Dean&Francis
ISSN 2959-6157

Application of Knowledge Distillation in Natural Language Processing

Xinyi Xu

Department of Computer Science and Technology, Huaihua University, Hunan, 111000, China
*Corresponding author: xuxuxinyi8@gmail.com

Abstract:

As the technology of AI progresses steadily, natural language processing (NLP) has become an essential field of study in computer science and AI. It includes technologies that allow computers to comprehend and analyze as well as create human language. The introduction of huge pre-trained language models like Bidirectional Encoder Representations from Transformers (BERT) and Generative Pre-trained Transformer (GPT) has boosted the performance of models to a great extent. However, these models face challenges such as their massive parameter counts, high computational costs, and incompatibility with resource-constrained embedded devices. As one of the effective model compression methods, knowledge distillation (KD) is a method of knowledge transfer between large models and lightweight models in the form of a teacher-student paradigm that has shown considerable performance-to-efficiency benefits. This paper is a review of the fundamental technical strategies in output layer distillation, feature layer distillation and multi-teacher assisted distillation. It synthesizes materials of pertinent research papers; therefore, providing a systematic account of the current situation in this technology. It briefly describes the common application cases and experimental findings of knowledge distillation in natural language processing such as sentiment analysis, text classification, multilingual processing, named entity recognition, and web filtering. This paper ends by summarizing existing knowledge distillation applications on natural language processing and future development perspectives.

Keywords: Knowledge distillation; Natural language processing; Teacher-student framework; Pre-trained language models.

ISSN 2959-6157

1. Introduction

One of the important subfields of artificial intelligence is natural language processing, whose fundamental aim is to have machines capable of comprehending and producing human language. Recently, language models based on transformers like Bidirectional Encoder Representations (BERT) and Generative Pre-trained Transformer (GPT) have set new world record results in applications to sentiment analysis, text classification, and named entity recognition. This is because they are of huge size in terms of their parameters, and they are trained on extensive data. These models however are often full of millions to billions of parameters. Their training and inference procedures incur such a huge amount of computational power, not to mention the ease of adaptation to resource-constrained situations, such as mobile hardware and small-form boards, which severely restricts their range of practical use.

Hinton et al. [1] introduced knowledge distillation, which establishes a teacher-student framework to transfer knowledge from complex teacher models to lightweight student models. This method greatly shortens the model parameters and the computation expenses and optimizes the performance of the teacher model. Over the past years, KD was widely studied in the context of NLP, with current representative distillation models, such as DistilBERT and TinyBERT [2], that improve their efficiency and achieve better results.

In the current paper, the paper will analyze the distinctions between different distillation techniques e.g. output layer distillation, feature layer distillation and multi-teacher assisted distillation by looking at corresponding research papers without touching on the same technical paths. By conducting a systematic review of common applications in the use of knowledge distillation in NLP, the paper is able to show that there is wide applicability and transferability of the study to NLP, which has successfully been used to make NLP efficient. Lastly, it gives an overview of the present situation in knowledge distillation in NLP and lays out the future challenges and opportunities.

2. Knowledge Distillation is a Core Technology in NLP

Based on the level of knowledge transfer, knowledge distillation techniques in the NLP field can be categorized into output layer distillation, feature layer distillation, and multi-teacher assisted distillation. Each type exhibits significant differences in implementation logic, typical models, and applicable scenarios.

2.1 Output Layer Distillation

Output layer distillation is the most fundamental distillation paradigm. Its core logic involves training the student model to learn the output distribution of the teacher model ('soft labels') by minimizing the difference between them using loss functions like KL divergence, thereby transferring the teacher model's decision knowledge [3]. For instance, in sentiment analysis tasks, Salmony et al. employed BERT as the teacher model, extracting its output layer's log probabilities as soft labels to train traditional machine learning models like Naive Bayes (NB) and Support Vector Machines (SVM). Experimental results on the IMDb movie review dataset demonstrated that the distilled NB model achieved an accuracy improvement from 78.7% to 80.3%, while SVM and logistic regression models saw approximately 1% accuracy gains, validating the performance enhancement of output layer distillation for lightweight models [4]. While straightforward to implement, such methods disregard semantic information in the teacher model's intermediate layers, leading to potential knowledge gaps in complex NLP tasks.

2.2 Feature Layer Distillation

Feature layer distillation brings about the similarity of features of teacher and student models at the intermediate level. Through the simulation of these intermediary features, the student model can have a better comprehension of the important information and thus better transfer of the knowledge. Such a method is appropriate when dealing with complicated NLP problems [3]. TinyBERT by Jiao et al. uses a two stage system of general distillation and task distillation. In pretraining it makes BERT identify with features of hidden layers followed by further optimization of task-relevant features in the process of fine-tuning. This effectively simplifies the complexity of a model and enhances it in terms of runtime performance [5].

Besides, MiniLM by Wang et al. pays attention to the compression of self-attention modules. It can compress by 175 parameters at greater than 95% performance on tasks, such as text classification and question-answering by aligning teacher-student model attention distributions through deep self-attention distillation.

Whereas feature-layer distillation only subtly preserves the semantic information, it involves creating a complex multi-layer loss function, which is more difficult to implement in comparison to output-layer distillation.

2.3 Multi-Teacher-Assisted Distillation

Multi-teacher distillation is an attempt to resolve the problem of knowledge transfer failure in models with large scale disparity between teacher and student models by

introducing several teacher models/intermediate assistant model(TA). The assistant model is initially trained on distillation with the teacher model and then the trained assistant model is again trained on the student model. Taking knowledge based on the teacher as a supportive activity according to Dong et al. TAKD (Teacher Assistant Knowledge Distillation) an intermediate model is developed between the teacher model (Bidirectional Encoder Representations from Transformers, BERT) and the student model (Convolutional Neural Network, CNN) so that secondary knowledge is conveyed in stages. This method was able to increase the accuracy of the CNN model on the IMDB dataset by a factor of 0.8222 to 0.87.86 [6]. Presenting an assistant model is an effective way of increasing the accuracy with which the student model is delivered when the problem of increasing the gap between teacher and student models is faced.

Multi-Teacher Collaborative Cyclic Knowledge Distillation (MTCCKD) is used in the case of Chinese named entity recognition tasks to use collaborative training on multiple teacher models. Its 3-layered student model also reports an F1 score of 91.0 with a 30 percent smaller model size than TinyBERT, which confirms the benefits of multi-teacher models on individual tasks[7]. Although these will improve the performance of distillation, the models need to be trained which adds to the computational cost.

3. Typical Application Scenarios of Knowledge Distillation in NLP

3.1 Sentiment Analysis

One of the basic jobs in NLP is sentiment analysis. Using the two-way decision-making skills of semantic understanding as shown by BERT, the system creates a chance to use natural language processing to decide how emotional the text is. This enables the models to have deep semantic associations in text, and they can effectively detect concealed emotional attitudes.

In addition to the research conducted by Salmony et al. [4], there is a sentiment analysis method presented by Ximing Dong et al. [6]. The knowledge distilled on BERT is initially used in training the teacher model. This is an educated teacher model which in turn directs the student model. Ximing Dong et al. experimented with the IMDB dataset and proved that TAKD is more accurate in comparison with Knowledge Distillation (KD). TAKD performed 5.5 percent better than Convolutional Neural Networks (CNNs). In cases where the capability difference between teacher and student models is not large, traditional knowledge distillation may serve as additional knowledge. But in the case when the gap is considerable, knowledge distillation cannot influence the accuracy of the student model significantly. Teacher-Assisted Knowledge Distillation is a useful method to transfer the knowledge acquired by the teacher to the student model.

3.2 Text Classification

Text classification is one of the basic tasks of NLP, which requires lightweight models and real-time operation. It deals with the process of attaching a text to classifications that are already established, which may include news classification, spam detection, sentiment analysis, among others. BERT with its strong bidirectional semantic understanding has enhanced the accuracy of the classification dramatically and is now the modern day mainstream solution in text classification. In their adaptive knowledge distillation model of text classification, Chen et al. dynamically adjust the weight of Imitation weights between the teacher model and the student model, using the cosine similarity to improve the F1 score of the TextCNN model to 92.5 up to 93.1[8]. The technique does not require parameter tuning by hand as it allows flexible, dynamic parameter adjustment. Its output is better than other knowledge distillation procedures that involve optimization of parameters in model construction, which increases model flexibility.

3.3 Multilingual Processing

In such a globalized world, it goes without saying that the need to communicate in more than one language has to be met. This has brought about multilingual trained models. MBERT turned out to be the most successful multilingual NLP model among the existing ones: the distillation model of multilingual BERT (mBERT). Cross-lingual pre-training creates a universal semantic representation framework that allows universal understanding across languages and downstream task adaptation. Lin was trained on the Europarl dataset of text of 21 languages, which was reduced to a 6-layer student model based on mBERT and showed an accuracy of 98.84 on multilingual datasets[9]. Expert findings prove the validity and the excellence of pre-trained mBERT in multilingual text categorization and can be transferred into the context of inadequately selected languages. Although the model is portrayed as effective on European languages, geographical constraints do limit its use on non-European geographical areas, making it beneficial to conduct additional studies on its applicability to other non-European areas.

ISSN 2959-6157

3.4 Named Entity Recognition

Chengqiong Ye et al. [10] used the dynamic operation as an automatic way to modify the loss function and use the intermediate Transformer layer of the BERT named entity recognition model as a part of the distillation.

On the Note4 dataset of named entity type recognition, it has been shown that the four dissimulation-based lightweight model knowledge types outperformed their six-layer pruning models, giving them better performance. The Micro-F1 score was 90.97, and the student model selected was effectively imitating the teacher model in its performance but the student model retained 98.84 of the F1 score which it had before. At the same time, student model had much smaller parameter size (only 15 percent of the teacher one) and was 7 times fast in inference.

Modest pruning on the model will cause a enormous loss of learning potentials and it will mark a sizeable decrease of performance. Also, the application of Transformer layers placed at the intermediate of the teacher model enhances the semantic data. The idea of compressing the model widthwise by reducing the dimension of the intermediate hidden layers was adopted to generate a lightweight student model with a simple structure, fewer parameters, and good performance. Knowledge distillation attains model lightweighting and to a large extent performance of the original model.

3.5 Webpage Filtering

The task in web filtering is to extract the webpage contents and identify the malicious URLs. T. Vornos et al. suggested implementing large language models (LLMs) that can be used to introduce the correct specifications and apply existing distillation methods to generate smaller, more specialized student models [11]. Experiments have shown that the teacher-student training procedure created by T. Vornos et al. reduces the size of the large language models by a factor of 175 times without affecting the accuracy. Based on the fine-tuning of the large language model, knowledge distillation is an effective way to increase the computational efficiency of the process of web content filtering and better adapt the model to the task of deployment to production. It is an effective way of detecting web content security.

4. Conclusions

The present paper mentions knowledge distillation as a fundamental technology in NLP. Depending on the tier of transfer, different knowledge distillation methods are generalized, including output-layer distillation, feature-layer distillation and multi-teacher distillation that constitute a multi-level technical framework. Common ways of applying knowledge distillation in natural language processing are overviewed, such as sentiment analysis, named entity recognition, multilingual web filtering and others. A review of the related literature indicates that distilled models are superior in compressing the parameters, accelerating the inference, and maintaining the performance, which is essential in supporting through deploying the edge devices and deploying NLP in a real-time setting. Nonetheless, with the increase in the distance between the teacher and student models, the traditional knowledge distillation has a hard time transferring knowledge, and the student models cannot learn effectively. Despite the suggested techniques such as teacher-assisted models, there is still a need to ensure that the learning efficiency of student models is improved. In the future, it can be expected that the lightweight development of NLP models will continue through improvements in dynamic distillation and multimodal distillation. In the modern world of growing customization, the concept of knowledge distillation is becoming more and more burdened with the need to obtain more adaptability among different specialized functions and provide more stable and correct functioning in the most extensive application conditions. At the same time, the increased robustness and interference resistance are essential in the case of working with huge datasets. Notable issues still exist in minimizing the computational complexity, the costs of computation, and the time complexity.

References

[1] Hinton G, Vinyals O, Dean J. Distilling the knowledge in a neural network. Computer Science, 2015, 14(7): 38–39.

[2] Vakili Y Z, Fallah A, Sajedi H. Distilled BERT model in natural language processing. 14th International Conference on Computer and Knowledge Engineering (ICCKE), Mashhad, Iran, 2024: 243–250.

[3] Mei T, Zi Y, Cheng X, Gao Z, Wang Q, Yang H. Efficiency optimization of large-scale language models based on deep learning in natural language processing tasks. 2nd International Conference on Sensors, Electronics and Computer Engineering (ICSECE), Jinzhou, China, 2024: 1231–1237.

[4] Salmony M Y A, Faridi A R. BERT distillation to enhance the performance of machine learning models for sentiment analysis on movie review data. 9th International Conference on Computing for Sustainable Global Development (INDIACom), New Delhi, India, 2022: 400–405.

[5] Jiao X, Yin Y, Shang L, Jiang X, Chen X, Liu L. TinyBERT: Distilling BERT for natural language understanding. arXiv:1909.10351, 2019.

[6] Dong X, Huang O, Thulasiraman P, Mahanti A. Improved knowledge distillation via teacher assistants for sentiment

analysis. IEEE Symposium Series on Computational Intelligence (SSCI), Mexico City, Mexico, 2023: 300–305.
[7] Jin C, Yang S. Named entity recognition method based on multi-teacher collaborative cyclical knowledge distillation. 27th International Conference on Computer Supported Cooperative Work in Design (CSCWD), Tianjin, China, 2024: 230–235.
[8] Chen Z, Hu T, Chen C, Ge J, Wu C, Cheng W. An adaptive knowledge distillation algorithm for text classification. IEEE International Conference on Emergency Science and Information Technology (ICESIT), Chongqing, China, 2021: 439–442.
[9] Lin L. Multilingual text classification based on deep learning models. 11th Joint International Information Technology and Artificial Intelligence Conference (ITAIC), Chongqing, China, 2023: 1202–1205.
[10] C. Ye and A. A. Hernandez, "Knowledge Distillation Scheme for Named Entity Recognition Model Based on BERT," *2023 5th International Conference on Machine Learning, Big Data and Business Intelligence,*Hangzhou, China, 2023, pp. 10-17, doi: 10.1109/MLBDBI60823.2023.10482264.
[11] Vörös T, Bergeron S P, Berlin K. Web content filtering through knowledge distillation of large language models. IEEE/WIC International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT), Venice, Italy, 2023: 357–361.