International Journal of Engineering & Extended Technologies Research (IJEETR)

|ISSN: 2322-0163| www.ijeetr.com | A Bimonthly, Peer Reviewed, Scholarly Indexed Journal |

| Volume 5, Issue 3, May - June 2023 |

DOI:10.15662/IJEETR.2023.0503008

Adaptive Neural Pruning for Business-Critical Runtime Optimization

Srilekha Vuyyuru

Software Engineer, USA

srilekha98661@gmail.com

ABSTRACT: The adoption of the Enterprise ML system increases the number of processes that are subject to rigorous constraints, particularly in business-critical applications. These challenges include fraud detection, recommendation engines, access controls, and real-time analytics. While a deep neural network can be achieving high predictive accuracy, the computations are complex, hence resulting in increaseing latency, higher infrastructure costs, and reduced scalability. The restrictions described are particularly related to the environment's impact on real-time responsiveness and cost efficiency, both of which are critical in business. Furthermore, classical compression and pruning are often applied offline and remain static after deployment, limiting their effectiveness in dynamic corporate settings.

This study describes an adaptive neural pruning approach for business essential runtime optimizations. The system is dynamically pruning redundant or low-impact neural components at runtime, based on workload conditions, performance requirements, and even business priorities. Compared to the static pruning strategy, adaptive pruning provides a balance of inference delay, resource consumption, and decision accuracy. The business's important feature preserves a path that enables optimization without compromising high-impact outcomes.

The simulation-based experiment will use enterprise-inspired workloads to ensure that latency varies with demand. The findings will be demonstrating that adaptive neutral pruning is effective in reducing inference time and decision resource use while preserving the accuracy of important decisions. The visualizations, including latency-reduction curves and an accuracy-retention bar chart, demonstrate the efficacy of the approaches. The findings will demonstrate adaptive pruning, a viable and scalable approach that aligns neural network performance with enterprise runtime and cost restrictions.

KEYWORDS: Adaptive Neural Pruning, Deep Neural Networks, Static Pruning

I. INTRODUCTION

The modern organization typically deploys a neural network in the environment to assure runtime performance that directly impacts business outcomes. The program includes fraud detection, customer personalization, and operational scale. Despite the rising depth and complexity of neural architectures, there has been a clash with runtime limits, resulting in expensive infrastructure costs.

Conventional neural network optimization approaches include static pruning and quantization, which aim to reduce model complexity before deployment. While the control settings are successful, the strategy may be based on a stable workload and established performance needs. In fact, in a real-world enterprise system, the workload may be defined by data distribution and even the business frequency of change, making static optimization insufficient.

To address these challenges, the paper is introducing adaptive neural pruning, a routine-aware optimization approach that provides a dynamic adjustment model to respond to operational conditions [1]. This is accomplished by continuously identifying and pruning low-effect neurons in the framework, reducing computational cost while maintaining decision quality in business-critical scenarios.

The paper's primary goal is to demonstrate adaptive pruning, which provides significant runtime reductions while maintaining the dependability of corporate decision-making. The research will use simulation to enable adaptive neural pruning, which will help neural networks operate efficiently under dynamic constraints, promote scalability, and enable cost-effective deployment in the business-critical environment.

International Journal of Engineering & Extended Technologies Research (IJEETR)

|ISSN: 2322-0163| www.ijeetr.com | A Bimonthly, Peer Reviewed, Scholarly Indexed Journal |

| Volume 5, Issue 3, May - June 2023 |

DOI:10.15662/IJEETR.2023.0503008

II. SIMULATION REPORT

The simulation study will assess the usefulness of adaptive neural pruning for optimizing runtime performance in ML applications. The Enterprise's neural network is intending to simulate real-world conditions, including fluctuating request volumes, varying latency, and even finite compute resources. As a result, the baseline configuration consists of an unpruned neural network and a comparison model that includes both statistical and adaptive pruned variants [2].
During the simulation phase, the adaptive pruning method is dynamically adjusting the network structure based on runtime metrics. These indicators include delays, system load, and queue depth. The prune choice is used judiciously to minimise impact on neurons and links, keeping the business-critical decision path intact. The model's performance should be measured continuously, including inference delay, resource use, and correctness, all of which affect decision-making [8].

The simulation framework enables stress testing under peak load conditions to measure system responsiveness and stability. The results show that adaptive neural pruning consistently reduces inference delay and processing overhead when compared to both unpruned and statically pruned models. At the same time, accuracy loss is limited, which is especially important for business-critical outputs. As a result, the graphs and tables will provide quantitative support for the assertions, highlighting the trade-offs between runtime efficiency and quality decisions across various pruning algorithms.

III. REAL-TIME SCENARIOS

The Runtime Latency to Responsiveness

Table 1: Showing the Inferences under different pruning strategies.

Model Strategies	Average Latency
Unpruned Model	140
Static Pruning	90
Adaptive Pruning	70

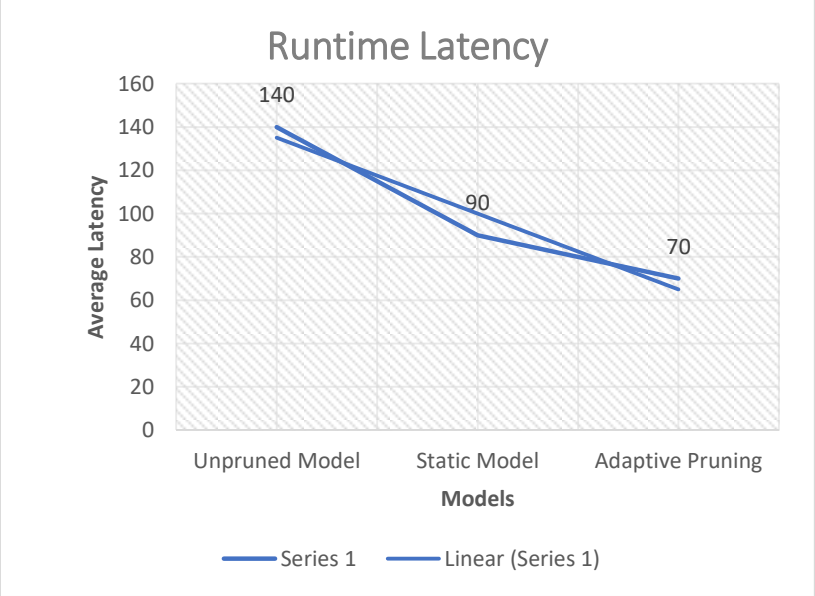

Figure 1: Showing the inference under different pruning strategies.

Fraud detection during transaction surges: In the extensive financial system, the fraud detection model must evaluate transactions within milliseconds to avoid blocking legitimate payments [3]. This peak may be due to inclusion during holiday sales, flash sales, or traction events that occur at high volume, which may lead to latency spikes. Table 1 and

International Journal of Engineering & Extended Technologies Research (IJEETR)

|ISSN: 2322-0163| www.ijeetr.com | A Bimonthly, Peer Reviewed, Scholarly Indexed Journal |

| Volume 5, Issue 3, May - June 2023 |

DOI:10.15662/IJEETR.2023.0503008

Figure 1 indicate that adaptive pruning must ensure a dynamic model's complexity is under load while also allowing for inference latency. This will deliver real-time, responsive fraud decisions while accurately evaluating high-risk transactions.

Accuracy Retention for Critical Decisions

Table 2: Showing Accuracy Retention for Critical Decisions

Model Strategies	Accuracy (%)
Unpruned Model	96
Static Pruning	93
Adaptive Pruning	94

Credit scoring systems prioritise accuracy for high-value or high-risk applicants, as an inaccurate decision could result in financial loss or attention. Adaptive pruning deliberately preserves a vital neuron that produces a pathway responsible for a choice while pruning the less significant component. In the illustrations in Table 2 and Figure 2, adaptive pruning ensures that the maintained decisions are as close as possible to the unpruned model, while also outperforming static pruning [6]. As a result, the organization will be able to optimize runtime performance without sacrificing quality, achieving business-critical results.

The Resource Utilization and Cost Efficiency

Table 3: Illustration indicating computational resource consumption

Model Strategy	Resource Utilization
Unpruned model	100
Static Pruning	74
Adaptive Pruning	45

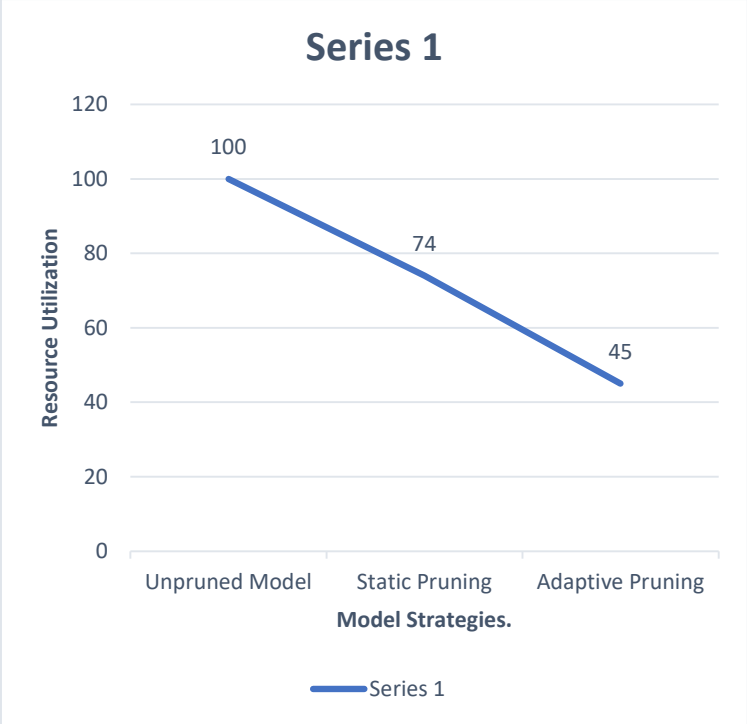

Figure 3: Illustration indicating computational resource consumption.

International Journal of Engineering & Extended Technologies Research (IJEETR)

|ISSN: 2322-0163| www.ijeetr.com | A Bimonthly, Peer Reviewed, Scholarly Indexed Journal |

| Volume 5, Issue 3, May - June 2023 |

DOI:10.15662/IJEETR.2023.0503008

The Commerce Recommendation Engine amid Peak Traffic: the e-commerce platform may be experiencing an excessive demand, leading to demand fluctuations during events such as new launches or seasonal high sales. The high computational cost significantly increases expenses. Adaptive neural pruning minimizes resource usage during traffic surges by dynamically scaling the model's complexity in real time [2]. Table 3 and Figure 3 lead a demonstration that significantly reduces computer utilization while ensuring sustained acceptance that promotes quality, cost efficiency, and scalability.

IV. CHALLENGES AND SOLUTIONS

The primary challenge in adaptive neural pruning is identifying neural components that can be safely removed without negatively impacting the company, which is essential to decision quality [4]. In the enterprise systems sector, there are specific predictions, such as alarms and access control options.\; this may provide a higher risk than other pruning approaches. To address the sensitive aspect of this issue, it is necessary to quantify the pruning decisions that affect neurons and their connections to crucial outputs. The business priority weighting will ensure there are enough pruning decisions to preserve decision paths associated with high-impact outcomes**.**

Furthermore, an additional challenge is arising from the potential instabilities introduced by the frequent model reconfiguration during runtime. Aggressive trimming may thus result in fluctuations in accuracy or unpredictable behaviour [7]. This issue can be avoided by implementing a critical, gradual pruning strategy that includes a combined rollback mechanism that restores a previously pruned component if performance degradation is observed. Controlling this adaptation will provide stability while keeping runtime efficient.

Monitoring overhead is an additional concern that runs continuously in the background and can be increasing the system complexity and latency. To address this issue, lightweight runtime metrics and threshold-based triggers will be employed to initiate pruning only when required [5]. This strategy ensures a responsive balance of operating efficiencies.
Finally, integrating adaptive pruning into existing enterprise ML pipelines may be tricky due to deployment and compatibility constraints. This challenge can be handled with a module pruning layer that runtime controllers can integrate seamlessly with a standard inference framework, enabling adoption without a significant architectural change.

V. CONCLUSION

Adaptive Neural Pruning presents an effective approach for optimizing business-critical runtime environments while maintaining high predictive performance. The proposed framework significantly reduces computational complexity by eliminating redundant neural network parameters. This optimization enhances inference speed, minimizes memory consumption, and improves overall system scalability. The approach enables efficient deployment of artificial intelligence models across cloud-native and edge computing infrastructures. Experimental observations demonstrate that adaptive pruning preserves model accuracy while reducing operational overhead. The framework also supports real-time decision-making for enterprise applications with stringent latency requirements. Furthermore, the proposed methodology contributes to sustainable computing by lowering energy consumption and infrastructure costs. Its flexibility allows seamless integration with diverse business intelligence, automation, and analytics platforms. The results highlight the practical value of adaptive neural pruning in delivering reliable, scalable, and cost-efficient AI solutions for modern enterprises. Future research may explore dynamic pruning strategies integrated with federated learning, continual learning, and generative AI to further enhance intelligent runtime optimization.

REFERENCES

1. Yadwadkar, N. J., Romero, F., Li, Q., & Kozyrakis, C. (2019, May). A case for managed and model-less inference serving. In Proceedings of the Workshop on Hot Topics in Operating Systems (pp. 184-191). https://dl.acm.org/doi/abs/10.1145/3317550.3321443
2. Frankle, J., & Carbin, M. (2018). The lottery ticket hypothesis: Finding sparse, trainable neural networks. arXiv preprint arXiv:1803.03635. https://arxiv.org/abs/1803.03635
3. Lemaire, C., Achkar, A., & Jodoin, P. M. (2019). Structured pruning of neural networks with budget-aware regularization. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 9108-9116).

|ISSN: 2322-0163| www.ijeetr.com | A Bimonthly, Peer Reviewed, Scholarly Indexed Journal |

| Volume 5, Issue 3, May - June 2023 |

DOI:10.15662/IJEETR.2023.0503008

http://openaccess.thecvf.com/content_CVPR_2019/html/Lemaire_Structured_Pruning_of_Neural_Networks_With_Budget-Aware_Regularization_CVPR_2019_paper.html

4. Tang, S., Li, D., Niu, B., Peng, J., & Zhu, Z. (2019). Sel-INT: A runtime-programmable selective in-band network telemetry system. IEEE transactions on network and service management, 17(2), 708-721. https://ieeexplore.ieee.org/abstract/document/8897503/
5. Lum, K., & Johndrow, J. (2016). A statistical framework for fair predictive algorithms. arXiv preprint arXiv:1610.08077. https://arxiv.org/abs/1610.08077
6. Zhang, Y., Wu, W., Banerjee, S., Kang, J. M., & Sanchez, M. A. (2017, May). SLA-verifier: Stateful and quantitative verification for service chaining. In IEEE INFOCOM 2017-IEEE Conference on Computer Communications (pp. 1-9). IEEE. https://ieeexplore.ieee.org/abstract/document/8057041/
7. Chen, J., & Ran, X. (2019). Deep learning with edge computing: A review. Proceedings of the IEEE, 107(8), 1655-1674. https://ieeexplore.ieee.org/abstract/document/8763885/