One-Teacher and Multiple-Student Knowledge Distillation on Sentiment Classification

Xiaoqin Chang[1] **Sophia Yat Mei Lee**[2] **Suyang Zhu**[1*]
Shoushan Li[1] **Guodong Zhou**[1]
[1]Natural Language Processing Lab, Soochow University, China
[2]Department of Chinese and Bilingual Studies, The Hong Kong Polytechnic University
xqchang@stu.suda.edu.cn
{syzhu, samlee, gdzhou}@suda.edu.cn
ym.lee@polyu.edu.hk

Abstract

Knowledge distillation is an effective method to transfer knowledge from a large pre-trained teacher model to a compacted student model. However, in previous studies, the distilled student models are still large and remain impractical in highly speed-sensitive systems (e.g., an IR system). In this study, we aim to distill a deep pre-trained model into an extremely compacted shallow model like CNN. Specifically, we propose a novel one-teacher and multiple-student knowledge distillation approach to distill a deep pre-trained teacher model into multiple shallow student models with ensemble learning[1]. Moreover, we leverage large-scale unlabeled data to improve the performance of students. Empirical studies on three sentiment classification tasks demonstrate that our approach achieves better results with much fewer parameters (0.9%-18%) and extremely high speedup ratios (100X-1000X).

1 Introduction

Sentiment classification is a task of classifying a text into sentimental orientation categories, such as *positive* and *negative*, and this task plays an important role in natural language processing (NLP) and benefits many real applications (Clavel and Callejas, 2016; Shen et al., 2018; Wang et al., 2019).

The past few years have witnessed the prevailing of deep pre-trained models on sentiment classification (Yu and Jiang, 2019; Ke et al., 2021; Chen et al., 2021). However, despite their significant improvements over non-pre-trained models like CNN and LSTM, their need for a large number of computing resources and relatively long inference time becomes a major bottleneck for real-world applications.

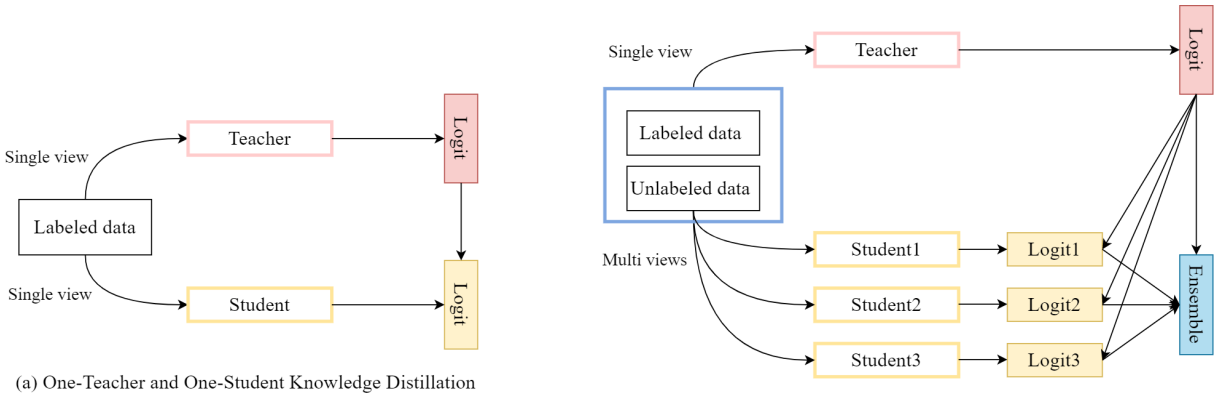
To solve this problem, knowledge distillation, which transfers knowledge from a large model (the teacher) to a smaller model (the student), is gaining popularity for reducing the computing and time costs. However, existing distilled models' parameters are still too large for some low-end devices and speed-sensitive applications. For instance, sentiment classification is usually an essential component of an information retrieval (IR) system (Paltoglou and Thelwall, 2010; Kauer and Moreira, 2016), in which users are highly speed-sensitive to the responding speed. Thus, improving the inference speed of the pre-trained model on sentiment classification becomes a critical element of applying the pre-trained sentiment classification model to IR applications. Motivated by the above, in this paper, we aim to propose a novel distillation technique that can distill a huge-parameterized pre-trained model like BERT (Devlin et al., 2019) into a minimal-parameterized non-pre-trained model like CNN or LSTM.

In principle, compared with traditional shallow models, deep pre-trained models have two major advantages. **First**, most pre-trained models are architectured in the manner of ensemble learning. In the literature, ensemble learning has been proven to be effective in performance boosting. For instance, BERT combines multiple Transformer layers under an ensemble architecture. Meanwhile, each Transformer layer contains the attention ensemble by using multiple self-attention heads (Vaswani et al., 2017).

Second, a pre-trained model highly benefits from the knowledge contained in unlabeled data. For instance, the BERT-base model is pre-trained using unlabeled data containing 3.3 billion words from Wikipedia and BooksCorpus. At least, the large scale of unlabeled data enables the pre-trained model to handle unknown words in a downstream task more easily. For instance, given a sentence *"Everything is awesome!"*, suppose that

*Corresponding author
[1]Our code is available at https://github.com/strive-hhh/OTMS-KD

Proceedings of the 29th International Conference on Computational Linguistics, pages 7042–7052
October 12–17, 2022.



(a) One-Teacher and One-Student Knowledge Distillation in Most Previous Studies

(b) One-Teacher and Multiple-Student Knowledge Distillation in Our Study

Figure 1: (a): The framework of most existing approaches; (b): The framework of our approach.

the word *"awesome"* is not observed in the training data in a sentiment classification task. Thus, non-pre-trained shallow models like CNN-based or LSTM-based classifiers cannot easily determine the sentiment orientation of *"awesome"*. On the contrary, however, pre-trained models like BERT can infer the meaning of *"awesome"*, which is close to the observed word *"excellent"*.

In this paper, inspired by the above, we propose a novel ensemble knowledge distillation approach by leveraging both ensemble learning and unlabeled data. Specifically, first, we use multiple shallow models, together with their ensemble model, as student models during distillation in order to take advantage of ensemble learning. Thanks to many previous studies on multi-view learning on sentiment classification, multiple student models could be easily obtained by using various kinds of multiple views, such as multiple types of embeddings (Ren et al., 2016) and multiple languages (Fei and Li, 2020). Second, we leverage large-scale unlabeled sentiment classification corpora during distillation. Different from most previous studies on one-teacher and one-student knowledge distillation, we propose a one-teacher and multiple-student ensemble distillation framework, which is illustrated in Figure 1.

Empirical studies on three sentiment classification tasks demonstrate that our approach outperforms the pre-trained teacher models with much fewer parameters (0.9%-18%) and extremely high speedup ratios (100X-1000X).

The remainder of this paper is organized as follows. Section 2 provides an overview of the related studies on sentiment analysis and knowledge distillation. Section 3 explains the details of our approach. Section 4 introduces the experimental settings and results. Section 5 states the conclusion and the future work.

2 Related Works

2.1 Sentiment Classification

In the last decade, the studies of sentiment classification have been dominated by neural network approaches. This line of research begins with developing sentiment classification models with shallow models, such as CNNs (Rakhlin, 2016; Johnson and Zhang, 2015), RNNs (Castellucci et al., 2014; Tang et al., 2015), and LSTM (Tai et al., 2015). Thereafter, some studies incorporate other methods into shallow models, such as attention methods (Yang et al., 2017; Liu and Zhang, 2017; Zeng et al., 2019) and graph neural networks, e.g., GCN (Marcheggiani and Titov, 2017; Vashishth et al., 2019).

Recently, deep pre-trained neural network models are becoming popular due to their highly promising performances. Large-scale pre-trained language models, such as ELMo (Peters et al., 2018), BERT (Devlin et al., 2019; Gao et al., 2019), and RoBERTa (Liu et al., 2019), have been shown to be rather effective for sentiment classification tasks with the learning paradigm of fine-tuning. More recently, pre-training models with the learning paradigm of prompt have become popular in some zero-shot or few-shot natural language processing tasks, among which sentiment classification is a classic and important task (Schick et al., 2020; Schick and Schütze, 2020; Gao et al., 2020).

2.2 Knowledge Distillation

In the field of natural language processing, the majority of the previous work on knowledge distillation has attempted to reduce the depth of BERT. For instance, Tang et al. (2019) propose compressing BERT models into a small LSTM model. Sun et al. (2019) introduce the Patient Knowledge Distillation approach to compress a large model into an equally effective lightweight shallow network. Sanh et al. (2019) develop a general-purpose pre-trained version of BERT called DistilBERT. Jiao et al. (2019) also propose a compact model called TinyBERT based on a new two-stage learning framework that captures both the general domain and task-specific knowledge in BERT. Zhou et al. (2021) suggest training a light named entity recognition using novel multi-grained knowledge distillation techniques. Instead of reducing the depth of BERT, Sun et al. (2020) attempt to reduce its width and develop a deep and thin model called MobileBERT. Unlike the existing fixed-size BERT compression models, Hou et al. (2020) introduce the DynaBERT model which can adjust the size and latency by selecting the sub-networks with different depth and width.

More recently, Reich et al. (2020) and Wu et al. (2021) propose multiple-teacher and one-student knowledge distillation frameworks for pre-trained language model compression. Besides, in the research field of computer vision, teacher-free ensemble distillation approaches, i.e., zero-teacher and multiple-student approaches, have been proposed in (Chen et al., 2020; Guo et al., 2020; Walawalkar et al., 2020; Li and Wang, 2019).

Different from the above studies, this paper introduces a novel ensemble knowledge distillation approach with the paradigm of one-teacher and multiple-student, harnessing the power of multiple shallow student models during distillation. To the best of our knowledge, this is the first work to research the ensemble knowledge distillation paradigm of the one-teacher and multiple-student model.

3 Methodology

In this section, we introduce the details of our approach.

3.1 Problem Description

Let D be a dataset and it contains both labeled data and unlabeled data where $D_l = \{x, y\}$ is labeled data and $D_u = \{x_u\}$ is unlabeled data. Our one-teacher and multiple-student knowledge distillation approach aims to distill the knowledge from a pre-trained teacher model $f(x; \theta^t)$ into an ensembled student model $g(x; \theta^s)$, where θ^t and θ^s are the model parameters of the teacher and the student respectively. In contrast to most existing studies, the parameters of the ensemble student model are much fewer than those of the teacher model, (i.e., $|\theta^s| << |\theta^t|$).

In this study, we apply our approach to three different types of sentiment classification tasks, i.e., supervised sentiment classification, zero-shot sentiment classification, and cross-lingual sentiment classification.

3.2 One Teacher Model

The teacher model is trained in different manners according to different types of sentiment classification tasks.

Supervised or Cross-lingual Sentiment Classification: In the supervised or cross-lingual sentiment classification task, we train the teacher model in a supervised manner. Let $\{x_i, y_i\}_{i=1}^N$ be a training set which contains N labeled training samples. x_i and y_i denote the ith input sample of the teacher and its gold label, respectively.

Following the work of Sun et al. (2019), the teacher model first computes the embedding $\mathbf{h}_i^t = f(x_i; \theta^t)$ of x_i where f represents the function of the teacher network. Then the teacher model feeds $\mathbf{h}_i^t$ into a linear layer and a softmax activation function to obtain the predicted sentiment label of x_i, i.e.,

$$\hat{y}_i = P^t(y_i|x_i) = softmax(\mathbf{W}^t \mathbf{h}_i^t) \quad (1)$$

where the superscript t means "teacher" model, $\mathbf{W}^t$ denotes the weight matrix to be learned in the linear layer. The tuned parameters of the teacher model can be represented as follows:

$$\hat{\theta}^t = arg \min_{\theta^t} \sum_{i=1}^{N} L_{CE}^t(x_i, y_i; [\theta^t, \mathbf{W}^t]) \quad (2)$$

where L_{CE}^t denotes the cross-entropy loss function applied in teacher's training. Then, the teacher model predicts samples in unlabeled data with soft labels, i.e.,

$$\begin{aligned} \hat{y}_u &= P^t(y_u|x_u) = softmax(\frac{\mathbf{W}^t \mathbf{h}_u^t}{T}) \\ &= softmax(\frac{\mathbf{W}^t f(x_u; \hat{\theta}^t)}{T}) \end{aligned} \quad (3)$$

where $P^t(\cdot|\cdot)$ denotes the prediction probability of the teacher. $\hat{\theta}^t$ denotes the updated parameters of the teacher. T denotes the temperature during distillation.

Zero-shot Sentiment Classification: In the zero-shot sentiment classification task, following the work of Gao et al. (2020), the teacher model is a prompt-based zero-shot learner. Let ${x_u}_{u=1}^{N'}$ be N' unlabeled data. Given an unlabeled sample x_u, a rewritten input through a manual prompt template is generated as follows:

$$x_u^{prompt} = [CLS]\ x_u\ It\ was\ [MASK].\ [SEP]$$

Let $\mathcal{M} : \mathcal{Y} \rightarrow \mathcal{V}$ be a mapping from the task label space to sentiment words. In our sentiment classification task, the sentimental labels $\{0, 1\}$ are mapped into two opinion words, i.e., {*terrible* (negative), *great* (positive)}. Then x_u^{prompt} is predicted by the teacher model through predicting the probability of filling $[MASK]$ with *terrible* (or *great*), which can be considered as the probability of predicting label. Specifically, the teacher model outputs the hidden representation of $[MASK]$:

$$\mathbf{h}_{[MASK]} = f(x_u^{prompt}; \theta^t) \tag{4}$$

where $\mathbf{h}_{[MASK]}$ denotes the hidden representation of $[MASK]$. Then the probability of predicting label (i.e., the soft label) is fetched through a linear layer and a softmax activation function, i.e.,

$$\begin{aligned} P^t(y_u|x_u) &= P^t([MASK] = \mathcal{M}(y)|x_u^{prompt}) \\ &= softmax(\frac{\mathbf{W}_{\mathcal{M}(y)} \cdot \mathbf{h}_{[MASK]}}{T}) \end{aligned} \tag{5}$$

where $\mathcal{M}(y) \in \{terrible, great\}$ denotes a certain sentiment word, and $\mathbf{W}_{\mathcal{M}(y)}$ denotes the pre-trained weight of the sentiment word.

3.3 Multiple Student Model

The ensembled student model consists of k student models. Let $\mathbf{A}_u$ be the sequential input (i.e., a matrix of word embeddings of a sentence) of the students model. $g_j(\mathbf{A}_u; \theta_j^s)$ represents the function of the jth student network, where $j \in [1, k]$ and θ_j^s denotes its parameters. Each student model first computes the vectorized representation $\mathbf{h}_{uj}^s = g_j(\mathbf{A}_u; \theta_j^s)$ of x_u. $\mathbf{h}_{uj}^s$ is then fed into a linear layer to obtain the prediction probability of the jth student model, i.e.,

$$\hat{y}_{uj} = P_j^s(y_u|\mathbf{A}_u) = \mathbf{W}_j^s \mathbf{h}_{uj}^s \tag{6}$$

where $\mathbf{W}_j^s$ is the weight matrix of the jth student model to be learned, and $P_j^s(\cdot|\cdot)$ denotes the prediction probability of the jth student model. The final ensembled probability is the weighted sum of all students' outputs, i.e.,

$$P_{ensemble}^s(y_u|\mathbf{A}_u) = softmax(\frac{\sum_{j=1}^k \alpha_j^s * \hat{y}_{uj}}{T}) \tag{7}$$

where $\alpha_j^s \in [0, 1]$ denotes the weight of the prediction probability of the jth student model and subjects to $\sum_{j=1}^k \alpha_j^s = 1$. In supervised or cross-lingual sentiment classification, the weights are learnable during model training with labeled data. But in zero-shot sentiment classification, since no labeled data is available, the weights are simply set to be the same (i.e., $\alpha_1^s = \alpha_2^s = ... = \alpha_k^s$).

3.4 Model Training

The objective loss function of a one-teacher and one-student distillation model is defined as follows:

$$L_{KD}(P^t, P^s) = \sum_{i=1}^{n} T^2 D_{KL}(P_i^t || P_i^s) \tag{8}$$

where n denotes the batch size, P_i^t and P_i^s denote the prediction probabilities of the ith sample outputted by the teacher and the student, respectively. D_{KL} is the KL divergence.

Different from the above, our approach applies an ensembled knowledge distillation loss function to train the multiple student models. Specifically, the ensembled *KD* loss is computed according to the predicted probabilities of student models as well as the final ensembled probability, i.e.,

$$\begin{aligned} Loss = &(\sum_{j=1}^{k} \lambda_j L_{KD}(P^t, P_j^s)) \\ &+ \lambda_e L_{KD}(P^t, P_{ensemble}^s) \end{aligned} \tag{9}$$

where λ_j denotes the weight of *KD* loss of the jth student and λ_e denotes the weight of *KD* loss of the ensembled students.

The learned ensembled student model is finally applied for evaluating the test set.

4 Experiments

In this section, we systematically evaluate our one-teacher and multiple-student knowledge distillation approach in three types of sentiment classification tasks, i.e., supervised sentiment classification, zero-shot sentiment classification, and cross-lingual sentiment classification.

4.1 Supervised Sentiment Classification

Dataset: The data of YELP (sentence-level) (Li et al., 2018), a widely used dataset for supervised sentiment classification, is used. Specifically, 3,000, 1,000, and 1,000 balanced samples are selected as training, development, and test data. An additional 100,000 samples are selected as unlabeled data which will be leveraged in the distillation process.

Evaluation Metrics: Standard *Accuracy* and *Macro-F1* are used to evaluate the performance of sentiment classification. Besides, the parameters and the inference time of per sample on CPU are applied to evaluate the operational performance of the distilled models.

Learning Models and Parameter Settings: All hyper-parameters are tuned according to the development set. The temperature T is set to 1.0. The batch size is set to 128. The teacher model is optimized by the AdamW (Loshchilov and Hutter, 2017) optimizer, where the initial learning rate is 2e-5 and weight decay is 1e-3. The student models are optimized by the Adam (Kingma and Ba, 2014) optimizer, where the initial learning rate is 1e-3 and weight decay is 1e-4 or 1e-5. In this task, the teacher model is the pre-trained 12-layer BERT-base, which has been the most frequently-researched teacher model in previous studies in the supervised learning setting. The student model is CNN with 100 kernels of 3 different sizes, in which the embedding size is 50 and the kernel size is 3x50, 4x50 and 5x50 respectively.

Multi-view Settings: Different types of word embeddings are used as multiple views to generate different student models. Specifically, we employ three different types of Glove embeddings (Pennington et al., 2014), i.e., Glove.6B.50d, Glove.twitter.27B.50d, and Glove.42B.300d.

Baselines: For comparison, we implement the following knowledge distillation approaches.

(1) DistilBERT (Sanh et al., 2019): This approach obtains a student model by transferring knowledge from the last layer of a pre-trained BERT in both the pre-training stage and optional fine-tuning stage. This is a one-teacher and one-student distillation approach and no unlabeled data is used.

(2) TinyBERT (Jiao et al., 2019): This approach obtains a student model by transferring knowledge from BERT with a novel transformer distillation method. This is a one-teacher and one-student distillation approach and no unlabeled data is used.

(3) MobileBERT (Sun et al., 2020): This approach obtains a student model by transferring knowledge from BERT-Large in the pre-training stage. This is a one-teacher and one-student distillation approach and no unlabeled data is used.

(4) XtremeDistil (Mukherjee and Awadallah, 2020): This approach obtains a student model by transferring knowledge from a multilingual pre-trained model, by leveraging teacher representations agnostic of its architecture and stage-wise optimization schedule. Moreover, this approach employs unlabeled data to boost performance. This is a one-teacher and one-student distillation approach and unlabeled data is used.

(5) MT-BERT (Wu et al., 2021): This approach obtains a student model with TinyBERT by transferring knowledge from multiple teachers, i.e., BERT, Roberta and UniLM. This is a multiple-teacher and one-student distillation approach and no unlabeled data is used.

(6) Distilled BiLSTM (Tang et al., 2019): This approach obtains a student model with a shallow neural network BiLSTM by transferring knowledge from BERT. This is a one-teacher and one-student distillation approach and no unlabeled data is used.

(7) Distilled Single CNN and Distilled Single CNN with more parameters: This approach distills a pre-trained BERT into a CNN with unlabeled data. The model with more parameters is obtained by leveraging 275 kernels and the embedding size of 300.

For reference, apart from distillation models, we also provide the results from models including CNN, BiLSTM, and Ensembled CNNs, which are trained with the labeled data only with no knowledge distillation.

Results: As shown in Table 1, compared with a single CNN, Ensembled CNNs improves very little (0.2%) when only training data are available. However, Distilled Ensembled CNNs with unlabeled data (our approach) achieves a 3.7% improvement in both *Accuracy* and *Macro-F1*. Moreover, our approach performs better than Distilled Single CNN with more parameters, which indicates that performance gain is more the result of ensemble distillation than only distilling a larger model. Even better, our approach achieves a slightly higher classification performance compared with the teacher model.

Methods	#Params	*Accuracy*	*Macro-F1*
BERT-base (Teacher) (Devlin et al., 2019)	109.48M	0.958	0.958
DistilBERT6 (Sanh et al., 2019)	65.78M	0.952	0.952
TinyBERT4 (Jiao et al., 2019)	14.35M	0.938	0.938
TinyBERT6 (Jiao et al., 2019)	66.96M	0.956	0.956
MobileBERT (Sun et al., 2020)	24.58M	0.947	0.947
XtremeDistil (Mukherjee and Awadallah, 2020) †	12.75M	0.960	0.960
MT-BERT4 (Wu et al., 2021)	14.35M	0.945	0.945
BiLSTM (Wang et al., 2018)	2.35M	0.919	0.919
CNN (Kim, 2014)	**0.15M**	0.927	0.927
Ensembled CNNs	1.20M	0.929	0.929
Distilled BiLSTM (Tang et al., 2019)	2.35M	0.920	0.920
Distilled Single CNN †	0.47M	0.958	0.958
Distilled Single CNN with more parameters †	3.74M	0.960	0.960
Distilled Ensembled CNNs (Our approach) †	3.74M	**0.964**	**0.964**

Table 1: Performances in supervised sentiment classification. "†" denotes that this model leverages unlabeled data during distillation.

Methods	#Params	Inf. time on CPU
Teacher	109.48M	10.61ms
Our approach	**3.74M**	**0.03ms**

Table 2: Operational performance of the teacher model and our approach in supervised sentiment classification.

Operational performance: The parameters and inference times of the teacher model and our approach are given in Table 2. The proposed model has a significantly smaller size (96.6% fewer parameters) and a notably faster inference speed (353 times faster) compared with the teacher model.

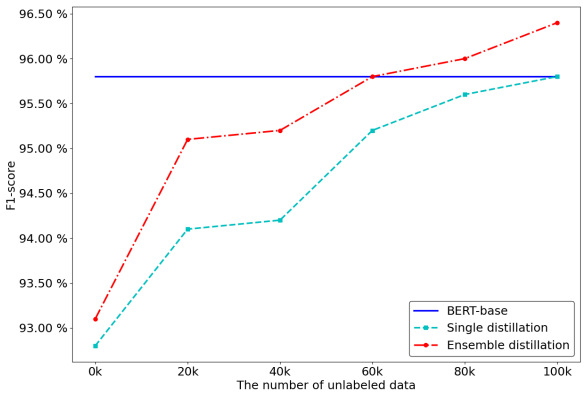


Figure 2: The influence of the scale of leveraged unlabeled data in supervised sentiment classification.

Influence of ensemble learning and leveraging unlabeled data: Figure 2 shows the influence of leveraging different scales of unlabeled data and applying ensemble learning. Both Distilled Single CNN and Distilled Ensembled CNNs perform much worse than the teacher model when no unlabeled data is available. However, Distilled Single CNN is able to achieve a highly similar performance compared with the teacher when 100k unlabeled data are leveraged. Furthermore, Distilled Ensembled CNNs surpasses the teacher when the scale of unlabeled data is over 80k.

4.2 Zero-shot Sentiment Classification

Dataset: The data of YELP (sentence-level) (Li et al., 2018) is used. Specifically, 1,000 balanced samples are selected as test data. An additional 100,000 samples are selected as unlabeled data, which will be leveraged in the distillation process. It is worthwhile to note that no training and development data is used in zero-shot sentiment classification.

Learning Models and Parameter Settings: The teacher model is the pre-trained 24-layer RoBERTa-large, which has been shown as an excellent model for zero-shot learning (Gao et al., 2020). Other parameter settings and multi-view settings are the same as supervised sentiment classification.

Baselines: Since few previous studies have conducted their research on knowledge distillation on zero-shot learning, we only implement the baseline approach of Distilled Single CNN with unlabeled data in this experiment.

Results: As shown in Table 3, Distilled Single

Methods	#Params	*Accuracy*	*Macro-F1*
RoBERTa-large (Teacher) (Liu et al., 2019)	408.98M	0.847	0.844
Distilled Single CNN †	**0.46M**	0.861	0.859
Distilled Single CNN with more parameters †	3.67M	0.872	0.874
Distilled Ensembled CNNs (Our approach) †	3.67M	**0.881**	**0.879**

Table 3: Performances in zero-shot sentiment classification. "†" denotes that this model leverages unlabeled data during distillation.

Methods	#Params	Inf. time on CPU
Teacher	408.98M	45.25ms
Our approach	**3.67M**	**0.03ms**

Table 4: Operational performance of the teacher model and our approach in zero-shot sentiment classification.

CNN outperforms the teacher model in both *Accuracy* and *Macro-F1* when 40k unlabeled samples are leveraged. Moreover, our approach outperforms the Distilled Single CNN with a 2.0% improvement in both *Accuracy* and *Macro-F1* and performs better than Distilled Single CNN with more parameters.

Operational performance: The parameters and inference times of the teacher model and our approach are given in Table 4. The proposed model has a significantly smaller size (99.1% fewer parameters) and a notably faster inference speed (1507 times faster) compared with the teacher model.

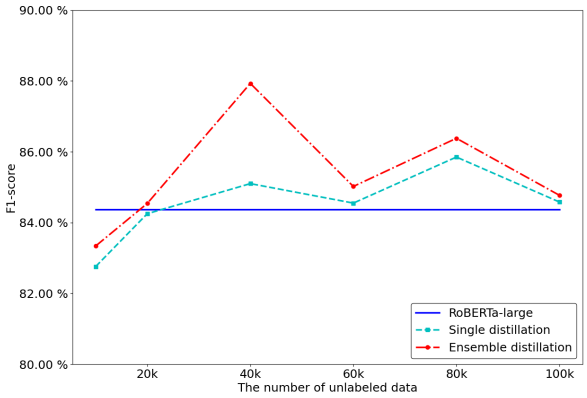


Figure 3: The influence of the scale of leveraged unlabeled data in zero-shot sentiment classification.

Influence of ensemble learning and leveraging unlabeled data: Figure 3 shows the influence of leveraging different scales of unlabeled data and applying ensemble learning. The zero-shot sentiment classification performance grows with the scale of leveraged unlabeled data when the size is less than 40k. However, the performance of our approach declines when the size of unlabeled data increases to over 60k and 100k. This might be due to the absence of a training set and validating set. Fortunately, the weaker performances of our approach are still better than those of the teacher model.

4.3 Cross-lingual Sentiment Classification

Datasets: In this experiment, Chinese is considered as the source language where labeled data is available and English is considered as the target language where only unlabeled data (with no labeled data) is available. 4,000 and 2,000 labeled Chinese samples in the document-level data of Hotel Review (Jie et al., 2016) is selected as the training and development data. 2,000 and 80,000 samples in the data of YELP (document-level) (Zhang et al., 2015) is selected as test and unlabeled data. All English (or Chinese) samples are translated into Chinese (or English) samples via Baidu Translator API.

Teacher Models: Several teacher models are designed to perform cross-lingual sentiment classification. As shown in Table 5, the first part is to employ BERT-Chinese-base and XLM-R-base(Chn) to train the teacher model with Chinese labeled data. Then, all English samples, i.e., unlabeled and test samples, are translated into Chinese for distillation and testing. The second part is to translate the Chinese labeled samples into English and then train the teacher model with BERT-English-base and XLM-R-base(Eng). Note that XLM-R (Conneau et al., 2020) is a state-of-the-art multilingual pre-training model. From Table 5, we can see that BERT-Chinese-base performs best and thus we choose it as the teacher model in the distillation experiment.

Multi-view Settings: English text, Chinese text, and together with their mixed text, are considered as three different views to generate three student models.

Baselines: Since few previous studies have con-

Methods	#Params	*Accuracy*	*Macro-F1*
BERT-Chinese-base (Devlin et al., 2019)	102.27M	**0.876**	**0.876**
XLM-R-base(Chn) (Conneau et al., 2020)	278.05M	0.855	0.854
BERT-English-base (Devlin et al., 2019)	109.48M	0.840	0.838
XLM-R-base(Eng) (Conneau et al., 2020)	278.05M	0.873	0.872

Table 5: Performances of different teacher models in cross-lingual sentiment classification.

Methods	#Params	*Accuracy*	*Macro-F1*
BERT-Chinese-base (Teacher) (Devlin et al., 2019)	102.27M	0.876	0.876
CNN (Chn) (Kim, 2014)	0.97M	0.691	0.691
CNN (Eng) (Kim, 2014)	**0.46M**	0.692	0.690
Ensembled CNNs	5.05M	0.714	0.714
Distilled Single CNN (Chn) †	1.31M	0.869	0.869
Distilled Single CNN (Eng) †	2.51M	0.868	0.868
Distilled Single CNN (Chn and Eng) †	15.36M	0.871	0.871
Distilled Ensembled CNNs (Our approach) †	18.88M	**0.878**	**0.878**

Table 6: Performances in cross-lingual sentiment classification. "†" denotes that this model leverages unlabeled data during distillation.

Methods	#Params	Inf. time on CPU
Teacher	102.27M	268.52ms
Our approach	**18.88M**	**1.22ms**

Table 7: Operational performance of the teacher model and our approach in cross-lingual sentiment classification.

ducted their research on knowledge distillation on cross-lingual learning, we only implement the baseline approach of CNN, Ensemble CNNs and Distilled Single CNN with unlabeled Data in this experiment.

Results: Table 6 shows the results of baselines and our approach in cross-lingual sentiment classification. CNN in both Chinese view and English view perform much more poorly than the teacher model, resulting in a significantly lower performance by Ensembled CNNs. However, by leveraging the unlabeled data, Distilled Single CNN in three views improve notably with an over 18% improvement in *Accuracy* and *Macro-F1* compared with a single CNN. Furthermore, our approach achieves higher *Accuracy* and *Macro-F1* than the teacher model.

Operational performance: The parameters and inference times of the teacher model and our approach are given in Table 7. The proposed model has a significantly smaller size (81.5% fewer parameters) and a notably faster inference speed (219 times faster) compared with the teacher model.

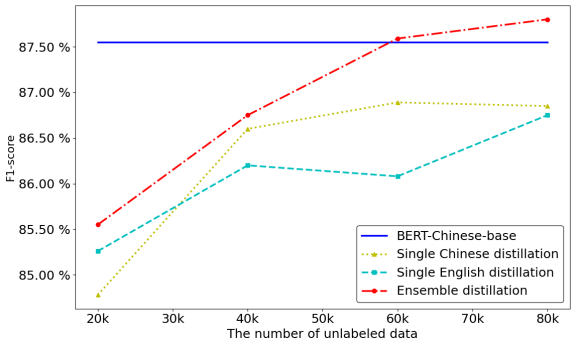


Figure 4: The influence of the scale of leveraged unlabeled data in cross-lingual sentiment classification.

Influence of ensemble learning and leveraging unlabeled data: Figure 4 shows the influence of leveraging different scales of unlabeled data and applying ensemble learning. From this figure, we can see that, in cross-lingual sentiment classification, our approach benefits greatly from unlabeled data. Moreover, distilling knowledge into ensembled CNNs results in consistently better performance than distilling knowledge into a single CNN.

5 Conclusion

In this study, we propose a novel approach of knowledge distillation, namely one-teacher and multiple-student knowledge distillation, in sentiment classification. Our approach is capable of compacting a large model into a minimal ensem-

ble model with both ensemble learning and unlabeled data. Empirical studies on three sentiment classification tasks demonstrate that the distilled model performs even better than the teacher model with much fewer parameters and a much better operational performance on CPU.

In our future work, we aim to improve our approach by carefully selecting a suitable number of unlabeled samples instead of using all of them. In addition, we would like to apply our approach to other NLP tasks besides sentiment classification.

Acknowledgements

We thank the anonymous reviewers for their insightful comments and suggestions. This research work was supported by a Major Project of Ministry of Science and Technology of China No.2020AAA0108604 and two NSFC grants, i.e., No.62076176 and No.62106166. This research work was also supported by a General Research Fund (GRF) project sponsored by the Research Grants Council, Hong Kong (Project No.15611021).

References

Giuseppe Castellucci, Simone Filice, Danilo Croce, and Roberto Basili. 2014. Unitor: Aspect based sentiment analysis with structured learning. In *Proceedings of the 8th international workshop on semantic evaluation (SemEval 2014)*, pages 761–767.

Defang Chen, Jian-Ping Mei, Can Wang, Yan Feng, and Chun Chen. 2020. Online knowledge distillation with diverse peers. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 3430–3437.

Yuezhe Chen, Lingyun Kong, Yang Wang, and Dezhi Kong. 2021. Multi-grained attention representation with albert for aspect-level sentiment classification. *IEEE Access*, 9:106703–106713.

Chloé Clavel and Zoraida Callejas. 2016. Sentiment analysis: From opinion mining to human-agent interaction. *IEEE Transactions on Affective Computing*, 7(1):74–93.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Hongliang Fei and Ping Li. 2020. Cross-lingual unsupervised sentiment classification with multi-view transfer learning. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5759–5771.

Tianyu Gao, Adam Fisch, and Danqi Chen. 2020. Making pre-trained language models better few-shot learners. *arXiv preprint arXiv:2012.15723*.

Zhengjie Gao, Ao Feng, Xinyu Song, and Xi Wu. 2019. Target-dependent sentiment classification with bert. *IEEE Access*, 7:154290–154299.

Qiushan Guo, Xinjiang Wang, Yichao Wu, Zhipeng Yu, Ding Liang, Xiaolin Hu, and Ping Luo. 2020. Online knowledge distillation via collaborative learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11020–11029.

Lu Hou, Zhiqi Huang, Lifeng Shang, Xin Jiang, Xiao Chen, and Qun Liu. 2020. Dynabert: Dynamic bert with adaptive width and depth. *arXiv preprint arXiv:2004.04037*.

Xiaoqi Jiao, Yichun Yin, Lifeng Shang, Xin Jiang, Xiao Chen, Linlin Li, Fang Wang, and Qun Liu. 2019. Tinybert: Distilling bert for natural language understanding. *arXiv preprint arXiv:1909.10351*.

HAO Jie, XIE Jun, SU Jingqiong, et al. 2016. An unsupervised approach for sentiment classification based on weighted latent dirichlet allocation. *CAAI Transactions on Intelligent Systems*, 11(4):539–545.

Rie Johnson and Tong Zhang. 2015. Semi-supervised convolutional neural networks for text categorization via region embedding. *Advances in neural information processing systems*, 28:919.

Anderson Uilian Kauer and Viviane P. Moreira. 2016. Using information retrieval for sentiment polarity prediction. *Expert Systems with Applications*, 61:282–289.

Zixuan Ke, Hu Xu, and Bing Liu. 2021. Adapting BERT for continual learning of a sequence of aspect sentiment classification tasks. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4746–4755, Online. Association for Computational Linguistics.

Yoon Kim. 2014. Convolutional neural networks for sentence classification.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Daliang Li and Junpu Wang. 2019. Fedmd: Heterogenous federated learning via model distillation. *arXiv preprint arXiv:1910.03581*.

Juncen Li, Robin Jia, He He, and Percy Liang. 2018. Delete, retrieve, generate: A simple approach to sentiment and style transfer. *arXiv preprint arXiv:1804.06437*.

Jiangming Liu and Yue Zhang. 2017. Attention modeling for targeted sentiment. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 572–577.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Ilya Loshchilov and Frank Hutter. 2017. Fixing weight decay regularization in adam. *CoRR*, abs/1711.05101.

Diego Marcheggiani and Ivan Titov. 2017. Encoding sentences with graph convolutional networks for semantic role labeling. *arXiv preprint arXiv:1703.04826*.

Subhabrata Mukherjee and Ahmed Awadallah. 2020. Xtremedistil: Multi-stage distillation for massive multilingual models. *arXiv preprint arXiv:2004.05686*.

Georgios Paltoglou and Mike Thelwall. 2010. A study of information retrieval weighting schemes for sentiment analysis. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 1386–1395, Uppsala, Sweden. Association for Computational Linguistics.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*.

A Rakhlin. 2016. Convolutional neural networks for sentence classification. *GitHub*.

Steven Reich, David Mueller, and Nicholas Andrews. 2020. Ensemble distillation for structured prediction: Calibrated, accurate, fast-choose three. *arXiv preprint arXiv:2010.06721*.

Yafeng Ren, Yue Zhang, Meishan Zhang, and Donghong Ji. 2016. Improving twitter sentiment classification using topic-enriched multi-prototype word embeddings. In *Thirtieth AAAI conference on artificial intelligence*.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Timo Schick, Helmut Schmid, and Hinrich Schütze. 2020. Automatically identifying words that can serve as labels for few-shot text classification. *arXiv preprint arXiv:2010.13641*.

Timo Schick and Hinrich Schütze. 2020. Exploiting cloze questions for few shot text classification and natural language inference. *arXiv preprint arXiv:2001.07676*.

Chenlin Shen, Changlong Sun, Jingjing Wang, Yangyang Kang, Shoushan Li, Xiaozhong Liu, Luo Si, Min Zhang, and Guodong Zhou. 2018. Sentiment classification towards question-answering with hierarchical matching network. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3654–3663, Brussels, Belgium. Association for Computational Linguistics.

Siqi Sun, Yu Cheng, Zhe Gan, and Jingjing Liu. 2019. Patient knowledge distillation for bert model compression. *arXiv preprint arXiv:1908.09355*.

Zhiqing Sun, Hongkun Yu, Xiaodan Song, Renjie Liu, Yiming Yang, and Denny Zhou. 2020. Mobilebert: a compact task-agnostic bert for resource-limited devices. *arXiv preprint arXiv:2004.02984*.

Kai Sheng Tai, Richard Socher, and Christopher D Manning. 2015. Improved semantic representations from tree-structured long short-term memory networks. *arXiv preprint arXiv:1503.00075*.

Duyu Tang, Bing Qin, Xiaocheng Feng, and Ting Liu. 2015. Effective lstms for target-dependent sentiment classification. *arXiv preprint arXiv:1512.01100*.

Raphael Tang, Yao Lu, Linqing Liu, Lili Mou, Olga Vechtomova, and Jimmy Lin. 2019. Distilling task-specific knowledge from bert into simple neural networks. *arXiv preprint arXiv:1903.12136*.

Shikhar Vashishth, Shib Sankar Dasgupta, Swayambhu Nath Ray, and Partha Talukdar. 2019. Dating documents using graph convolution networks. *arXiv preprint arXiv:1902.00175*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Devesh Walawalkar, Zhiqiang Shen, and Marios Savvides. 2020. Online ensemble model compression using knowledge distillation. In *European Conference on Computer Vision*, pages 18–35. Springer.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. 2018. Glue: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461*.

Jingjing Wang, Changlong Sun, Shoushan Li, Xiaozhong Liu, Luo Si, Min Zhang, and Guodong Zhou. 2019. Aspect sentiment classification towards question-answering with reinforced bidirectional attention network. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3548–3557, Florence, Italy. Association for Computational Linguistics.

Chuhan Wu, Fangzhao Wu, and Yongfeng Huang. 2021. One teacher is enough? pre-trained language model distillation from multiple teachers. *arXiv preprint arXiv:2106.01023*.

Min Yang, Wenting Tu, Jingxuan Wang, Fei Xu, and Xiaojun Chen. 2017. Attention based lstm for target dependent sentiment classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31.

Jianfei Yu and Jing Jiang. 2019. Adapting bert for target-oriented multimodal sentiment classification. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pages 5408–5414. International Joint Conferences on Artificial Intelligence Organization.

Jiangfeng Zeng, Xiao Ma, and Ke Zhou. 2019. Enhancing attention-based lstm with position context for aspect-level sentiment classification. *IEEE Access*, 7:20462–20471.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. *Advances in neural information processing systems*, 28:649–657.

Xuan Zhou, Xiao Zhang, Chenyang Tao, Junya Chen, Bing Xu, Wei Wang, and Jing Xiao. 2021. Multi-grained knowledge distillation for named entity recognition. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5704–5716.