IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 11 February 2024, accepted 26 February 2024, date of publication 1 March 2024, date of current version 7 March 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3372568

RESEARCH ARTICLE

Heterogeneous Student Knowledge Distillation From BERT Using a Lightweight Ensemble Framework

CHING-SHENG LIN[1], CHUNG-NAN TSAI[2], JUNG-SING JWO[1,3], CHENG-HSIUNG LEE[1], AND XIN WANG[4], (Senior Member, IEEE)

[1]Master Program of Digital Innovation, Tunghai University, Taichung 40704, Taiwan
[2]Lam Research Japan GK, Yokohama, Kanagawa 222-0033, Japan
[3]Department of Computer Science, Tunghai University, Taichung 40704, Taiwan
[4]Department of Epidemiology and Biostatistics, School of Public Health, University at Albany, State University of New York, Rensselaer, NY 12144, USA

Corresponding author: Ching-Sheng Lin (cslin612@thu.edu.tw)

This work was supported in part by the National Science and Technology Council (NSTC) of Taiwan under Grant 112-2221-E-029-019.

ABSTRACT Deep learning models have demonstrated their effectiveness in capturing complex relationships between input features and target outputs across many different application domains. These models, however, often come with considerable memory and computational demands, posing challenges for deployment on resource-constrained edge devices. Knowledge distillation is a prominent technique for transferring the expertise from an advanced yet heavy teacher model to a more efficient leaner student model. As ensemble methods have exhibited notable enhancements in model generalization and have achieved state-of-the-art performance in various machine learning tasks, we adopt ensemble techniques to perform knowledge distillation from BERT using multiple lightweight student models. Our approach applies lean architectural paradigms of spatial and sequential networks including LSTM, CNN and their fusion to perform data processing from distinct perspectives. Instead of using contextual word representations which require more space in natural language processing applications, we take advantage of a single static pre-trained and low-dimensional word embedding space to be shared among student models. Empirical studies are conducted on the sentiment classification problem and our model outperforms not only other existing techniques but also the teacher model.

INDEX TERMS Knowledge distillation, ensemble methods, BERT, LSTM, CNN, contextual word representations, pre-trained and low-dimensional word embedding space, sentiment classification problem.

I. INTRODUCTION

Over the past decade, significant advancements in artificial intelligence have transformed the world and reshaped our lives, largely attributed to the progress of neural networks through deep learning. This revolution has greatly enhanced the capacity of computers to perceive, listen, and comprehend their environments, leading to remarkable advancement in the integration of AI across various scientific disciplines and other aspects of human achievement [1]. Especially in recent years, large language models including the BERT [2] and GPT series (GPT-1 [3], GPT-2 [4], GPT-3 [5] and ChatGPT) have achieved significant success in many natural language processing tasks. Nevertheless, the computing costs, substantial space requirements and increased inference time associated with large neural networks impose limitations on their deployment on edge devices for different downstream applications. With recognition of these drawbacks, in both industry and academia, there has been a particular focus on researching and developing models that can be efficiently operated on resource-limited devices.

Various approaches have been suggested to resolve this problem such as low-rank networks [6], [7], efficient convolutional neural networks [8], [9], and pruning based methods [10], [11]. Among those efforts to build more efficient models by taking advantages of large networks, knowledge

The associate editor coordinating the review of this manuscript and approving it for publication was Maria Chiara Caschera.

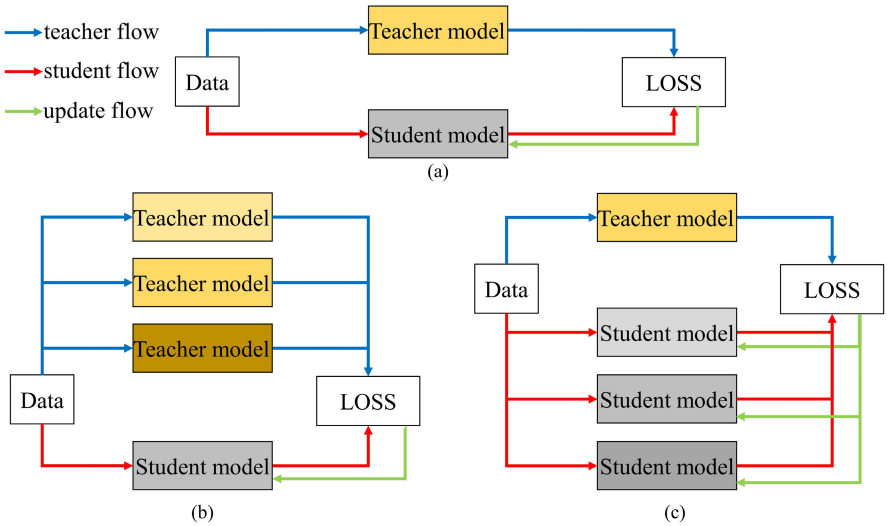


FIGURE 1. The classification of knowledge distillation: (a) one-teacher and one-student model; (b) multiple-teacher and one-student model; (c) one-teacher and multiple-student model.

distillation is a popular approach where a complex teacher model transfers its knowledge to a simpler student model for helping the student model imitate the teacher's behavior and making it more computationally efficient [12]. In general, knowledge distillation can be divided into three key categories: feature-based, response-based and relation-based distillation methods [13]. Feature-based knowledge distillation concentrates on incorporating the internal representations and features learned by the teacher model into the student model [14], [15], response-based knowledge focuses on transferring output responses generated by the teacher model to guide the student model's predictions [16], [17], and relation-based knowledge emphasizes on providing essential structural correlation information existing within the teacher model to the student model [18], [19].

Ensemble learning is a simple yet effective learning strategy that involves combining multiple models to alleviate the problem of overfitting and improve higher accuracy [20]. Many ensemble methods have consistently yielded top-tier results in numerous machine learning competitions [21]. Moreover, they also work well for knowledge distillation because of the capability of leveraging multiple and diverse models to capture different aspects of the data, resulting in a more robust and generalized student model [22].

Sentiment classification is a task of Natural Language Processing (NLP) which focuses on categorizing text into sentiment ratings. It has been applied in real-world scenarios to assess opinions and attitudes [23]. However, the state-of-the-art methods to solve this problem often adopt large language models, which are resource and time-consuming, and are difficult to implement within the constraints of real-world settings [24]. In this paper, we investigate knowledge distillation based on ensemble learning to improve model efficiency and sentiment classification performance. The traditional knowledge distillation model employs the one-teacher and one-student architecture, while the multiple-teacher and one-student model introduces collaboration among several teachers to enhance the learning process. Additionally, the one-teacher and multiple-student model facilitates the simultaneous transfer of knowledge from a single teacher to multiple students (Figure 1). Unlike prior research which explores different types of word embeddings on top of the same CNN student model structure [25], our approach makes an effort to minimize model complexity via sharing a single static pre-trained word embedding space among distinct student model architectures. By enabling knowledge transfer starting from the same compact semantic space, our multi-view fusion scheme allows complementary spatial (CNN) and sequential (LSTM) feature extractions without inflating resource demands.

The innovation and primary contributions of our proposed approach can be summarized as follows.

- Compared to large language models such as BERT which contains over 100 million parameters, our distilled ensemble model adopts lean architectural paradigms of neural networks including LSTM, CNN, and their fusion to significantly reduce parameter usage owing to their simple topology. Moreover, we employ a single static pre-trained word embedding space with only 50-dimension features to be shared among student models rather than using contextual word representations which require large parameter spaces to capture word semantics.
- Empirical studies are conducted on YELP dataset for the sentiment analysis problem. Results demonstrate that our proposed network, with substantially reduced model size, achieves the highest accuracy of 97.3% compared to other models. Specifically, our model contains only 1.69M parameters, which is about 65 times smaller than the teacher BERT model with 109 M parameters.

Our approach ensures efficient resource utilization while maintaining competitive performance.

The rest of this paper is organized as follows. In section II, we discuss the research works and corresponding applications related to knowledge distillation. Section III describes our proposed model in detail. Section IV presents the experimental evaluation and the results. Finally, we summarize our findings and suggest potential future research avenues in Section V.

II. RELATED WORK

In this section, we review several research fields relevant to our work including the knowledge distillation, ensemble learning and sentiment classification.

A. KNOWLEDGE DISTILLATION

Response-based, feature-based, and relation-based knowledge are three common types of knowledge distillation techniques used in machine learning. The fundamental idea of response-based methods is to directly emulate the final output of the teacher network. The soft target distribution, which is the probabilistic and continuous output generated by the teacher model, is proposed to guide the training of the student model by minimizing the temperature cross entropy [26]. Due to the teacher model's occasional incorrect predictions, providing misguided guidance may result in suboptimal performance for the student model. Conditional teacher-student learning is then proposed to selectively learn from the teacher model or ground truth labels, depending on the teacher's ability to predict the truth [27]. To reduce the performance gap between the teacher and student models, WSLD proposes the use of weighted soft labels for distillation from a bias-variance trade-off perspective for the purpose of increasing bias and decreasing variance [17].

Since response-based knowledge distillation overlooks intermediate-level supervision for thorough guidance, feature-based knowledge distillation concentrates on investigating intermediate feature maps and the corresponding information to offer better supervised training for the student models. FitNet, the first feature-based method, is introduced to align intermediate representations layer by layer between the teacher and student models, aiming to enhance the student's performance. While this approach is simple and intuitive, it may face challenges related to convergence and performance due to the lack of high-level knowledge and the capacity gap between the two networks [28], [29]. A novel Exclusivity-Consistency regularized Knowledge Distillation (EC-KD) introduces a position-aware exclusivity strategy to enhance diversity among filters within the same layer, alleviate the limitations of student models and combine weight exclusivity and feature consistency in one unified framework [30]. To avoid semantic misalignment between specific teacher-student layer combinations, Semantic Calibration for Cross-layer Knowledge Distillation (SemCKD) employs an attention mechanism to automatically assign suitable target layers from the teacher model to each student layer [31].

While response-based and feature-based knowledge distillation involve using the outputs of specific layers in the teacher model, relation-based methods go one step further to examine the cross-sample and cross-layer relationships as valuable knowledge [32]. A Flow of Solution Process (FSP) explores how features evolve across layers to encourage the student model to emulate the flow of the teacher model using the Gram matrix [33]. A novel Instance Relationship Graph (IRG) approach is proposed to model the knowledge of a single network layer by treating instance features and instance relationships as vertices and edges, followed by feature space transformation across multiple layers [34]. Probabilistic Knowledge Transfer (PKT) instructs the student model by aligning the probability distributions of the teacher model, and it leverages feature representations to capture instance-level relationships as probabilistic distributions during the training process [35].

B. ENSEMBLE LEARNING

Ensemble learning is a highly effective method for enhancing the performance of deep learning models by averaging the outputs of a small set of independently trained neural networks with identical architectures to improve the prediction accuracy compared to individual models [22]. Two most well-known ensemble approaches are bagging and boosting [36]. Recently, ensemble learning has been employed in knowledge distillation to augment model generalization and enhance the robustness of the student model. MT-BERT is a multi-teacher knowledge distillation framework by incorporating multiple pre-trained language models to learn a higher-quality student model [37]. On the contrary, the one-teacher and multiple-student is used to leverage the collective strength of several shallow student models of the same architecture throughout the distillation process [25].

C. SENTIMENT CLASSIFICATION

Sentiment classification, a way to identify the subjective information for the given context, has a wide range of applications across various industries and domains such as social media monitoring [38], customer feedback analysis [39], and financial forecasting [40]. Traditional supervised machine learning approaches have been widely investigated in this research area. For example, K-Nearest Neighbor (KNN) and Naive Bayes (NB) are used to detect the sentiments expressed in Twitter messages and subsequently categorize them into four categories (Happy-Active, Happy-Inactive, Unhappy-Active and Unhappy-Inactive) [41]. A Fisher function method based on probabilistic latent semantic analysis is introduced to enhance the kernel function of support vector machine where the experiment is conducted on Twitter sentiment corpus and the average accuracy is 87.20% [42]. In the past years, the research focus has shifted

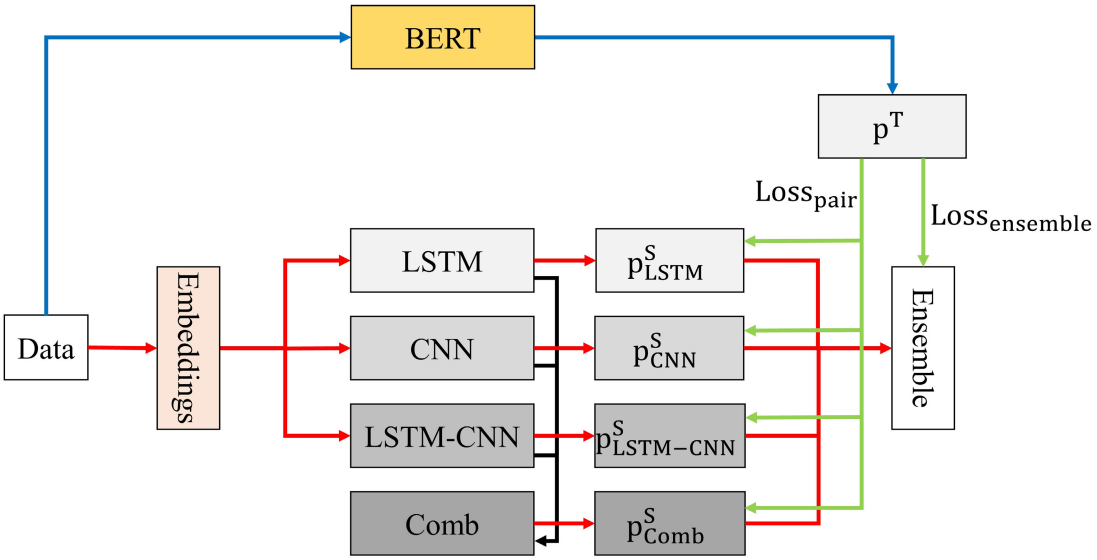


FIGURE 2. The architecture of our proposed model.

towards neural network methods which include the traditional network architectures and transformer-based model. In traditional networks such as CNN and LSTM models, word embeddings (e.g., Word2Vec, GloVe) are used as text representations where each word is characterized by a fixed pre-trained word vector. Research on sentiment analysis by the use of traditional network architectures with static word embeddings, including CNN [43], LSTM [44], CNN-LSTM [45], and attention mechanisms [46], [47], has been extensively explored and has led to significant advancements. As these static word embeddings do not consider the context of the surrounding words for the given sentence, they are not able to capture contextual information effectively. Recently, large-scale pre-trained language models based on Transformer architecture employ contextual embeddings that are generated by considering the entire context of the sentence. BERT [48] and its variants such as ALBERT [49], RoBERTa [50] and DistilBERT [51] have proven to be highly effective for sentiment classification tasks with the paradigm of fine-tuning PLMs.

III. PROPOSED METHOD

Given the training data samples, there exists a teacher model T with trainable parameters θ^t. The distilled student model S with parameters θ^s is trained by the model T and training data samples. The objective is to produce a simpler S with less parameters and S is able to perform even better than T on the testing data. Our proposed method is a kind of one-teacher and multiple-student architecture (shown in Figure 2). Our system architecture employs a single lightweight pre-trained word embedding as input, shared among our ensemble system consisting of an LSTM, a CNN, and their fusion to extract both sequential and spatial features. The distillation objective incorporates two loss functions to facilitate knowledge transfer from the teacher model (BERT) to our student models. The details will be discussed in the following sections.

A. THE TEACHER MODEL

A good teacher model should achieve high accuracy in the target task, enabling the student model to learn effectively. For training the teacher model, we select BERT model to perform the sentiment classification due to its great success in NLP tasks. It consists of 12 layers, 768 embedding size, 12 multi-head attentions and about 110M parameters. With the prevalence of the pretraining and fine-tuning paradigm, BERT model plays the role of the pre-trained foundation model and an additional linear layer with a softmax activation function is attached to make the final prediction. The input text x_i is first sent to the BERT model to generate the embeddings h_i^T, followed by feeding to the softmax classifier to predict the probability of label y_i:

$$p^T(y_i|x_i) = \text{softmax}(W^T h_i^T) \tag{1}$$

where superscript T denotes the teacher model and W^T is the matrix for sentiment classification parameters of the linear layer. The loss function of the teacher model is defined as the cross entropy loss and the optimization objective is to minimize the loss:

$$\text{Loss}^T = \sum_{i=1}^{N} L_{CE}(y_i, \hat{y}_i^T) \tag{2}$$

where L_{CE} means the cross entropy function, $\hat{y}_i^T$ is the prediction of teacher model, and y_i is the true label.

B. THE STUDENT MODEL

In knowledge distillation, the student model is a smaller architecture that aims to learn from a knowledgeable teacher

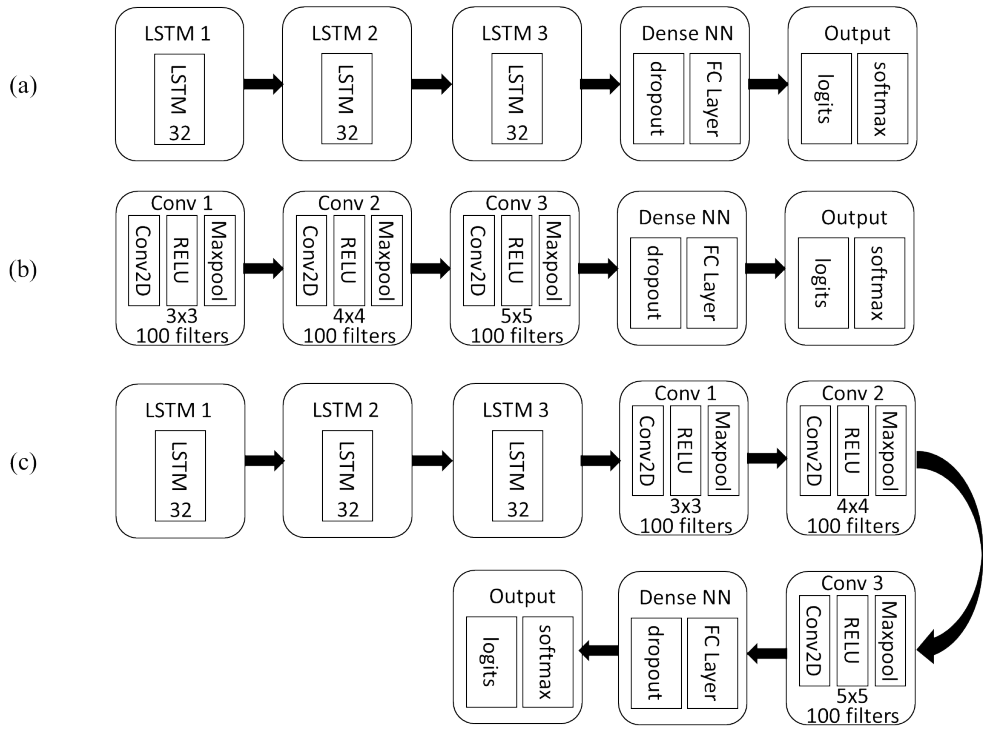


FIGURE 3. Three base student models (a) LSTM model; (b) CNN model; (c) LSTM-CNN model.

model to improve its performance on the sentiment classification task. In this research, we apply a heterogeneous ensemble-based approach using lean architectural paradigms of neural networks including LSTM, CNN and their combination to reduce the parameter usage. The three models, LSTM, CNN and LSTM-CNN, are used separately to learn on the same input data and make the final prediction based on the weighted predictions of each individual model. The input x_i is processed by a lookup layer, which is an embedding weight matrix, to generate the embeddings of each token and we concatenate all token embeddings to obtain the final embedding representation $h_i \in R^{n\times d}$ of x_i where n is the number of words and d is the embedding dimension. We use Glove.6B.50d GLOVE as pre-trained word embeddings [52]. It is worth to notice that we use the same input h_i in our three base student models.

The LSTM network is a special type of Recurrent Neural Network (RNN) designed to process sequential data [53]. One of its major components is memory cells that store and update information over time, allowing it to capture long-term dependencies in sequences. The main advantage of LSTM is the ability to address the vanishing/exploding gradient problem, making it particularly effective in tasks involving natural language processing and time series analysis, where preserving and learning from past contextual information is important [54]. We use the LSTM model as the first base model and the details of the architecture are shown in Figure 3 (a). There are three LSTM layers with hidden size 32. The embedding representation h_i is sent to all LSTM layers to obtain the encoding Enc^{LSTM} for the whole sentence. Subsequently, the encoded data is forwarded to a dropout layer and fully connected network to obtain logits $\text{logit}^{\text{LSTM}}$.

The CNN is a special type of feed-forward artificial neural networks with the ability to process structured grid data through convolutional and pooling layers. The convolutional operator scans input with filters of varying sizes to capture local patterns and hierarchies of features at different scales while the pooling operator is employed to identify important features and reduce the computation cost [55], [56]. We use the CNN structure as our second student model and the diagram of the architecture is shown in Figure 3 (b). Our proposed CNN model is composed of three convolutional blocks, each containing 100 filters. We apply filters of various sizes including 3×3, 4×4, and 5×5, to extract features across different spatial dimensions from the input. Within each convolutional block, a ReLU activation function is used to introduce non-linearity in the network, and a max-pooling operation is then adopted to capture the important features of the sentence denoted as Enc^{CNN}. Subsequently, a fully connected network is added to the network to perform high-level feature aggregation and produce the logits $\text{logit}^{\text{CNN}}$.

In addition to the above two base models, we present another two hybrid modules where the first one is a LSTM-CNN model and the second one is the combination of all aforementioned representations to form a new classifier. The LSTM-CNN model involves a two-stage process, with the first stage employing LSTM for sequential feature extraction from the input, and the second stage utilizing CNN

for spatial feature extraction from the output of the LSTM. Subsequently, a fully connected network is incorporated for high-level feature aggregation and the final prediction. Our implementation of the LSTM-CNN shown in Figure 3 (c) cascades the first LSTM student model with the second CNN student model. After the process of CNN model, the representation is denoted as $\text{Enc}^{\text{LSTM-CNN}}$ and is further supplied to a fully connected layer for obtaining the logits $\text{logit}^{\text{LSTM-CNN}}$. The second hybrid model is a combinatorial structure that concatenates the representations of the previous three student models (Enc^{LSTM}, Enc^{CNN} and $\text{Enc}^{\text{LSTM-CNN}}$) to generate the encoding Enc^{Comb}, which is then passed to the fully connected layer for producing the logits $\text{logit}^{\text{Comb}}$.

For each student model, one final softmax activation is used to compute the probability distribution of labels.

$$\text{p}^{\text{S}}_{\text{LSTM}}(\text{y}_\text{i}|\text{x}_\text{i}) = \text{softmax}(\text{logit}^{\text{LSTM}}) \tag{3}$$

$$\text{p}^{\text{S}}_{\text{CNN}}(\text{y}_\text{i}|\text{x}_\text{i}) = \text{softmax}(\text{logit}^{\text{CNN}}) \tag{4}$$

$$\text{p}^{\text{S}}_{\text{LSTM-CNN}}(\text{y}_\text{i}|\text{x}_\text{i}) = \text{softmax}(\text{logit}^{\text{LSTM-CNN}}) \tag{5}$$

$$\text{p}^{\text{S}}_{\text{Comb}}(\text{y}_\text{i}|\text{x}_\text{i}) = \text{softmax}(\text{logit}^{\text{Comb}}) \tag{6}$$

C. TRAINING STRATEGY

In the training stage, the distillation loss consists of two loss functions for transferring knowledge from the teacher network to student networks. The first loss is called the pair loss ($\text{Loss}_{\text{pair}}$) to measure KL divergence (Div_{KL}) between each student with the teacher model forcing the individual student model to generate similar predicted distributions of the teacher model.

$$\begin{aligned}\text{Loss}_{\text{pair}} = {} & \alpha_{\text{LSTM}} \times \text{Div}_{\text{KL}}\left(\text{p}^{\text{T}}, \text{p}^{\text{S}}_{\text{LSTM}}\right) + \alpha_{\text{CNN}} \\ & \times \text{Div}_{\text{KL}}\left(\text{p}^{\text{T}}, \text{p}^{\text{S}}_{\text{CNN}}\right) + \alpha_{\text{LSTM-CNN}} \\ & \times \text{Div}_{\text{KL}}\left(\text{p}^{\text{T}}, \text{p}^{\text{S}}_{\text{LSTM-CNN}}\right) + \alpha_{\text{Comb}} \\ & \times \text{Div}_{\text{KL}}\left(\text{p}^{\text{T}}, \text{p}^{\text{S}}_{\text{Comb}}\right)\end{aligned} \tag{7}$$

The second one is the ensemble loss ($\text{Loss}_{\text{ensemble}}$) to better leverage complementary information among student models. We first calculate the class probability distribution $\text{p}^{\text{S}}_{\text{ensemble}}$ by applying the softmax function to the weighted sum of the logits from the four student models. Then, the KL divergence is utilized to calculate the similarity between p^{T} and $\text{p}^{\text{S}}_{ensemble}$.

$$\begin{aligned}\text{p}^{\text{S}}_{\text{ensemble}} = {} & \text{softmax}(\beta_{\text{LSTM}} \times \text{logit}^{\text{LSTM}} + \beta_{\text{CNN}} \\ & \times \text{logit}^{\text{CNN}} + \beta_{\text{LSTM-CNN}} \\ & \times \text{logit}^{\text{LSTM-CNN}} + \beta_{\text{Comb}} \times \text{logit}^{\text{Comb}})\end{aligned} \tag{8}$$

$$\text{Loss}_{\text{ensemble}} = \text{Div}_{\text{KL}}\left(\text{p}^{\text{T}}, \text{p}^{\text{S}}_{ensemble}\right) \tag{9}$$

The final loss is defined as the combination of the pair loss and the ensemble loss.

$$\text{Loss}_{\text{total}} = \delta_{\text{pair}} \times \text{Loss}_{\text{pair}} + \delta_{\text{ensemble}} \times \text{Loss}_{\text{ensemble}} \tag{10}$$

We present the training pseudocode of our model in Algorithm 1. The output distribution of the teacher model BERT (T) is in line 4 and the output distribution of each student model is in lines 5-6. The pair-wise loss ($\text{Loss}_{\text{pair}}$) between the teacher model output and each student model is in line 7. The ensemble loss ($\text{Loss}_{\text{ensemble}}$) between the teacher model output and ensemble output of student models is in lines 8-9. The final loss and the update are in lines 10-11.

Algorithm 1 Heterogeneous Student Knowledge Distillation Model

Input: The training dataset D{X, Y}, a word embeddings Glove.6B.50d GLOVE (Emb) and a pre-trained BERT teacher model (T)

Output: θ^{S} for all the trainable weights in our proposed model

1: Randomly initialize θ^{S}
2: **repeat**:
3: **for** each mini-batch {x, y} in D:
4: Calculate the output distribution of the T in Eq. (1) for {x}
5: Input {x} to Emb to obtain {h}
6: Pass {h} to each student model S_i to obtain the corresponding output distribution in Eq. (3), (4), (5) and (6)
7: Compute the $\text{Loss}_{\text{pair}}$ in Eq. (7) based on the distributions of T and each S_i
8: Calculate the ensemble distribution by combing the logits of each S_i based on the expression of Eq. (8)
9: Compute the $\text{Loss}_{\text{ensemble}}$ in Eq. (9) using KL divergence between the ensemble distribution and the distribution of T
10: Compute the total loss $\text{Loss}_{\text{total}}$ in Eq. (10)
11: Update the weight θ^{S} to minimize the $\text{Loss}_{\text{total}}$
12: **end for**
13: **until** convergence

IV. EXPERIMENTS AND RESULTS

In this section, we conduct comprehensive empirical studies to evaluate the effectiveness of the proposed approach in knowledge distillation for the sentiment classification task. The discussion within this section covers the following essential parts: the sentiment dataset utilized in our experiments, the formulation of evaluation criteria, performance comparisons with existing methodologies along with detailed analysis and ablation studies aimed at investigating the impact of the primary components.

A. EXPERIMENTAL SETTINGS

Yelp is a famous and popular online platform where people are allowed to search and review the businesses in different industries. Those review data have been widely used for various NLP tasks such as sentiment analysis, recommendation systems and text classification, etc. In our sentiment classification task, we utilize the same dataset configuration as prior works in order to make a fair comparison [25], [57]. The sizes of the training, validation, and testing sets are 3000, 1000, and 1000, respectively. Each sample corresponds to sentences from a review and has been labelled as either positive or negative sentiment. Figure 4 displays examples of the sample data where 1 means positive review and 0 means negative review.

1	they also have a nice wine list .
0	we got up and walked out and will never be back .
0	on a scale of bad to good , it was kind of meh .
1	location is great as well , right across the movie theater .
0	i wan na love bar louie , i do .
1	i 'll definitely order pizza from here again .
1	they 're professional yet very approachable and reasonable prices .
1	great music , fun crowd , pretty decent prices on drinks .
1	i had filet and crab cake and it was delicious .

FIGURE 4. Data examples in yelp.

The teacher model is the BERT-base network pre-trained with 12 hidden layers. To evaluate the effectiveness of our approach, we choose the non-distillation based and distillation based methods to compare with our solution for the sentiment classification. Non-distillation based methods include BiLSTM, CNN, and Ensemble CNNs. Distillation based methods consist of TinyBERT [58], MT-BERT [37] and Distilled Ensemble CNNs [25]. TinyBERT is a two-stage learning framework which encourages the linguistic knowledge transferring from a large teacher BERT to a smaller student BERT, and is a one-teacher and one-student architecture. MT-BERT is a multiple-teacher and one-student distillation approach using a multi-teacher co-finetuning framework that aligns the output hidden spaces of multiple teacher networks to enhance collaborative student teaching through shared pooling and prediction modules. Distilled Ensemble CNNs is a one-teacher and multiple-student model to distill knowledge from a large pre-trained teacher model into multiple shallow CNN student models by ensemble learning. We consider two evaluation metrics in this work where the accuracy is used to assess the performance in classification and the number of parameters is used to measure the model size.

To optimize training quality of our model, we adjust hyper-parameters through empirical processes and the chosen hyper-parameter values are shown in Table 1. The experimental comparison is executed in a Windows 10 environment and is run on a desktop computer which is equipped with an Intel Core i9 CPU, 128 GB of memory and an NVIDIA GeForce RTX 3090 24 GB GPU with CUDA 11.1.

B. EVALUATION RESULTS

We compare the Yelp sentiment classification results of prior methods with our distillation approach shown in Table 2. The first row is the teacher model that we aim to distill, and it is a BERT-base network with an accuracy of 95.8% and a large parameter size of 109.48M.

Traditional non-distillation techniques devoid of resource-intensive large language models have fallen short of matching teacher network capabilities, despite notably reduced model complexity. Ensemble CNN model has attained an accuracy of 92.9% using 1.20M, exhibiting inferior performance compared to the teacher BERT model under significant architectural compression. Accordingly, by incorporating knowledge distillation, accuracy has improved markedly to surpass the teacher model. Distilled Ensemble CNN exceeds the BERT model with a 1.3% accuracy boost. However, model complexity still demands substantial parameters (4.15M).

Building upon the foundation of previous research, our approach achieves outstanding evaluation performance of 97.3% with a modest parameter budget of 1.69M. In comparison to previous state-of-the-art methods, our unified model demonstrates superior representational capabilities while utilizing only a fraction of the parameters, with an increase in size of only 1.54M relative to the most economical non-distillation alternative (CNN). Simultaneously improving accuracy and efficiency across all key benchmarks, our compact ensemble design sets new standards in achieving a balance between predictive accuracy and architectural simplicity, showcasing significant advancements in model design.

In conclusion, our model attains the highest accuracy among all methods. Regarding the reduction of the parameter

TABLE 1. Hyper-parameter configuration.

Parameter	Value
batch size	128
epoch	20
learning rate	0.001
weight_decay	0.00001
LSTM – hidden size	32
LSTM – number of layers	3
CNN – filter size	3x3 4x4 5x5
CNN - number of filters	100

TABLE 2. Experimental results. We implement Distilled Ensemble CNN based on the released codes and other results are directly from [25]. In all methods, we underline the best performance-performing result, while in the distillation based approaches, we use bold font to highlight the top performance.

Category	Models	# of paras	Accuracy
Teacher	BERT-base	109.48M	95.8%
Non-Distillation	BiLSTM	2.35M	91.9%
	CNN	<u>0.15M</u>	92.7%
	Ensemble CNN	1.20M	92.9%
Distillation	TinyBERT	66.96M	95.6%
	MT-BERT	14.35M	94.5%
	Distilled Ensemble CNN	4.15M	97.1%
	Our Model	**1.69M**	**<u>97.3%</u>**

TABLE 3. Ablation experimental results.

Component	Model	Accuracy
Teacher	Our Model	97.3%
student model	LSTM and CNN	96.7%
	LSTM andLSTM-CNN	96.6%
	CNN and LSTM-CNN	97.0%
loss function	$\text{Loss}_{\text{pair}}$	96.8%
	$\text{Loss}_{\text{ensemble}}$	97.0%
embeddings	embeddings from scratch	96.9%

size, we rank third in overall performance but we perform best among distillation based methods.

C. ABLATION STUDIES

To better understand the influence of each contributing component of our method on effectiveness, we conduct ablation studies including the student model, different loss functions and the usage of pre-trained word embeddings. The ablation results are shown in Table 3.

1) CHOICE OF THE STUDENT MODEL

We evaluate the performance of our distillation approach on different combination of student models. As shown in Table 3, it can be concluded that our approach combining all student models outperforms those methods using only two student models.

2) EFFECT OF THE LOSS FUNCTION

To evaluate the contribution of each loss function, we remove each term from the objective function defined in Eq. 10 and make the following two observations. First, it can be seen that the accuracy of $\text{Loss}_{\text{ensemble}}$ is 0.2% higher than that of $\text{Loss}_{\text{pair}}$. In other words, the one-teacher-versus-all-student loss strategy is superior to the one-teacher-versus-one-student loss strategy. Second, our approach performs better than models using a single loss function. These findings indicate the effectiveness of joint loss ($\text{Loss}_{\text{total}}$), as similarly illustrated in the concurrent research [59].

3) IMPACT OF WORD EMBEDDINGS

In order to comprehend how well the pre-trained language model performs, we conduct training for an embedding model starting from scratch. The embedding model is initialized randomly with a uniform distribution $(-1, 1)$, resulting in an accuracy of 96.9%. The pre-trained embedding model (Glove.6B.50d in this research) performs better than the one trained from scratch.

Based on the above ablation studies conducted, we conclude that variations in the student model, diverse loss functions, and the utilization of pre-trained word embeddings all impact the effectiveness of our knowledge distillation model. These findings highlight the importance of carefully considering these factors when implementing our knowledge distillation technique in academic research or practical applications.

V. CONCLUSION

In this study, we apply multiple lightweight student models and a low-dimensional pre-trained word embedding model to address the knowledge distillation conducted in the sentiment classification task. Our model achieves the state-of-the-art accuracy results and also outperforms the teacher model. Moreover, our model size is only 1.69M and has significantly decreased the number of parameters compared to the existing models.

In our future work, we plan to build upon this paper with a particular focus on the following aspects. First, as our current approach is based on weighted ensemble learning, we will investigate to enhance the interaction between student models to learn more informative representations for maximizing the generalization performance. Second, we aim to apply our model to more complex NLP classification tasks and learn from more advanced teacher models. Being able to reduce parameters in state-of-the-art models will bring significant benefits to and have a great impact on enterprises.

REFERENCES

[1] J. Dean, "A golden decade of deep learning: Computing systems & applications," *Daedalus*, vol. 151, no. 2, pp. 58–74, May 2022.
[2] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," 2018, *arXiv:1810.04805*.
[3] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever. (2018). *Improving Language Understanding by Generative Pre-Training*. [Online]. Available: https://s3-us-west-2.amazonaws.com/openai-assets/research-covers/language-unsupervised/language_understanding_paper.pdf
[4] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language models are unsupervised multi-task learners," *OpenAI Blog*, vol. 1, no. 8, p. 9, 2019.
[5] T. B. Brown et al., "Language models are few-shot learners," 2020, *arXiv:2005.14165*.
[6] J. Pan, P. Zhu, K. Zhang, B. Cao, Y. Wang, D. Zhang, J. Han, and Q. Hu, "Learning self-supervised low-rank network for single-stage weakly and semi-supervised semantic segmentation," *Int. J. Comput. Vis.*, vol. 130, no. 5, pp. 1181–1195, May 2022.
[7] M. Eo, S. Kang, and W. Rhee, "An effective low-rank compression with a joint rank selection followed by a compression-friendly training," *Neural Netw.*, vol. 161, pp. 165–177, Apr. 2023.
[8] D. Ghimire, D. Kil, and S.-H. Kim, "A survey on efficient convolutional neural networks and hardware acceleration," *Electronics*, vol. 11, no. 6, p. 945, Mar. 2022.
[9] J. Lee, L. Mukhanov, A. S. Molahosseini, U. Minhas, Y. Hua, J. M. del Rincon, K. Dichev, C.-H. Hong, and H. Vandierendonck, "Resource-efficient convolutional networks: A survey on model-, arithmetic-, and implementation-level techniques," *ACM Comput. Surv.*, vol. 55, no. 13, pp. 1–36, Jul. 2023.
[10] Y.-J. Zheng, S.-B. Chen, C. H. Q. Ding, and B. Luo, "Model compression based on differentiable network channel pruning," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 34, no. 12, pp. 10203–10212, Dec. 2023.
[11] G. Fang, X. Ma, M. Song, M. Bi Mi, and X. Wang, "DepGraph: Towards any structural pruning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 16091–16101.
[12] H. Jeon, S. Park, J.-G. Kim, and U. Kang, "PET: Parameter-efficient knowledge distillation on transformer," *PLoS One*, vol. 18, no. 7, Jul. 2023, Art. no. e0288060.
[13] J. Gou, B. Yu, S. J. Maybank, and D. Tao, "Knowledge distillation: A survey," *Int. J. Comput. Vis.*, vol. 129, no. 6, pp. 1789–1819, Jun. 2021.

[14] S. Zagoruyko and N. Komodakis, "Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer," 2016, *arXiv:1612.03928*.

[15] B. Heo, M. Lee, S. Yun, and J. Y. Choi, "Knowledge transfer via distillation of activation boundaries formed by hidden neurons," in *Proc. AAAI Conf. Artif. Intell. (AAAI)*, 2019, vol. 33, no. 1, pp. 3779–3787.

[16] X. Jin, B. Peng, Y. Wu, Y. Liu, J. Liu, D. Liang, J. Yan, and X. Hu, "Knowledge distillation via route constrained optimization," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 1345–1354.

[17] H. Zhou, L. Song, J. Chen, Y. Zhou, G. Wang, J. Yuan, and Q. Zhang, "Rethinking soft labels for knowledge distillation: A bias-variance tradeoff perspective," 2021, *arXiv:2102.00650*.

[18] W. Park, D. Kim, Y. Lu, and M. Cho, "Relational knowledge distillation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3962–3971.

[19] C. Yang, H. Zhou, Z. An, X. Jiang, Y. Xu, and Q. Zhang, "Cross-image relational knowledge distillation for semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 12309–12318.

[20] I. D. Mienye and Y. Sun, "A survey of ensemble learning: Concepts, algorithms, applications, and prospects," *IEEE Access*, vol. 10, pp. 99129–99149, 2022.

[21] U. Asif, J. Tang, and S. Harrer, "Ensemble knowledge distillation for learning improved and efficient networks," in *Proc. Eur. Conf. Artif. Intell.*, vol. 325, 2020, pp. 953–960.

[22] Z. Allen-Zhu and Y. Li, "Towards understanding ensemble, knowledge distillation and self-distillation in deep learning," 2020, *arXiv:2012.09816*.

[23] M. Bordoloi and S. K. Biswas, "Sentiment analysis: A survey on design framework, applications and future scopes," *Artif. Intell. Rev.*, vol. 56, no. 11, pp. 12505–12560, Nov. 2023.

[24] A. Bello, S.-C. Ng, and M.-F. Leung, "A BERT framework to sentiment analysis of tweets," *Sensors*, vol. 23, no. 1, p. 506, Jan. 2023.

[25] X. Chang, S. Y. M. Lee, S. Zhu, S. Li, and G. Zhou, "One-teacher and multiple-student knowledge distillation on sentiment classification," in *Proc. 29th Int. Conf. Comput. Linguistics*. Gyeongju, Republic of Korea: International Committee on Computational Linguistics, Oct. 2022, pp. 7042–7052.

[26] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," in *Proc. Deep Learn. Represent. Learn. Workshop NIPS*, 2014, pp. 1–9.

[27] Z. Meng, J. Li, Y. Zhao, and Y. Gong, "Conditional teacher–student learning," in *Proc. ICASSP - IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2019, pp. 6445–6449.

[28] A. Romero, N. Ballas, S. E. Kahou, A. Chassang, C. Gatta, and Y. Bengio, "FitNets: Hints for thin deep nets," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2015, pp. 1–13.

[29] J. Kang and J. Gwak, "Ensemble learning of lightweight deep learning models using knowledge distillation for image classification," *Mathematics*, vol. 8, no. 10, p. 1652, Sep. 2020.

[30] X. Wang, T. Fu, S. Liao, S. Wang, Z. Lei, and T. Mei, "Exclusivity-consistency regularized knowledge distillation for face recognition," in *Proc. 16th Eur. Conf. Comput. Vis. (ECCV)*, vol. 12369, Aug. 2020, p. 325.

[31] D. Chen, J. P. Mei, Y. Zhang, C. Wang, Z. Wang, Y. Feng, and C. Chen, "Cross-layer distillation with semantic calibration," in *Proc. AAAI*, 2021, pp. 7028–7036.

[32] C. Yang, X. Yu, Z. An, and Y. Xu, "Categories of response-based, feature-based, and relation-based knowledge distillation," in *Advancements in Knowledge Distillation: Towards New Horizons of Intelligent Systems*. Cham, Switzerland: Springer, 2023, pp. 1–32.

[33] J. Yim, D. Joo, J. Bae, and J. Kim, "A gift from knowledge distillation: Fast optimization, network minimization and transfer learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 7130–7138.

[34] Y. Liu, J. Cao, B. Li, C. Yuan, W. Hu, Y. Li, and Y. Duan, "Knowledge distillation via instance relationship graph," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 7089–7097.

[35] N. Passalis, M. Tzelepi, and A. Tefas, "Probabilistic knowledge transfer for lightweight deep representation learning," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 5, pp. 2030–2039, May 2021.

[36] W.-C. Kao, H.-X. Xie, C.-Y. Lin, and W.-H. Cheng, "Specific expert learning: Enriching ensemble diversity via knowledge distillation," *IEEE Trans. Cybern.*, vol. 53, no. 4, pp. 2494–2505, Apr. 2023.

[37] C. Wu, F. Wu, and Y. Huang, "One teacher is enough? Pre-trained language model distillation from multiple teachers," 2021, *arXiv:2106.01023*.

[38] Z. Wei, W. Liu, G. Zhu, S. Zhang, and M.-Y. Hsieh, "Sentiment classification of Chinese Weibo based on extended sentiment dictionary and organisational structure of comments," *Connection Sci.*, vol. 34, no. 1, pp. 409–428, Dec. 2022.

[39] P. Bhuvaneshwari, A. N. Rao, Y. H. Robinson, and M. N. Thippeswamy, "Sentiment analysis for user reviews using bi-LSTM self-attention based CNN model," *Multimedia Tools Appl.*, vol. 81, no. 9, pp. 12405–12419, Apr. 2022.

[40] B. Fazlija and P. Harder, "Using financial news sentiment for stock price direction prediction," *Mathematics*, vol. 10, no. 13, p. 2156, Jun. 2022.

[41] M. Suhasini and B. Srinivasu, "Emotion detection framework for Twitter data using supervised classifiers," in *Proc. 3rd Data Eng. Commun. Technol.*, 2020, pp. 565–576.

[42] K.-X. Han, W. Chien, C.-C. Chiu, and Y.-T. Cheng, "Application of support vector machine (SVM) in the sentiment analysis of Twitter dataset," *Appl. Sci.*, vol. 10, no. 3, p. 1125, Feb. 2020.

[43] X. Ouyang, P. Zhou, C. H. Li, and L. Liu, "Sentiment analysis using convolutional neural network," in *Proc. IEEE Int. Conf. Comput. Inf. Technol., Ubiquitous Comput. Commun., Dependable, Autonomic Secure Comput., Pervasive Intell. Comput.*, Oct. 2015, pp. 2359–2364.

[44] D. Li and J. Qian, "Text sentiment analysis based on long short-term memory," in *Proc. 1st IEEE Int. Conf. Comput. Commun. Internet (ICCCI)*, Oct. 2016, pp. 471–475.

[45] L. Khan, A. Amjad, K. M. Afaq, and H.-T. Chang, "Deep sentiment analysis using CNN-LSTM architecture of English and Roman Urdu text shared in social media," *Appl. Sci.*, vol. 12, no. 5, p. 2694, Mar. 2022.

[46] Y. Feng and Y. Cheng, "Short text sentiment analysis based on multi-channel CNN with multi-head attention mechanism," *IEEE Access*, vol. 9, pp. 19854–19863, 2021.

[47] W. Li, F. Qi, M. Tang, and Z. Yu, "Bidirectional LSTM with self-attention mechanism and multi-channel features for sentiment classification," *Neurocomputing*, vol. 387, pp. 63–77, Apr. 2020.

[48] M. Hoang, O. A. Bihorac, and J. Rouces, "Aspect-based sentiment analysis using BERT," in *Proc. 22nd Nordic Conf. Comput. Linguistics*, Sep./Oct. 2019, pp. 187–196.

[49] Z. Ding, Y. Qi, and D. Lin, "Albert-based sentiment analysis of movie review," in *Proc. 4th Int. Conf. Adv. Electron. Mater., Comput. Softw. Eng. (AEMCSE)*, Mar. 2021, pp. 1243–1246.

[50] K. L. Tan, C. P. Lee, K. S. M. Anbananthen, and K. M. Lim, "RoBERTa-LSTM: A hybrid model for sentiment analysis with transformer and recurrent neural network," *IEEE Access*, vol. 10, pp. 21517–21525, 2022.

[51] V. Dogra, A. Singh, S. Verma, Kavita, N. Z. Jhanjhi, and M. N. Talib, "Analyzing DistilBERT for sentiment classification of banking financial news," in *Intelligent Computing and Innovation on Data Science*. Singapore: Springer, 2021, p. 582.

[52] J. Pennington, R. Socher, and C. Manning, "Glove: Global vectors for word representation," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2014, pp. 1532–1543.

[53] P. Li, W. Li, Z. He, X. Wang, Y. Cao, J. Zhou, and W. Xu, "Dataset and neural recurrent sequence labeling model for open-domain factoid question answering," 2016, *arXiv:1607.06275*.

[54] X. Yang, Y. Zhang, and M. Chi, "Time-aware subgroup matrix decomposition: Imputing missing data using forecasting events," in *Proc. IEEE Int. Conf. Big Data (Big Data)*, Dec. 2018, pp. 1524–1533.

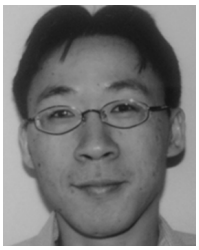
[55] N. Kalchbrenner, E. Grefenstette, and P. Blunsom, "A convolutional neural network for modelling sentences," 2014, *arXiv:1404.2188*.

[56] K. Fukami, R. Maulik, N. Ramachandra, K. Fukagata, and K. Taira, "Global field reconstruction from sparse sensors with Voronoi tessellation-assisted deep learning," *Nature Mach. Intell.*, vol. 3, no. 11, pp. 945–951, Oct. 2021.

[57] J. Li, R. Jia, H. He, and P. Liang, "Delete, retrieve, generate: A simple approach to sentiment and style transfer," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Human Lang. Technol.*, vol. 1, Jun. 2018, pp. 1865–1874.

[58] X. Jiao, Y. Yin, L. Shang, X. Jiang, X. Chen, L. Li, F. Wang, and Q. Liu, "TinyBERT: Distilling BERT for natural language understanding," 2019, *arXiv:1909.10351*.

[59] J. Gou, L. Sun, B. Yu, L. Du, K. Ramamohanarao, and D. Tao, "Collaborative knowledge distillation via multiknowledge transfer," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, pp. 1–13, Oct. 20, 2022.



CHING-SHENG LIN received the B.S. and M.S. degrees in applied mathematics from National Chung-Hsing University, Taiwan, in 1998 and 2001, respectively, and the Ph.D. degree in computer science from the State University of New York at Albany, in 2016.

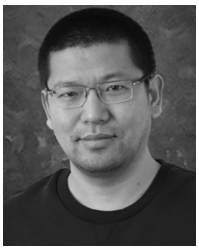
He is currently an Associate Professor with the Master Program of Digital Innovation, Tunghai University, Taiwan. His research interests include natural language processing, artificial intelligence, and machine learning.

CHUNG-NAN TSAI received the M.S. degree in electrical and computer engineering from Auburn University, Auburn, AL, USA, in 2012, and the M.S. degree in electrical and computer engineering from Oregon State University, Corvallis, OR, USA, in 2016. He is currently a Senior Data Scientist with Lam Research Japan GK, Japan. His research interests include machine learning, natural language processing, virtual metrology, and smart manufacturing.

JUNG-SING JWO received the B.E. degree in mechanical engineering from National Taiwan University, Taiwan, in 1984, and the M.S. and Ph.D. degrees in computer science from The University of Oklahoma, Norman, OK, USA, in 1991. Currently, he is a Professor with the Department of Computer Science and the Chair of the Master Program of Digital Innovation, Tunghai University, Taiwan. He served as the Director of the Board of Governors, Software Engineering Association of Taiwan (SEAT), from 2014 to 2017. He is the author of three books and holds 12 software patents. His research interests include business AI, smart manufacturing, software engineering, and user experience design.

CHENG-HSIUNG LEE received the B.I.M. and M.I.M. degrees in information management from Chaoyang University of Technology, in 2002 and 2004, respectively, and the Ph.D. degree in computer science and engineering from National Chung-Hsing University, Taichung, Taiwan, in 2013.

From 2017 to 2018, he was a Postdoctoral Research Fellow with the Center of Intelligent and Innovation Manufacturing System, Tunghai University, Taiwan. Currently, he is an Associate Professor with the Master Program of Digital Innovation, Tunghai University. His current research interests include machine learning, big data analytics, computer vision, and smart manufacturing.



XIN WANG (Senior Member, IEEE) received the Ph.D. degree in computer science from the University at Albany, State University of New York (SUNY), in 2015. He is currently an Assistant Professor with the Department of Epidemiology and Biostatistics, School of Public Health, University at Albany, SUNY. He is also a Faculty Member of the UAlbany AI Institute. His research interests include artificial intelligence, reinforcement learning, deep learning, and their applications.