Clustering-Based Knowledge Distillation with Sentence Pruning Processing for Efficient Student Model Training

Anonymous ACL submission

Abstract

Large-scale pre-trained language models, such as BERT, RoBERTa, and GPT, achieve state-of-the-art performance across various NLP tasks but face significant deployment challenges in resource-constrained environments due to high computational demands. To address this, we propose Clustering-Based Knowledge Distillation with Sentence Pruning Processing, a novel framework that enhances knowledge transfer by integrating multiple teacher models and refining sentence-level representations. Our approach employs cosine similarity measurement to identify cluster centers based on the strongest edge weights, ensuring that the most informative sentences are preserved. Additionally, clustering-based pruning with dynamic thresholds, guided by TF-IDF-based importance scores, effectively removes redundant information while retaining critical knowledge. By aggregating diverse knowledge from multiple teachers, Clustering-Based Knowledge Distillation with Sentence Pruning Processing enhances model robustness and generalization. Experimental results on multiple NLP benchmarks demonstrate that our method outperforms existing knowledge distillation techniques, achieving higher accuracy while significantly reducing computational overhead and inference time. Notably, our framework achieves up to 23× speedup on SST-2 and improves RTE accuracy by 4.5%, demonstrating its effectiveness for efficient NLP model deployment.

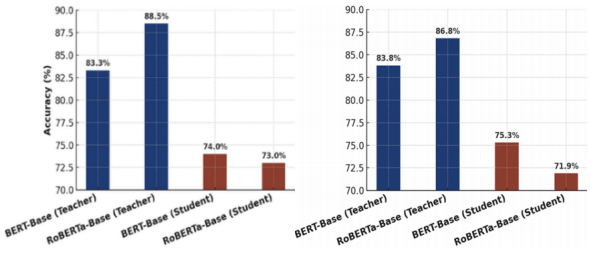
1 Introduction

Large-scale pre-trained language models, such as **BERT**, **RoBERTa**, and **GPT**, have established new benchmarks across various NLP tasks, achieving state-of-the-art performance(Koroteev, 2021; Delobelle et al., 2020; Achiam et al., 2023). However, their substantial computational requirements pose challenges for real-world deployment, particularly in low-power and constrained computing environments(Jiao et al., 2020). To address this challenge, **Knowledge Distillation (KD)** has been widely adopted as an effective model compression technique that transfers knowledge from a large **teacher model** to a smaller **student model**, enabling efficient inference while maintaining high performance.

Despite its effectiveness, conventional KD methods exhibit several limitations. Traditional KD techniques primarily focus on aligning the output distributions of teacher and student models, often neglecting **structured representations**, such as inter-sentence relationships and token-level importance(Wei et al., 2024). This oversimplified knowledge transfer may lead to suboptimal learning in the student model. Furthermore, most KD approaches rely on a **single teacher model**, which restricts the diversity of knowledge imparted to the student model, potentially limiting its generalization ability(Pham et al., 2023). Additionally, transferring knowledge directly from a complex teacher model may introduce **irrelevant information**, thereby reducing the overall learning efficiency of the student model(Yuan et al., 2024).

A key observation motivating our study is that different teacher models exhibit varying levels of effectiveness depending on the downstream task. As shown in Figure 1, RoBERTa-Base generally achieves higher accuracy than BERT-Base on MRPC and MNLI-mm tasks, indicating that it encodes richer knowledge. However, when transferring knowledge from these teacher models to a smaller student model (e.g., BERT3), we observe that the student model trained with BERT-Base performs better than the one trained with RoBERTa-Base. This suggests that using a stronger teacher model does not always result in better student performance, as excessively complex teacher models may introduce hard-to-learn knowledge, making student model training less effective.

To overcome these limitations, we propose **clustering-Based Knowledge Distillation with**



(a) MRPC dataset (b) MNLI-mm dataset

Figure 1: Comparison of Teacher and Student Model Performance on MRPC and MNLI-mm tasks. While RoBERTa-Base outperforms BERT-Base as a standalone model, its knowledge transfer to student models does not always lead to better performance, highlighting the importance of structured knowledge distillation.

Sentence Pruning Processing, a novel framework that enhances knowledge transfer by integrating multiple teacher models while refining the input representation through sentence-level pruning. Our method utilizes **Clustering-Based Knowledge Distillation with Sentence Pruning Processing**, which aggregates knowledge from multiple teacher models to enhance robustness and diversity while modeling inter-sentence relationships through clustering-based representation. This approach effectively retains essential information, optimizing the student model's learning process.

This study investigates the following key research objectives:

- **How can multiple teacher models be effectively integrated to enhance knowledge transfer?**
- **How does sentence-level pruning improve the efficiency of student model training?**
- **What are the effects of our method on the student model's performance, particularly when applied to novel tasks or datasets?**

The main contributions of this work are summarized as follows. We introduce a **clustering-based saliency-driven pruning mechanism** that effectively compresses sentence representations while preserving essential information, improving student model efficiency. Finally, we conduct comprehensive experiments on **benchmark NLP datasets** to validate the effectiveness of the proposed method, demonstrating superior performance compared to conventional KD approaches.

2 Related Work

2.1 Knowledge Distillation

Knowledge Distillation (KD) has been widely studied as an effective model compression technique that enables a smaller student model to inherit the knowledge of a larger teacher model(Gu et al., 2024).The conventional KD approach primarily focuses on minimizing the divergence between the teacher's and student's output distributions, typically using soft probability distributions(Gao, 2023). introduced temperature scaling to soften the logits, facilitating smoother knowledge transfer. While KD has proven effective in reducing model size without significant performance degradation, most traditional approaches concentrate on aligning output distributions rather than capturing richer structural knowledge(Zhang et al., 2024).

To mitigate this limitation, recent studies have incorporated intermediate-layer feature matching, where the student model learns from multiple hidden layers of the teacher (Haidar et al., 2021; Zhang et al., 2024). However, these methods still struggle to effectively utilize inter-sentence relationships, which are crucial for NLP tasks. *Motivated models* have attempted to improve KD by introducing reinforcement learning-based knowledge selection, adaptive teacher-student interaction, and structured representation learning. Despite these advancements, existing approaches remain limited in handling noisy knowledge, often leading to inefficient student model training(Song et al., 2022; Xu et al., 2020).

2.2 Ensemble-Based Knowledge Distillation

To enhance the quality of knowledge transfer, a knowledge distillation based on ensembles has been proposed, where multiple teacher models are used instead of a single teacher. The key motivation behind this approach is to improve generalization by exposing the student model to diverse knowledge representations from different teachers (Yuan et al., 2021; Wu et al., 2022; Gou et al., 2021). One common strategy is to average the logits from multiple teachers, thereby creating a more robust probability distribution for the student model to learn from.

Several studies have explored adaptive weighting mechanisms to balance the influence of each teacher, optimizing the relevance of transferred knowledge(Du et al., 2020). *Motivated models*, such as RL-KD ensembles, employ reinforcement

learning to dynamically select teachers based on reward mechanisms, improving the effectiveness of knowledge transfer (Qiu et al., 2022; Hong et al., 2021). However, while ensemble-based KD improves robustness, a major challenge is that direct aggregation of multiple teacher outputs can introduce conflicting information, making student training inefficient(Shao and Chen, 2023).

Recent research has attempted to refine teacher selection through reinforcement learning and adaptive weighting, but these methods do not fully address the problem of **noisy**, particularly in scenarios where the student model has limited capacity to process overly complex teacher outputs(Fan et al., 2021; Yuan et al., 2021). This limitation highlights the need for a more structured approach to selecting and distilling knowledge from multiple teachers.

2.3 Limitations of Existing Approaches

Although KD has been instrumental in compressing large-scale NLP models, existing methods still have notable limitations. **Single-teacher KD** restricts the diversity of knowledge transfer, whereas **ensemble KD** methods often suffer from computational inefficiencies(Wu et al., 2021; Wang et al., 2022). *Motivated models* using reinforcement learning for teacher selection have improved knowledge transfer, but they still fail to efficiently filter out **irrelevant knowledge**, leading to suboptimal student model learning.

Additionally, most KD approaches **overlook the importance of structured representations**, leading to inefficient training of student models. Clustering-based sentence selection methods offer a promising solution to address these issues by refining knowledge representations before distillation(Sadeghi et al., 2024).

To overcome these challenges, our work combines **ensemble knowledge distillation with sentence clustering-based pruning using cosine similarity measurement** to transfer structured and diverse knowledge while filtering out irrelevant content. By leveraging **multiple teacher models and clustering-based pruning, which applies dynamic thresholds to reflect TF-IDF-based sentence importance**, our approach enhances the efficiency of student model training while preserving essential information. Our experimental results demonstrate that integrating **sentence clustering-based pruning using cosine similarity measurement** with **ensemble KD** leads to superior performance compared to previous models, achieving a better trade-off between efficiency and accuracy.

3 Method

To address the limitations of conventional knowledge distillation, we propose an **Clustering-Based Knowledge Distillation with Sentence Pruning Processing** framework. This approach enhances knowledge transfer efficiency by integrating multiple teacher models and refining knowledge representations through a clustering-based sentence pruning mechanism. As illustrated in Figure 2, the framework consists of key component: The framework consists of **Clustering-Based Knowledge Distillation with Sentence Pruning Processing**, which aggregates knowledge from multiple teacher models to enhance diversity and robustness while removing irrelevant information through clustering-based sentence pruning to optimize the student model's learning process.

3.1 Ensemble Knowledge Distillation

Traditional knowledge distillation methods primarily rely on a single teacher model, limiting the diversity of knowledge transferred to the student model(Amirkhani et al., 2021). To mitigate this issue, our approach employs an **Clustering-Based Knowledge Distillation with Sentence Pruning Processing** strategy that integrates multiple teacher models. By combining the outputs of multiple teachers, the student model benefits from a more comprehensive and diverse set of representations, thereby enhancing generalization performance and reducing the risk of overfitting to a specific teacher model.

As shown in Figure 2, each teacher model produces sentence embeddings, which are aggregated to form a unified representation computed as $\mathbf{E}_{ensemble} = \frac{1}{N}\sum_{i=1}^{N}\mathbf{E}_{T_i}$, where $\mathbf{E}_{T_i}$ denotes the embedding from the i-th teacher model and N is the total number of teacher models.

To refine the knowledge transfer, the student model is trained to minimize the discrepancy between its output distribution and the ensemble teacher distribution using **Kullback-Leibler Divergence(KL) loss**:

$$\mathcal{L}_{KL} = \frac{1}{M}\sum_{j=1}^{M}\sum_{i}\mathbf{P}^{j}_{ensemble,i}\log\frac{\mathbf{P}^{j}_{ensemble,i}}{\mathbf{P}^{j}_{student,i}} \quad (1)$$

where $\mathbf{P}^{j}_{student}$ and $\mathbf{P}^{j}_{ensemble}$ denote the probability distributions of the student model and the ensemble teacher for the j-th sentence, respectively.

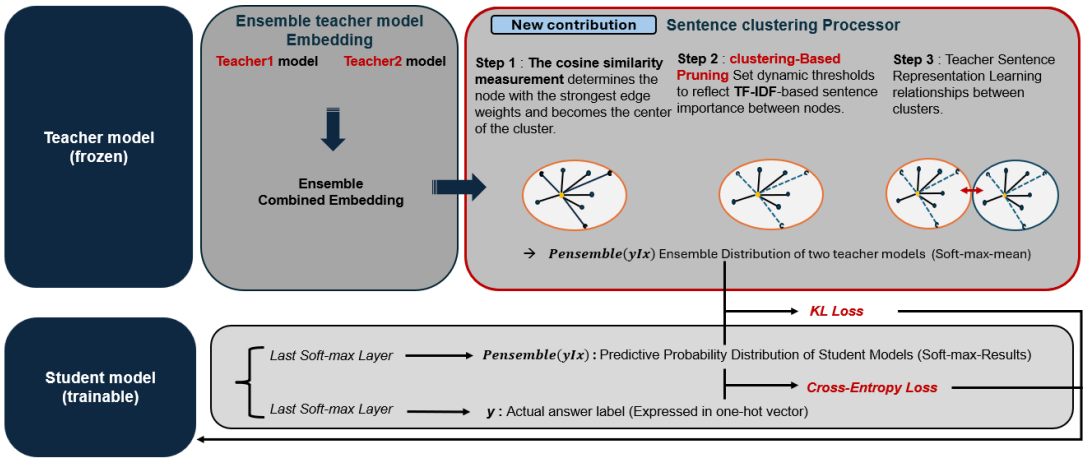


Figure 2: Overview of the Ensemble Knowledge Distillation Framework with Sentence clustering Processing

3.2 Clustering-based sentence pruning

Ensemble distillation provides a more comprehensive and nuanced knowledge representation; however, directly utilizing multiple teacher outputs often introduces redundancy, leading to increased computational overhead and decreased training efficiency for the student model. To address this issue, we propose a **clustering-based sentence pruning** mechanism that systematically filters out less informative sentences while preserving key semantic information, thereby improving both efficiency and effectiveness in knowledge transfer.

As illustrated in Figure 2, the pruning process begins by constructing a clustering representation of inter-sentence relationships(Li et al., 2019; Onan et al., 2017). Each sentence is treated as a node in a graph, with edges established based on semantic similarity measures. The weight of the edge between two sentences, w_{ij}, is computed using **cosine similarity**, formulated as:

$$w_{ij} = \cos(\mathbf{E}_{v_i}, \mathbf{E}_{v_j}) = \frac{\mathbf{E}_{v_i} \cdot \mathbf{E}_{v_j}}{\|\mathbf{E}_{v_i}\| \|\mathbf{E}_{v_j}\|} \tag{2}$$

where $\mathbf{E}_{v_i}$ and $\mathbf{E}_{v_j}$ denote the embeddings of sentences v_i and v_j, respectively.

The proposed pruning mechanism follows a structured multi-step approach. First, sentences are clustered based on their semantic similarities, where central nodes in the clustering graph represent the most contextually significant sentences. Next, a **clustering-based pruning strategy** is applied to remove less informative sentences. Specifically, a dynamic thresholding mechanism leveraging **TF-IDF scores** is used to quantify sentence importance. The importance score of each sentence is defined as:

$$I(v_i) = \frac{TF(v_i) - \min(TF)}{\max(TF) - \min(TF)} \tag{3}$$

To determine the pruning threshold τ, we adopt a **two-stage filtering strategy**. Initially, **percentile-based filtering** is applied, where sentences with importance scores below the 80th percentile are removed. Subsequently, an **adaptive thresholding** method refines the selection process by computing τ as:

$$\tau = \mu + 0.5\sigma \tag{4}$$

where μ and σ represent the mean and standard deviation of the remaining sentence scores, respectively. Sentences with importance scores below this threshold are further pruned:

$$\mathcal{V}_{pruned} = \{v_i \in \mathcal{V} \mid I(v_i) \geq \tau\} \tag{5}$$

This combined percentile-based and adaptive thresholding approach ensures that only semantically meaningful and contextually significant sentences are retained while filtering out less informative ones.

we employ an ensemble method that combines the predictive probabilities of multiple teacher models (Teacher1, Teacher2) by averaging their outputs. This approach provides a more generalized supervisory signal to the student model by merging the probability distributions predicted by individual teacher models for the same input x.

Each teacher model generates a softmax output for the input x, denoted as $P_{teacher1}(y|x)$ and $P_{teacher2}(y|x)$. These distributions are averaged to create a new probability distribution:

$$P_{ensemble}(y|x) = \frac{P_{teacher1}(y|x) + P_{teacher2}(y|x)}{2} \tag{6}$$

Following the pruning process, **sentence representation learning** is performed to enhance the structural coherence of the remaining sentences. The preserved sentences are optimized to maintain their contextual integrity while maximizing the efficiency of knowledge transfer to the student model. By leveraging inter-sentence relationships within clusters, the proposed method refines sentence embeddings to facilitate effective knowledge distillation. Through this approach, the student model benefits from a more compact yet informative representation, significantly improving both training efficiency and model performance.

3.3 Student Model Training

Averaging reduces individual model bias, enabling the student model to learn from a more stable probability distribution.

The student model is trained using both soft labels from the teacher ensemble and hard labels from ground truth. The training objective combines Cross-Entropy (CE) Loss for accurate classification and Kullback-Leibler (KL) Divergence Loss to align the student's predictions with the teacher ensemble. A weighting parameter λ balances these losses to optimize knowledge transfer.

By integrating multi-teacher ensemble learning with clustering-based sentence pruning, our framework enhances knowledge transfer efficiency while maintaining computational efficiency and model accuracy.

4 Experiments

4.1 Experimental Setup and Data Statistics

We evaluate the proposed approach on six NLP tasks from the GLUE benchmark(Wang et al., 2018), covering a diverse range of language understanding challenges. **RTE** involves textual entailment, determining whether a premise entails a hypothesis. **QQP**(Wang et al., 2018) assesses paraphrase detection, identifying whether two Quora questions are semantically equivalent. **QNLI**, reformulated from SQuAD(Rajpurkar et al., 2016), evaluates whether a context passage contains the answer to a given question. **SST-2**(Socher et al., 2013), from the Stanford Sentiment Treebank, is a binary sentiment classification task. **MNLI-m**(Williams et al., 2017) is a subset of MNLI that evaluates textual entailment across multiple genres, where test samples match the domain of the training data. **MRPC**(Dolan and Brockett, 2005) is a paraphrase detection task assessing whether two sentences express the same meaning despite wording differences. These datasets encompass textual entailment, paraphrase detection, sentiment classification, and question-answering inference, providing a robust framework for evaluating the generalization of the proposed method. The dataset statistics are summarized in Table 1.

Dataset	#Train	#Dev	#Test
RTE	2,490	277	3,000
QQP	363,849	40,430	390,965
QNLI	104,743	5,463	5,463
SST-2	67,349	872	1,821
MNLI-m	392,702	9,815	9,796
MRPC	3,668	408	1,725

Table 1: Statistics of the datasets used in the experiments.

4.2 Baseline Models and Implementation Details

For evaluating our approach, we compared it against multiple baseline methods. Vanilla Knowledge Distillation (V-KD) (Hao et al., 2023) trains student models using a single teacher, such as BERT_{12} or RoBERTa_{12}. U-Ensemble Teacher(Yang et al., 2020), averages the outputs of all teacher models by assigning them equal weights. Rand-Single-Ensemble Teacher(Fukuda et al., 2017), randomly selects a teacher model for each mini-batch to generate soft targets for student training. W-Ensemble Teacher(Chebotar and Waters, 2016), applies pre-determined, fixed weights to each teacher model. In addition to these baselines, we proposed two improved ensemble strategies. LR-Ensemble Teacher employs a Logistic Regression-based approach to adaptively compute the optimal weights for teacher models. Depending on whether the weights are learned from the training set or the development set, the method is referred to as LR-Train-Ensemble and LR-Dev-Ensemble, respectively. For the teacher models, we fine-tuned widely-used transformer architectures, including BERT_{12} and RoBERTa_{12}, where the subscript 12 denotes that each model consists of 12 transformer layers. To construct student models, we utilized simplified versions of BERT, incorporating 4 and 6 transformer layers, denoted as BERT_4 and BERT_6, respectively. This aligns with the methodology presented in Patient KD (Sun et al., 2019).

4.3 Experimental Setup

Our experiments followed the Patient KD framework (Sun et al., 2019). The student models, BERT_4 and BERT_6, were initialized using the bottom 4 and 6 layers of BERT-Base. Their distillation process involved tuning hyperparameters such as temperature T values $\{5, 10, 20\}$, loss balance coefficients α $\{0.2, 0.5, 0.7\}$, and γ values $\{0.3, 0.5, 0.7, 0.9\}$, optimized based on the development set.

For fine-tuning the teacher models, we utilized publicly available pre-trained weights from BERT_{12} and RoBERTa. The training setup included learning rates of $\{1e-5, 2e-5, 5e-5\}$, a batch size of 32, a sequence length of 128, and 4 training epochs. The best-performing model was selected based on accuracy on the development set.

To enhance the distillation process, a logistic regression-based policy function was employed for teacher selection, optimized using Monte Carlo policy gradients (Williams, 1992). During knowledge distillation pretraining, the student model was initialized with pre-trained BERT weights and trained further using an average ensemble of teacher outputs.

4.4 Main Results

Following pretraining, Knowledge Distillation (KD) and Teacher Selection (TS) models (Ye et al., 2020; Amara et al., 2022; Lee et al., 2023) were trained iteratively in an alternating manner.

Table 2 provides general performance metrics for various natural language processing (NLP) models and is included to facilitate the understanding of Table 3. Table 2 presents a comparison of different models in terms of the number of parameters, computational cost (FLOPs), speedup, and task-specific performance.

Table 3 evaluates the impact of different teacher configurations on student model performance. Multi-teacher distillation with Sentence clustering processor generally improves accuracy across tasks. For RTE, using both BERT-Base and RoBERTa achieves the highest accuracy (0.69), indicating that diverse pre-trained models enhance generalization. SST-2 reaches 0.95 accuracy with BERT-Base and BERT-Large, highlighting the effectiveness of multi-teacher distillation for sentiment classification. QNLI shows mixed results: accuracy improves from 0.83 to 0.84 with RoBERTa as a teacher but drops to 0.59 when using a smaller DistilBERT-4 student, suggesting that relational reasoning tasks may require more structured knowledge transfer.

Table 4 compares various knowledge distillation methods, showing that Our Method achieves 87.17 accuracy on MNLI-m and 95.4 on SST-2, surpassing conventional approaches and demonstrating strong generalization in sentiment analysis tasks. However, performance on MRPC (70.9) and RTE (60.7) is slightly lower than RL-KD-based methods, indicating that additional optimization may be needed for low-resource datasets due to limited reward signals. Overall, Our Method excels in large-scale NLU tasks, highlighting the effectiveness of reinforcement learning-based distillation. Teacher Sentence Representation Learning, the final stage of the Sentence clustering Processor, enhances knowledge transfer by capturing contextual relationships, though improvements in structural representation may be required for tasks like RTE.

The inference time was measured both before and after applying sentence pruning to evaluate its impact on computational efficiency. the model processed all input sentences, and the inference time (T_{base}) was measured. After pruning, a sentence selection mechanism was applied, where less important sentences were removed based on similarity and saliency scores. The remaining sentences were then passed through the model and the pruned inference time (T_{pruned}) was measured using the same method. The speed-up factor (S) was calculated as the ratio of the baseline to the pruned inference time ($S = T_{base}/T_{pruned}$), quantifying the reduction in computational cost achieved through pruning.

Table 5 presents the impact of sentence pruning on accuracy and F1 score across three GLUE tasks: SST-2, RTE, and QNLI. The pruning process led to varying effects on model performance, with accuracy retention differing across tasks. In the SST-2 dataset, the pruning rate was 5.7%, resulting in a marginal decrease of 0. 50% in precision and 0. 34% in the F1 score, indicating that the model remained relatively robust to pruning. Conversely, in the RTE dataset, pruning led to a significant improvement in accuracy, increasing from 64.29% to 68.75% (+4.5%), with a corresponding F1 score increase of +2.6%. This suggests that pruning effectively removed non-informative sentences, thereby enhancing model performance. In contrast, for QNLI, which had a pruning rate of 31.7%, the accuracy decreased slightly by 0.62%, and the F1 score was reduced by 0.35%. These results indicate that

System	#Params	#FLOPs	Speedup	RTE	QQP	QNLI	SST-2	Avg
BERT_BASE (Teacher)	109M	22.5B	1.0x	67.0	71.1	90.9	93.4	**80.6**
DistilBERT_6 (Student)	67.0M	11.3B	2.0x	58.4	70.1	88.9	92.5	**77.5**
DistilBERT_4 (Student)	52.2M	7.6B	3.0x	54.1	68.5	85.2	91.4	**74.8**
BERT_LARGE (Teacher)	340M	110B	-	70.4	70.4	92.3	93.2	**81.6**
RoBERTa (Teacher)	125M	40B	-	86.6	86.6	94.7	96.4	**91.1**

Table 2: Comparison of different models in terms of parameter count, FLOPs, speedup, and performance across NLP tasks. The best accuracy for each task is highlighted in **bold**.

Task	Teacher1	Teacher2	Student	Accuracy	Trend
RTE	bert-base-uncased	bert-large-uncased	bert-base-uncased	0.68	↑
QQP	bert-base-uncased	bert-large-uncased	bert-base-uncased	0.85	↑
QNLI	bert-base-uncased	bert-large-uncased	bert-base-uncased	0.83	↓
SST-2	bert-base-uncased	bert-large-uncased	bert-base-uncased	**0.94**	↑
Avg	-	-	-	**0.82**	↑
RTE	**bert-base-uncased**	**roberta-base**	**distilbert-base-uncased_6**	0.69	↑
QQP	**bert-base-uncased**	**roberta-base**	**distilbert-base-uncased_6**	0.86	↑
QNLI	**bert-base-uncased**	**roberta-base**	**distilbert-base-uncased_6**	0.84	↓
SST-2	bert-base-uncased	roberta-base	distilbert-base-uncased_6	**0.93**	↑
Avg	-	-	-	**0.83**	↑
RTE	bert-base-uncased	bert-large-uncased	distilbert-base-uncased_4	0.61	↑
QQP	bert-base-uncased	bert-large-uncased	distilbert-base-uncased_4	0.72	↑
QNLI	bert-base-uncased	bert-large-uncased	distilbert-base-uncased_4	0.59	↓
SST-2	**bert-base-uncased**	**bert-large-uncased**	**distilbert-base-uncased_4**	**0.95**	↑
Avg	-	-	-	**0.72**	↓

Table 3: Performance comparison of different student models trained with various teacher configurations. The best accuracy per task is highlighted in **bold**, and performance changes are indicated with arrows.

while pruning improves computational efficiency, its impact on accuracy is task-dependent.

Table 6 evaluates the inference time and computational speed-up achieved through pruning. The inference time was measured by starting a timer immediately before passing the input data to the model. For the Baseline model, the full input text was provided without any pruning, while the Pruned model first applied a clustering-based sentence pruning mechanism, retaining only the most informative sentences before feeding them into the model. The timer was stopped as soon as the model produced the output, ensuring that only the forward pass execution time was recorded.

All experiments were conducted using NVIDIA RTX 6000 Ada GPUs (0 to 7), leveraging multi-GPU parallelism to optimize inference efficiency.

The SST-2 task, despite a relatively low pruning rate (5.7%), exhibited a substantial speed-up of 23×, with inference time reducing from 1.15s to 0.05s. Similarly, RTE achieved a 2.5× reduction in inference time, dropping from 0.05s to 0.02s, demonstrating that pruning significantly reduces computational overhead while maintaining or even improving accuracy. For QNLI, the inference time improved from 0.37s to 0.23s, corresponding to a 1.6× speed-up. While the absolute speed gains vary by task, the results highlight the efficiency gains

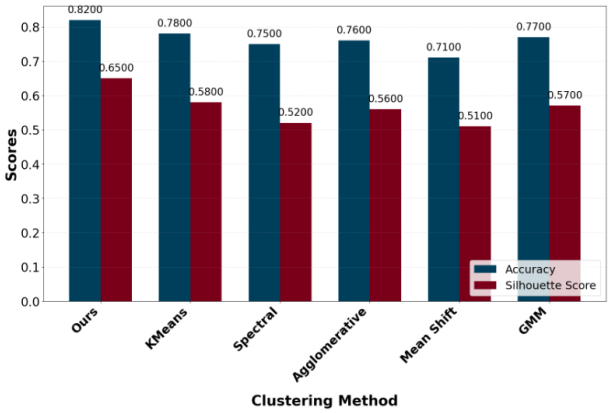


Figure 3: Clustering Method Performance: Comparison of Accuracy and Silhouette Score across different clustering methods.(Using MNLI-m Dataset)

achievable through pruning, particularly in tasks where less important information can be effectively removed without degrading performance.

These findings collectively indicate that sentence pruning can enhance inference efficiency with minimal impact on accuracy, and in some cases, even improve performance by reducing noise. However, the extent of performance retention is highly task-dependent, suggesting that pruning strategies should be carefully designed based on task characteristics.

Figure 3 compares the performance of various

Teacher	Student	Strategy	MNLI-m (Acc.)	MRPC (Acc.)	RTE (Acc.)	SST-2 (Acc.)
Rand-Single-Ensemble	BERT6	V-KD	80.7	77.7	61.7	90.6
W-Ensemble	BERT6	V-KD	77.2	81.1	62.1	90.6
LR-Dev-Ensemble	BERT6	V-KD	81.1	80.6	64.6	90.8
Best-Single-Ensemble	BERT6	V-KD	80.5	80.4	66.1	90.3
MT-BERT-Ensemble	BERT6	RL-KD	-	-	**75.7**	94.6
RL-KD (reward1)	BERT6	RL-KD	82.0	82.8	67.1	91.7
RL-KD (reward2)	BERT6	RL-KD	82.1	82.1	67.2	91.4
RL-KD (reward3)	BERT6	RL-KD	81.6	**83.3**	68.2	92.3
Our Method	BERT6	RL-KD	**87.17**	70.9	60.7	**95.4**

Table 4: Performance comparison of different knowledge distillation methods using BERT6 as the student model. **Rand-Single-Ensemble** selects a single teacher randomly, while W-Ensemble applies weighted averaging of multiple teachers. **LR-Dev-Ensemble** employs logistic regression for teacher selection, and **Best-Single-Ensemble** chooses the best-performing teacher. **RL-KD (reward1, reward2, reward3)** uses reinforcement learning with different reward settings for knowledge transfer. **MT-BERT** utilizes multi-teacher co-finetuning with shared pooling and multi-teacher losses.

Task	Prune	Acc		ΔAcc.	F1		ΔF1
	Rate (%)	Base	Pruned	(%)	Base	Pruned	(%)
SST-2	5.7	51.72	51.22	-0.50	39.27	38.93	-0.34
RTE	**32.8**	64.29	**68.75**	**+4.5**	53.46	**56.02**	**+2.6**
QNLI	31.7	44.32	43.70	-0.62	39.09	38.74	-0.35

Table 5: Impact of Sentence Pruning on Accuracy and F1 Score.

Task	Inf. Time (s) Base	Inf. Time (s) Pruned	Speed-up (×)
SST-2	1.15	**0.05**	**23×**
RTE	0.05	**0.02**	**2.5×**
QNLI	0.37	**0.23**	**1.6×**

Table 6: Impact of Sentence Pruning on Inference Time and Speed-up.

clustering methods: Sentence Clustering Processor (Ours), KMeans, Spectral, Agglomerative, Mean Shift, and GMM. The blue bars represent accuracy, and the red bars indicate silhouette scores. Sentence Clustering Processor (Ours) achieves the highest accuracy (0.82) and a competitive silhouette score (0.65), demonstrating its effectiveness.

KMeans, a centroid-based method, partitions data into K clusters by minimizing the distance between points and cluster centers, showing high accuracy (0.78) but a lower silhouette score (0.58). Spectral Clustering, which utilizes graph theory and eigenvectors of similarity matrices, performs moderately due to its reliance on pairwise similarities. Agglomerative Clustering, a hierarchical method merging clusters based on proximity, shows stable but average performance. Mean Shift, which iteratively shifts points to high-density regions, exhibits moderate scores. GMM (Gaussian Mixture Model), a probabilistic method modeling data as a mixture of Gaussian distributions, achieves a balance between accuracy (0.77) and silhouette score (0.57). Overall, our Sentence Clustering Processor outperforms traditional techniques in both accuracy and cluster quality.

5 Conclusion

In this study, we propose a **Clustering-Based Knowledge Distillation with Sentence Pruning** framework to enhance student model training efficiency. Our approach integrates multiple teacher models for improved knowledge diversity and employs a **clustering-based sentence pruning mechanism** to remove less important content from input representations. Experiments on **SST-2, RTE, and QNLI** benchmarks show that our method retains high accuracy while significantly **reducing inference time and computational cost**. Notably, pruning improved RTE accuracy by **4.5%** and achieved up to **23× inference speedup** on SST-2, demonstrating its effectiveness for **resource-constrained environments**. Furthermore, **multi-teacher knowledge distillation** combined with **task-aware sentence pruning** enhances student model performance by filtering irrelevant information, addressing the redundancy often found in traditional knowledge distillation methods.

6 Limitations

Sentence pruning significantly enhances inference speed, achieving up to 23× acceleration, but introduces a trade-off between efficiency and accuracy. Tasks relying on rich contextual information may suffer from accuracy degradation, necessitating adaptive pruning strategies that adjust based on task complexity.

References

Josh Achiam et al. 2023. Gpt-4 technical report. *arXiv preprint*, arXiv:2303.08774.

Ibtihel Amara et al. 2022. Ces-kd: curriculum-based expert selection for guided knowledge distillation. In *2022 26th International Conference on Pattern Recognition (ICPR)*, pages 1901–1907. IEEE.

Abdollah Amirkhani et al. 2021. Robust semantic segmentation with multi-teacher knowledge distillation. *IEEE Access*, 9:119049–119066.

Yevgen Chebotar and Austin Waters. 2016. Distilling knowledge from ensembles of neural networks for speech recognition. In *Interspeech*, pages 3439–3443.

Pieter Delobelle, Thomas Winters, and Bettina Berendt. 2020. Robbert: A dutch roberta-based language model. *arXiv preprint arXiv:2001.06286*.

Bill Dolan and Chris Brockett. 2005. Automatically constructing a corpus of sentential paraphrases. In *Third International Workshop on Paraphrasing (IWP2005)*.

Shangchen Du et al. 2020. Agree to disagree: Adaptive ensemble knowledge distillation in gradient space. In *Advances in Neural Information Processing Systems*, volume 33, pages 12345–12355.

Yang Fan et al. 2021. Learning to reweight with deep interactions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 7385–7393.

Takashi Fukuda et al. 2017. Efficient knowledge distillation from an ensemble of teachers. In *Interspeech*, pages 3697–3701.

Minghong Gao. 2023. A survey on recent teacher-student learning studies. *arXiv preprint*, arXiv:2304.04615.

Jianping Gou et al. 2021. Knowledge distillation: A survey. *International Journal of Computer Vision*, 129(6):1789–1819.

Yuxian Gu et al. 2024. Minillm: Knowledge distillation of large language models. In *The Twelfth International Conference on Learning Representations*.

Md Akmal Haidar et al. 2021. Rail-kd: Random intermediate layer mapping for knowledge distillation. *arXiv preprint*, arXiv:2109.10164.

Zhiwei Hao et al. 2023. Vanillakd: Revisit the power of vanilla knowledge distillation from small scale to large scale. *arXiv preprint*, arXiv:2305.15781.

Zhang-Wei Hong, Prabhat Nagarajan, and Guilherme Maeda. 2021. Periodic intra-ensemble knowledge distillation for reinforcement learning. In *Machine Learning and Knowledge Discovery in Databases. Research Track: European Conference, ECML PKDD 2021, Bilbao, Spain, September 13–17, 2021, Proceedings, Part I*, pages 87–103. Springer International Publishing.

Xiaoqi Jiao, Yichun Yin, Lifeng Shang, Xin Jiang, Xiao Chen, Linlin Li, Fang Wang, and Qun Liu. 2020. Tinybert: Distilling bert for natural language understanding. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4163–4174.

Mikhail V. Koroteev. 2021. Bert: A review of applications in natural language processing and understanding. *arXiv preprint*, arXiv:2103.11943.

Hayeon Lee et al. 2023. A study on knowledge distillation from weak teacher for scaling up pre-trained language models. *arXiv preprint*, arXiv:2305.18239.

Lianqiang Li, Jie Zhu, and Ming-Ting Sun. 2019. A spectral clustering based filter-level pruning method for convolutional neural networks. *IEICE TRANSACTIONS on Information and Systems*, 102(12):2624–2627.

Aytuğ Onan, Serdar Korukoğlu, and Hasan Bulut. 2017. A hybrid ensemble pruning approach based on consensus clustering and multi-objective evolutionary algorithm for sentiment classification. *Information Processing & Management*, 53(4):814–833.

Cuong Pham, Tuan Hoang, and Thanh-Toan Do. 2023. Collaborative multi-teacher knowledge distillation for learning low bit-width deep neural networks. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 6435–6443.

Zengyu Qiu et al. 2022. Better teacher better student: Dynamic prior knowledge for knowledge distillation. *arXiv preprint arXiv:2206.06067*.

Pranav Rajpurkar et al. 2016. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint*, arXiv:1606.05250.

Mohammadreza Sadeghi, Zihan Wang, and Narges Armanfard. 2024. Forward-backward knowledge distillation for continual clustering. *arXiv preprint arXiv:2405.19234*.

Baitan Shao and Ying Chen. 2023. Decoupled knowledge with ensemble learning for online distillation. *arXiv preprint arXiv:2312.11218*.

Richard Socher et al. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1631–1642.

Jie Song et al. 2022. Spot-adaptive knowledge distillation. *IEEE Transactions on Image Processing*, 31:3359–3370.

Siqi Sun et al. 2019. Patient knowledge distillation for bert model compression. *arXiv preprint*, arXiv:1908.09355.

Alex Wang et al. 2018. Glue: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint*, arXiv:1804.07461.

Chaofei Wang et al. 2022. Learn from the past: Experience ensemble knowledge distillation. In *2022 26th International Conference on Pattern Recognition (ICPR)*, pages 4736–4743. IEEE.

Jingxuan Wei, Yifan Gao, Rui Wang, Kehai Chen, Masao Utiyama, Eiichiro Sumita, and Tiejun Zhao. 2024. Sentence-level or token-level? a comprehensive study on knowledge distillation. *arXiv preprint*, arXiv:2404.14827.

Adina Williams, Nikita Nangia, and Samuel R. Bowman. 2017. A broad-coverage challenge corpus for sentence understanding through inference. *arXiv preprint arXiv:1704.05426*.

Ronald J. Williams. 1992. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*, 8:229–256.

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

Chuhan Wu et al. 2022. Unified and effective ensemble knowledge distillation. *arXiv preprint arXiv:2204.00548*.

Guodong Xu et al. 2020. Knowledge distillation meets self-supervision. In *European Conference on Computer Vision*, pages 588–604, Cham. Springer International Publishing.

Ze Yang et al. 2020. Model compression with two-stage multi-teacher knowledge distillation for web question answering system. In *Proceedings of the 13th International Conference on Web Search and Data Mining (WSDM)*, pages 690–698.

Han-Jia Ye, Su Lu, and De-Chuan Zhan. 2020. Distilling cross-task knowledge via relationship matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12396–12405.

Fei Yuan et al. 2021. Reinforced multi-teacher selection for knowledge distillation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 14284–14291.

Mengyang Yuan, Bo Lang, and Fengnan Quan. 2024. Student-friendly knowledge distillation. *Knowledge-Based Systems*, 296:111915.

Shuoxi Zhang, Zijian Song, and Kun He. 2024. Neural collapse inspired knowledge distillation. *arXiv preprint arXiv:2412.11788*.