing the LT issue in many supervised-learning based approaches. Since tail labels often are associated with under-represented users or niche products/messages, addressing this issue has become important for both business and social reasons. From my talk, I hope that the audience can obtain follows: (1) a deeper understanding on the LT issue and its impacts, (2) knowing latest solutions on addressing this issue, and (3) learning a promising solution based on contrastive learning and two-stage training strategy.

	Automotive		Electronics		Beauty	
	$F1_w$	$F1_m \uparrow$	$F1_w$	$F1_m \uparrow$	$F1_w$	$F1_m \uparrow$
BERT-CE	**78.03**	63.95	**67.68**	52.94	71.44	56.64
cRT [9]	77.85	63.72	67.54	52.99	71.55	55.88
KCL [6, 8]	76.87	65.17	65.18	53.39	71.44	58.26
+FNA	76.54	64.65	66.08	**53.69**	**71.65**	**58.31**
+FNE	77.96	**65.82**	65.73	53.67	71.43	57.95

Table 1: IC performance on the three Amazon product data sets: Automotive, Electronics and Beauty. The best results are highlighted using bold fonts. $F1_w$ and $F1_m$ denote the weighted F1 and macro F1.

REFERENCES

[1] Kaidi Cao, Colin Wei, Adrien Gaidon, Nikos Arechiga, and Tengyu Ma. 2019. Learning imbalanced datasets with label-distribution-aware margin loss. *arXiv preprint arXiv:1906.07413* (2019).

[2] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. 2002. SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research* 16 (2002), 321–357.

[3] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*. PMLR, 1597–1607.

[4] Tsai-Shien Chen, Wei-Chih Hung, Hung-Yu Tseng, Shao-Yi Chien, and Ming-Hsuan Yang. 2021. Incremental False Negative Detection for Contrastive Learning. *arXiv preprint arXiv:2106.03719* (2021).

[5] Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. 2019. Class-balanced loss based on effective number of samples. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 9268–9277.

[6] Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple Contrastive Learning of Sentence Embeddings. In *Empirical Methods in Natural Language Processing (EMNLP)*.

[7] Tri Huynh, Simon Kornblith, Matthew R Walter, Michael Maire, and Maryam Khademi. 2020. Boosting contrastive self-supervised learning with false negative cancellation. *arXiv preprint arXiv:2011.11765* (2020).

[8] Bingyi Kang, Yu Li, Sa Xie, Zehuan Yuan, and Jiashi Feng. 2021. Exploring Balanced Feature Spaces for Representation Learning. In *International Conference on Learning Representations*. https://openreview.net/forum?id=OqtLIabPTit

[9] Bingyi Kang, Saining Xie, Marcus Rohrbach, Zhicheng Yan, Albert Gordo, Jiashi Feng, and Yannis Kalantidis. 2019. Decoupling representation and classifier for long-tailed recognition. *arXiv preprint arXiv:1910.09217* (2019).

[10] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. 2020. Supervised contrastive learning. *arXiv preprint arXiv:2004.11362* (2020).

[11] Mengmeng Li, Tian Gan, Meng Liu, Zhiyong Cheng, Jianhua Yin, and Liqiang Nie. 2019. Long-tail hashtag recommendation for micro-videos with graph convolutional network. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*. 509–518.

[12] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*. 2980–2988.

[13] Ziwei Liu, Zhongqi Miao, Xiaohang Zhan, Jiayun Wang, Boqing Gong, and Stella X. Yu. 2019. Large-Scale Long-Tailed Recognition in an Open World. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[14] Lin Xiao, Xiangliang Zhang, Liping Jing, Chi Huang, and Mingyang Song. 2021. Does Head Label Help for Long-Tailed Multi-Label Text Classification. *arXiv preprint arXiv:2101.09704* (2021).

[15] Yuzhe Yang and Zhi Xu. 2020. Rethinking the value of labels for improving class-imbalanced learning. *arXiv preprint arXiv:2006.07529* (2020).

[16] Boyan Zhou, Quan Cui, Xiu-Shen Wei, and Zhao-Min Chen. 2020. Bbn: Bilateral-branch network with cumulative learning for long-tailed visual recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 9719–9728.

[17] Xiangzeng Zhou, Pan Pan, Yun Zheng, Yinghui Xu, and Rong Jin. 2020. Large scale long-tailed product recognition system at alibaba. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*. 3353–3356.