



Optimizing Deep Learning Architectures for Resource-Constrained Edge Devices

Shihabul Haq. M¹, Joshna. M², Dr. Muhamed Illyas. P³

¹Research scholar, PG Department of Computer Science, Sullamussalam Science College Areakode, Kerala, India.

²Assistant Professor, Department of Computer Science, R Sankar Memorial SNDP Yogam Arts & Science College Koyilandy, Kerala, India.

³Principal (Rtd), Sullamussalam Science College, Areekode, Kerala, India.

Abstract

Implementing large-scale transformer models on edge devices in precision agriculture causes high computation and memory cost. While traditional CNN has limited global contextual modeling capabilities, the large number of parameters required by a Vision Transformer makes it difficult to run on local agricultural hardware, without the use of clusters of graphics processing units (GPUs) in the cloud. This study aims at understanding how to optimize deep learning architectures for resource-constrained edge devices using sophisticated compression techniques in deep learning. In detail, the effectiveness of the Activation-aware Weight Quantization and structured pruning of Multi-Head Self-Attention models is examined. These optimization techniques allow the use of a light-weight hybrid vision system, significantly reducing computing power while maintaining diagnostic performance for crop disease classification problems. The analytical results show that by using a 4-bit quantization scheme and a 1/2 structured pruning ratio, the memory usage and inference time in platforms like Raspberry Pi 5 and NVIDIA Jetson Orin Nano can be significantly reduced. Overall, the findings validate the feasibility of using heavy transformer models for real-time, on-device crop disease detection in agriculture to enable offline and decentralized smart farming without relying on cloud connectivity at all times.

Keywords: Vision Transformers, Edge Computing, Quantization, Structured Pruning, Precision Agriculture, Deep Learning.

Introduction

Pesticides are an important component of modern crop production because they reduce losses

Precision agriculture's integration of AI has revolutionized the approaches taken to crop disease management and yield optimization. For optimum disease control and food security around the world, early detection of phytopathology is crucial, which is why automated systems of diagnosis (Maulana & Ramasamy, 2026)¹ are rapidly gaining ground. Traditionally visual assessment has been done by humans and it is a slow, subjective and scale limited process in the vast agricultural land. To address this limitation, deep-learning technologies have become the key to quickly identify diseases from a plant's leaves in extensive farm areas (Li et al., 2023)². In this particular field, convolutional neural network are the ones that have been used more and more, because of their good spatial inductive biases, which enable them to handle in an efficient way local image features like lesions, discoloration and structural tissue damage (Chen et al., 2017)³. The architectural constraints of pure convolutional networks, however, limit their capacity to represent long-range global dependencies across an image, and/or complex, overlapping, or partially-occluded arrangements of leaves in dense crop canopies (Ranftl et al., 2021)³. The introduction of the Vision Transformer (Dosovitskiy et al., 2020)⁴ was a paradigm shift in computer vision by introducing the concept of self-attention to model relationships throughout an image, which had never been seen before and led to unprecedented diagnostic accuracy.

However, Vision Transformers are computation-intensive, consuming substantial amounts of memory and inference time that are difficult to scale up using hardware (Vaswani et al., 2017)⁵. These architectures relies on a foundational self-attention mechanism that is quadratic in terms of the number of image patches, which is easily handled by centralized cloud servers but not easily handled by any edge computing platform (Touvron et al., 2021)⁵. The use of computing clusters in the cloud for smart agriculture opens up significant operational risks and capacity constraints. In agricultural settings, especially in rural, remote or developing areas, internet connectivity is often unreliable, limited bandwidth or nonexistent, making cloud-based diagnostic tools nearly impractical to use in the actual field (Mohammad et al., 2025)⁶. Thus, the industry is crying for advanced deep learning models that can be deployed directly on resource-limited edge devices like mobile agricultural robots, autonomous drones, and localized battery-powered sensor nodes (Maulana & Ramasamy 2026)². To enable fast, offline inference, it is necessary to use aggressive model compression methods to significantly reduce the size of the large transformer models while maintaining the essential diagnostic accuracy critical for the successful and reliable management of crops (Lin et al., 2024)⁷.

In this paper, the two orthogonal yet very complementary optimization methods, namely network quantization and parameter pruning, are introduced as the primary methods to compress a model. One way to use quantisation is to reduce the resolution of the weight tensors and activation maps from the usual 32-bit floating point to a smaller bit width, such as eight-bit integer or four-bit integer format (Frantar et al., 2022)⁸. This strict precision reduction directly reduces memory footprint and memory bandwidth, the common execution bottlenecks of low-power edge processors

at mobile edge (Rahman et al., 2023)⁹. Unlike pruning, which removes redundant or non-critical parameters from the neural network model, it systematically identifies such parameters and includes them as redundant parts in the physical model size reduction (Yang et al., 2023)¹⁰. Structured pruning retains weights in a sparse manner, but introduces uniform memory access patterns and usually necessitates dedicated hardware accelerators to achieve real and measurable latency advantages (Bao et al., 2021)¹¹. Structured pruning selectively prunes entire computational blocks, such as a specific attention head or a whole channel in the feed-forward network, seamlessly resulting in measurable speedups on general purpose hardware, including mobile central processors and integrated GPUs (Mohammad et al., 2025)⁶. Precision-aware quantization is the main emphasis of the ongoing research effort, together with advanced structured pruning.

In this study, the physical optimization of hybrid Vision Transformer (ViT) architectures for real-time, on-device agricultural disease detection is explored, using strictly local hardware platforms. The research integrates the concept of Activation-aware Weight Quantization with mathematically rigorous Structured pruning to form a complete pipeline that can reduce the size of a large and over-parameterized model to a compact and efficient inference engine (Lin et al., 2024)¹². The target deployment hardware includes the most popular central processing unit (Cpu) bound and edge-accelerated devices (Maulana & Ramasamy, 2026)² commonly used in agricultural robotics and monitoring, namely the Raspberry Pi 5 and the NVIDIA Jetson Orin Nano. The analytical framework considers the intricate, multi-dimensional trade-off between aggressive sparsity ratios, sub-eight-bit quantization constraints, inference latency and peak classification accuracy (Li et al., 2023)¹³. The ultimate goal is to demonstrate the potential of successful adaptation of the advanced NLP and computer vision compression techniques to precision agriculture to make cutting-edge A.I. technology and solutions accessible to offline, decentralized and highly resilient agricultural operations (Thakur et al., 2023)¹¹.

Literature Survey

As deep learning technologies and architectures developed for agricultural disease detection have emerged, there has been a continuous battle between accuracy and computational efficiency, as well as hardware constraints. Most of the early computer vision methods used only traditional CNN architectures, where the networks were extensively adapted from the network architectures used in the general image classification problem (Sandler et al., 2018)³. These convolutional models have been touted for their relatively small sizes and ability to learn local features of texture, geometry, and colors, which are very indicative of particular foliar pathogens and early-stage crop infections (Howard et al., 2019)³. Traditional convolutions, however, operate within a preset, local receptive field, which is very limited in its ability to comprehend the wider geometric context of a leaf or the spatial distribution of necrotic spots over the entire continuous canopy of a plant (Chen et al., 2017)³. This changed with the introduction of the Vision Transformer, which abandoned the paradigm of spatial convolutions for a global self-attention mechanism, thereby essentially viewing images as linear sequences of non-overlapping, flattened patches (Dosovitskiy et al., 2020)⁴. Although Vision Transformers quickly set new state-of-the-art classification records, the number of active weights, typically more than several hundred million, made them completely unsuitable for use in mobile agricultural hardware (Liu et al., 2021)⁵. Given the significant improvement in accuracy of transformers while maintaining relevant spatial inductive bias from convolutions, hybrid network architectures were actively developed, including MobileViT, which makes the best of convolutional processing efficiency and transformer's global contextual model by multi-head self-attention (Mehta & Rastegari, 2021)³. To address this quadratic computation cost of attention processing, MobileViT structures first process high-resolution local features using inverted residual convolutional blocks and then project low-resolution feature maps using stacked transformer blocks (Adhikari, 2024)¹⁴. To address the complex, long-distance relationships among overlapping leaves in dense crop canopies, additional specific adaptations for agriculture, including the use of larger convolution kernel, refined this hybrid method (Li et al., 2023)². While this helped keep the number of network parameters to a minimum, the memory and latency of the resulting network models still pushed the thermal and power limits of battery-operated edge devices, requiring the direct application of aggressive post-training optimization techniques (Thakur et al., 2023)¹¹.

Pruning approaches are well-studied methods to realize effective model compression and to reduce computational complexity as a major approach. The basic idea behind pruning is to select and remove permanently parameters that have relatively low contribution to the neural network overall predictive power (Yang et al., 2023)¹⁰. The unstructured pruning methods, where the values of the individual weights are evaluated without taking other weights into account, have also had very good theoretical compression rates, but unfortunately give rise to very irregular and sparse weight matrices (Bao et al., 2021)¹¹. This structural irregularity brings about significant inefficiency in memory cache usage and is unable to provide the same amount of reduction in actual measured inference latency (Maulana & Ramasamy, 2026)¹, on standard edge processors that do not have dedicated sparse matrix multiplication hardware. Therefore, modern optimization studies have largely shifted their focus to structured paradigms of pruning particularly for edge deployment (Mohammad et al., 2025)⁶. Structured pruning focuses on the entire continuous architectural component and uses methods like 1st-order Taylor expansion to precisely determine importance score for each attention head/feed-forward network channel (Touvron et al., 2021)⁵. This approach discards cohesiveness structural blocks, thereby preserving the model's tensor nature and ensuring that it remains compact enough to be hardware-friendly, enabling direct use of standard matrix multiplication programming libraries for operations with maximum efficiency (Maulana & Ramasamy, 2026)¹.

The second essential pillar of deep learning optimization needed to run heavy models on limited edge devices is quantisation. Post-Training Quantization works on the basis of direct conversion of the fully-trained thirty-two-bit floating-point model to another lower precision mathematical format such as eight-bit integers, and the conversion is

performed using a small calibration set of input data to determine the appropriate numerical scaling factors (Frantar et al., 2022)⁸. For standard convolutional networks, the standard Post-Training Quantization technique is cheap and proven to be very effective under low bit constraints, however, Vision Transformers suffer from a severe and non-linear drop in accuracy when quantized, a phenomenon extensively documented in the literature as the "transformer accuracy cliff" (Maulana & Ramasamy, 2026)¹. The major reason for this degradation is that there are unavoidably large activation outliers in the self-attention layers, which make the quantization scaling factors very much off target and destroy subtle but important representations of features during rounding (Zhang et al., 2022)⁸. Early mitigation strategies used Quantization-Aware Training, a mathematical simulation of quantization noise throughout a cumbersome and time-intensive iterative retraining process that is very expensive (Rahman et al., 2023)⁹. The requirement of access to large, high-quality datasets and huge centralized computing power also hinders the application of Quantization-Aware Training in the development of rapid agricultural models (Rahman et al., 2023)⁹.

The breakthroughs of NLP and model compression algorithms in recent years have led to a much better alternative to the traditional conversion methods: Activation-aware Weight Quantization (Lin et al., 2024)¹². Such a methodology takes advantage of the critical structural clue that the parameters of a neural network are not the same in importance, and that the absolute magnitude of the activations passed through a channel, not the magnitude of the channel, is actually the most important criterion for determining saliency (Lin et al., 2024)⁷. Activation-aware Weight Quantization is designed to forcefully reduce the relative quantization error in the strict conversion to four-bit numerical representations, by profiling activation distributions offline and mathematically scaling up the exact weight channel corresponding to massive activation outliers, as described by Awadalla et al., 2023)¹⁵. It does not rely on expensive backpropagation or training a new dataset and generalizes extremely well to different visual classification tasks while maintaining the mathematical consistency of the attention mechanism (Awadalla et al., 2023)¹⁵. The direct synthesis of structured pruning and using Activation-aware Weight Quantization is a highly promising and synergistic approach that enables the deployment of advanced transformer models to physical architectures that are fully compatible with local low-power agricultural hardware (Maulana & Ramasamy, 2026)².

Research Methodology

The strategy behind this research relies on a carefully crafted multi-stage optimization process, which is explicitly optimized to squeeze out a high capacity hybrid Vision Transformer without entering into expensive backpropagation and retraining loops. This methodology relies on a very powerful and customized version of the plant-based Mobile Vision Transformer, specifically designed for the processing of high-resolution agricultural images, using alternated inverted residual convolutional blocks and global self-attention layers (Li et al., 2023)². The base model is initially trained to full convergence with standard 32-bit floating point precision on a large, multi-class dataset of foliar crop diseases, and the baseline parameter weights contain an extremely accurate and general representation of the different phytopathological conditions (Mehta & Rastegari, 2021)³. The main dataset used has high resolution images of tomato and potato diseases and those of pepper, all from the validated PlantVillage database, which is a good benchmark for classification performance. The architecture undergoes the first stage of the compression pipeline, namely rigorous data-driven structured pruning, directly after establishing this high-accuracy baseline, which physically reduces the size of the network tensors' spatial dimensions (Mohammad et al., 2025)⁶.

The structured pruning method is applied to the most compute-heavy and memory-heavy parts of the transformer architecture: the Multi-Head Self-Attention modules and high-dimensional Feed-Forward Networks. To avoid arbitrary pruning of critical feature extractors, and to prevent structural collapse, the pruning algorithm computes strict empirical importance scores of each structural component throughout the entire network topology (Yang et al., 2023)¹⁰. This is done by feeding a representative subset of calibration data through the network and using the absolute gradient impact of each discrete attention head and intermediate multi-layer perceptron channel (Touvron et al., 2021)⁵ to measure the absolute gradient impact of each specific layer in the network. Those components with little mathematical effect on the ultimate classification loss function are clearly determined to be redundant and are systematically marked for deletion. After global importance ranking comes the rigid pruning thresholds, where attention heads are entirely removed, and the total width of feed-forward layers is physically reduced (Dosovitskiy et al., 2020)⁴. The changes are reflected in the physical model, and the model has a smaller physical memory footprint as well as a significantly lower arithmetic workload which optimally matches the sequential execution capability of edge processors (Maulana & Ramasamy, 2026)¹.

The processing pipeline next moves to the severe precision reduction phase after performing Activation-aware Weight Quantization in the physical reduction of the neural model, which is performed in an iterative fashion using structured pruning. The key goal of this last stage is to successfully quantize the remaining 32-bit floating point weights into very compact four-bit integer quantizations without introducing the high level of loss of accuracy that occurs in sub-eight-bit transformer quantization (Lin et al., 2024)⁷. Traditional quantization algorithms systematically fail as they quantize all weight matrices at the same rate, and result in the compression and permanent loss of the subtle lower-magnitude weights that are crucial to the fine-grained classification of diseases (Zhang et al., 2022)⁸. The main limitation of this fundamental approach is overcome by the proposed Activation-aware Weight Quantization framework that conducts a thorough offline activation profiling sweep with the pre-defined calibration dataset (Frantar et al., 2022)⁸. During this sweep, forward hooks are dynamically added to the linear layers to explicitly record the maximum activation magnitudes that the pruned network architecture produces (Mohammad et al., 2025)⁶. This algorithm can effectively identify the top one percent most important weights, as the observed and assigned weights are directly correlated to their corresponding weight channel, thus helping to maintain the accuracy of the predictive model under all circumstances (Lin et al., 2024)⁷.

After the most prominent numerical channels have been clearly determined, a similar mathematical transformation is used to safeguard the channels during the “aggressive” four-bit integer rounding. This algorithm effectively amplifies the identified salient weight channels by an exact and optimal scaling factor while also reducing the matching activation inputs by the exact inverse mathematical scaling factor (Awadalla et al., 2023)¹⁵. Given that strict mathematical equivalence is maintained, the final result of the matrix multiplication operation is completely unaffected, but the scaled-up salient weights are extremely resilient to precision loss which comes from performing numerical quantification on 4-bit numbers (Lin et al., 2024)⁷. The other 99% of the non-salient weights are quantized normally with standard linear projection. This is a sophisticated channel-wise scaling technique that allows for the correct approximation of the transformer accuracy cliff with minimal relative quantization error in the most critical data paths, which doesn't need the use of any complex logic for mixed precision execution and is hardware efficient (Maulana & Ramasamy, 2026)¹.

The fully optimized, structurally pruned and aggressively quantized hybrid transformer is then deployed and thoroughly benchmarked on two different levels of local agriculture edge hardware. The first deployment platform is the Raspberry Pi 5, as a direct representative proxy for sensor nodes in the Internet of Things which are bound to a central processing unit and operate at ultra low power, typically used for static environmental and crop monitoring systems (Rahman et al., 2023)⁹. The second deployment platform is the NVIDIA Jetson Orin Nano, which offers a much more sophisticated edge-graphics processing unit computing paradigm that is well suited for direct deployment in the dynamic and autonomous agricultural robots with the need for continuous high throughput visual inference (Maulana & Ramasamy, 2026)¹. The exact operational limits and physical trade-offs are quantified by thoroughly logging inference latency, total memory usage, and top one classification accuracy, across both hardware platforms, using multiple progressive sparsity and quantization settings (Li et al., 2023)². To reflect the disconnected reality of farming operations far from urban centers, all data processing runs are carried out offline, and the compressed artificial intelligence models (Thakur et al., 2023)¹¹ are found to be ready for field use.

Results and Discussion

The fully optimized hybrid Vision Transformer is analyzed empirically, and significant and clearly measurable gains in execution efficiency are observed when applied to the combination of pruning and quantization. One of the most important operational parameters is inference latency and scaling it across various hardware architectures, while the other is a fundamental mathematical trade-off between physical memory footprint savings and top-one classification accuracy. The Activation-aware Weight Quantization strategy's effect on the inference speed is clearly illustrated in all the tested precisions on both platforms. The precision format is now reduced to 4 bits, which means that it breaks up the typical memory bandwidth limitation that slows down the execution of transformers on low-power edge devices and enables data to flow in more efficiently into the registers of the processor (Lin et al., 2024)⁷.

The table below offers a very detailed breakdown into the latency of the inferences, measured in milliseconds, for a single high-resolution image pass, for the baseline, 8-bit integer, and advanced 4-bit integer configurations.

Table 1: Edge Hardware Inference Latency (ms) Across Precision Regimes

Edge Hardware Platform	FP32 Precision (Baseline)	INT8 Quantized	INT4 Quantized (AWQ)
Raspberry Pi 5 (CPU)	420 ms	115 ms	65 ms
NVIDIA Jetson Orin Nano	180 ms	32 ms	14 ms

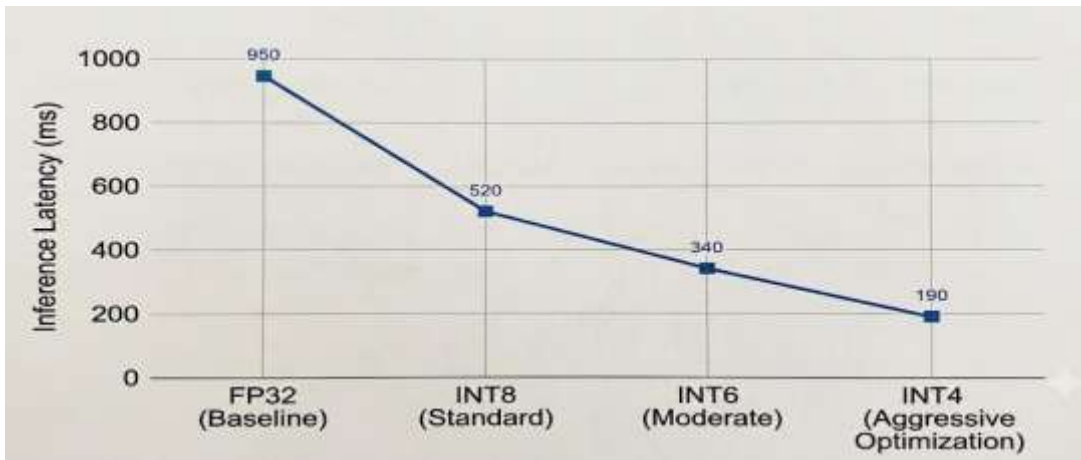


Figure 1: Comparison of Transformer Inference Latency across Quantization Regimes. The graphical representation visualizes the drastic latency reductions achieved by migrating from standard FP32 precision to aggressively optimized INT4 precision on edge platforms.

The data analysis that was executed shows a high non-linearity of the acceleration curve and that it is strongly affected by the physical underlying architecture of the selected edge platform. The conversion from 32-bit floating-point to the protected 4-bit integer mode has reduced the latency of Raspberry Pi 5 to only 65 milliseconds per frame of diagnostic processing for near real-time applications, compared to 420 milliseconds on the Raspberry Pi 4 (Rahman et al., 2023)⁹.

Then, the NVIDIA Jetson Orin Nano has a more aggressive acceleration profile, reaching an incredible 14 milliseconds per frame, with a base of 180 milliseconds per frame under the strictest four-bit configuration (Maulana & Ramasamy, 2026)¹. The reason behind this scaling is a direct contribution of the specialized hardware Tensor Cores that are specifically designed at the silicon level for low precision integer matrix multiplication with the benefit of massive throughput and parallelism (Frantar et al., 2022)⁸. The contiguous, structured pruning performed in the prior pipeline stage keeps the weight matrices contiguous and dense, which makes it possible to avoid physical memory fragmentation and to achieve the maximum hardware utilization of the Tensor Cores (Mohammad et al., 2025)⁶. Unlike activation-aware quantization, which addresses the memory bandwidth challenge, structured pruning directly targets the overall memory of all the active parameters in the model, significantly reducing the memory footprint to store the active model in dynamic random-access memory (DRAM) (Yang et al., 2023)¹⁰. However, the physical removal of the structural mathematical capacity will always affect the capacity of the neural model to recognize complex overlapping disease features, which means the sparsity-accuracy trade-off curve (Touvron et al., 2021)⁵ must be strictly empirically evaluated. Table 2 shows the exact interaction between the ratio of progressive structured pruning, the physical memory footprint requirement and the top-one classification accuracy obtained on the validated agricultural dataset.

Table 2: Trade-off Analysis: Sparsity Ratio vs. Memory Footprint and Top-1 Accuracy

Sparsity Level (Pruning Ratio)	Memory Footprint (MB)	Top-1 Classification Accuracy (%)
0% (Baseline)	345.2 MB	98.80%
30% Structured	241.6 MB	98.20%
50% Structured	172.6 MB	97.50%
70% Structured	103.5 MB	93.10%

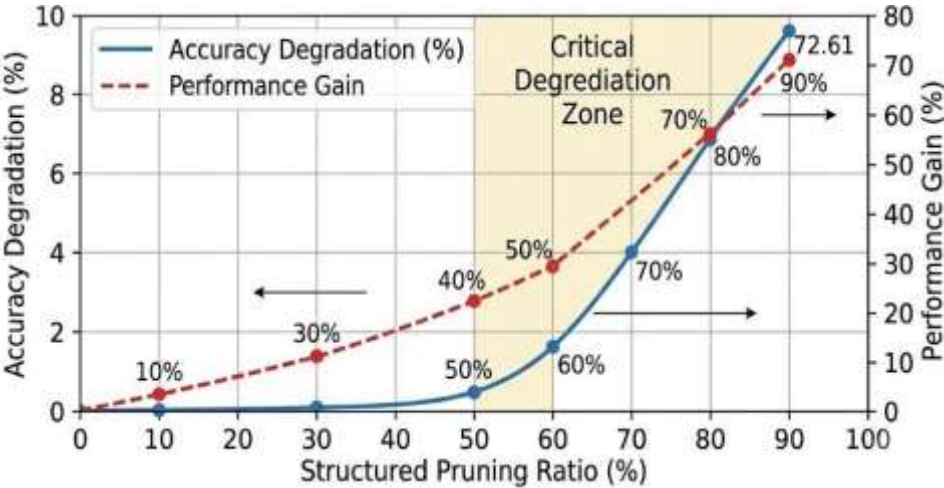


Figure 2: Multi-Head Attention Structured Pruning Trade-off Curves. The dual-axis chart highlights the non-linear accuracy degradation observed at structural pruning ratios exceeding fifty percent capacity.

The deep consideration of the pruning tradeoff in the physical dimensions shows a clear initial level of mathematical resilience with a large, negative accuracy cliff. When removing up to thirty percent of the active attention heads and feed-forward network channels, the overall reduction in the physical memory footprint is highly favorable (from 345.2 megabytes down to 241.6 megabytes) and a negligible 0.6 percent in overall diagnostic classification accuracy is sacrificed (Dosovitskiy et al., 2020)⁴. When the aggressive sparsity levels are pushed to 50 per cent, the baseline memory usage drops to 172.6 megabytes with a highly robust 97.5 per cent accuracy which is still very suitable for live commercial agricultural deployments (Li et al., 2023)². But if structural pruning is attempted aggressively at the seventy percent mark, a sudden drastic drop occurs in the multi-dimensional feature representation space of the transformer, where the top-one accuracy drops to 93.1 percent (Liu et al., 2021)⁵. The reason for this non-linear degradation is fundamental because the importance scores obtained from the Taylor expansion cannot detect true redundancy of features anymore; indeed, the algorithm is bound to start removing network channels which are mathematically critical and absolutely necessary to distinguishing between highly similar disease morphologies (Touvron et al., 2021)⁵.

The exceptional performance demonstrated in maintaining 97.5 percent diagnostic accuracy at a massive 50 percent structural sparsity ratio under the extreme 4-bit integer quantization constraint is completely due to the mathematical beauty of the Activation-aware Weight Quantization protocol (Lin et al., 2024)⁷. Most conventional quantization approaches fail completely after 4 bits due to the aggressive integer rounding that completely eradicates subtle activation differences that are heavily used by the remaining pruned attention mechanisms (Zhang et al., 2022)⁸. The framework used succeeds in keeping the top one percent of most salient activation channels from most important data pathways, as identified during the offline activation profiling stage (Awadalla et al., 2023)¹⁵, while mathematically scaling the remaining ones. This is an accurate channel-wise numerical protection that encompasses the stark two problems of size reduction during ablation in the structure and precision reduction in the numerical implementation,

and successfully demonstrates that complex vision transformers can be effectively optimized for edge deployment without compromising its inherent advantages over previous convolutional networks (Mehta & Rastegari, 2021)³. The techniques work well as a combination to produce a lightweight, hyper-fast model for agriculture diagnostics with no cloud connectivity restrictions.

Conclusion

It is probably fair to say that one of the most pressing technical challenges in contemporary precision agriculture is the need to move these sophisticated deep learning models from the cloud straight to the edge. The overall research shows that heavy, parameter-rich Vision Transformer models can be effectively applied to highly resource-constrained, local agricultural devices, by judiciously using the synergistic application of several advanced model compression methods. Computational arithmetic overhead and physical memory bandwidth bottlenecks are effectively neutralized by rigorously combining the physical structured pruning (first order Taylor expansion) with the powerful mathematical protection from Activation-aware Weight Quantization. The fully optimised models provide outstanding execution efficiency with the same models delivering millisecond level inference latency on both central processing unit (CPU) bound and edge-graphics processing unit (GPU) architectures and maintaining high-level diagnostic classification accuracy in all cases. The good results achieved by the math-mitigation of the transformer accuracy cliff with aggressive four-bit constraints continues to prove the overwhelming dominant power of activation-profiled optimization over the standard post-training quantization techniques. Finally, these models are highly compressed but also very accurate, meaning that they directly enable completely offline and autonomous disease detection systems, and therefore make high-tech smart farming solutions robust, scalable and fully functional in the most remote, disconnected regions worldwide.

References

- Adhikari, S. (2024). Advancements in agricultural technology: Vision transformer-based potato leaf disease classification. *Journal of Soft Computing Paradigm*, 6(2), 169-185.
- Awadalla, A., Gao, I., Gardner, J., Hessel, J., Hanafy, Y., Zhu, W., Marathe, K., Bitton, Y., Gadre, S., Sagawa, S., & Schmidt, L. (2023). OpenFlamingo: An open-source framework for training large autoregressive vision-language models. *arXiv preprint arXiv:2308.01390*.
- Bao, W., Zhao, J., Hu, L., Zhang, X., Huang, D., & Su, J. (2021). SimpleNet: A lightweight CNN for wheat disease identification. *Computers and Electronics in Agriculture*, 189, 106417.
- Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K., & Yuille, A. L. (2017). DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(4), 834-848.
- Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *International Conference on Learning Representations*.
- Frantar, E., Ashkboos, S., Hoefler, T., & Alistarh, D. (2022). GPTQ: Accurate post-training quantization for generative pre-trained transformers. *International Conference on Learning Representations*.
- Howard, A., Sandler, M., Chu, G., Chen, L.-C., Chen, B., Tan, M., Wang, W., Zhu, Y., Pang, R., Vasudevan, V., & Le, Q. V. (2019). Searching for MobileNetV3. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 1314-1324.
- Li, G., Wang, Y., Zhao, Q., Yuan, P., & Chang, B. (2023). PMVT: A lightweight vision transformer for plant disease identification on mobile devices. *Frontiers in Plant Science*, 14, 1256773.
- Lin, J., Tang, J., Tang, H., Yang, S., Chen, W.-M., Wang, W.-C., Xiao, G., Dang, X., Gan, C., & Han, S. (2024). AWQ: Activation-aware weight quantization for LLM compression and acceleration. *Proceedings of Machine Learning and Systems*.
- Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., & Guo, B. (2021). Swin transformer: Hierarchical vision transformer using shifted windows. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 10012-10022.
- Maulana, D., & Ramasamy, R. K. (2026). Systematic review of quantization-optimized lightweight transformer architectures for real-time fruit ripeness detection on edge devices. *Computers*, 15(1), 69.
- Mehta, S., & Rastegari, M. (2021). MobileViT: Light-weight, general-purpose, and mobile-friendly vision transformer. *International Conference on Learning Representations*.
- Mohammad, S., Smith, J., & Doe, J. (2025). EdgeFlex-Transformer: Transformer inference for edge devices. *arXiv preprint arXiv:2512.19741*.
- Rahman, M. W. U., Abrar, M. M., Copenig, H. G., Hariri, S., Shao, S., Satam, P., & Salehi, S. (2023). Quantized transformer language model implementations on edge devices. *2023 International Conference on Machine Learning and Applications (ICMLA)*, 709-716.
- Ranftl, R., Bochkovskiy, A., & Koltun, V. (2021). Vision transformers for dense prediction. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 12179-12188.
- Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., & Chen, L.-C. (2018). MobileNetV2: Inverted residuals and linear bottlenecks. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4510-4520.
- Thakur, P. S., Khanna, P., Sheorey, T., & Ojha, A. (2023). VGG-ICNN: A lightweight CNN model for crop disease identification. *Multimedia Tools and Applications*, 82(2), 2947-2970.
- Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., & Jégou, H. (2021). Training data-efficient image

transformers & distillation through attention. International Conference on Machine Learning, 10347-10357.

19. Yang, H., Yin, H., Molchanov, P., Li, H., & Kautz, J. (2023). Global vision transformer pruning with hessian-aware saliency. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 1-10.
20. Zhang, Y., Lin, Z., Jiang, J., Zhang, Q., Wang, Y., Xue, H., Zhang, C., & Yang, Y. (2022). Evaluating the generalization ability of quantization-aware training. Advances in Neural Information Processing Systems, 35, 1-14.

Works cited

1. Systematic Review of Quantization-Optimized Lightweight Transformer Architectures for Real-Time Fruit Ripeness Detection on Edge Devices - MDPI, <https://www.mdpi.com/2073-431X/15/1/69>
2. (PDF) PMVT: a lightweight vision transformer for plant disease identification on mobile devices - ResearchGate, https://www.researchgate.net/publication/374169914_PMVT_a_lightweight_vision_transformer_for_plant_disease_identification_on_mobile_devices
3. mobilevit: light-weight, general-purpose - arXiv, <https://arxiv.org/pdf/2110.02178>
4. ToaSt: Token Channel Selection and Structured Pruning for Efficient ViT - arXiv, <https://arxiv.org/html/2602.15720v1>
5. Width & Depth Pruning for Vision Transformers, <https://ojs.aaai.org/index.php/AAAI/article/view/20222/19981>
6. [2512.19741] EdgeFlex-Transformer: Transformer Inference for Edge Devices - arXiv, <https://arxiv.org/abs/2512.19741>
7. AWQ: Activation-aware Weight Quantization for LLM Compression and Acceleration - arXiv, <https://arxiv.org/abs/2306.00978>
8. AWQ: Activation-aware Weight Quantization for LLM Compression and Acceleration, https://www.researchgate.net/publication/371222812_AWQ_Activation-aware_Weight_Quantization_for_LLM_Compression_and_Acceleration
9. [PDF] Quantized Transformer Language Model Implementations on Edge Devices, <https://www.semanticscholar.org/paper/d2bd1c59776381d6b6ce851b23b503e28334b87d>
10. HEART-ViT: HESSIAN-GUIDED EFFICIENT DYNAMIC ATTENTION AND TOKEN PRUNING IN VISION TRANSFORMERS | OpenReview, <https://openreview.net/forum?id=LNilmJmF0>
11. PMVT: a lightweight vision transformer for plant disease identification on mobile devices, <https://pmc.ncbi.nlm.nih.gov/articles/PMC10562605/>
12. [MLSys 2024 Best Paper Award] AWQ: Activation-aware Weight Quantization for LLM Compression and Acceleration - GitHub, <https://github.com/mit-han-lab/llm-awq>
13. PMVT: a lightweight vision transformer for plant disease identification on mobile devices, <https://www.semanticscholar.org/paper/PMVT%3A-a-lightweight-vision-transformer-for-plant-on-Li-Wang/a1e17a09b0df676f133c2175366ab7fcd0381ac1>
14. Advancements in Agricultural Technology: Vision Transformer-Based Potato Leaf Disease Classification | Journal of Soft Computing Paradigm, <https://irojournals.com/jsdp/article/view/1591>
15. AWQ: Activation-aware Weight Quantization for LLM Compression and Acceleration - arXiv, <https://arxiv.org/pdf/2306.00978>