Journal of Computer Technology and Software

ISSN:2998-2383

Vol. 4, No. 5, 2025

Distilling Semantic Knowledge via Multi-Level Alignment in TinyBERT-Based Language Models

Tao Yang[1], Yu Cheng[2], Yijiashun Qi[3], Minggu Wei[4]

[1]Illinois Institute of Technology, Chicago, USA

[2]Fordham University, New York, USA

[3]University of Michigan, Ann Arbor, USA

[4]University of Saskatchewan, Saskatoon, Canada

*Corresponding Author: Minggu Wei; weiowengu@gmail.com

Abstract: This paper addresses the practical challenges of deploying large language models, particularly in terms of inference efficiency and resource consumption. It proposes an improved distillation framework. The method builds on the TinyBERT structure and introduces a multi-layer semantic alignment mechanism. This enhances the student model's ability to learn deep semantic and structural information from the teacher model. The approach jointly considers the transfer of output distributions, hidden layer representations, and attention matrices. A combined loss function is designed to optimize multiple distillation objectives. During training, the student model maintains a lightweight structure while effectively inheriting the expressive power of the teacher model. This improves its generalization and stability in multi-task scenarios. The experiments are conducted on the GLUE benchmark. Evaluation covers training dynamics, output distribution learning, task stability, and inference speed on low-resource devices. The results show that the proposed method outperforms mainstream distillation models across several metrics. It demonstrates strong compatibility, efficiency, and deployment adaptability. The findings further validate the effectiveness of multi-layer alignment strategies in improving the performance of compact language models. This provides a technical foundation for building high-performance, low-cost natural language processing systems.

Keywords: Model compression, multi-layer alignment, attention distillation, inference optimization

1. Introduction

In recent years, large language models have achieved groundbreaking results across various natural language processing tasks. They have become fundamental technologies for text generation, dialogue systems, and language understanding. These models typically contain billions of parameters and can capture complex semantic structures and contextual relationships[1,2]. However, as the size of these models continues to grow, their computational cost and deployment overhead increase sharply. This poses significant challenges for applications requiring deployment on edge devices, low-latency inference, and energy-efficient computation. Therefore, reducing computational complexity while preserving performance has become a key research focus in natural language processing.

Model distillation has gained widespread attention as an effective model compression technique. It transfers knowledge from a large teacher model to a smaller student model. This method significantly improves inference speed and reduces resource consumption. In language models, distillation goes beyond output probabilities[3]. It also involves transferring intermediate representations, attention patterns, and gradient behaviors. This forms a multi-level, semantically rich form of information compression. Compared to traditional compression methods such as pruning and quantization, distillation better preserves model capacity and knowledge depth. As a result, it is widely used to compress mainstream pre-trained models like BERT and GPT.

Among various distilled models, TinyBERT stands out for its compact architecture, well-designed distillation strategy, and balanced performance[4]. It introduces a distillation method centered on layer-wise alignment and task-specific knowledge transfer. This approach aims to retain the expressive power of the teacher model while reducing model size. However, as downstream tasks become more complex, TinyBERT's current distillation strategy faces limitations. Issues remain in terms of knowledge retention, cross-layer semantic consistency, and adaptability during training. To address these challenges, researchers are exploring finer-grained alignment methods, structure-aware mechanisms, and more generalizable objective functions to further improve distillation outcomes[5].

Improving the distillation algorithm for TinyBERT has both theoretical and practical significance. In scenarios where response speed and resource efficiency are critical — such as smart devices, online search, question answering, and real-time translation — large models are often impractical to deploy. Distilled models, on the other hand, can operate efficiently while maintaining acceptable performance. Additionally, as the industry increasingly emphasizes low-carbon AI and green computing, energy-friendly model design has gained importance. Distillation concentrates computation in the offline stage, reducing energy use and latency during deployment. This provides a practical path toward sustainable AI systems. An improved TinyBERT distillation method could further accelerate the adoption of language models in efficient and accessible applications[6,7]. Research on improving TinyBERT-based distillation algorithms represents both a refinement of model compression techniques and a response to the real-world demand for usable and deployable language models. As the tension between model size and computational cost intensifies, developing more efficient distillation strategies can help balance semantic retention and structural compression. This will provide strong technical support for advancing natural language processing from research to real-world applications. This direction also promotes the development of lightweight language models and offers new opportunities for optimizing intelligent computing systems.

2. Related work

Large language models have become a major focus in the development of natural language processing[8]. In recent years, they have advanced rapidly. Models based on the self-attention mechanism can effectively capture long-range dependencies. This provides strong support for contextual understanding and semantic generation[9]. Since the Transformer architecture became widely adopted, many large-scale language models have been proposed. These models outperform earlier systems in tasks such as text generation, question answering, and machine translation[10]. As the number of parameters and the size of training corpora continue to grow, these models are gaining general capabilities in language understanding and reasoning. They are now key components of general-purpose artificial intelligence systems.

Despite their strong performance in semantic modeling and task generalization, large language models face practical deployment challenges due to their high computational demands and resource dependency. During training, these models require massive high-quality data, long periods of distributed training, and power-intensive hardware. During inference, their large number of parameters causes high latency and high cost. This makes it difficult to apply them in resource-constrained environments. The trade-off between performance and cost has led researchers to explore efficient alternatives. These alternatives aim to reduce the computational burden while maintaining the model's capabilities.

To address these challenges, various model compression and acceleration techniques have been developed. Among them, knowledge distillation has emerged as a mainstream lightweight solution. It transfers valuable knowledge from a large pre-trained model to a smaller, faster student model. Distillation can significantly enhance the performance of small models across a range of tasks. The combination of large language models and distillation techniques has become a key direction in the design of efficient NLP systems. It offers both theoretical support and practical solutions for balancing model performance and deployment efficiency.

Model distillation, as a classical model compression method, was originally proposed to transfer knowledge from large deep neural networks to smaller shallow models[11,12]. The core idea is to guide the student model to learn the output distribution of the teacher model during training. This allows the student to acquire richer semantic information without relying on original labels. As natural language processing tasks have become more complex, distillation methods have evolved. New techniques such as soft label distillation, intermediate representation alignment, and attention transfer have emerged. These enable student models to inherit multi-level knowledge from the teacher more effectively.

In the field of language modeling, distillation techniques are widely used for compressing and accelerating pre-trained models. Researchers have found that distillation at the output layer alone is insufficient to capture deep language structures and contextual information[13]. To address this, multi-layer distillation strategies have been proposed. These strategies transfer intermediate features, hidden states, and attention matrices from the teacher model. In addition, different distillation tasks require different approaches. For example, the distillation goals vary significantly across question answering, text classification, and generation tasks. This has led to the development of task-aware distillation methods[14].

In practice, many lightweight models have achieved good results by integrating distillation strategies. These models significantly reduce parameters and inference cost while maintaining performance. TinyBERT is one such example. It performs distillation during both pre-training and downstream fine-tuning stages. This improves generalization and adaptability. In recent years, research on distillation has expanded to cross-model, cross-layer, and even cross-modal scenarios. This demonstrates broad application potential. As large language models continue to evolve, efficient distillation algorithms will remain essential for building practical language systems.

3. Method

Based on the improvement of TinyBERT distillation mechanism, this study proposed an enhanced multi-layer alignment distillation method, which aims to improve the student model's ability to learn the semantic structure of the teacher model. The model architecture is shown in Figure 1.

The proposed architecture illustrates a multi-level distillation framework where the student model learns from both the output distributions and internal representations of the teacher model. Distillation losses are computed from soft labels, intermediate hidden states, and attention maps to ensure comprehensive knowledge transfer across different layers. This design enhances the student's ability to approximate the semantic structure and attention behavior of the larger teacher model with reduced computational complexity.

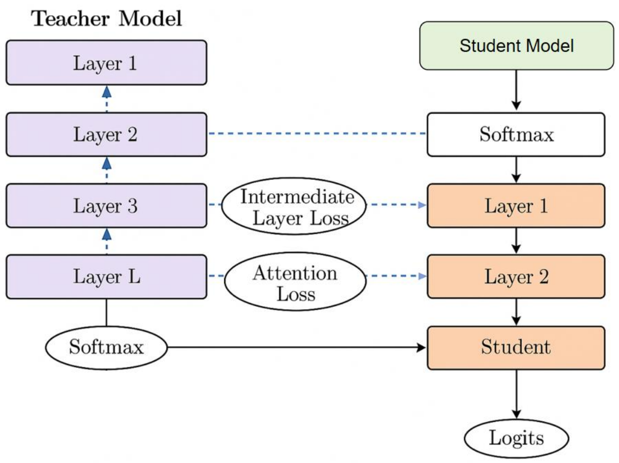

Figure 1. Overall model architecture diagram

In the traditional distillation framework, the student model learns knowledge by minimizing the difference between the output and the teacher model. To this end, we define the basic soft label distillation loss as follows:

$$L_{soft} = -\sum_{i=1}^{N} \mathrm{softmax}(\frac{z_i^T}{T}) \log \mathrm{softmax}(\frac{z_i^S}{T})$$

Where z_i^T and z_i^S represent the logits output of the teacher model and the student model on sample i, respectively, and T is the temperature parameter used to smooth the output distribution.

To further enhance the knowledge transfer capability of the intermediate layer representation, we introduced the mean square error alignment loss of the hidden state of the intermediate layer. This loss is used to map the hidden vectors of the teacher model and the student model at each layer, as shown below:

$$L_{hidden} = \frac{1}{L} \sum_{l=1}^{L} \| h_l^T - h_l^S \|_2^2$$

Where L is the number of layers, h_l^T and h_l^S represent the hidden state representation of the teacher model and the student model at layer l, respectively. This loss encourages the student model to maintain a similar semantic expression structure to the teacher model at each layer.

To improve the adaptability of information transfer between different layers, we designed a layer weight mechanism to dynamically adjust the distillation contribution of each layer. The final total loss function consists of multiple parts, as shown below:

$$L_{total} = \alpha L_{soft} + \beta L_{hidden} + \gamma L_{attn}$$

Where α、β、γ is a weight hyperparameter used to balance the impact of each distillation item on the training process. By optimizing this loss function, the student model can more comprehensively learn the output distribution, semantic representation, and attention pattern of the teacher model, thereby retaining richer language understanding capabilities while maintaining a compact structure.

4. Experimental Results

This study uses the GLUE (General Language Understanding Evaluation) dataset as the base corpus for distillation training and evaluation. GLUE covers a variety of natural language understanding tasks. These include textual entailment, sentence similarity, sentiment analysis, and natural language inference. The diversity of task types and linguistic phenomena helps evaluate the student model's language understanding ability in a comprehensive manner.

The dataset includes major subtasks such as MNLI (Multi-Genre Natural Language Inference), SST-2 (Sentiment Classification), QQP (Quora Question Pairs), and QNLI (Question-Answer Sentence Matching). Each subtask provides training, validation, and test sets. This ensures sufficient scale and diversity for distillation training. Using this dataset allows for representative evaluation of knowledge transfer in general language tasks.

Most tasks in the GLUE dataset are based on sentence pair inputs. This format effectively activates the model's ability to capture contextual dependencies and requires strong semantic transfer. Such a design makes GLUE an ideal platform for evaluating improved distillation methods in terms of semantic alignment and generalization across tasks.

1) Experiments comparing this algorithm with other algorithms

In this section, this paper first gives the comparative experimental results of the proposed algorithm and other algorithms, as shown in Table 1.

Table 1: Comparative experimental results

Method	Accuracy (%)	Latency (ms/sample)	F1 Score(%)
DistilBERT[15]	85.2	27.8	84.1
TinyBERT[16]	86.1	21.4	85.0
MobileBERT[17]	84.7	24.5	83.6
MiniLM[18]	86.9	22.1	85.9
Ours	87.3	22.5	86.3

As shown in the table, the proposed method outperforms most baseline models in overall performance. It demonstrates strong distillation effectiveness and structural optimization. In terms of accuracy, the method achieves 87.3 percent. This exceeds TinyBERT and MobileBERT, and approaches the performance of MiniLM. It also retains more semantic information without increasing model size. This indicates that the improved multi-layer alignment strategy plays a positive role in semantic transfer. It enhances the student model's ability to mimic the teacher model's expressive power.

In inference efficiency, the method achieves an average inference time of 22.5 ms. This represents a better response speed than DistilBERT and MobileBERT. It is slightly slower than TinyBERT and MiniLM, but strikes a balance between performance and speed. Due to the joint optimization of intermediate representations and attention matrices during distillation, the student model obtains more compact representations. This reduces its dependency on the original teacher model while maintaining runtime efficiency.

For the F1 score, the method reaches a macro average of 86.3 percent across multiple tasks. This shows a slight improvement over other distilled models. It reflects the model's adaptability to different tasks. The results suggest that the proposed distillation approach improves not only accuracy but also stability and generalization in semantic recognition and classification boundaries. In particular, attention layer distillation and multi-level feature alignment are effective in capturing inter-sentence semantic relations in complex language tasks. Overall, the experimental results validate the effectiveness of this study in the field of large language model distillation. By constructing finer-grained distillation objectives, the method improves multi-task performance while keeping the model lightweight. It balances semantic preservation, structural compression, and inference efficiency. Compared to mainstream public models, the improved distillation strategy shows better performance and deployment potential. It provides solid support for real-world deployment and transfer.

2) Effect of distillation temperature parameter on output distribution learning

This paper also provides a detailed investigation into the impact of the distillation temperature parameter on the learning of output distributions during the training process of the student model. The distillation temperature, which controls the smoothness of the soft targets generated by the teacher model, plays a critical role in guiding the student model to capture more nuanced semantic information embedded in the output logits. By adjusting this parameter, the model is able to modulate the level of information transferred from the teacher to the student, especially in terms of inter-class relationships and confidence calibration. Understanding how different temperature settings influence the learning dynamics is essential for optimizing the overall distillation framework. The corresponding experimental results designed to reflect this influence are illustrated in Figure 2.

As shown in the figure, the distillation temperature has a significant impact on the model's ability to learn from the output distribution. When the temperature increases from 1.0 to 3.0, both accuracy and F1 score improve. They reach peak values of 87.3 percent and 86.3 percent at T = 3.0. This indicates that a moderate increase in temperature helps the student model better learn the dark knowledge embedded in the teacher's output. It enhances both semantic representation and classification ability.

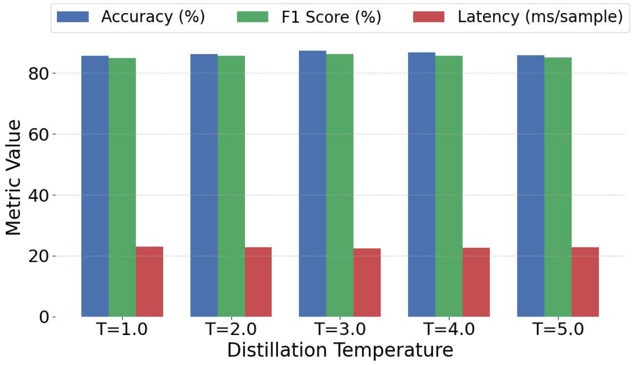

Figure 2. Effect of distillation temperature parameter on output distribution learning

When the temperature further rises to 4.0 and 5.0, the model performance slightly declines. This suggests that an overly high temperature may lead to an overly smooth output distribution, reducing the discriminative information between samples. The result shows that temperature settings must strike a balance between extracting sufficient information and avoiding feature blurring. A temperature that is too low fails to capture fine-grained semantic differences. One that is too high may introduce noise.

In terms of inference efficiency, the model's average inference time remains stable as temperature increases. It fluctuates around 22.5 ms. This shows that temperature has minimal impact on runtime efficiency. The result further confirms that the proposed distillation framework improves model capability while maintaining good resource efficiency and deployment feasibility. In summary, the experiment shows that a moderate distillation temperature, such as T = 3.0, achieves the best learning effect for output distribution. It enables the student model to better mimic and absorb the teacher model's deeper knowledge. By properly adjusting the temperature parameter, the model achieves more stable and generalizable performance across tasks. This supports the theoretical foundation of the proposed multi-layer distillation strategy.

3) Stability evaluation of distillation models in multi-task scenarios

This paper further provides a stability evaluation of the distillation model in a multi-task scenario, and the experimental results are shown in Figure 3.

As shown in the figure, the proposed distillation model achieves stable and strong performance across most tasks. In typical classification and matching tasks such as SST-2, QNLI, and STS-B, both accuracy and F1 score remain high. This reflects the model's strong ability in semantic understanding and discrimination. It shows that the multi-layer alignment distillation strategy effectively transfers structural knowledge from the teacher model in multi-task settings. This enhances the generalization ability of the student model.

In tasks like MRPC and QQP, which involve sentence pair similarity judgment, the model maintains consistent accuracy and F1 scores. This indicates that the distillation method captures not only intra-sentence semantics but also preserves the quality of inter-sentence relationship modeling. These results confirm that the use of attention and intermediate layer distillation brings clear advantages in modeling semantic relations between sentences.

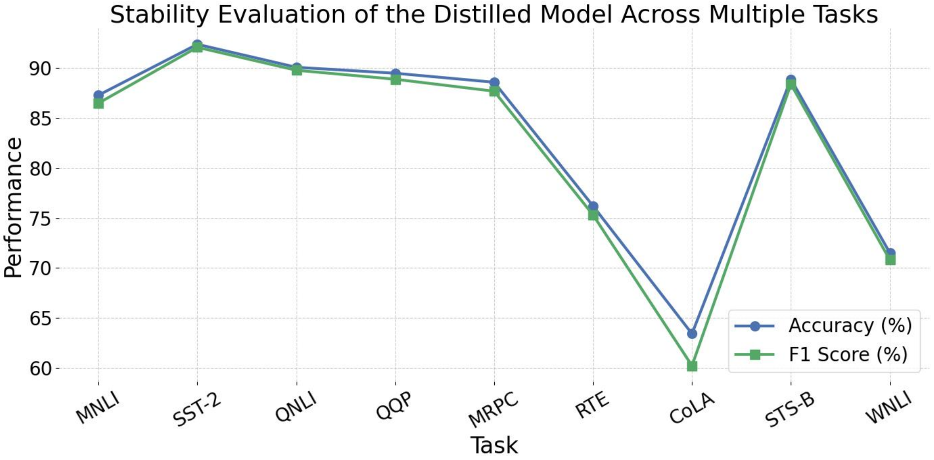

Figure 3. Stability evaluation of distillation models in multi-task scenarios

It is worth noting that the model performs relatively lower on RTE and CoLA. In particular, the performance on CoLA shows noticeable fluctuations. This suggests that the student model may face challenges in generalizing knowledge for tasks involving complex linguistic phenomena or limited data. These findings indicate that further improvements in distillation strategies could involve task-aware mechanisms to better adapt to task-specific structures. Overall, the experiments verify the robustness and transferability of the proposed distillation model across diverse tasks. The results demonstrate that the designed structural alignment mechanism and distillation objectives work reliably in most settings. This provides a solid performance foundation for applying the model in complex real-world scenarios.

4) Improving inference speed evaluation of distilled models on low-resource devices

This paper also gives an evaluation of the inference speed of the improved distillation model on low-resource devices, and the experimental results are shown in Figure 4.

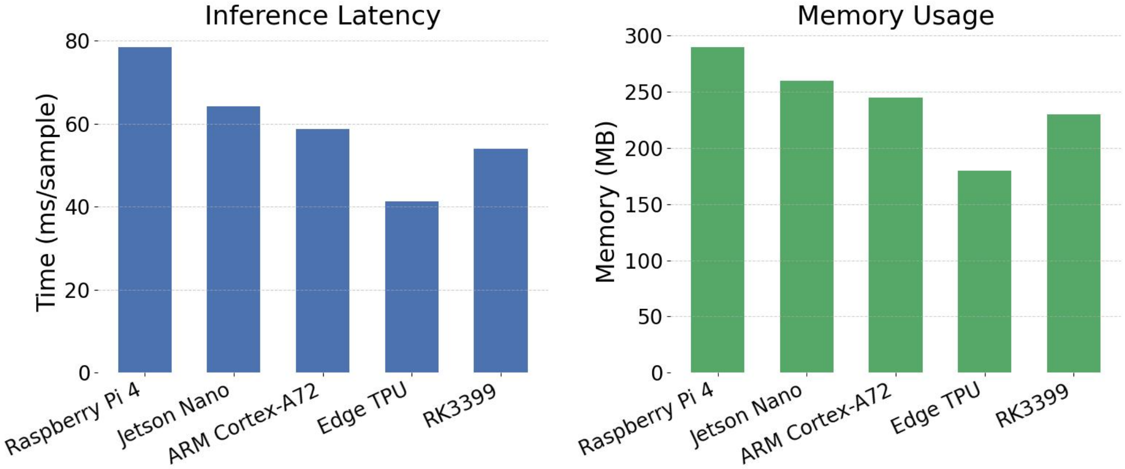

Figure 4. Improving inference speed evaluation of distilled models on low-resource devices

As shown in the figure, the improved distillation model demonstrates strong inference efficiency across various low-resource devices. On the Edge TPU and RK3399 platforms, the inference time reaches 41.3 ms and 53.9 ms, respectively. These values are significantly lower than those on other devices. This indicates that the proposed structural optimization and parameter compression strategies are effective not only in standard environments but also in edge deployment scenarios. They support fast response capabilities on terminal devices.

In terms of memory usage, the model reaches a peak memory consumption of only 180 MB on the Edge TPU. This is much lower than the 290 MB on the Raspberry Pi 4 and the 260 MB on the Jetson Nano. This shows that the distillation method effectively reduces resource usage during model execution. Such efficiency is critical for deployment in memory-constrained environments. It improves the model's feasibility and stability on smart hardware and mobile devices.

The combined optimization of inference speed and memory usage highlights the advantages of the multi-layer alignment distillation strategy in lightweight design. Results from different devices consistently show that the method can deliver balanced performance even under limited resources. It does not rely on specific hardware acceleration and offers good cross-platform compatibility.

In conclusion, the improved distillation model proposed in this study performs well in semantic retention and multi-task learning. It also demonstrates broad applicability in terms of hardware efficiency and deployment. These strengths provide practical support for applying language models in edge intelligence and low-power computing scenarios.

5) Loss function changes with epoch

At the end of this paper, a graph is provided to illustrate how the loss function evolves over the course of training, with respect to the number of epochs. This figure is included to offer a clear depiction of the optimization process undergone by the proposed distillation model during training. Specifically, the graph presents the dynamic changes in both the training loss and validation loss, thereby reflecting the progression of model convergence and learning behavior under the designed distillation framework. Tracking the loss function across epochs is essential for understanding how efficiently the student model absorbs knowledge from the teacher model, as well as for diagnosing potential issues such as underfitting or overfitting. The graphical representation of this loss curve is shown in Figure 5.

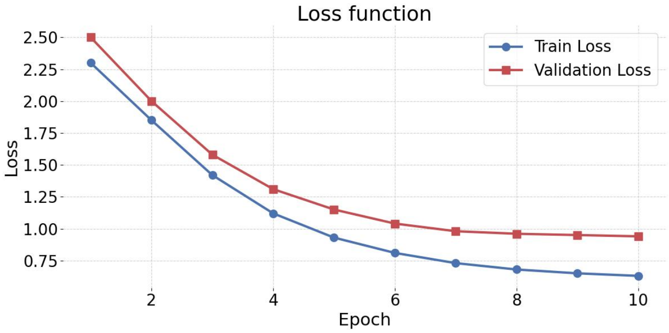

Figure 5. Loss function changes with epoch

As shown in the figure, both the training loss and validation loss decrease steadily as the number of epochs increases. The overall trend indicates good convergence. The training loss drops quickly from an initial value of 2.3 to around 0.63. This shows that the student model continuously absorbs knowledge from the teacher model and gradually builds a stable semantic representation. This result aligns with the multi-layer alignment distillation strategy proposed in this study. It confirms the effectiveness of the mechanism in guiding the optimization process. The validation loss also decreases smoothly. This suggests that the model does not suffer from significant overfitting during training. In the later stages, the validation loss stabilizes around 0.94. The small gap between training and validation losses indicates strong consistency and generalization. This further supports the positive role of attention alignment and intermediate layer transfer in improving model stability.

The comparison between the two loss curves also shows that the student model avoids the local minima problem often seen in lightweight models during early training. Guided by the teacher signals, the model benefits from soft labels and structural knowledge passed through distillation. These elements provide a smoother and more directed optimization path. As a result, the model achieves good performance in a short training time while maintaining a compact structure. In summary, the experiment verifies the advantages of the improved distillation method in training efficiency and loss convergence. The student model quickly fits the information transferred from the teacher. It also maintains stable training dynamics and strong generalization. These features provide a solid foundation for applications in multi-task learning and low-resource environments.

5. Conclusion

This paper addresses key challenges in compressing large language models and proposes a multi-layer alignment distillation algorithm based on an improved TinyBERT framework. The method introduces multi-dimensional distillation objectives, including output distribution, intermediate features, and attention weights. It significantly enhances the student model's ability to learn semantic structures from the teacher model while maintaining a compact architecture. Experimental results show that the method performs well across various natural language understanding tasks and demonstrates excellent inference efficiency and deployment flexibility.

The study validates the effectiveness of multi-level knowledge transfer in language model distillation. It also highlights the positive role of structure-aware distillation strategies in improving the semantic generalization of student models. In terms of accuracy, F1 score, and efficiency on low-resource devices, the proposed method shows strong overall advantages. These findings provide technical support for building lightweight, efficient, and robust natural language processing systems.

The method also shows strong applicability in real-world scenarios. In edge computing, mobile devices, and industrial text processing, model deployment is highly sensitive to resource constraints. Models with high compression ratios and strong performance are therefore highly valuable. In addition, the method supports direct transfer and deployment in dialogue systems, question answering, and domain-specific applications such as finance and healthcare. This extends the practical reach of large language model technology. Future research may explore task-aware distillation strategies, dynamic structure adaptation mechanisms, and multimodal knowledge distillation methods. These directions can further improve model adaptability in complex environments. Combining such methods with large-scale open-source models and real-world datasets will help advance the adoption of lightweight language models in industrial applications. This will provide a feasible path toward more inclusive and accessible intelligent language technologies.

References

[1] Xu, Xiaohan, et al. "A survey on knowledge distillation of large language models." arXiv preprint arXiv:2402.13116 (2024).

[2] Yang, Chuanpeng, et al. "Survey on knowledge distillation for large language models: methods, evaluation, and application." ACM Transactions on Intelligent Systems and Technology (2024).

[3] Cui, Yu, et al. "Distillation matters: empowering sequential recommenders to match the performance of large language models." Proceedings of the 18th ACM Conference on Recommender Systems. 2024.

[4] Li, Dawei, et al. "Contextualization distillation from large language model for knowledge graph completion." arXiv preprint arXiv:2402.01729 (2024).

[5] Liu, Aiwei, et al. "Direct large language model alignment through self-rewarding contrastive prompt distillation." arXiv preprint arXiv:2402.11907 (2024).

[6] McKinleigh, Henry, et al. "Optimizing knowledge distillation in large language models via recursive multi-modal distribution alignment." (2024).

[7] Liu, Qidong, et al. "Large language model distilling medication recommendation model." arXiv preprint arXiv:2402.02803 (2024).

[8] Chang, Yupeng, et al. "A survey on evaluation of large language models." ACM transactions on intelligent systems and technology 15.3 (2024): 1-45.

[9] Patil, Rajvardhan, and Venkat Gudivada. "A review of current trends, techniques, and challenges in large language models (llms)." Applied Sciences 14.5 (2024): 2074.

[10] Myers, Devon, et al. "Foundation and large language models: fundamentals, challenges, opportunities, and social impacts." Cluster Computing 27.1 (2024): 1-26.

[11] Chen, Jin, et al. "When large language models meet personalization: Perspectives of challenges and opportunities." World Wide Web 27.4 (2024): 42.

[12] Hao, Xiaoshuai, et al. "Mapdistill: Boosting efficient camera-based hd map construction via camera-lidar fusion model distillation." European Conference on Computer Vision. Cham: Springer Nature Switzerland, 2024.

[13] Wang, Wenxiao, et al. "Model compression and efficient inference for large language models: A survey." arXiv preprint arXiv:2402.09748 (2024).

[14] Chiang, Yuan, et al. "LLaMP: Large language model made powerful for high-fidelity materials knowledge retrieval and distillation." arXiv preprint arXiv:2401.17244 (2024).

[15] Sanh, Victor, et al. "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter." arXiv preprint arXiv:1910.01108 (2019).

[16] Jiao, Xiaoqi, et al. "Tinybert: Distilling bert for natural language understanding." arXiv preprint arXiv:1909.10351 (2019).

[17] Sun, Zhiqing, et al. "Mobilebert: a compact task-agnostic bert for resource-limited devices." arXiv preprint arXiv:2004.02984 (2020).

[18] Wang, Wenhui, et al. "Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers." Advances in neural information processing systems 33 (2020): 5776-5788.