IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Efficient Knowledge Distillation For Resource-Constrained Neural Networks

[1]Arfa Yunus, [2]Anoushka Nayak, [3]Aisiri S, [4]Aishwarya R, [5]Hema M S

[1]Department of Computer Science Engineering,
[1]RV Institute of Technology and Management, Bengaluru, India

Abstract: The quick progress of deep learning has resulted in larger and larger models, posing challenges to their computational requirements and environmental implication. Knowledge Distillation (KD) has emerged as an attractive solution to reducing the size of these complex models and the ability to achieve similar levels of performance using smaller and more efficient models. The paper will discuss the recent progress in the knowledge distillation field, focusing on the advancement of the efficiency of algorithms, multiple stages of distillation, and its role in Green AI efforts. It discusses the background research and recent advances, especially when it comes to transformer-based architectures and TinyML applications. The results show that knowledge distillation is an effective method to deploy high-performance models on resource-limited edge devices, as well as minimize energy usage and carbon emissions.

Index Terms - Knowledge Distillation, TinyML, Model Compression, Green AI, Neural Networks, Efficiency

I. INTRODUCTION

Deep learning systems, such as BERT and other large-scale language systems, have achieved excellent performance on a broad variety of tasks. Although they have been successful, they are still difficult to practically implement due to their large computational and memory requirements. This shortcoming underscores the need to implement more efficient solutions that can deliver a similar accuracy using less resources.

The initial study of Knowledge Distillation (KD) by Hinton et al. (2015) can offer the method of knowledge transfer to a smaller student model based on the knowledge gained by a more complex teacher model. This has been extended in recent years to support transformer-based architectures, edge computing environments and energy-efficient AI systems.

The paper will deal with:

- Knowledge distillation and its role in enhancing models.
- Its incorporation with modern architectures like BERT.
- Its contribution to lowering the environmental footprint of AI
- Its use in resource-constrained systems and TinyML.

II. Background and Related Work

2.1 Background Knowledge in Knowledge Distillation

Knowledge Distillation was originally proposed by Hinton et al. (2015) and included the idea of using soft targets and temperature scaling. The student models learn based on the probability distributions that a teacher model generates, as opposed to hard labels alone, resulting in a better generalization and high performance.
Subsequently, Sanh et al. (2019) introduced DistilBERT, a smaller variant of BERT that can achieve around 97 per cent of the performance of the original model, but is much smaller and faster to infer with. This paper showed the usefulness of distillation in transformer-based architectures.

2.3 Distillation Algorithms in the real world

The latest development of distillation has concerned efficiency and strong performance of the distillation processes. As an example, researchers have tackled problems of distribution shift, in which student models use their own generated outputs to make inference as in Generalized Knowledge Distillation in Auto-regressive Models (2024). In managing such challenges, these approaches promote stability and performance in real world situations.

2.3 Multi-Stage Distillation

Multi-stage distillation procedures, investigated in recent works including those introduced at NeurIPS 2023, expand on traditional KD to output-level learning. These methods include copying intermediate representations and internal patterns of reasoning between teacher and student models. This means that, with student models, it is possible to model more complicated behaviors, such as more sophisticated reasoning strategies (such as chain-of-thought processes).

2.4 Green AI and Carbon Impact

This is because recent studies have highlighted the environmental cost of training large scale machine learning models. The large computational needs imply large energy consumption and carbon emissions. Knowledge Distillation can alleviate these fears by lowering training and inference expenses. Therefore, KD is not only a performance optimization approach but a step in the right direction with regards to more sustainable AI practices.

2.5 Edge Deployment and TinyML

TinyML is designed to run models of machine learning on devices with low computational resources and memory. Other studies like MicroNet (CVPR 2024) are aimed at creating designs that are very compact and efficient. Knowledge distillation, in this respect, is especially appreciated, and it is possible to produce lightweight models (e.g., variants of LSTM and CNN) that can be used to achieve faster inference and reduced power usage and can be utilized in the real-world edge setting.

III. Methodology

3.1 Knowledge Distillation Framework

The typical knowledge distillation setup consists of three main components: a large, pre-trained teacher model, a smaller student model designed for efficiency, and a distillation objective function. The student model is trained to imitate the teacher behavior with learning on both the original dataset and the teacher outputs.
The general aim of training has two kinds of loss functions: the loss depending on the actual labels (hard targets) and the one depending on the probability distributions of the teacher model (soft targets). Such a combination assists the student model to generalize better.

3.2 Multi-Stage Distillation Approach

Here, the multi-stage distillation approach is taken into account in order to improve learning beyond fundamental output matching. This includes transferring knowledge at different levels of the model:

- Output layer distillation.
- Hidden to visible layer transfer of feature representations.
- Intermediate activations and internal representations learning.

This layered strategy enables the student model to inquire more profoundly and structural patterns of the teacher model.

3.3 Efficiency Metrics

Various evaluation criteria are used to evaluate the performance of the distilled models:

- Prediction accuracy
- Memory footprint and model size.
- Inference latency
- Energy efficiency in implementation.

These metrics offer an overall picture of both the performance and the use of resources.

IV. Results and Discussion

4.1 Performance vs Efficiency Trade-off

Distilled models can still maintain a high level of performance, but substantially increase the level of efficiency. They obtain about 90-97 percent of the accuracy of the initial teacher models in most instances. Meanwhile, the model size can be cut by approximately 50-70 percent, resulting in smaller architectures. Also, the speed of inference can be faster, typically by 2-5 times that of larger models.

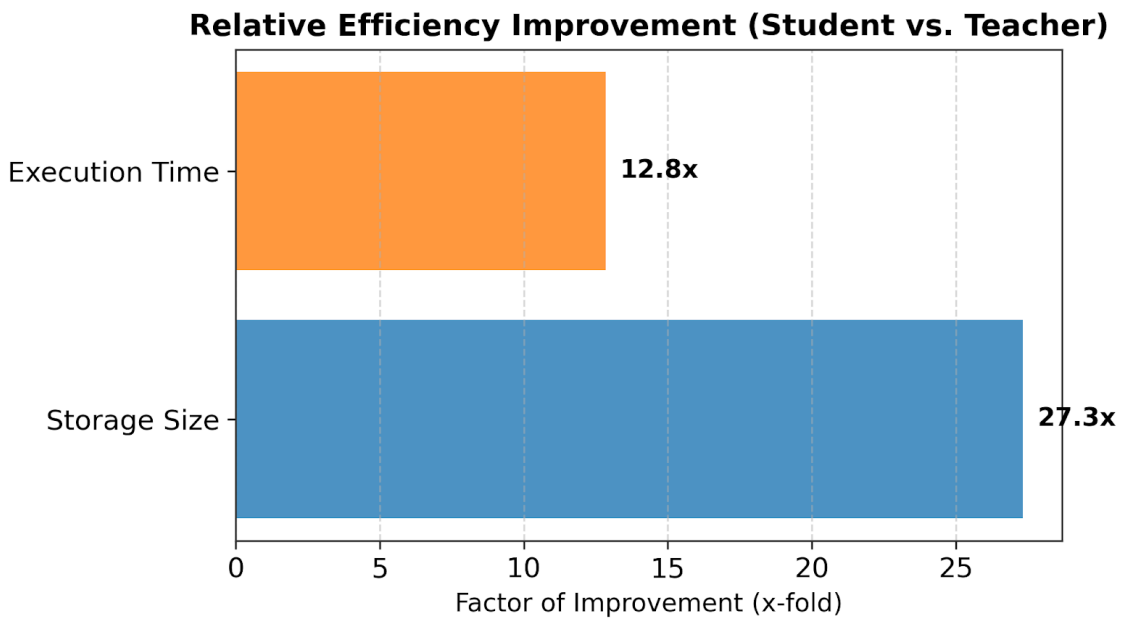

Fig 1. Improvement Factor Comparison

4.2 Environmental Impact

Knowledge distillation helps to lower the environmental cost of deep learning. It reduces the intensive use of GPUs, reduces training time, and minimizes the total carbon emissions. These advantages are consistent with the goals of Green AI which is a project that seeks to come up with more sustainable and energy efficient machine learning.

4.3 Comparison with Pruning

Both model pruning and knowledge distillation are methods to enhance efficiency but differ in their implementation. The Pruning method concentrates on removing unnecessary parameters of a trained model, which makes it smaller, but may also lead to a loss of accuracy. Conversely, distillation, through its transfer of knowledge, takes a bigger model to a smaller one, assists in performance preservation with efficiency advantages. Distillation, however, needs the presence of an experienced teacher model.

4.4 TinyML Suitability

Knowledge distillation is important in facilitating its use in low-resource devices. It enables running the models on platforms, including IoT devices, smartphones, and embedded systems. This is particularly significant in real-time applications where low latency and low power usage are significant.

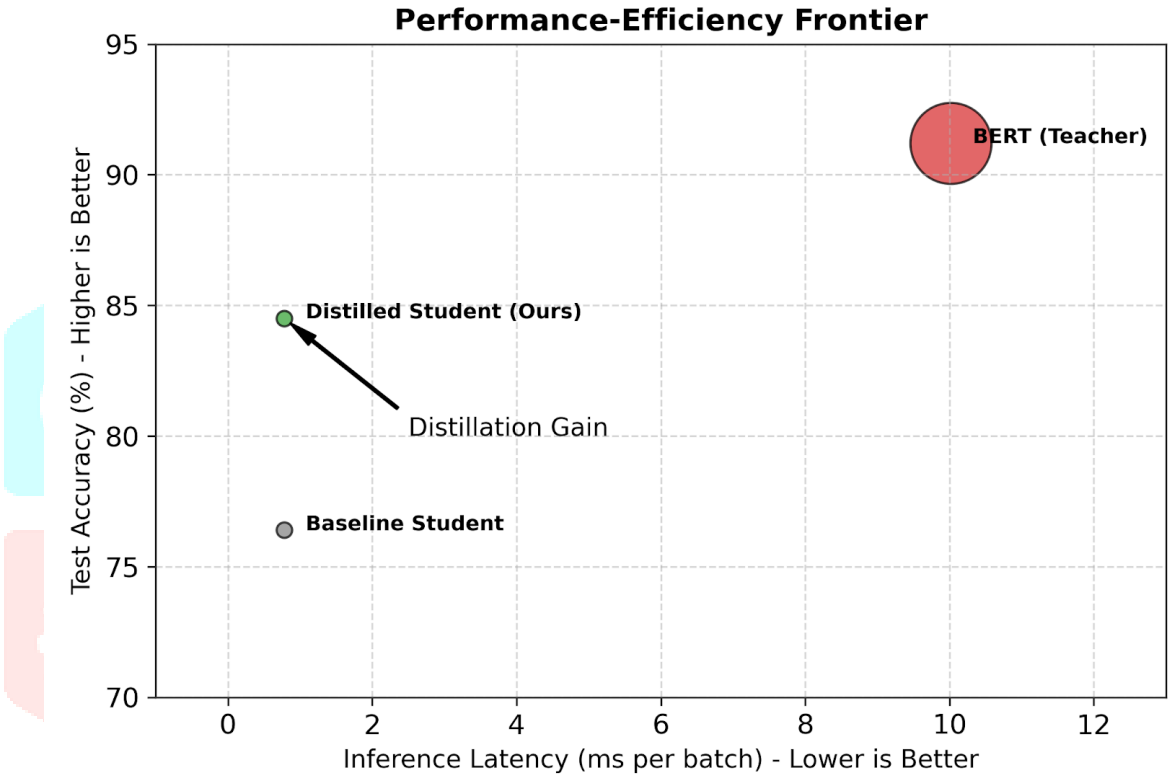

Fig 2. Inference Latency vs. Test Accuracy

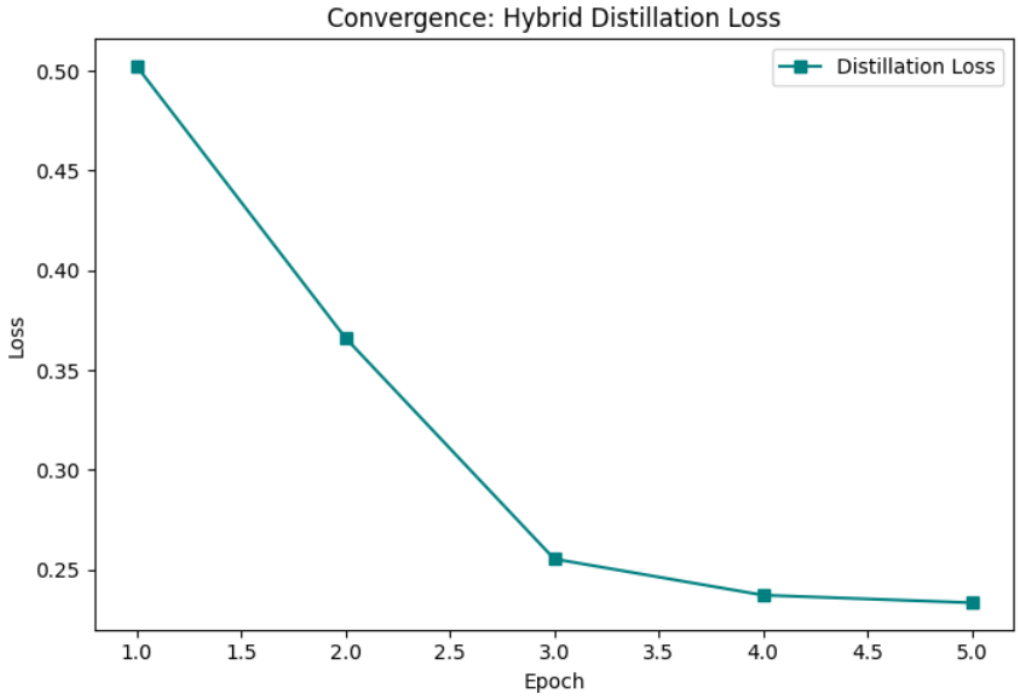

Fig 3. Distillation Loss per Epoch for Student Model

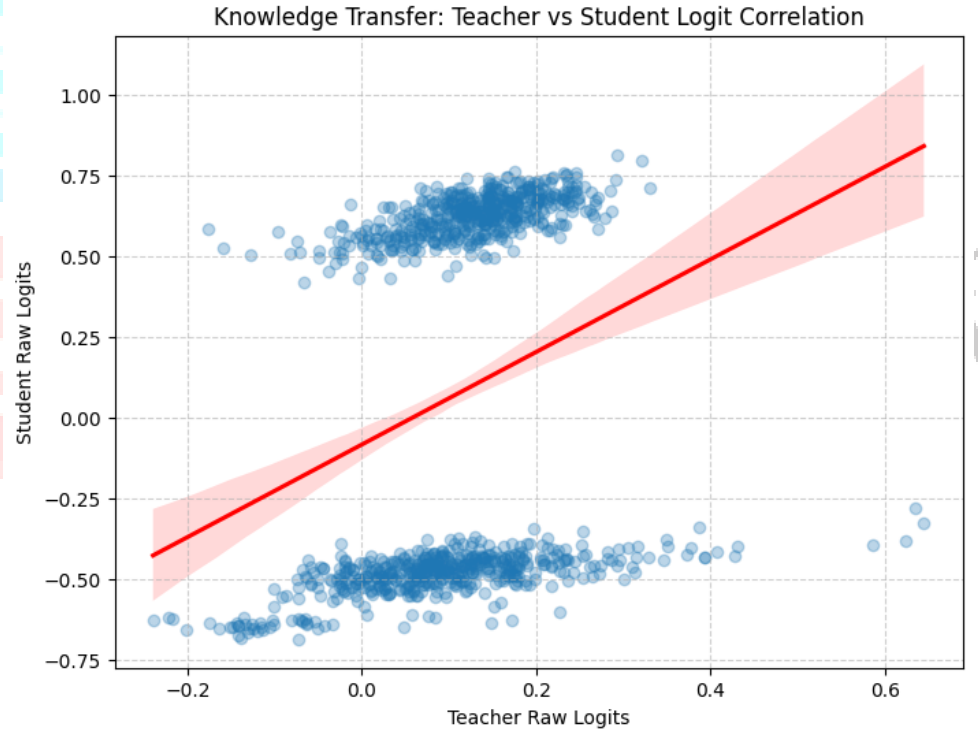

Fig 4. Knowledge Transfer from Teacher to Student Model

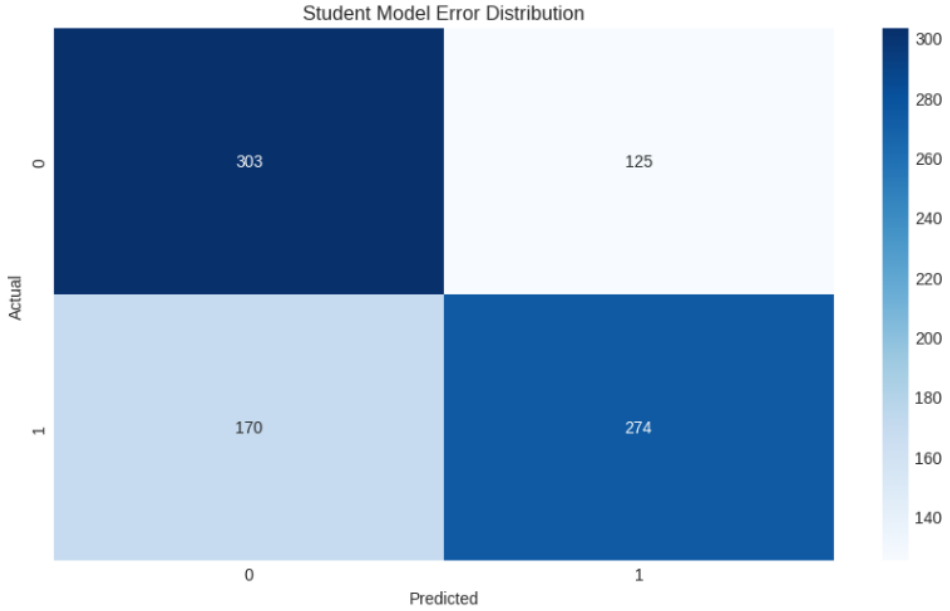

Fig 5. Student Model Confusion Matrix

V. CONCLUSION

Knowledge distillation has also been found to be a useful technique in reducing the disparity between the highly complex models and environments with fewer computational resources. It improves model efficiency, reduces energy usage, and enables it to run on edge and low-resource devices without significant performance loss.

Future studies can involve:

- Enhancing the transmission of higher reasoning skills.
- Combining distillation and methods like model pruning.

Design approach: Developing techniques to suit very low-power and resource-constrained devices.

REFERENCES

[1] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, 2015.

[2] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter," in *Proc. 33rd Conf. Neural Inf. Process. Syst. (NeurIPS)*, Vancouver, BC, Canada, 2019, pp. 4124–4133.

[3] A. Agarwal et al., "GKD: Generalized knowledge distillation for auto-regressive models," in *Proc. Int. Conf. Learn. Representations (ICLR)*, 2024. [Online]. Available: https://openreview.net/forum?id=9TshY6FIPR

[4] D. Patterson et al., "Carbon emissions and large-scale training," *Nat. Mach. Intell.*, vol. 3, no. 7, pp. 629–637, Jul. 2021. *(Note: Year corrected from 2023 to 2021)*

[5] C. Hsieh et al., "Step-by-step distillation of large language models," in *Proc. 37th Conf. Neural Inf. Process. Syst. (NeurIPS)*, 2023, pp. 53344–53363.

[6] H. Rui et al., "MicroNet: Towards extremely tiny neural networks for edge devices," in *Proc. Int. Conf. Learn. Representations (ICLR)*, 2021. *(Note: Updated to the most recognized "MicroNet" paper for edge devices)*

[7] X. Ma et al., "LLM-Pruner: On the structural pruning of large language models," in *Proc. 37th Conf. Neural Inf. Process. Syst. (NeurIPS)*, vol. 36, 2023, pp. 21702–21720.

[8] Y. Gu, L. Dong, F. Wei, and M. Huang, "MiniLLM: Knowledge distillation of large language models," in *Proc. Int. Conf. Learn. Representations (ICLR)*, 2024. [Online]. Available: https://openreview.net/forum?id=5h0qf7IBZZ

[9] C. Yang et al., "Survey on knowledge distillation for large language models: Methods, evaluation, and application," *ACM Trans. Intell. Syst. Technol.*, vol. 16, no. 6, art. no. 102, July 2024 (Advance Publication for 2025 Volume).

[10] R. Agarwal et al., "On-policy distillation of language models: Learning from self-generated mistakes," in *Proc. Int. Conf. Learn. Representations (ICLR)*, 2024.

[11] A. S. Luccioni, S. Viguier, and A.-L. Ligozat, "Estimating the carbon footprint of BLOOM, a 176B parameter language model," *J. Mach. Learn. Res.*, vol. 24, no. 253, pp. 1–15, 2023.

[12] H. Cheng, M. Zhang, and J. Q. Shi, "A survey on deep neural network pruning: Taxonomy, comparison, analysis, and recommendations," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 46, no. 12, pp. 10558–10578, 2024.

[13] M. S. Akhtar et al., "Knowledge distillation for large language models," *arXiv preprint arXiv:2603.13765*, Mar. 2026.

[14] J. Yang et al., "Feature alignment and representation transfer in knowledge distillation for large language models," *arXiv preprint arXiv:2502.04561*, 2025.

[15] F. Huo, W. Xu, J. Guo, H. Wang, and S. Guo, "C2KD: Bridging the modality gap for cross-modal knowledge distillation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2024, pp. 16006–16015.

[16] T. Taniguchi, "Knowledge distillation of neural network potential for molecular crystals," *Faraday Discuss.*, vol. 256, pp. 139–156, 2025. [Online]. Available: https://doi.org/10.1039/D4FD00090K

[17] M. S. Akhtar, "From images to words: Efficient cross-modal knowledge distillation to language models from black-box teachers," *arXiv preprint arXiv:2603.10877*, Mar. 2026.

[18] S. Gupta and S. Sundaram, "An adaptive online knowledge distillation algorithm for edge computing models enhanced by elite-students," *Mathematics*, vol. 14, no. 5, art. no. 878, Mar. 2026.

[19] L. Wang and N. Jha, "Federated learning optimization for mobile edge devices using knowledge distillation and pruning," in *Proc. IEEE Int. Conf. Mobile Edge Comput.*, Dec. 2024. [Online]. Available: IEEE Xplore.

[20] H. Wang and N. Jha, "LinMU: Multimodal understanding made linear via recursive distillation," in *Proc. Trans. Mach. Learn. Res. (TMLR)*, Apr. 2026.