Deep Active Learning for Text Classification with Diverse Interpretations

Qiang Liu[1,2], Yanqiao Zhu[1,2], Zhaocheng Liu[3], Yufeng Zhang[1], and Shu Wu[1,2,*]
[1]Center for Research on Intelligent Perception and Computing, Institute of Automation, Chinese Academy of Sciences
[2]School of Artificial Intelligence, University of Chinese Academy of Sciences [3]RealAI
{qiang.liu, shu.wu}@nlpr.ia.ac.cn, {yanqiao.zhu, yufeng.zhang}@cripac.ia.ac.cn, lio.h.zen@gmail.com

ABSTRACT

Recently, Deep Neural Networks (DNNs) have made remarkable progress for text classification, which, however, still require a large number of labeled data. To train high-performing models with the minimal annotation cost, active learning is proposed to select and label the most informative samples, yet it is still challenging to measure informativeness of samples used in DNNs. In this paper, inspired by piece-wise linear interpretability of DNNs, we propose a novel Active Learning with DivErse iNterpretations (ALDEN) approach. With local interpretations in DNNs, ALDEN identifies linearly separable regions of samples. Then, it selects samples according to their diversity of local interpretations and queries their labels. To tackle the text classification problem, we choose the word with the most diverse interpretations to represent the whole sentence. Extensive experiments demonstrate that ALDEN consistently outperforms several state-of-the-art deep active learning methods.

CCS CONCEPTS

• **Computing methodologies** → **Natural language processing**; **Active learning settings**; *Neural networks.*

KEYWORDS

Active learning; text classification; diverse interpretations

ACM Reference Format:
Qiang Liu, Yanqiao Zhu, Zhaocheng Liu, Yufeng Zhang, and Shu Wu. 2021. Deep Active Learning for Text Classification with Diverse Interpretations. In *Proceedings of the 30th ACM International Conference on Information and Knowledge Management (CIKM '21), November 1–5, 2021, Virtual Event, QLD, Australia.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3459637.3482080

1 INTRODUCTION

In recent years, Deep Neural Networks (DNNs) have achieved the state-of-the-art supervised performance in numerous research tasks. Among them, a typical task in natural language processing is text classification, where deep models such as Convolutional Neural Networks (CNNs) [14] and Recurrent Neural Networks (RNNs) [28] are often adopted. However, such deep models require a large number of labeled samples, which are expensive and labor-consuming to obtain in real-world applications. Fortunately, active learning, which aims to identify and label the most informative samples from a pool of unlabeled data to train deep models with limited labels, is a promising approach to relieve this problem [1, 3, 29, 33].

Existing works on active learning mainly select samples based on uncertainty and diversity. Taking Expected Gradient Length (EGL) [12] as an example, it computes the sample uncertainty as the norms of gradients of losses with respect to the model parameters. Following EGL, EGL-Word [33] selects the word with the largest EGL among all samples to query its label so as to maximize the model performance for text classification. In addition, Bayesian Active Learning by Disagreement (BALD) [6] measures the uncertainty according to the probabilistic distribution of the model output via Bayesian inference, where an approximation by dropout is usually incorporated [5]. On the other hand, to measure the diversity of samples, some works define the active learning task as a CORESET problem [24] and uses the embedding of the last layer in deep models as the representation of samples. There are also attempts to trade off between uncertainty and diversity [13, 29]. For example, Batch Active learning by Diverse Gradient Embeddings (BADGE) [1] can be viewed as a combination of EGL and CORESET. Meanwhile, there are empirical experiments to evaluate above approaches on text classification [3, 21, 26, 30].

Recently, the interpretability of DNNs has received increasingly attention, among which most works focus on local piece-wise interpretability [2, 22]. To be specific, previous works [2, 10, 17] investigate the local interpretability of DNNs and show that a deep model with piece-wise linear activations, e.g., Maxout [8] and the family of ReLU [7, 18], can be regarded as a set of numerous local linear classifiers. The linear separable regions corresponding to these linear classifiers can be determined by the local piece-wise interpretations in DNNs that are calculated via gradient backpropagation [15, 23, 27, 32] or feature perturbation [4, 9]. In other words, samples used in a DNN could be divided into numerous linearly separable regions according to their local interpretations and samples in the same linearly separable region are classified by the same local linear classifier [2]. Therefore, fitting a DNN model is roughly equivalent to fitting all the linear classifiers in different linearly separable regions. Inspired by this, we propose to actively select samples in different linearly separable regions with the maximally diverse local interpretations, so that linear classifiers in different linearly separable regions can be all well trained.

In this paper, we propose a novel Active Learning with DivErse iNterpretations (ALDEN) approach for text classification. In our

*To whom correspondence should be addressed.

CIKM '21, November 1–5, 2021, Virtual Event, QLD, Australia

ACM ISBN 978-1-4503-8446-9/21/11…$15.00
https://doi.org/10.1145/3459637.3482080