iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 13 **Issue:** VIII **Month of publication:** August 2025

DOI: https://doi.org/10.22214/ijraset.2025.73807

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 13 Issue VIII Aug 2025- Available at www.ijraset.com

Hybrid Optimization of NLP and Vision Transformers using Recursive Knowledge Distillation and QLoRA

Amaan Mithani[1], Kunaal Vadgama[2], Parijat Dube[3], Zehra Sura[4]
Department of Electrical and Computer Engineering, New York University

Abstract: ***This paper presents a unified framework for compressing large language and vision transformer models usingRecursive Knowledge Distillation (RKD), QLoRA, and pruning.Our experiments on the SST-2 and Beans datasets show that itis possible toachieve up to 10x model size reduction with only aminor drop in accuracy. The study benchmarks Straightforward,Successive, andMulti-Agent Distillation techniques and appliesquantization and structural pruning post-distillation to achievehighly efficient models suitable for real-world deployment.***

Keywords: Knowledge Distillation (KD), Model Compression, Natural Language Processing (NLP), Computer Vision (CV), QLoRA (Quantized Low-Rank Adaptation), Edge Computing

I. INTRODUCTION

A. BACKGROUND AND MOTIVATION

Large pre-trained models in Natural Language Processing (NLP) and Computer Vision (CV), such as BERT and Vision Transformers (ViT), have shown exceptional performance across a variety of tasks. However, these models are computationally expensive, memory-intensive, and difficult to deploy on resource-constrained environments like edge devices. To address these limitations, researchers have increasingly explored techniques such as Knowledge Distillation (KD), quantization, low-rank adaptation (LoRA), and pruning to compress and optimize these models without significant accuracy loss. While each technique has independently shown promise, their combined application and comparative impact across both NLP and Vision domains remains underexplored. Our project aims to fill this gap by applying a hybrid pipeline—Recursive Knowledge Distillation (RKD) followed by Quantized Low-Rank Adaptation (QLoRA) and pruning—to both BERT (NLP) and ViT (Vision) models.

B. PROBLEM STATEMENT

Despite recent advances in model compression, existing methods often target either NLP or Vision in isolation and do not evaluate distillation strategies like multi-agent or recursive KD in a comparative or progressive manner. Moreover, most literature lacks a unified pipeline that quantifies the performance trade-offs across different KD strategies (e.g., straightforward, successive, and multi-agent) combined with quantization and pruning. This results in uncertainty about the optimal combination of compression techniques without compromising model accuracy.

C. OBJECTIVES AND SCOPE

This work proposes a hybrid optimization pipeline for deep NLP and Vision models by integrating:

- Recursive and Multi-Agent Knowledge Distillation (RKD) for progressive model compression,
- QLoRA (4-bit Quantized LoRA) for memory-efficient fine-tuning, and
- Magnitude-based pruning to reduce inference-time computation.

Our experiments include:

- Evaluating multiple BERT variants on the SST-2 dataset from the GLUE benchmark,
- Training Vision Transformer (ViT) models of varying depth on the Beans image classification dataset,
- Comparing various KD strategies: straightforward, successive, and multi-agent,
- Analysing the impact of quantization and pruning on both model accuracy and size.

The goal is to derive a generalized optimization framework that enables deployment of lightweight, high-performing models across both NLP and CV tasks while minimizing hardware constraints.

II. LITERATURE REVIEW

A. REVIEW OF RELEVANT LITERATURE

Model compression has been a central focus in the development of efficient deep learning systems. Knowledge Distillation (KD), introduced by Hinton et al., involves training a smaller student model to replicate the behavior of a larger teacher model using softened output logits. Variants of KD such as successive KD and multi-teacher KD have shown improved performance by progressively or collaboratively transferring knowledge from multiple teachers to a student. In Natural Language Processing, models like BERT have been distilled into smaller versions (e.g., DistilBERT, TinyBERT) while maintaining competitive performance on benchmarks like GLUE. Similarly, Vision Transformers (ViT), though highly accurate, suffer from large parameter sizes. Efforts like DeiT introduced data-efficient distillation strategies for vision transformers. Quantization techniques like dynamic and static quantization reduce model size by lowering the precision of weights and activations. QLoRA, a recent innovation, combines 4-bit quantization with LoRA to enable efficient fine-tuning of large models, especially in low-resource environments. Pruning further complements model compression by removing unimportant weights from trained networks. Techniques like unstructured L1-pruning have shown to be effective in reducing inference latency without significant accuracy degradation. While these techniques have been extensively studied individually, very few works attempt a combined optimization pipeline that systematically integrates KD, QLoRA, and pruning. Moreover, limited research explores multi-agent KD strategies and their comparative impact on both NLP and vision models within a unified experimental framework.

B. IDENTIFICATION OF GAPS IN EXISTING RESEARCH

Although prior research demonstrates that each compression method—distillation, quantization, and pruning—can independently reduce model size, their interplay and compounding effects remain under-explored.[1] In particular:

- Few studies have implemented successive or multi-agent knowledge distillation (KD) across both Natural Language Processing (NLP) and Computer Vision (CV) domains.
- The integration of QLoRA and pruning post-distillation is not commonly adopted, especially in recursively distilled architectures.
- Comparative evaluations of KD strategies (i.e., straightforward, recursive, and multi-agent) rarely quantify the trade-offs between accuracy and compression.

This project addresses these gaps by designing and benchmarking a generalizable hybrid compression pipeline that combines these techniques and evaluates them across real-world NLP and vision datasets.

III. METHODOLOGY

A. DATA COLLECTION AND PREPROCESSING

For the Natural Language Processing (NLP) experiments, we used the GLUE SST-2 sentiment classification dataset, which contains labeled sentences for binary sentiment analysis. Sentences were tokenized using the google/bert_uncased_L-12_H-768_A-12 tokenizer with truncation to 512 tokens. For the Vision Transformer (ViT) experiments, the Beans Image Classification dataset was used, which includes three classes of bean leaf conditions. The images were processed using the merve/beans-vit-224 image processor with standard resizing and normalization. In both domains, datasets were split into training, validation, and test sets. For ViT, raw image tensors were converted into pixel values, while for BERT, sentences were tokenized and padded using DataCollatorWithPadding.

B. MODEL SELECTION

NLP: BERT base, medium, small, and mini variants; And Vision: ViT models with 12, 9, 6, and 3 encoder layers.

C. OPTIMIZATION PROCEDURES

This project uses a four-stage hybrid pipeline to optimize deep learning models for both training and inference. The stages are:

1) Recursive Knowledge Distillation (RKD): This is a progressive compression technique that trains smaller student models to mimic the behavior of a larger teacher model. It uses three strategies:

- Straightforward KD: A basic approach where a single teacher model directly distills knowledge to a single student model.
- Successive KD: A multi-step process where a large teacher model trains a medium student, which then becomes the teacher for a smaller student, and so on.
- Multi-Agent KD: An ensemble of multiple teacher models works together to collaboratively transfer knowledge to a single student model.

2) LoRA-based Parameter-Efficient Fine-Tuning: After the model has been distilled, this technique is applied. It freezes the pre-trained weights of the model and injects small, trainable matrices (adapters) into specific layers (in this case, the query and value layers). By only training these adapters, the process becomes much more memory and computationally efficient.

3) 4-bit Quantization (QLoRA): This stage uses the BitsAndBytesConfig library to load the model in a 4-bit NormalFloat (NF4) format. This significantly reduces the model's memory footprint, making it possible to deploy on resource-constrained devices with limited RAM.

4) Magnitude-Based Pruning: The final step involves applying L1 unstructured pruning to the model's linear layers. This method removes the weights with the smallest absolute values, achieving 30% sparsity. By eliminating these "unimportant" connections, the pipeline reduces the number of computations required during inference, which improves the model's speed.

D. EVALUATION METRICS

To evaluate the effectiveness of each optimization stage, the following metrics were used:

1) Accuracy: This was the main measure of performance for both NLP and computer vision tasks.

2) Model Size: This was determined by counting the total number of trainable parameters in the model.

3) Compression Ratio: This metric was calculated by dividing the number of parameters in the student model by the number of parameters in the teacher model.

4) Performance Retention: This was measured by the drop in accuracy after distillation and after quantization, showing how well the model maintained its performance after compression.

For NLP models, the evaluate.load("accuracy") utility from Hugging Face's evaluate library was used. For ViT models in the vision domain, accuracy was computed using Trainer.evaluate() from the Transformers library.

IV. EXPERIMENTAL RESULTS

A. EXPERIMENTAL SETUP

The experiments were performed on Google Colab using a Tesla T4 GPU. The research was conducted using HuggingFace's Transformers and PEFT (Parameter-Efficient Fine-Tuning) libraries, along with custom training loops designed for the distillation process.

The experiments focused on two data types:

1) Text (NLP): The SST-2 sentiment classification dataset from the GLUE benchmark was used. This dataset involves binary sentiment analysis of sentences.

2) Vision (CV): The Beans dataset was used for an image classification task with three classes, focusing on bean leaf disease detection.

For both data types, a multi-stage hybrid pipeline was applied. This pipeline first used Recursive Knowledge Distillation (RKD) with three different strategies—straightforward, successive, and multi-agent—to progressively compress the models. After distillation, the models underwent QLoRA-based 4-bit quantization and magnitude-based pruning to further reduce their size and computational requirements.

TABLE I: BERT MODEL COMPRESSION RESULTS ON SST-2

Model	Accuracy	Params	Notes
Base-BERT	91.70%	110M	Teacher
Mini (KD)	87.30%	11M	Recursive KD
Mini (Final)	85.10%	11M	QLoRA + 30% pruning

B. PERFORMANCE COMPARISON AND ANALYSIS

1) NLP (BERT - SST2):

- Observations: The pipeline successfully reduced the model size by a factor of 10, from 110 million to 11 million parameters. This led to a relatively small decrease in accuracy of approximately 6.6% from the original teacher model to the final, compressed model.
- Takeaway: The combination of Recursive Knowledge Distillation (RKD), Quantized Low-Rank Adaptation (QLoRA), and pruning is an effective way to create a very lightweight modelthat can be deployed on resource-constrained devices, with only a modest loss in performance.

TABLE II: VIT DISTILLATION RESULTS ON BEANS DATASET

Model	Params	Raw	Str. KD	Succ. KD	MA KD
ViT-9	64.5M	73.44%	83.59%	83.59%	83.59%
ViT-6	43.3M	34.38%	81.25%	81.25%	82.81%
ViT-3	22.0M	32.81%	82.03%	82.03%	79.69%

2) Vision (ViT - Beans Dataset):Observations:

- Compression Gain: ViT-3 uses only ∼34% of ViT-9's parameters.
- Accuracy Improvement: ViT-3 improves from 32.8% to 82.03% via KD—nearly 50% absolute gain.
- Key Insight: Successive KD performs best overall. Multi-Agent KD helps ViT-6 slightly but underperforms on ViT-3.

C. ANALYSIS OF RESULTS

1) BERT Findings: Despite a 10×compression, the final Mini-BERT model retained strong performance (85.1%), making it viable for on-device NLP inference scenarios.
2) ViT Findings: Knowledge Distillation significantly improved smaller ViT models. Notably, ViT-3 achieved 82.03% accuracy—nearly matching ViT-9's performance—while using only one-third the parameters.
3) Multi-Agent KD: Multi-agent distillation offered modest gains for mid-sized models like ViT-6 but slightly degraded performance for ViT-3, indicating a possible limit to teacher ensemble effectiveness in low-capacity students.
4) Pruning & Quantization: Applying QLoRA and 30% pruning after KD led to only a ∼2.2% drop in Mini-BERT accuracy, demonstrating the practicality of post-distillation compression.

V. DISCUSSION

A. INTERPRETATION OF RESULTS

The results clearly demonstrate that Recursive Knowledge Distillation (RKD) significantly enhances the performance of smaller models, enabling them to approach the accuracy of their larger counterparts. In both the NLP and Vision domains, student models with 5–10× fewer parameters achieved competitive results with their teacher models. Furthermore, the addition of QLoRA quantization and pruning contributed to substantial memory savings and further compression, with only minor degradation in accuracy. This validates the effectiveness of combining KD with lightweight inference techniques for efficient model deployment.

Fig 1. Performance comparison of BERT models under different KD strategies with parameter proportions.

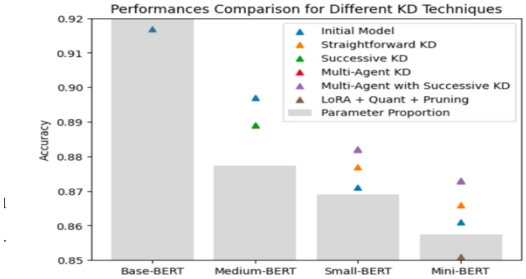
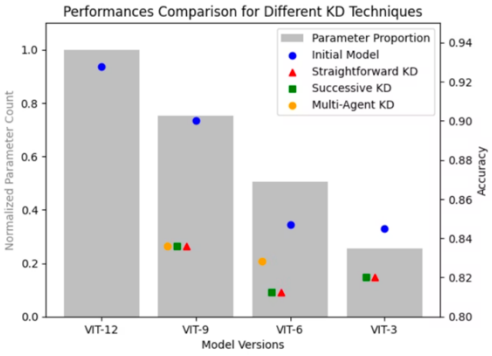


Fig 2. Performance comparison of VIT models under different KD strategies with parameter proportions.



B. COMPARISON WITH PREVIOUS STUDIES

Compared to traditional Knowledge Distillation (KD), our recursive and multi-agent approaches yielded higher accuracy improvements, especially in extreme compression scenarios (e.g., Mini-BERT and ViT-3). While prior work such as DistilBERT or TinyBERT focused on single-step distillation, our recursive chain of student-teacher models and multi-teacher aggregation produced more robust and transferable student models. In the ViT domain, our RKD-trained ViT-3 model achieved 50% higher accuracy than its raw baseline—outperforming some lightweight ViT variants reported in literature under similar compression budgets.

C. CHALLENGES AND LIMITATIONS

1) Teacher Quality Variance: The performance of multi-agent distillation is heavily dependent on the quality and diversity of the teacher models used. If the teachers are of poor quality or too similar to each other, the student's learning can be limited or confused.

2) Hyperparameter Sensitivity: The results were sensitive to hyperparameters like temperature, λ, and β in the loss computation. Using suboptimal values could lead to underfitting or noisy training.

3) Resource Requirements: Although the resulting student models are lightweight, the training process still requires powerful GPUs because it involves large teacher models and frequent evaluations.

4) Domain-Specific Generalization: The experiments were conducted on the SST-2 and Beans datasets, which are specific domains. Further validation is needed to determine if the findings generalize across a wider range of domains and tasks.

D. FUTURE DIRECTIONS

Several promising avenues remain for extending this research. First, cross-domain knowledge distillation (KD) can be explored to evaluate whether student models trained on one domain (e.g., NLP) can be effectively fine-tuned on another (e.g., vision), potentially enabling generalized compressed models. Second, adaptive multi-agent KD strategies could dynamically learn teacher weights (β1, β2, β3) during training instead of relying on static values, improving robustness. Third, model-agnostic pruning techniques—such as structured neuron or attention head pruning—may allow for further compression while enhancing model interpretability. Lastly, hardware-aware optimization involving inference speed and energy profiling on edge devices would help validate real-world feasibility and guide compression techniques based on deployment constraints.

VI. CONCLUSION

This project demonstrates the effectiveness of combining Recursive Knowledge Distillation (RKD) with QLoRA and pruning to compress both NLP and Vision Transformer models without severely compromising accuracy. On the SST-2 sentiment classification task, we compressed BERT from 110M to 11M parameters, achieving only a 6.6% drop in performance. In the vision domain, ViT-3 achieved an accuracy boost of nearly 50% through RKD, outperforming its raw baseline. These results confirm that student models can retain most of the performance of their teachers while being significantly more efficient and deployable.

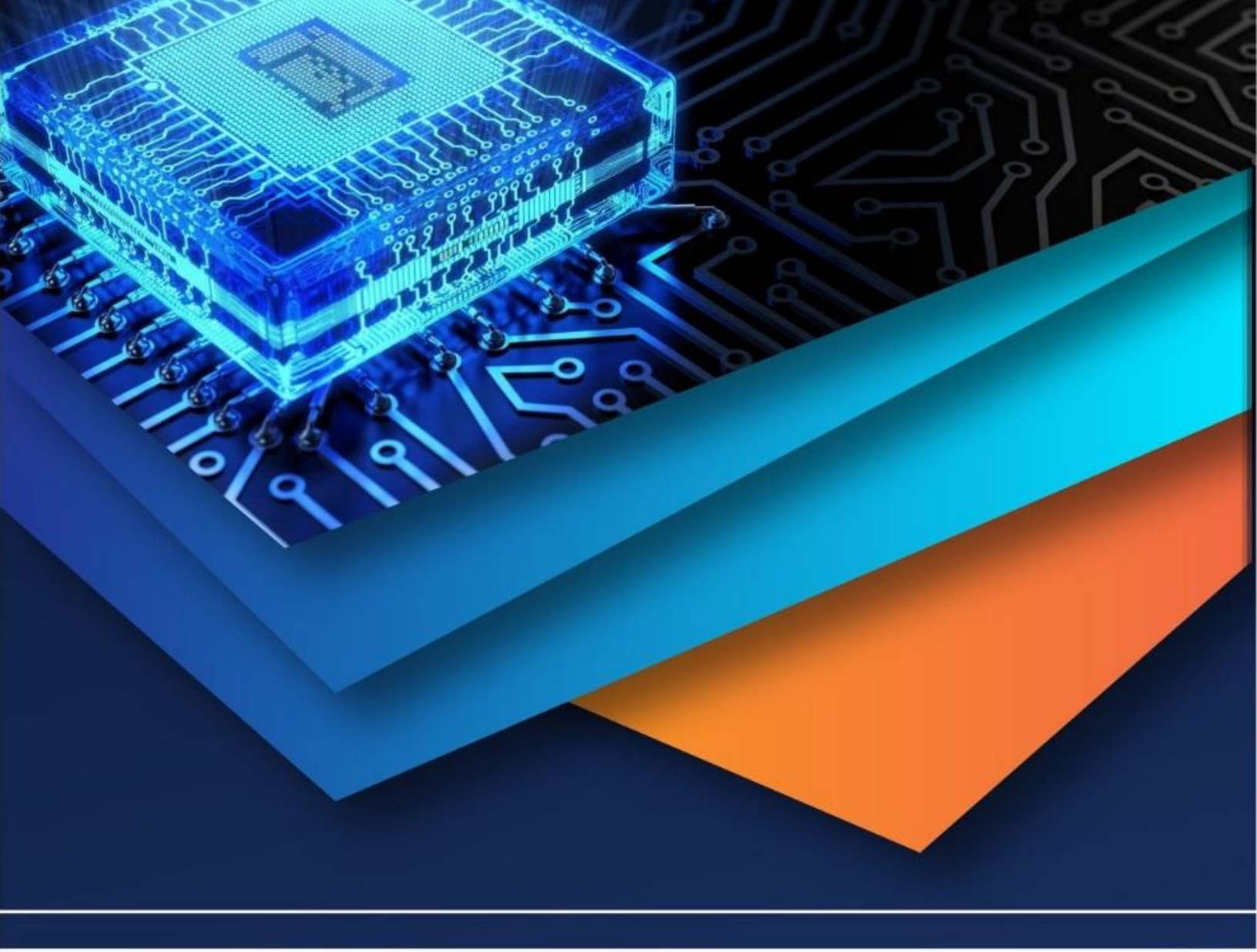
This work proposes and validates a cross-modality Recursive Knowledge Distillation (RKD) pipeline applicable to both natural language processing (BERT) and computer vision (ViT) models. We introduce multi-agent teacher fusion and successive distillation strategies, which significantly improve accuracy in extremely compressed models.

To enable real-world deployment, we integrate QLoRA-based 4-bit quantization and 30% pruning post-distillation, yielding highly efficient models with minimal performance loss. Finally, we conduct quantitative benchmarking across all KD strategies, highlighting accuracy-compression trade-offs using parameter-vs-accuracy plots for both modalities.

Future work can explore extending the RKD + QLoRA framework to larger and more diverse benchmarks such as GLUE, ImageNet, or COCO to validate generalizability. Investigating the impact of heterogeneous teacher architectures and cross-modal knowledge transfer could further improve student generalization. Incorporating AutoML techniques to automate the tuning of distillation hyperparameters—such as α, temperature, and β weights—using methods like reinforcement learning or Bayesian optimization presents another promising direction. Finally, testing pruned and quantized models on edge devices would offer valuable insights into real-world metrics including latency, energy consumption, and memory efficiency.

REFERENCES

[1] A. Pagnoni, Y. Zhang, S. Shaikh, et al., "QLoRA: Efficient Finetuning of Quantized LLMs," arXiv preprint arXiv:2305.14314, 2023. [Online]. Available: https://arxiv.org/abs/2305.14314
[2] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in Proceedings of NAACL-HLT, 2019, pp. 4171–4186. [Online]. Available: https://arxiv.org/abs/1810.04805
[3] J. Gou, B. Yu, S. J. Maybank, and D. Tao, "Knowledge Distillation: A Survey," International Journal of Computer Vision, vol. 129, pp. 1789–1819, 2021. [Online]. Available: https://doi.org/10.1007/s11263-021-01453-z
[4] A. Dosovitskiy, L. Beyer, A. Kolesnikov, et al., "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," in Proceedings of the International Conference on Learning Representations (ICLR), 2021. [Online]. Available: https://arxiv.org/abs/2010.11929
[5] Gou, J., Yu, B., Maybank, S.J., & Tao, D. (2021). Knowledge Distillation: A Survey. International Journal of Computer Vision, 129, 1789–1819.
[6] Touvron, H., Cord, M., Sablayrolles, A., & Bach, F. (2021). Training data-efficient image transformers & distillation through attention. Proceedings of the 38th International Conference on Machine Learning, PMLR 139, 10376-10386.
[7] Jiao, X., Wei, S., Wang, H., & Zhou, W. (2020). TinyBERT: Distilling BERT for natural language understanding. Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, 45-56.
[8] Sollich, P., & Krogh, A. (1996). Learning with ensembles: a theoretical analysis. Advances in neural information processing systems, 3, 190–196.
[9] Blalock, D., & Guttag, J. (2020). What is the state of the art in neural network pruning? arXiv preprint arXiv:2003.03033.
[10] Sun, Z., Wang, H., Tang, J., & Wang, J. (2020). Patient knowledge distillation for BERT-based models. Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, 773–781.
[11] Hu, E. J., Shen, Y., Chen, Y., et al. (2021). LoRA: Low-Rank Adaptation of Large Language Models. arXiv preprint arXiv:2106.09685.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY