Control False Negative Instances In Contrastive Learning To Improve Long-tailed Item Categorization

Anonymous ACL submission

Abstract

Item categorization (IC) is an important core technology in e-commerce natural language processing (NLP). Given category labels' long-tailed distribution, IC performances on tail labels tend to be poor due to sporadic supervision. To address the long-tail issue in classification, an increasing number of methods have been proposed in the computer vision domain. In this paper, we adopted a new method, which consists of decoupling the entire classification task into (a) learning representations in a k-positive contrastive learning (KCL) way 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 instances (FN) may harm the representation learning step. After eliminating FN instances, IC performance (measured by macro-F1) has been further improved.

1 Introduction

Item categorization (IC) is to classify a product into a node in a category taxonomy. It is a fundamental task in e-commerce and the basis of personal recommendations, query understanding and so on. One of the challenges to building a highly effective IC system is products' long-tailed (LT) label distribution, where only a few head classes have a lot of samples, while the other large volume of tail classes only consists of a few samples. Consequently, sporadic supervision on these tail labels tends to cause unsatisfactory IC performance.

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

In this paper, we propose to utilize contrastive learning to address the LT challenge in the IC task. The proposed framework uses unsupervised *SimCSE* (Gao et al., 2021) for data augmentation and *K-positive contrastive loss (KCL)* (Kang et al., 2021) to learn feature embeddings in balanced feature space. 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 cancellation can further improve CL-based LT-addressing method. Our main contributions can be summarized as:

- We apply contrastive learning to address the LT challenge in the IC text classification.
- We recognize the false negative sample issue in K-positive contrastive loss and apply a false negative cancellation strategy to mitigate its negative impact.

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.,Class-balanced loss (Cui et al., 2019), Label-Distribution-Aware Margin loss (LDAM) (Cao et al., 2019) and so on. 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. For example, in computer vision, SimCLR (Chen et al., 2020) uses the consistence between an anchor image and its transformed version and the in-consistence between the anchor and other instances in a batch (in-batch negative instances) to guide encoder training. If any in-batch negative instance shares the label carried by the anchor image, such an instance is called *false negative* (FN). FN samples are found to be harmful to CL methods and corresponding mitigation methods are proposed (Huynh et al., 2020; Chen et al., 2021). Inspired by the success of SimCLR in computer vision, CL-based text representation learning has been a hot research topic in NLP. SimCSE (Gao et al., 2021) uses dropout operations existing in Transformer to be an effective text augmentation and can learn effective text representations. In the LT-addressing two-stage method, self-learning 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 long tail 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.

3 Methodology

Let x denote the title of a product and y is its label. The IC can be formulated as a text classification task and can be described as: given a product title x, IC needs to predict the category label y.

3.1 Unsupervised SimCSE

Recently, unsupervised SimCSE (Gao et al., 2021) is proposed to learn sentence embeddings using a self-supervised contrastive learning method. The unsupervised SimCSE maximizes the agreement of the representations of a positive pair by using the InfoNCE loss represented in Eq. 1.

$$\mathcal{L}(h, h^+, H^-) = -log\frac{e^{\frac{sim(h,h^+)}{\tau}}}{\sum_{h^-\in H^-} e^{\frac{sim(h,h^-)}{\tau}}} \tag{1}$$

where h, h^+ and H^- are the representations of the anchor sample x, a positive instance x^+ and the set of negative instances. In the unsupervised SimCSE, the positive instance is the same as the anchor sample (i.e., $x^+ = x$). The negative sample set consists of the set of all other samples in the same batch as the anchor sample.

The anchor sample x and its positive sample $x^+ = x$ are encoded using two different BERT (Devlin et al., 2018) based encoders which share the same architecture but use different random dropout masks. The encoder can be represented as:

$$\begin{aligned} h &= tanh(MLP(BERT(x,z))) \\ h^+ &= tanh(MLP(BERT(x^+,z^+))) \end{aligned} \tag{2}$$

where $BERT(x,z)$ denotes the BERT encoder using a random dropout mask. MLP is a one-layer fully connected layer and $tanh$ represents the hyperbolic tangent activation function. z and z^+ are two different random dropout masks in BERT at rate of 0.1.

3.2 SimCSE with K-positive Contrastive Loss

To use important supervision signals provided by the labels, we propose the SimCSE-KCL framework illustrated in Fig. 1(a) to incorporate the K-positive contrastive loss (KCL) into the SimCSE framework. Compared with the unsupervised SimCSE, the SimCSE-KCL uses K more positive instances randomly sampled from the batch containing the anchor. The KCL can be represented as:

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

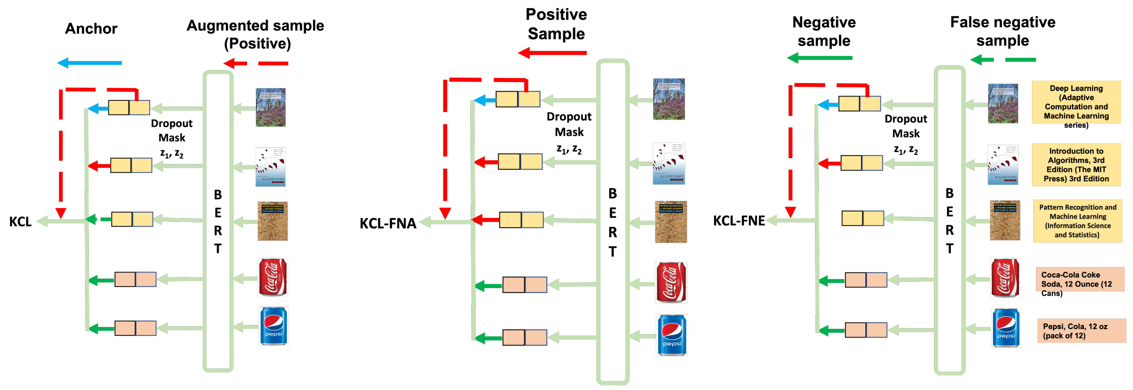
where h' represents the self-augmented representation of x and H_K^+ represents the representation set of the of sampled K positive samples from the batch. H^- denotes the corresponding negative sample representation set given the anchor and positive sample. K is the hyper-parameter representing the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

By incorporating the KCL into the SimCSE framework, the SimCSE-KCL can both take advantage of the contrastive loss to learn the balanced features and improve the semantic discrimination ability from the learned features.

3.3 False Negative Cancellation

A drawback of the SimCSE-KCL is some positive samples will be considered as negative if there are more than $K + 1$ samples belonging to the same class in a batch. As shown in Fig 1(a), when $K = 1$, the third sample is false negative and excluded from the positive set in SimCSE-KCL. The occurrence of such false negative samples may degrade the quality of the learned embeddings and further hurt the classification performance.

To alleviate the influence of the false negative samples, we propose two frameworks: SimCSE-KCL-FNA and SimCSE-KCL-FNE, which utilizes the


Anchor
Augmented sample (Positive)
Dropout Mask z_1, z_2
KCL
B E R T
Positive Sample
Dropout Mask z_1, z_2
KCL-FNA
B E R T
Negative sample
False negative sample
Dropout Mask z_1, z_2
KCL-FNE
B E R T
Deep Learning (Adaptive Computation and Machine Learning series)
Introduction to Algorithms, 3rd Edition (The MIT Press) 3rd Edition
Pattern Recognition and Machine Learning (Information Science and Statistics)
Coca-Cola Coke Soda, 12 Ounce (12 Cans)
Pepsi, Cola, 12 oz (pack of 12)

(a) SimCSE-KCL

(b) SimCSE-KCL-FNA

(c) SimCSE-KCL-FNE

Figure 1: Illustration of the proposed frameworks: SimCSE-KCL, SimCSE-KCL-FNA and SimCSE-KCL-FNE, when $K = 1$. Blue, red and green arrows denote the anchor, positive and negative instances correspondingly. Specifically, dashed red arrow is the augmented sample and dashed green arrow represents the false negative sample.

attraction and elimination strategy as in (Grill et al., 2020) to cancel false negative samples.

The SimCSE-KCL-FNA uses an attraction strategy, where all positive samples rather than K sampled positive instances are included in the positive set. As shown in Fig. 1(b), the third sample considered as a negative sample in SimCSE-KCL is included in the positive set in SimCSE-KCL-FNA. When the ground-truth labels are known, the loss in SimCSE-KCL-FNA is the same as supervised contrastive loss (Khosla et al., 2020). It uses more labels and contains no noise compared with SimCSE-KCL.

The SimCSE-KCL-FNE uses the elimination strategy shown in Fig. 1(c), which ignores the false negative samples in calculating the contrastive loss by neither including it in the positive sample set nor in the negative sample set. We can find in Fig. 1(c) the third sample is not used to calculate the loss. Despite the less information used in SimCSE-KCL-FNE, removing the noise in the data can boost the performance of the models.

4 Experiment

Datasets: The experiments are performed on the three categories of Amazon product (McAuley et al., 2015; He and McAuley, 2016) datasets: Automotive, Beauty, and Electronics. Each sample has a title and a category label. All three datasets show long tail characteristics[1]

Experimental Setup: We compare the three proposed frameworks with BERT using cross-entropy loss (BERT-CE), cRT (Kang et al., 2019), and unsupervised SimCSE (SimCSE$_{us}$). For both cRT and SimCSE based models, 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 using the best models on the dev set measured by macro F1.

False negative sample rate: Following (Chen et al., 2021), we calculate the false negative rate in SimCSE-KCL for the three datasets. The calculated false negative rates are 0.036 (Automotive), 0.068 (Electronics) and 0.102 (Beauty), showing that there are significant number of false negative samples when using KCL.

Performance with long-tailed IC: The experimental results are shown in the left part in Table 1. We can observe that all contrastive learning-based models outperform BERT-CE and cRT in terms of macro F1, which suggests the effectiveness of contrastive learning to address the long tail issue in IC. Although cRT also uses two-stage training and show success in some computer vision tasks, its performance on IC is not as expected.

When comparing the SimCSE$_{us}$ with the three supervised contrastive methods, SimCSE-KCL,

[1] The details of statistics and label frequency are in Appendix.

[2] The code will be available.

	Automotive		Electronics		Beauty		Auto$_H$	Auto$_M$	Auto$_T$
	$F1_w$	$F1_m$↑	$F1_w$	$F1_m$↑	$F1_w$	$F1_m$ ↑	$F1_m$↑	$F1_m$↑	$F1_m$↑
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
SimCSE-KCL	76.87	65.17	65.18	53.39	71.44	58.26	74.99	65.06	55.36
SimCSE-KCL-FNA	76.54	64.65	66.08	**53.69**	**71.65**	**58.31**	74.46	64.88	54.53
SimCSE-KCL-FNE	77.96	**65.82**	65.73	53.67	71.43	57.95	**75.97**	**65.78**	**55.61**

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.

SimCSE-KCL-FNA, and SimCSE-KCL-FNE, we can find at least one supervised contrastive methods can beat the SimCSE$_{us}$ and in most cases SimCSE$_{us}$ is the worst model. It illustrates that introducing semantics information can boost the model performance. However, the way of introducing the semantics information should be carefully chosen.

Moreover, we can observe that the false negative cancelling contrastive loss outperforms all baselines including SimCSE-KCL in terms of macro F1. This pattern suggests the necessity and effectiveness to eliminate the influence of the false negative samples in SimCSE-KCL. When comparing the two false negative cancelling strategies, we can find the SimCSE-KCL-FNE works better on Automotive and Electronics datasets, while SimCSE-KCL-FNA works better on the Beauty dataset. One possible reason is that the different false negative rates of the three datasets. The beauty dataset is much larger than it of other two datasets. Therefore the SimCSE-KCL-FNE loss will eliminate too many samples and further degrade the performance rather than improve it.

Performance on Subsets of Automotive: To investigate the performance of the models on the classes with different label frequencies, we split the whole Automotive dataset into three subsets: Auto$_H$, Auto$_M$ and Auto$_T$ and evaluate the models by macro F1. Auto$_H$ consists of $132,590$ samples in the most frequent 318 head classes. Auto$_T$ is the subsets including $7,855$ samples in the least frequent 317 tail classes. Auto$_M$ includes the remaining $20,280$ samples in the 318 medium classes.

As shown in the right part in Table. 1, SimCSE-KCL-FNE outperforms all other models on all three subsets and the improvement is more significant in the tail classes, showing that the false negative elimination and contrastive learning do address the long tail issue. In addition, the performance decreases as the decrease of the label frequencies for all the models, illustrating the lacking of samples limits the model performance.

5 Conclusion

In modern large-scaled item categorization tasks, category labels are naturally distributed in a long tail pattern. This issue challenges the tail labels' classification performance 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 two cancellation strategies to reduce the corresponding influences. The experimental results prove that the proposed method helps improve the performance on long-tailed data and the false negative cancellation can help boost the performance compared with KCL in IC.

For future research, there are several possible directions: (1) more sophisticated text augmentation in the CL stage, (2) more useful negative samples, e.g., focusing on *hard negative samples*, and (3) applications to more e-commerce NLP tasks, e.g., product attribute extraction.

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)*.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, et al. 2020. Bootstrap your own latent: A new approach to self-supervised learning. *arXiv preprint arXiv:2006.07733*.

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.

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*.

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.

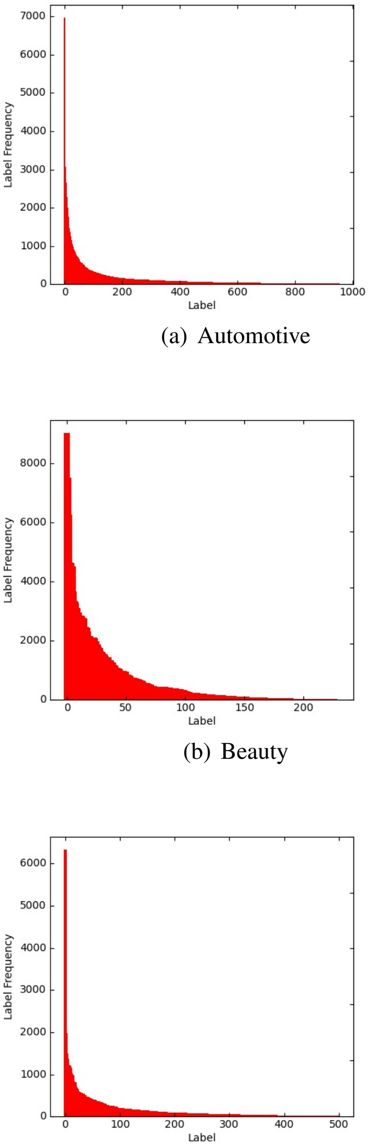
Yuzhe Yang and Zhi Xu. 2020. Rethinking the value of labels for improving class-imbalanced learning. *arXiv preprint arXiv:2006.07529*.

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 Data Statistics and Label Frequency Plots

	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 1: Statistics of Datasets



(a) Automotive

(b) Beauty

(c) Electronics

Figure 1: Label Frequency Histogram of Automotive, Beauty and Electronics Dataset

The data statistics of the three datasets are shown in Table. 1. In Fig. 1, the histogram of the label frequencies of the three datasets are shown. All the three datasets have the long-tailed issue.

B False Negative Calculation

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.

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.