Self-Evolution Learning for Mixup: Enhance Data Augmentation on Few-Shot Text Classification Tasks

Haoqi Zheng[1*], **Qihuang Zhong**[2*], **Liang Ding**[3], **Zhiliang Tian**[1†],
Xin Niu[1†], **Changjian Wang**[1] , **Dongsheng Li**[1], **Dacheng Tao**[4]
[1]College of Computer, National University of Defense Technology
[2]School of Computer Science, Wuhan University [3]JD Explore Academy [4]University of Sydney

Abstract

Text classification tasks often encounter few-shot scenarios with limited labeled data, and addressing data scarcity is crucial. Data augmentation with mixup merges sample pairs to generate new pseudos, which can relieve the data deficiency issue in text classification. However, the quality of pseudo-samples generated by mixup exhibits significant variations. Most of the mixup methods fail to consider the varying degree of learning difficulty in different stages of training. And mixup generates new samples with one-hot labels, which encourages the model to produce a high prediction score for the correct class that is much larger than other classes, resulting in the model's over-confidence. In this paper, we propose a self-evolution learning (SE) based mixup approach for data augmentation in text classification, which can generate more adaptive and model-friendly pseudo samples for the model training. SE caters to the growth of the model learning ability and adapts to the ability when generating training samples. To alleviate the model over-confidence, we introduce an instance-specific label smoothing regularization approach, which linearly interpolates the model's output and one-hot labels of the original samples to generate new soft labels for label mixing up. Through experimental analysis, experiments show that our SE brings consistent and significant improvements upon different mixup methods. In-depth analyses demonstrate that SE enhances the model's generalization ability.

1 Introduction

Recently, generative large language models (LLMs) have won great popularity in natural language processing (NLP), and have achieved impressive performance on various NLP tasks (Kocoń et al., 2023; Peng et al., 2023; Lu et al., 2023c). However, empirical studies (Zhong et al., 2023) suggest that LLMs do not always outperform BERT in some language understanding tasks. Hence, employing BERT is still a viable option in some applications. Text classification tasks often encounter few shot scenarios (e.g. NLI and Paraphrase tasks), where there are limited suitable labeled data available for training. Data augmentation (DA) generates new data by changing the original data through various methods, which enlarges the training dataset to alleviate the issue of data scarcity.

In text classification tasks, DA methods can be divided into two categories: DA methods like EDA (Wei and Zou, 2019), Back-Translation (Kobayashi, 2018), and others based on synthesis such as mixup. The first category conducts DA by altering only the inputs. These methods only alter the inputs to generate new data while maintaining the original labels. These methods are easy to implement, but the input only changes a little thus leading to augmented inputs with limited diversity, which may reduce model generalization. The second category of DA methods modify both inputs and labels, which changes the input samples in a certain way and simultaneously changes the corresponding labels to compose a new sample. These methods tend to generate samples more distinct from the original samples.

Mixup is a DA method that modifies both inputs and labels. It mixes up inputs of samples and their labels, where labels are commonly represented with one-hot encoding. Most of these methods mix up inputs of two samples on their input text (Yun et al., 2019) or hidden-level representations (Verma et al., 2019). However, the pseudo sample, simply combined with two samples, may not be adaptive to the model's learning ability and friendly to the model training. Recently, some work (Sawhney et al., 2022; Park and Caragea, 2022)

[*] Haoqi Zheng and Qihuang Zhong contribute equally to this work.
[†] Corresponding Authors

Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, pages 8964–8974
December 6-10, 2023 ©2023 Association for Computational Linguistics