Multi-Label text classification (MTC)

1st Danlei Zhu
BNU-HKBU United International College, Zhuhai, China
2130201617

2nd Siyu Teng
BNU-HKBU United International College, Zhuhai, China
2130201702

3rd Junhao Huang
BNU-HKBU United International College, Zhuhai, China
2130201710

Abstract—Multi-Label text classification (MTC) is a well-known NLP task that can be applied in many fields, for example, news annotation, web page tagging, and item categorization. Because of the unequal distribution of datasets, MTC is difficult to get better results. As in the dataset we used in NLP WorkShop5, there are significantly more positive reviews with scores of 4 and 5 than negative reviews with scores of 1, 2 and 3. The imbalance between positive and negative data is a great difficulty for MTC, and the smaller amount of data is called ”Conner Cases”. In order to solve the Extreme multi-label text classification (XMTC) task, we use a new label tree-based deep learning model for XMTC. First, a multi-label attention mechanism, enabling raw text as input, is used to capture the most relevant part of text to each label. Second, a shallow and wide probabilistic label tree (PLT) is adopted to handle millions of labels, which can improve the performance on ”tail labels”.

Index Terms—Text classification, NLP, deep learning, multi-label.

I. Introduction

Natural Language Processing (NLP), as a multidisciplinary field of linguistics, computer science, and artificial intelligence, has received extensive research attention due to the rapid development of machine learning methods recently. Compared to the traditional statistical NLP, deep neural network shows impressive suitability in NLP, which can achieve state-of-the-art results in many natural language tasks, especially for extreme multi-label text classification (XMTC) task. Due to recent development of the Internet, the application scenarios of MTC are prevalent in our daily life, such as item categorization in Taobao, news annotation, recommendation system, website tagging, and so on. The objective of XMTC is to label a given text with multiple possible labels from an extremely large-scale label set. This is distinct from the multi-class classification, which only tag each given text one single label.

During the era of big data, the data scale of the information systems has experienced massive growth. The Internet service providers have to handle millions of labels and samples. Text classification for those extreme scale datasets becomes an extremely difficult NLP task. The difficulties of XMTC mainly comes from two parts. First, the training phrase of XMTC for extreme datasets requires significant computational challenges, which brings many troubles in developing effective deep learning model for developers. Besides, millions of labels (tail labels) experience the lack of positive samples, thus decreasing the training effectiveness on these tail labels.

According to the representation of the input datasets and labels, we can classify the existing works into two categories: (1) Utilize the semantic features of the labels or input datasets as the training datasets. This type of strategy requires extra effort to extra semantic features such as bag-of-words from the input texts. (2) Directly use the raw input text as the training dataset. This method directly feed the training model with sparse vectors of the texts.

In order to improve the training accuracy for tail labels, many works [1]–[4] proposed label partitions or probabilistic label trees (PLTs) strategies to exploit the relations within labels.

II. Related work

Many works have achieved better performance on Extreme Multi-label. Generally, the approaches can be broadly categorized into five types, one-vs-all OVA approach, tree-based approach, embedding-based approach, deep learning approach and transformer approach.

a) One-vs-all OVA Approach: One-vs-all (OVA) Approach is a heuristic technique for multi-class classification utilizing binary classification algorithms that categorizes multi-class dataset into multiple binary problem. The meaningful contribution is prediction accuracy improvement. But the framework computational size and speed are the major challenges. Several works have tried overcome above disadvantages. The prior research DiSoEMC(Distributed Sparse Machines) [5] is the first work to attempts scaling up one-to-one paradigm in extreme multi-label classification problems by parallel training speed-up. ProXML [6] modified DiSEMC and focus on long-tail label prediction. Similarity, PPDSparse [7] and PD-Sparse [8] use dual sparsity to accelerate training and prediction.

b) Tree-based Approach: Tree based method is a solution in computational issue. In addition, the well known advantages are less training and prediction time by recursively splitting the labels or features. For instance, FastXML is a classifier for extreme multi-label classification that uses a nDCG-based ranking loss function [9]. It learns a hierarchy structure over the feature space rather than the label space. Similarity, the label partition can use Gini index to evaluate the performance [10]. The Parabel [1] recursively splitting the labels into two balanced groups produces each label tree. If nodes have less than M labels, they become leaves and are not partitioned further. The leaf nodes have linear 1-vsAll classifiers, one for each label in the leaf, that have only been trained on samples that have at least one leaf node label. Negative examples for training a label's classifier come from other labels in the same leaf as the provided label.

c) Embedding-based Approach: Embedding models employ a low-rank representation for the label matrix in order to a low-dimensional search for label similarity. Embedding-based approaches suppose the label matrix space can be represented by a low-dimensional latent space with comparable latent representations for related labels. For example, SLEEC (Sparse Local Embedding for Extreme Classification) [11] reduce the number of labels by embedding labels onto low dimensional space. Typically, SLEEC uses k-nearest neighbor clusters labels into small groups. In addition, The [12] introduces a standard empirical risk minimization framework by using various loss functions and regularization. However, embedding-based models generally perform worse than sparse one-vs-all techniques like PPDSparse [7] to achieve equivalent computational speedups, which might be attributed to the inefficiency of the label representation structure.

d) Deep Learning Approach: With the development of neural network architecture, many deep learning model have shown better improvement in XMC problems. XML-CNN [13] is the first work to implement deep neural network in XMC. It learns text representation by forwarding training to CNN networks. XML-CNN also contains a hidden layer to project text features onto low dimensional space in order to reduce the model computational size. However, unlike basic multi-label classification, XML-CNN only utilizes a simple fully connected layer to score all labels with binary entropy loss, making it difficult to deal with big label sets. AttentionXML [4] uses a probabilistic label tree (PLT) that can deal millions of labels instead of a simple fully connected layer for label scoring. For a single dataset, it must train on many models. AttentionXML solves this problem by multiplying the weight of the current layer model by the weight of its upper layer model, allowing the model to converge fast.

e) Transformer Approach: The NLP proposes a new concept: pre-training then fine-tuning. BERT [14] is one of the pioneering efforts whose pre-training targets include token prediction and following sentence prediction tasks. On the other hand, Transformer model outperforms existing state-of-the-art after pre-training on large-scale unsupervised corpora like Wikipedia and BookCorpus. X-transformer [15] is the first pre-trained model implementation on XMC problems. Compared with AttentionXML, X-transformer has higher accuracy. But there are two major shortcomings, firstly, the computational size of model. Secondly, negative labels sampling reduces the prediction accuracy. X-BERT [16] learns the label representations from the label and the input text. The process for fine-tuning BERT models to capture the contextual link between input text and the generated label clusters is the important component of X-BERT. An ensemble of multiple BERT models trained on heterogeneous label clusters yields best final model.

III. Methodology

IV. Results

V. Conclusions

References

[1] Y. Prabhu, A. Kag, S. Harsola, R. Agrawal, and M. Varma, "Parabel: Partitioned label trees for extreme classification with application to dynamic search advertising," in Proceedings of the 2018 World Wide Web Conference, 2018, pp. 993–1002.

[2] S. Khandagale, H. Xiao, and R. Babbar, "Bonsai: diverse and shallow trees for extreme multi-label classification," Mach. Learn., vol. 109, no. 11, pp. 2099–2119, 2020. [Online]. Available: https://doi.org/10.1007/s10994-020-05888-2

[3] H. Yu, K. Zhong, and I. S. Dhillon, "PECOS: prediction for enormous and correlated output spaces," CoRR, vol. abs/2010.05878, 2020. [Online]. Available: https://arxiv.org/abs/2010.05878

[4] R. You, Z. Zhang, Z. Wang, S. Dai, H. Mamitsuka, and S. Zhu, "Attentionxml: Label tree-based attention-aware deep model for high-performance extreme multi-label text classification," Advances in Neural Information Processing Systems, vol. 32, 2019.

[5] R. Babbar and B. Schölkopf, "Dismec: Distributed sparse machines for extreme multi-label classification," in Proceedings of the tenth ACM international conference on web search and data mining, 2017, pp. 721–729.

[6] ——, "Data scarcity, robustness and extreme multi-label classification," Machine Learning, vol. 108, no. 8, pp. 1329–1351, 2019.

[7] I. E. Yen, X. Huang, W. Dai, P. Ravikumar, I. Dhillon, and E. Xing, "Ppdsparse: A parallel primal-dual sparse method for extreme classification," in Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2017, pp. 545–553.

[8] I. E.-H. Yen, X. Huang, P. Ravikumar, K. Zhong, and I. Dhillon, "Pd-sparse: A primal and dual sparse approach to extreme multiclass and multilabel classification," in International conference on machine learning. PMLR, 2016, pp. 3069–3077.

[9] Y. Prabhu and M. Varma, "Fastxml: A fast, accurate and stable tree-classifier for extreme multi-label learning," in Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining, 2014, pp. 263–272.

[10] J. Weston, A. Makadia, and H. Yee, "Label partitioning for sublinear ranking," in International conference on machine learning. PMLR, 2013, pp. 181–189.

[11] K. Bhatia, H. Jain, P. Kar, M. Varma, and P. Jain, "Sparse local embeddings for extreme multi-label classification," Advances in neural information processing systems, vol. 28, 2015.

[12] H.-F. Yu, P. Jain, P. Kar, and I. Dhillon, "Large-scale multi-label learning with missing labels," in International conference on machine learning. PMLR, 2014, pp. 593–601.

[13] J. Liu, W.-C. Chang, Y. Wu, and Y. Yang, "Deep learning for extreme multi-label text classification," in Proceedings of the 40th international ACM SIGIR conference on research and development in information retrieval, 2017, pp. 115–124.

[14] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," arXiv preprint arXiv:1810.04805, 2018.

[15] W.-C. Chang, H.-F. Yu, K. Zhong, Y. Yang, and I. S. Dhillon, "Taming pretrained transformers for extreme multi-label text classification," in Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2020, pp. 3163–3171.

[16] W.-C. Chang, H.-F. Yu, K. Zhong, Y. Yang, and I. Dhillon, "X-bert: extreme multi-label text classification with using bidirectional encoder representations from transformers," arXiv preprint arXiv:1905.02331, 2019.