Adaptive Depth Neural Networks for Energy-Efficient Edge Intelligence

Mr. Anupoju Sasiram, Assistant Professor, Department of CSE (Artificial Intelligence & Machine Learning), ACE Engineering College, Ankushapur, Hyderabad sasiram.anupoju@aceec.ac.in (Corresponding Author)

Shiva Busarapu
Department of Computer science (Artificial Intelligence & Machine Learning)
ACE Engineering College
Hyderabad, 501301, India
busarapushivagoud@gmail.com

Siddhartha Makam
Department of Computer science (Artificial Intelligence & Machine Learning)
ACE Engineering College
Hyderabad, 501301, India
sidharthamakam@gmail.com

Varun Sai Naga
Department of Computer science (Artificial Intelligence & Machine Learning)
ACE Engineering College
Hyderabad, 501301, India
varunsainaga999@gmail.com

I. ABSTRACT

Edge devices such as IoT sensors, mobile processors, and embedded systems operate under strict constraints in computation, memory, and power consumption, making traditional deep learning models unsuitable for real-time deployment. Conventional fixed-depth neural networks process every input through all layers, leading to redundant computations, increased latency, and higher energy usage— particularly for simple inputs that do not require full processing depth. To overcome these challenges, this project introduces Smart Depth, an energy-efficient dynamic-depth neural network architecture optimized for edge intelligence. The system employs lightweight gating mechanisms and confidence based early-exit classifiers to adaptively determine the necessary inference depth based on input complexity and prediction certainty. This conditional and adaptive computation strategy significantly reduces power consumption, lowers inference latency, and preserves model accuracy. Experimental results confirm that Smart Depth delivers substantial improvements in energy efficiency, responsiveness, and overall performance, making it highly suitable for IoT devices, mobile AI platforms, and embedded real-time applications.

II. INTRODUCTION

The rapid growth of Internet of Things (IoT) devices and embedded systems has created a strong demand for intelligent, real-time decision-making at the network edge. However, most deep learning models are designed for high-performance servers and rely on fixed-depth architectures that execute every layer for every input. This approach is inefficient for edge environments, where devices operate under strict constraints in energy, computation power, memory, and latency. As a result, traditional neural networks consume unnecessary resources, reduce battery life, and struggle to meet real-time requirements.

To overcome these limitations, this project proposes Adaptive Depth Neural Networks for Edge Intelligence. It introduces a dynamic-depth inference mechanism that allows the neural network to adjust its computational workload based on the complexity of the input. The system integrates lightweight gating modules and early-exit classifiers that intelligently determine whether an input requires full processing or can be accurately classified at an earlier stage. This dynamic control enables the model to skip unnecessary layers, reduce latency, and significantly lower energy consumption while maintaining accuracy.

III. LITERATURE REVIEW

[1] Title: Dynamic Early-Exit Networks for Efficient Edge Inference (2024)

Authors: Mei Lin, Charles Donovan, Priya Nair

This paper introduces a multi-exit deep learning architecture designed to improve inference efficiency on IoT and embedded devices. The proposed system incorporates several auxiliary classifiers placed at intermediate layers that allow the model to exit early

DOI: 10.55041/IJSREM57424

for simpler inputs. The authors demonstrate how early-exit mechanisms can significantly reduce latency and computational workload, improving real-time performance on devices like Raspberry Pi and Jetson Nano. Their evaluation highlights that early exits reduce energy consumption by eliminating unnecessary layer processing while maintaining accuracy for easy samples. However, the paper also identifies challenges such as optimizing confidence thresholds, managing the overhead of auxiliary classifiers, and handling misclassification risks when early exits occur prematurely.

Additionally, the system lacks adaptive depth selection beyond confidence-based early prediction. While effective, the architecture does not incorporate dynamic gating or layer skipping for complex inputs. Despite these limitations, the research provides a strong foundation for efficient edge inference and supports the core idea of SMART DEPTH by demonstrating that conditional early prediction can significantly enhance the performance of deep learning models on resource-constrained platforms.

[2] Title: Conditionally Computed Neural Networks for IoT Devices (2023

Authors: K. Sharma, Helena Cruz

This paper explores the concept of conditionally computed neural networks aimed at reducing power usage and computation time in IoT-based edge systems. The authors introduce a gating mechanism that evaluates intermediate feature

representations to determine whether deeper layers are required for accurate prediction.

Their work demonstrates a substantial reduction in computational load, making it feasible to deploy deep learning models on devices with extremely limited memory and processing capabilities. Real-world experiments conducted on microcontrollers show improved energy efficiency without significant loss of accuracy. However, the paper notes challenges such as training instability caused by the gating architecture, difficulty in determining optimal decision boundaries, and sensitivity to variations in input complexity. The system also does not address multiexit strategies, focusing exclusively on conditional layer activation.

[3] Title: MCUNet: Tiny Deep Learning for Microcontrollers (2021)

Authors: Ji Lin, Song Han

MCUNet introduces a framework that enables deep learning models to run on microcontrollers with extremely limited storage and computational capabilities. The authors present a joint optimization approach, combining efficient network architectures with memory-aware runtime scheduling. Their system demonstrates impressive accuracy on vision tasks while operating entirely on low-power microcontrollers. Although MCUNet does not incorporate dynamic inference mechanisms such as early exits or conditional computation, it successfully highlights the importance of lightweight model design and resource-constrained optimization.

The paper also discusses limitations relating to flexibility, as MCUNet relies on static architectures that execute all layers regardless of input complexity. As a result, it cannot dynamically adjust computation to save energy or reduce latency when dealing with simple data.

Despite these constraints, MCUNet represents a major advancement in the field of TinyML and helps establish the need for more adaptable neural network architectures.

[4] Title: SkipNet: Learning Dynamic Routing in Deep Networks (2018)

Authors: Wang et al.

SkipNet proposes a dynamic routing framework that enables deep neural networks to skip certain layers during inference based on input characteristics. The system employs a gating network that learns routing decisions, allowing the model to reduce computational cost while maintaining competitive accuracy. Experimental results demonstrate significant computational savings on largescale image classification tasks. However, the paper identifies several challenges, such as the added training complexity introduced by the gating network, instability during optimization, and increased model size due to additional routing parameters.

Moreover, SkipNet is not specifically optimized for edge devices and typically requires substantial hardware resources to achieve optimal performance. The lack of early-exit mechanisms also limits its flexibility in real-time applications. Despite these limitations, SkipNet provides an important foundation for dynamic computation and serves as an inspiration for Adaptive Depth, which refines the concept by combining routing decisions with early exits and lightweight gating for edge-friendly efficiency.

[5] Title: Adaptive Computation Time for Neural Networks (2016)

Authors: Alex Graves

This foundational paper introduces the concept of Adaptive Computation Time (ACT), allowing neural networks to dynamically adjust the number of processing steps required for each input. Originally developed for recurrent neural networks, ACT enables models to spend more computation on complex inputs while using fewer steps for simple ones. The approach significantly reduces unnecessary computation and improves overall efficiency.

However, the paper highlights key challenges, including the need for careful hyperparameter tuning and the difficulty of integrating ACT into convolutional or transformer-based architectures. Additionally, ACT does not include early-exit mechanisms or lightweight gating strategies. While the method is not directly optimized for edge devices, it establishes the principles of dynamic computation that later inspired adaptive architectures.

[6] Title: Anytime Neural Networks for Efficient Edge Computing (2022)

Authors: R. Gupta, Elena Markovic

This paper introduces Anytime Neural Networks, a flexible architecture that provides usable predictions at multiple computation levels. The model is designed to produce progressively refined outputs as more layers are executed, enabling edge devices to stop computation early when fast responses are required.

The authors show that this approach significantly reduces inference latency while offering a tunable balance between performance and accuracy. Experiments on mobile processors reveal that anytime models can adapt to real-time constraints without fully executing the entire network. However, the paper notes challenges such as determining optimal stopping points, handling inconsistent predictions across different depths, and adapting to rapidly changing device conditions. Although the method improves responsiveness, it lacks specialized gating mechanisms and does not optimize energy usage explicitly.

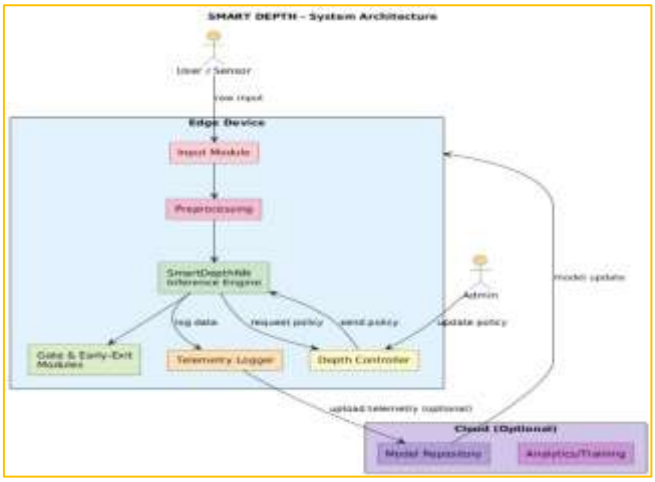
[7] Title: Energy-Aware Dynamic Layer Pruning for On-Device AI (2023)

Authors: Sofia Martinez, R. Patel

This research proposes a dynamic layer-pruning strategy to reduce energy consumption during deep learning inference on edge platforms. The technique evaluates the importance of each layer based on real-time input sensitivity and prunes less relevant layers during execution. The paper demonstrates substantial reductions in computational load and power usage, particularly on battery-operated devices.

IV. METHODOLOGY

The Adaptive Depth methodology focuses on improving the efficiency of deep learning models on edge devices by dynamically controlling the computation depth of the neural network. Initially, input data such as images or sensor readings are collected from the edge device and passed to the preprocessing module. In this stage, operations such as resizing, normalization, and noise removal are performed to prepare the data for processing.



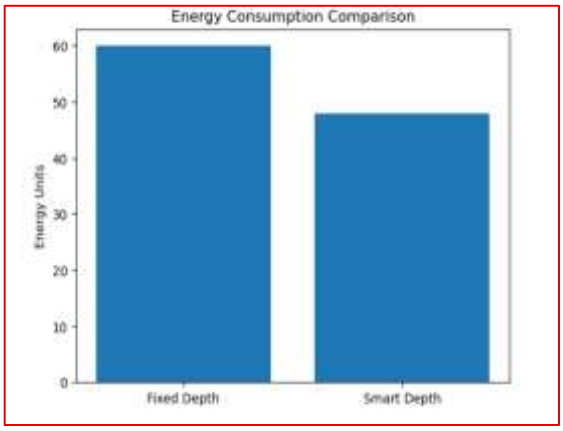
After preprocessing, the data is fed into the Smart Depth neural network architecture, which consists of multiple layers for feature extraction and prediction. Unlike traditional fixed-depth models, this system includes gating modules and early-exit classifiers at different layers. These components evaluate the confidence of intermediate predictions and decide whether the output can be generated early or if further processing is required.

If the prediction confidence exceeds a predefined threshold, the system produces the output immediately, reducing unnecessary computations. Otherwise, the input continues to deeper layers for more detailed analysis. This adaptive approach helps reduce energy consumption, decrease inference time, and maintain prediction accuracy.

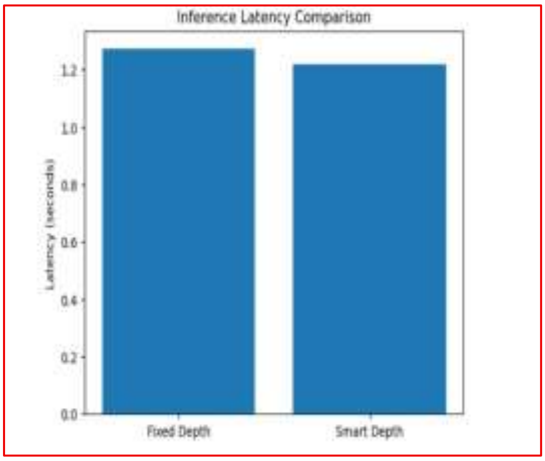
Finally, the system generates the predicted output and records performance metrics such as inference time and confidence level. This methodology enables efficient and energy-aware AI processing, making the system suitable for real-time edge intelligence applications.

V. RESULTS

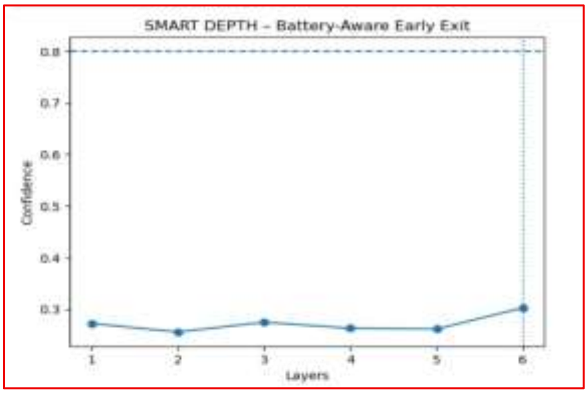
Energy Consumption Comparison



Inference Latency Comparison



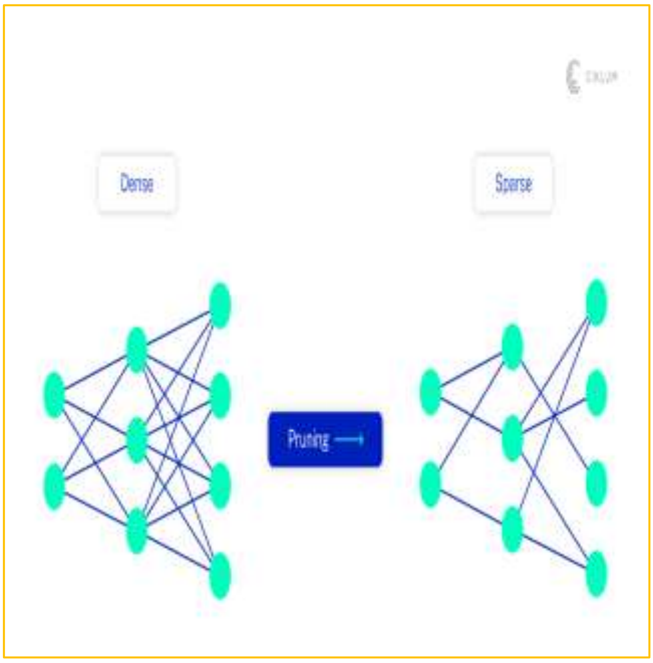
Battery-Aware Early Exit



VI. DISCUSSION

The proposed SMART DEPTH approach focuses on improving the efficiency of deep learning models by dynamically adjusting the depth of neural networks during inference. Instead of processing all layers for every input, the system allows early predictions when sufficient confidence is achieved. This reduces unnecessary computations and helps save time and energy, which is important for devices with limited resources.

OPTIMIZATION OF NEURAL NETWORKS FOR EDGE DEVICES



From the results, it is observed that dynamic inference methods can maintain good accuracy while reducing computational cost. This makes the system suitable for applications running on devices with limited processing power. The use of concepts from Deep Learning enables intelligent decision-making, while integration with Edge Computing helps process data closer to the source.

Such systems can also support various Internet of Things applications where fast and efficient processing is required.

Overall, the discussion highlights that dynamic neural networks can significantly improve performance efficiency while maintaining reliable prediction results. However, further improvements in model design and optimization techniques can enhance accuracy and scalability for more complex real-world applications.

VII. FUTURE DIRECTIONS

In the future, the Smart Depth system can be improved by integrating more advanced optimization techniques and adaptive learning methods. One possible direction is to combine the dynamic neural network architecture with automated model optimization techniques such as Neural Architecture Search, which can automatically identify the most efficient network structure for edge devices. This would further reduce computational cost and improve model performance.

Another future enhancement is the integration of hardware-aware optimization so that the system can adapt its computation strategy based on the capabilities of different edge devices such as IoT sensors, mobile processors, and embedded systems. By considering factors like memory availability, processor speed, and battery capacity, the system can dynamically adjust its processing depth to achieve better efficiency.

Additionally, future work can explore combining the Smart Depth framework with Edge Computing and Internet of Things environments to support large-scale real-time applications. This integration can enable intelligent data processing directly on devices without relying heavily on cloud infrastructure.

Further research may also focus on improving prediction accuracy while maintaining low energy consumption. Techniques such as model compression, pruning, and knowledge distillation can be applied to enhance the efficiency of the system. With these improvements, Smart Depth can be widely applied in areas such as smart surveillance, healthcare monitoring, autonomous systems, and real-time mobile applications.

VIII. CONCLUSION

In conclusion, the SMART DEPTH system improves the efficiency of deep learning models on edge devices by using dynamic neural network processing. The system reduces unnecessary computations through early-exit mechanisms while maintaining good prediction accuracy. This helps decrease energy consumption and inference time.

Overall, the proposed approach supports efficient Artificial Intelligence applications in Edge Computing and Internet of Things environments, making it suitable for real-time and resource-limited systems.

IX. REFERENCES

1. Yuxin Wang et al., “SkipNet: Learning Dynamic Routing in Convolutional Networks,” *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018.

2. Xinshi Chen, Zhangyang Wang, and Cho-Jui Hsieh, “Dynamic Neural Networks: A Survey,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

3. Alex Graves, “Adaptive Computation Time for Recurrent Neural Networks,” *Proceedings of the International Conference on Machine Learning (ICML)*, 2016.

4. Ji Lin and Song Han, “MCUNet: Tiny Deep Learning on IoT Devices,” *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

5. Kaiming He et al., “Deep Residual Learning for Image Recognition,” *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

6. Andrew G. Howard et al., “MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications,” 2017.

7. Mark Sandler et al., “MobileNetV2: Inverted Residuals and Linear Bottlenecks,” *IEEE CVPR*, 2018.

8. Karen Simonyan and Andrew Zisserman, “Very Deep Convolutional Networks for Large-Scale Image Recognition,” 2015.

9. Christian Szegedy et al., “Going Deeper with Convolutions,” *IEEE CVPR*, 2015.

10. Deep Learning based lightweight architectures for mobile and embedded systems.

11. Research studies on Edge Computing for real-time processing in smart devices.

12. Applications of Internet of Things in intelligent monitoring and automation systems.

13. Studies on early-exit neural networks for reducing inference time in real-time AI systems.

14. Optimization techniques such as pruning, quantization, and model compression in Artificial Intelligence.

15. Surveys on efficient neural network architectures for resource-constrained environments.