A Distilled BERT with Hidden State and Soft Label Learning for Sentiment Classification

Shuyong Wei1,*, Defa Yu1, Chenguo Lv1*

1Department of Computer Science and Technology, Heilongjiang University, Harbin 150080, China

Email of all the authors: weisy6@163.com, weishuyong@21cn.com, ydf_mirco@163.com, cgl257@163.com

*Corresponding Author: Chenguo Lv; email: cgl257@163.com

Abstract: BERT is a pre-trained language model. Although the model is proven to be highly performant in a variety of natural language understanding tasks, its large size makes it hard to implement in practical situation where computing resource is limited. In order to improve the model efficiency of BERT for sentiment analysis task, we propose a novel distilled version of BERT. It distills knowledge from the full-size BERT model, which serves as the teacher model. The distilled model efficiently learns the last hidden state and soft label of the teacher model, which are different from previous models. We use distillation learning objective that is able to effectively transfer knowledge from the original big model to the compact model. Our model reduces BERT model size by ~40%, but retains ~98.2% of performance in sentiment classification task. Our model achieves promising results in SST-2 sentiment analysis, and outperforms previous distilled model.

1. Introduction

Pre-trained language models such as BERT[1], RoBERTa[2] and XLNet[3], have made a great deal of contribution to the performance in a variety of Natural Language Processing (NLP) tasks. The pre-trained language models are trained with unlabeled corpus and then fine-tuned on target task datasets. The general domain knowledge from pre-training dataset such as Wikipedia, gives advantages to the model and leads to gains in the downstream tasks, compared to models lacking a mechanism to learn from the general knowledge. However, the drawback is the resultant large model size, in hundred million, which makes it hard to implement BERT in practical situation with limited computing resource.

Knowledge distillation (KD)[4,5] is a technique for model compression. It has one teacher and one student models. The original model is the large one to be compressed, which is denoted as teacher model. The smaller model is the desired model with reduced model size, which is denoted as student model. The student model learns from the teacher model by a learning objective or a combination of learning objectives. Distilled Version of BERT has been proposed, such as DistilBERT[6] and tinyBERT[7].

In this paper, we use a novel learning objective to distill knowledge from BERT. The learning function has two parts. One part is learning the last hidden state of BERT, namely the output of the last transformer. The second one is learning the soft label of the BERT model. Our contributions are:

- We use a distillation learning objective that is able to more effectively transfer knowledge from the original big model to the compact model.
Our model reduces BERT model size by ~40%, but retains ~98.2% of performance in sentiment classification task.

2. Material and Methods

2.1 Dataset
SST-2 is Stanford Sentiment Treebank dataset[8]. The dataset has polarity labels showing whether the reviews are positive or negative. We use the dataset for fine-tuning the teacher model, for knowledge distillation and for fine-tuning the student model.

2.2 Models
BERT model fine-tuned by SST-2 is used as the teacher model in our work. The pre-trained model, BERT-base-cased, is downloaded from Hugging face website[9]. The models used in this study, and the training procedures are illustrated in Figure 1.

The learning objective is denoted as $L_{total}$. It is the sum of two learning objectives, denoted as $L_{hidden}$ and $L_{soft}$, respectively. $L_{hidden}$ is Kullback–Leibler divergence between $softmax\left(\frac{z_t}{T}\right)$ and $softmax\left(\frac{z_s}{T}\right)$, where $z_t$ and $z_s$ represent the outputs of the last transformer of the teacher and the student model, respectively, namely the last hidden state of them. $T$ represents temperature, which affects the smoothness of the outputs. In our study, we set $T = 1$ and use it as the standard softmax function, because our preliminary results show that other $T$ values lead to similar results. $L_{soft}$ is Kullback–Leibler divergence between $softmax\left(\frac{s_t}{T}\right)$ and $softmax\left(\frac{s_s}{T}\right)$, where $s_t$ and $s_s$ are the soft labels (logits) of the teacher and the student model, respectively.

The overall learning objective for knowledge distillation is denoted by their sum:

$$L_{total} = L_{hidden} + L_{soft}$$

where $L_{total}$ is used for compressing BERT model to a desired small compact model. Our distillation objective is different from the learning objective in DistilBERT[6], since we include the soft label learning objective for the target task. By contrast, DistilBERT uses a linear combination of three losses, namely distillation loss, masked language loss and a cosine embedding loss. In addition, we use the task dataset for both distillation and fine-tuning, rather than general-domain dataset for distillation and task.
dataset for fine-tuning as in DistilBERT. In the experimental section, it will show that our learning objective is more advantageous for the target task.

2.3 Training
The student model structure follows the general architecture and hyper-parameters of BERT and transformer [1,10]. The hidden size is 768 and maximum sequence length is 128. Firstly, we train the teacher models. We use SST-2 training set to fine-tune the 12-layer BERT model. A softmax layer is added to the top, which outputs the probability of each category, namely the possibility of polarity of the review examples. A total of 3 teacher models are trained, which are fine-tuned by 1, 2 or 4 epochs. For fine-tuning of teacher models, batch size of 32 and learning rate of 2e-5 are used.

Secondly, we use the learning objective $L_{\text{total}}$ described in the last subsection to distill BERT into a small 6-layer model. For the embedding and softmax layers, the student models use the parameters of teacher models. For the transformer layers, the student is initialized by the even layer of the teacher model. That is, the 0 to 5-th transformer layer of the student model use the parameters of the 0, 2, 4, 6, 8, 10-th layer of BERT-base-cased model. Each teacher model is distilled for 1, 2, 4, 8 or 16 epochs. The resultant small student models are saved for next-stage fine-tuning. For knowledge distillation, batch size of 128 and learning rate of 2e-5 are used.

Thirdly, we fine-tune the student models. SST-2 is used to fine-tune the 6-layer student model for 5 epochs. Early stopping is used. Three runs with random seeds are performed. The average accuracy is reported. For fine-tuning of student models, batch size of 32 and learning rate of 2e-5 are used.

3. Results and Discussion
We perform the procedures as described in the method section and evaluate the performance of the compressed model, namely the 6-layer student model.

3.1 Accuracy of distilled models
Table 1. Accuracy (%) of the compressed models on SST-2. All the student models are fine-tuned for 5 epochs and early stopping is used.

| Teacher BERT | # Fine-tuning epochs for teacher model | Distillation epochs | 1 epoch | 2 epochs | 4 epochs | 8 epochs | 16 epochs |
|--------------|---------------------------------------|--------------------|---------|----------|----------|----------|----------|
| Teacher 1    | 1 epoch                               |                    | 91.3    | 91.7     | 91.7     | 91.8     | 91.9     |
| Teacher 2    | 2 epochs                              |                    | 91.4    | 91.3     | 91.4     | 91.5     | 91.5     |
| Teacher 3    | 4 epochs                              |                    | 91.4    | 91.4     | 91.4     | 91.5     | 91.6     |

We explore the effectives of the student models with different fine-tuned teacher models and various numbers of distillation epochs. The student models are obtained through three steps: (1) teacher model fine-tuning, (2) knowledge distillation and (3) student model fine-tuning. Three teacher models are used, which are full-size BERT models fine-tuned for 1, 2, or 4 epochs, respectively. Then, each teacher model is distilled for 1, 2, 4, 8 or 16 epochs. The resultant student models are fine-tuned for 5 epochs and evaluated on the dev set of SST-2. Results are shown in Table 1. Overall, as the increase of distillation epochs, increased performances of the student models are observed for all the three teacher models. The performance gains range from 0.1% to 0.6%. Interestingly, the teacher model fine-tuned with 1 epoch has more advantage than those with more than 1 epochs. Nearly all the distilled student models learning from teacher 1 perform better than those learn from teacher 2 or 3. A possible reason is that the teacher model is overfitting for knowledge distillation if the number of epochs is greater than 1. Among all the settings we tested in Table 1, a strategy with one-epoch fine-tuning for teacher and 16-epoch distillation achieves the best performance.
3.2 More distillation epochs

Table 2. Average accuracy (%) of the compressed models. The teacher BERT is firstly fine-tuned for 1 epoch and distilled for 16, 32 or 64 epochs. Student models are then fine-tuned for 5 epochs.

| Number of epochs | Dev set accuracy | Test set accuracy |
|------------------|------------------|-------------------|
| 16               | 91.9             | 91.9              |
| 32               | 91.9             | 92.0              |
| 64               | 92.0             | 91.9              |

It is possible that more distillation epochs give rise to more performance gains. To explore whether 16 epochs are enough, we also investigate higher number of epochs for knowledge distillation. The BERT model fine-tuned for only 1 epoch is chosen as the teacher model, since it achieves best performance as shown in Table 1. We try 32 and 64 epochs. Results are shown in Table 2. Distillation for more epochs only leads to marginal increase to the performance, indicating that distillation for 16 or 32 epochs may be enough. Although distillation for 64 epochs also achieves similar performance, it is not recommended, since it is time-consuming and has the possibility of overfitting.

3.3 Comparison to previous models

Table 3. Performances of the original and compressed models on SST-2. Slash represents not reporting by previous study.

| Type              | Model             | Dev set | Test set | Number of layers | Number of parameters |
|-------------------|-------------------|---------|----------|------------------|---------------------|
| Full-size         | BERT-base[1]      | 92.7    | 93.5     | 12               | 110 million         |
| Compressed        | DistilBERT[6]     | 91.3    | /        | 6                | 66 million          |
| Compressed        | Ours              | 91.9    | 91.9     | 6                | 66 million          |

We compared our model to the original teacher model and previous distilled BERT model. Results are shown in Table 3. DistilBERT is a lighter and smaller BERT model. It distills knowledge from 12-layer BERT model with corpus of Wikipedia and Toronto ebooks. It is a general domain compact BERT model. After fine-tuning on SST-2 dataset, its validation accuracy is 91.3% on the development set. By contrast, our model uses distillation objective different from those of DistilBERT, and performs better than DistilBERT. On the dev set of SST-2 dataset, a gain of 0.7% is observed between our model and DistilBERT, demonstrating the effectiveness of our distillation objective, which is proposed in the method section. Compared to the full-size 12-layer BERT, our model reduces the model resize by about 40% (from 110 to 66 million of parameters), but retains 98.2% of performance in sentiment classification task.

4. Conclusions

We propose distillation learning objective that is able to effectively transfer knowledge from the original big model to the compact model for sentiment classification task. Our model reduces BERT model size by ~40%, but retains ~98.2% of performance in sentiment classification task.

Acknowledgement

This work is financially supported by Heilongjiang University Interdisciplinary Innovation Project for Basic Research of Heilongjiang Province (HDJCCX-201607).

References

[1] Devlin J, Chang M-W, Lee K, Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. 2018;(Mlm). http://arxiv.org/abs/1810.04805
[2] Liu Y, Ott M, Goyal N, et al. RoBERTa: A Robustly Optimized BERT Pretraining Approach. 2019;(1).
[3] Yang Z, Dai Z, Yang Y, Carbonell J, Salakhutdinov R, Le Q V. XLNet: Generalized Autoregressive Pretraining for Language Understanding. Published online 2019:1-18. http://arxiv.org/abs/1906.08237

[4] Hinton G, Vinyals O, Dean J. Distilling the Knowledge in a Neural Network. Published online 2015:1-9. http://arxiv.org/abs/1503.02531

[5] Romero A, Ballas N, Kahou SE, Chassang A, Gatta C, Bengio Y. FitNets: Hints for thin deep nets. 3rd Int Conf Learn Represent ICLR 2015 - Conf Track Proc. Published online 2015:1-13.

[6] Sanh V, Debut L, Chaumond J, Wolf T. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. Published online 2019:2-6. http://arxiv.org/abs/1910.01108

[7] Jiao X, Yin Y, Shang L, et al. TinyBERT: Distilling BERT for Natural Language Understanding. Published online 2019:1-14. http://arxiv.org/abs/1909.10351

[8] Socher R. Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank. Empir Methods Nat Lang Process. 2004;(October):1631-1642. doi: 10.1371/journal.pone.0073791

[9] Wolf T, Debut L, Sanh V, et al. HuggingFace’s Transformers: State-of-the-art Natural Language Processing. Published online 2019. http://arxiv.org/abs/1910.03771

[10] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. Adv Neural Inf Process Syst. 2017;2017-Decem(Nips):5999-6009.