Weighted Contrastive Learning With False Negative Control to Help Long-tailed Product Classification

Tianqi Wang	Lei Chen	Xiaodan Zhu		
University of Buffalo	Rakuten Institute of Technology	Queen's University		
Buffalo, NY	Boston, MA	Kingston, ON, Canada		
twang47@buffalo.edu	lei.a.chen@rakuten.com	xiaodan.zhu@queensu.edu		

Younghun Lee and Jing Gao

Purdue University, West Lafayette, IN

{younghun,jinggao}@purdue.edu

Abstract

Item categorization (IC) aims to classify product descriptions into leaf nodes in a categorical taxonomy, which is a key technology used in a wide range of applications. Along with the fact that most datasets often has a long-tailed distribution, classification performances on tail labels tend to be poor due to scarce supervision, causing many issues in real-life applications. To address IC task's long-tail issue, K-positive contrastive loss (KCL) is proposed on image classification task and can be applied on the IC task when using text-based contrastive learning, e.g., SimCSE. However, one shortcoming of using KCL has been neglected in previous research: false negative (FN) instances may harm the KCL's representation learning. To address the FN issue in the KCL, we proposed to reweight the positive pairs in the KCL loss with a regularization that the sum of weights should be constrained to K+1 as close as possible. After controlling FN instances with the proposed method, IC performance has been further improved and is superior to other LT-addressing methods.

1 Introduction

Item categorization (IC) aims to classify a product into a node of a taxonomy hierarchy. The textual descriptions of the products are used as the input and thus the task can be formulated as a text classification problem. IC is a fundamental task in e-commerce and the base for many applications such as personal recommendation and query understanding. One of the major challenges in building a highly effective real-life IC system is the serious long-tailed (LT) problem-A few head classes have the majority of the product items, while each of the remaining (large number of) tail classes contains only a few items. Consequently, the scarce supervision available for these tail classes tends to cause unsatisfactory classification performance. In the most recent years, several novel

LT-addressing methods, e.g., methods utilizing selfsupervision (Yang and Xu, 2020) and contrastive learning (CL) (Kang et al., 2021), have emerged in computer vision. However, the related research on natural language processing (NLP) tasks is still limited.

However, when utilizing unsupervised contrastive learning, e.g., K-positive contrastive loss (KCL) (Kang et al., 2021), the issue of appearing False Negative (FN) samples hurts model learning. Figure 1 shows an example of the FN issue and the performance impact reported in (Huynh et al., 2020). In this paper, to build IC models and use KCL to solve the LT issue, we propose a novel method to control the FN issue, which entails assigning different weights for each positive sample in the KCL loss and tries to keep the sum of these weights equal to a predefined value. The experimental results on the three Amazon product category datasets show that the proposed contrastive learning methods help on improving the model performance on tail classes and the FN controlling can further improve CL-based LT-addressing method. Our main contributions can be summarized as:



Figure 1: False Negative (FN) samples may appear when applying contrastive learning. Addressing the FN issue can improve the learned model's downstream performance. Result is from the Table 2 in (Huynh et al., 2020) regarding image classification on the ImageNet data using the model learned by SimCLR (Chen et al., 2020)

- We recognize the false negative sample issue in K-positive contrastive loss.
- We propose a novel false negative controlling method to mitigate the its negative impact and show the effectiveness of proposed model.
- To the best of our knowledge, we are the first to apply contrastive learning to address the LT challenge in the IC text classification.

2 Related Work

Many methods have been proposed to address the LT issue. One category of those methods resamples the data to balance the label distribution, e.g., SMOTE (Chawla et al., 2002). Another category of methods assign different weights to samples based on their label frequencies, e.g., Focal loss (Lin et al., 2017) Class-balanced loss (Cui et al., 2019), Label-Distribution-Aware Margin loss (LDAM) (Cao et al., 2019) and so on. Some of those loss-balanced methods are also applied to the NLP domain in (Huang et al., 2019) and transfer learning (Xiao et al., 2021) methods are also proposed for long tail classification.

Recently, a *two-stage* training strategy (exampled in (Kang et al., 2019; Zhou et al., 2020)), 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.

Contrastive learning (CL) has been found to be effective in providing high-quality encoders in a simple self-learning fashion. The CL-based text representation learning has become a hot research topic in NLP. SimCSE (Gao et al., 2021) uses dropout operations to be an effective text augmentation and can learn effective text representations.

In the LT-addressing two-stage method, selflearning which discards the influence of label distribution has been used in its representation learning stage, e.g., (Yang and Xu, 2020; Kang et al., 2021). Besides simply using self-supervision, including the supervision signal from existing labels can improve the representation learning (Khosla et al., 2020). However, introducing semantics information may suffer from the LT issue and hurt the performance of tail classes. To address this issue, K-positive contrastive loss (Kang et al., 2021) 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 sample. FN samples are found to be harmful to CL methods and corresponding mitigation methods are proposed (Huynh et al., 2020; Chen et al., 2021).

3 Methodology

Let x denote the title of a product and y its category label. Then IC can be formulated as a text classification problem: given a product title x, we a model to predict the class label y, where h and h^+ are the representations of the anchor sample x and its corresponding positive sample x^+ , respectively. $H^$ is the representations of negative sample set X^- of the given the anchor sample x and positive sample x^+ . $h^- \in H^-$ is the representation of the negative sample x^- in X^- where X^- is the negative sample set given the anchor sample x and its positive sample x^+ . τ is a temperature hyper-parameter and $sim(\cdot, \cdot)$ denotes the cosine similarity of the two vectors.

3.1 Recap of KCL

The KCL is a state-of-the-art model that learns balanced feature representations for long-tailed label distribution. It defines a positive sample set by sampling K samples belonging to the same class as the anchor if such samples are more than K in existing mini-batch. The KCL can be calculated by the Eq. 1.

$$\mathcal{L}_{KCL} = \frac{1}{(K+1)} \sum_{h \in H} \sum_{h^+ \in \{h'\} \cup H_K^+} \mathcal{L}(h, h^+)$$
$$\mathcal{L}(h, h^+) = -\log \frac{e^{sim(h, h^+)/\tau}}{\sum_{h_i \in H - \{h\}} e^{sim(h, h_i)/\tau}}$$
(1)

where h is the anchor sample representation and h' the self-augmented representation of h. H_K^+ represents the representation set of sampled K positive samples from the batch. H denotes the samples in the same batch. K is the hyper-parameter representing the defined number of positive pairs.

In NLP, to generate h', we propose to use the SimCSE (Gao et al., 2021) method. In particular, an anchor sample x is encoded using a BERT (Devlin et al., 2018) model with varying dropout masks. The encoded representations can be represented as:

$$h = tanh(MLP(BERT(x, z)))$$

$$h' = tanh(MLP(BERT(x, z')))$$
 (2)

where BERT(x, z) denotes the BERT encoder using a random dropout mask. For MLP, we use a layer of fully connected network with the tanhactivation, while z and z' are two different random dropout masks in BERT at rate of 0.1.



Figure 2: KCL applied on a batch with five samples, P0 to P4. Color on input blocks shows labels. P0 serves as an anchor and its positive pair is obtained by Sim-CSE with a different dropout mask. When K = 1, P1 is added into the positive set H^+ . P2, however, is assigned into the negative set H^- and a false negative case appears and is marked as a red cross.

Fig. 2 illustrate how the KCL represented in Eq. 1 is applied on a mini-batch with five samples. Since P2 and P0 share the same label, their encoded representations are expected to be close. However, when running KCL with K = 1, P2 will be wrongly pushed away from P0 being a FN sample.

3.2 False Negative Control

As shown in Fig 2, a significant drawback of KCL is that some positive samples can be treated as negative if there are more than K + 1 samples belonging to the same class in a batch. The occurrence of the false negative instances will degrade the quality of the learned representations and further hurt the classification performance. To alleviate the false negative, we propose a novel method named false negative control (FNC), which assigns varying weights to positive samples in the KCL loss based on the embeddings of the anchor sample and the positive samples, represented as:

$$w_{h,h^+} = ReLU(MLP(h \oplus h^+)) \tag{3}$$

where w_{h,h^+} is the weight for the positive sample h^+ with respect to the anchor sample h. \oplus is the vector concatenation operation.

With the learned weights, we propose the weighted contrastive loss by re-weighting the pos-

itive samples in the original InfoNCE loss, which can be defined as:

$$L_w(h, h^+) = -\log \frac{e^{sim(h, h^+)/\tau}}{\sum_{h_i^+ \in H^+} w_i e^{sim(h, h_i^+)/\tau} + \sum_{h_i^- \in H^-} e^{sim(h, h_i^-)/\tau}}$$
(4)

To satisfy the property of balancing the feature space for classes with different frequencies in KCL while controlling the FN issue at the same time, we propose the KCL-FNC loss with the aforementioned defined weighted contrastive loss with the constraint that the summation of these weights is as close to a predefined value (K + 1) as possible. The KCL-FNC loss is defined as:

$$\mathcal{L}_{KCL-FNC} = \sum_{h \in H} \sum_{\substack{h^+ \in \{h'\} \cup H_K^+ \\ \mathcal{L}_w(h,h^+) + \lambda \mathcal{L}_{reg}(h,H^+) \\ (K+1)}}$$
(5)

where λ is the balanced parameter and $\mathcal{L}_{reg}(h, H^+)$ is the regularization loss denoted as:

$$\mathcal{L}_{reg}(h, H^+) = |\sum_{h^+ \in H^+} w_{h, h^+} - K - 1| \quad (6)$$

The advantages of the proposed KCL-FNC loss over exiting FN controlling methods are two-folds: (1) learning balanced feature representations and (2) applying as much available information as possible. The attraction strategy (FNA) in (Huynh et al., 2020; Chen et al., 2021), which include all positive samples rather that K sampled positive samples, makes the KCL roll back to the supervised contrastive loss when the ground-truth labels are known, and therefore destroys the KCL's property of learning balanced representations. The elimination (FNE) strategy, which ignores the FN samples in calculating the contrastive loss, loses valuable information and further degrades the representationlearning performance, especially the number of such instances are large, such as for head labels in a imbalanced data set.

4 Experiments

Datasets. The experiments are performed on the public Amazon dataset (McAuley et al., 2015; He and McAuley, 2016) which is a widely used benchmark. Following (Tayal et al., 2020), we use three categories of Amazon product datasets: Automotive, Beauty, and Electronics. Each sample in the

	Autor	notive	Elect	ronics	Be	auty	Auto _H	Auto_M	Auto_T
	$F1_w$	$F1_m\uparrow$	$F1_w$	$F1_m\uparrow$	$F1_w$	$F1_m \uparrow$	$ F1_m\uparrow$	$F1_m\uparrow$	$F1_m\uparrow$
BERT-CE	78.03	63.95	67.68	52.94	71.44	56.64	75.42	64.51	51.78
cRT	77.85	63.72	67.54	52.99	71.55	55.88	75.20	63.99	51.78
$SimCSE_{us}$	76.36	64.25	65.82	53.30	70.99	58.06	74.16	64.92	54.65
KCL	76.87	65.17	65.18	53.39	71.44	58.26	74.99	65.06	55.36
KCL-FNA	76.54	64.65	66.08	53.69	71.65	58.31	74.46	64.88	54.53
KCL-FNE	77.96	65.82	65.73	53.67	71.43	57.95	75.97	65.78	55.61
KCL-FNC	78.05	66.20	66.24	54.02	71.84	58.56	76.02	66.50	56.07

Table 1: Model Performance on Long-tailed IC. The left part of the table shows the performance on the three datasets: Automotive, Electronics and Beauty. The right part shows the results on the three subsets of the Automotive dataset, where Auto_H, Auto_M and Auto_T consist of the head, medium and tail classes in Automotive. The best results are highlighted using bold fonts. $F1_w$ and $F1_m$ denote the weighted F1 and macro-F1.

datasets has a title and a category label. All three datasets have long-tail issue.¹

Overall Performance. We compare the proposed method with the following models: BERT with cross-entropy loss (BERT-CE), cRT (Kang et al., 2019), unsupervised SimCSE (SimCSE_{us}), and controlling false negative instances with attraction (KCL-FNA) and elimination (KCL-FNE). Note that except for BERT-CE, all other baseline models use the two-stage approach to address the long tail issue, in which the classifier is trained using a balanced data set.

The experimental results are shown in Table 1. We can observe that all contrastive learning-based models outperform BERT-CE and cRT in terms of macro-F1, which demonstrates the effectiveness of contrastive learning in addressing the long tail issue in IC. The calculated FN sample rates are 0.036 (Automotive), 0.068 (Electronics) and 0.102 (Beauty), showing that there are significant number of FN samples when using KCL¹. When comparing the false negative controlling methods with the KCL, we observed that those false negative controlling methods achieved better performance in terms of macro-F1. The results demonstrate the necessity of controlling the FN issue in KCL. Among those false negative controlling methods, the proposed method outperforms all other methods, showing its advantage over existing methods.

Performance on Subsets. To investigate the performance of the models on the classes with different label frequencies, we split the Automotive dataset into three subsets according to the label frequency: Head, Median and Tail and evaluate the models by macro-F1 on the subsets¹. The model performance on the subsets are included in Table 1. We can see that the performance decreases along with the decrease of the label frequencies for each single model, illustrating the lacking of samples limits the model performance. Moreover, the methods based on KCL outperforms all other baselines. The proposed FN controlling method achieves the best performance on all subsets which demonstrates the false negative controlling method can help address the long tail issue in IC task without hurting the overall performance. The details can be found in the Appendix.

5 Conclusion

In large-scaled item categorization tasks, category labels are naturally distributed in a long tail pattern, which challenges the performance on tail classes due to severe supervision missing. To address this challenge, we adopt a two-stage LT-addressing method that was originally proposed in the image classification task. To make this method work on our text classification task, we use the recently proposed SimCSE (Gao et al., 2021) to do an effective text transformation and KCL loss in the representation learning stage. Furthermore, we recognize there are false negative samples caused by using the KCL loss and propose a novel controlling method to reduce the corresponding negative influences. The experimental results prove that the proposed method helps improve the performance on long-tailed data and the false negative controlling can further help boost the performance when using KCL. While we worked on item classification in this paper, we will extend the model to other problems.

6 Limitations

A major limitation of this research work is only item classification, one specific type of NLU tasks, is used in our experiments. To better evaluate our proposed KCL-FNC method, an expanded testing task set will provide more convincing power. In addition, we only used cross-entropy (CE) loss when training models, in both representation and classifier learning stages. It will be interesting to see the compound effect when applying our proposed method together with some advanced loss types, such as LDAM (Cao et al., 2019).

References

- 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.
- 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*, 16:321–357.
- 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*, pages 1597–1607. PMLR.
- 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*.
- 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*, pages 9268–9277.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.
- Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple contrastive learning of sentence embeddings. In *Empirical Methods in Natural Language Processing (EMNLP).*
- Ruining He and Julian McAuley. 2016. Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering. In *proceedings of the 25th international conference on world wide web*, pages 507–517.
- Yi Huang, Buse Giledereli, Abdullatif Köksal, Arzucan Özgür, and Elif Ozkirimli. 2021. Balancing methods for multi-label text classification with long-tailed class distribution. *arXiv preprint arXiv:2109.04712*.

- 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*.
- 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*.
- 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*.
- 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*.
- 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*, pages 2980–2988.
- 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).*
- Julian McAuley, Christopher Targett, Qinfeng Shi, and Anton Van Den Hengel. 2015. Image-based recommendations on styles and substitutes. In *Proceedings* of the 38th international ACM SIGIR conference on research and development in information retrieval, pages 43–52.
- Kshitij Tayal, Rahul Ghosh, and Vipin Kumar. 2020. Model-agnostic methods for text classification with inherent noise. In *Proceedings of the 28th International Conference on Computational Linguistics: Industry Track*, pages 202–213.
- 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*.
- Yuzhe Yang and Zhi Xu. 2020. Rethinking the value of labels for improving class-imbalanced learning. *arXiv preprint arXiv:2006.07529*.
- 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*, pages 9719–9728.

A Dataset Statistics

The statistics of the three data sets: Automotive, Beauty and Electronics are shown in Table A. 1 and the histogram of the label frequencies of the three data sets are shown in Figure A. 1.

	Labels	Samples	Title Length
Automotive	953	160,725	9.90 ± 5.51
Beauty	229	159,805	10.26 ± 5.61
Electronics	500	86,357	14.90 ± 9.56

Table A. 1: Statistics of Datasets

B Implementation Details

For all the models except for the BERT, we follow the two-stage training protocol in (Kang et al., 2019). The batch size is set to 32 and initial learning rate is 1e - 5 with a linear decay. The datasets are preprocessed following (Tayal et al., 2020). We split the training datasets into two subsets: train vs. dev that is used to select hyperparameters and validate the performance 2 . The models are evaluated using two metrics: macro-F1 (F1_m) and weighted F1 (F1 $_w$). Note that macro-F1 is frequently used in evaluating LT-addressing methods. Since it calculates the F1 for each class and averages them, it is significantly influenced by the performance of tail classes. We report the results on the test set using the best models on the dev set measured by macro-F1.

C False Negative Rate Calculation

To calculate the false negative rate, we use the obtained embeddings of SimCSE-KCL in the first stage after 10 epoch and report the average of five runs. We calculate the false negative rate of those three datasets where the batch size is set to 32 and K is set to 1. Following (Chen et al., 2021), we calculate the false negative rate in SimCSE-KCL for the three datasets. The false negative rate fnr is the number of false negative samples among top 25% the most similar samples of the anchor in a batch, which can be represented as:

$$fnr = \frac{\sum_{i=1}^{N} \sum_{x_j \in B_i} max(0, |B_i^j| - (K+1))}{\sum_{i=1}^{N} (0.25 \times |B_i| \times (|B_i| - 1))}$$

N is the number of batches. B_i is the set of samples in batch i and $|B_i|$ is the number of samples in batch *i*. $|B_i^j|$ is the number of samples belonging to the same class as x_j in the 25% most similar samples with the sample x_j .

D Data Statistics of the Subsets

The classes are sorted based on their frequencies. The Head subset consists of the samples in the most frequent $\frac{1}{3}$ classes and the Tail subset includes the samples belonging to the least frequent $\frac{1}{3}$ classes. The rest samples belonging to the $\frac{1}{3}$ median frequent classes consists of the Median subset. The statistics of the subsets are shown in the Table D. 1.

	Automotive		
	Head	Median	Tail
Label	318	318	317
Sample	132,590	20,318	7,817

Table D. 1: Data statistics of the subsets of the three original training data sets based on the label frequencies.

²The code will be available after acceptance.



Figure A. 1: Label Frequency Histogram of Automotive, Beauty and Electronics Datasets