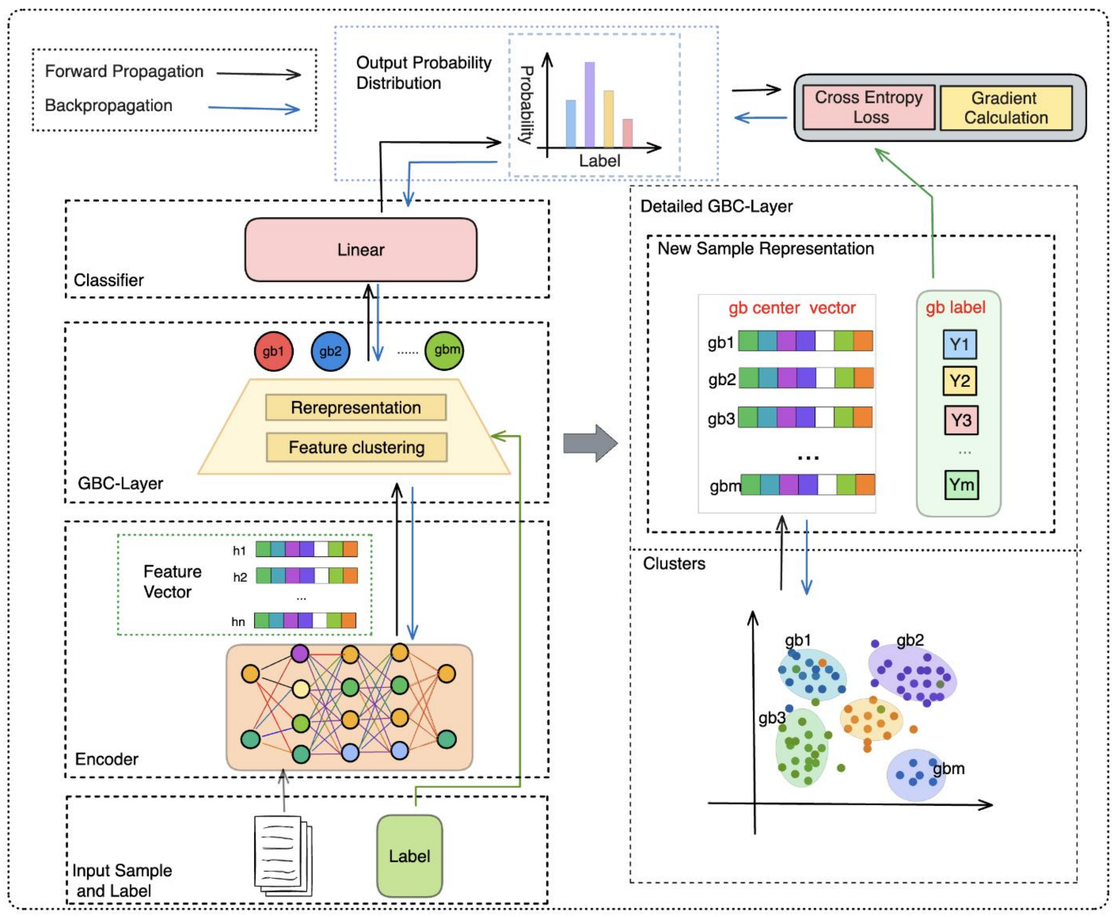
GB_RAIN: Combating Textual Label Noise by Granular-ball based Robust Training

Zeli Wang, Tuo Zhang, Shuyin Xia, Longlong Lin, Guoyin Wang

• ABSTRACT:

Most natural language processing tasks rely on massive labeled data to train an outstanding neural network model. However, the label noise (i.e., wrong label) is inevitably introduced when annotating large-scale text datasets, which significantly degrades the performance of neural network models. To overcome this dilemma, we propose a novel Granular-Ball based t$_{RAIN}$ing framework, named GB_{RAIN}, to realize robust coarse-grained representation learning, thus combating label noises in diverse text tasks. Specifically, considering that most samples in the dataset are precisely labeled, GB_{RAIN} first proposes a dynamic granular-ball clustering algorithm to blend seamlessly into the traditional neural network model. A striking feature of the clustering algorithm is that it can adaptively group the embedding vectors of similar data into the same set (hereafter referred to as a granular-ball). The embedding vectors and labels of all samples from the same set will be coarse-grainedly represented by the center vector and the label of the granular-ball, respectively. Consequently, noise labels can be rectified through the labels of most of the labeled data. Moreover, we introduce a new gradient backpropagation mechanism compatible with our framework, which can help optimize coarse-grained embedding vectors with iterative training. Empirical results on text classification and name entity recognition tasks demonstrate that our proposal GB_{RAIN} is indeed effective in contrast to the state-of-the-art baselines.

• OVERVIEW:

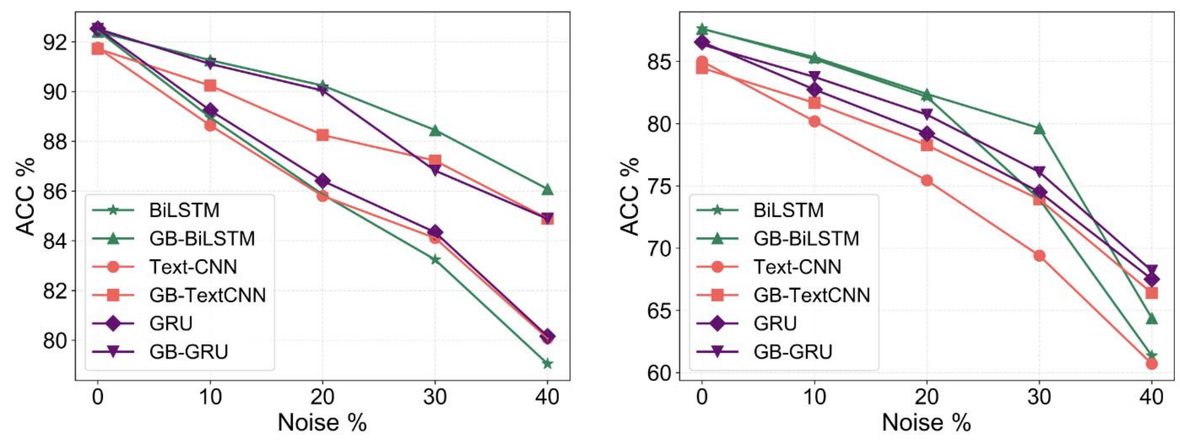


Innovative points

- A novel coarse-grained representation learning method is proposed, which enhances the model's representation capability by clustering similar data into granular balls and using the center vector and label of each granular ball to roughly represent all the samples.
- A new gradient backpropagation mechanism is proposed, compatible with this framework, which can help optimize coarse-grained embedding vectors through iterative training, thereby further enhancing the model's performance.

Result

- This article conducted relevant experiments in two mainstream NLP tasks, random noise experiments on text classification tasks, and remote supervision experiments on named entity recognition tasks. Compared with the baseline model, the method proposed in this article achieves the best results.



(a) Accuracy rate in Agnews dataset (b) Accurate rate in IMDB dataset

Summary/Conclusion:

In this paper, we propose a new Granular-Ball based t_RAINing framework, named GB_RAIN to mitigate the impact of label noises on the performance of neural network models. The main advantage of GB_RAIN is its ability to learn coarse-grainedly the embedding vectors for all samples and then adaptively rectify their labels. On top of that, GB_RAIN also introduces a new backpropagation mechanism to further refine the coarse-grained embedding vectors with iterative training. Experimental results on entity recognition and text classification tasks show that our proposed solutions significantly reduce the impact of label noises on the traditional neural network model when contrasted with the state-of-the-art baselines.

Chongqing University of Posts and Telecommunications