Research on Lightweight and NPU Hardware Acceleration of ViT Model Based on Pruning-Distillation-Quantization

Meilin Deng
Shenyang University of Technology, Shenyang, Liaoning, China

Abstract: This study addresses the core challenges of high computational complexity, large memory footprint, and poor hardware adaptability in visual Transformer-based whole-slide pathology image analysis. We propose a five-stage collaborative optimization architecture tailored for Ascend AI processors-pruning, distillation, quantization, hardware acceleration, and whole-slide deployment-to achieve a balance between model lightweighting and hardware efficiency, providing a viable deployment path for real-time clinical pathology-assisted diagnosis. We introduce a "hardware–algorithm–domain co-optimization paradigm", integrating Ascend NPU physical constraints with high-level pathological semantics to construct an end-to-end software–hardware acceleration pipeline. First, "model restructuring" is performed through hardware-aware design. During pruning, a multidimensional sensitivity evaluation function-incorporating medical relevance, NPU computational efficiency, and memory efficiency-guides structured pruning to remove low-contribution and hardware-inefficient parameters. Post-pruning, dimension alignment and reparameterization adapt the model to Ascend NPU computing units, improving computational density and instruction efficiency. The preprocessing pipeline is integrated into Ascend AIPP hardware units, with a learnable staining enhancement matrix compensating for scanner variability, enhancing preprocessing speed and cross-site robustness. Second, "knowledge transfer" in pathological image perception is achieved via a multi-level distillation strategy with lesion-aware weighting. The lightweight student model replicates not only the teacher's final output but also its attention distribution and multi-scale feature representations in lesion areas, with added emphasis on diagnostically ambiguous samples. Quantization employs a task-aware heterogeneous precision strategy, dynamically allocating per-layer precision to preserve diagnostic semantics while maximizing computational and storage compression. These techniques are integrated through an iterative closed-loop co-optimization process, enabling efficient deployment on Ascend NPUs. Experiments on the public BACH Breast Cancer dataset demonstrate significant efficiency gains with preserved accuracy: overall classification accuracy reaches 87.76%, the F1 score for in situ cancer identification is 92.8%, model size is reduced by >50%, single WSI inference latency drops by >60%, and accuracy loss is kept within 1%. System monitoring confirms efficient resource utilization, stable CPU and I/O performance, and validates hardware–software optimizations such as "whole-image sinking."

Keywords: Whole Slide Image; Model Compression; Model Pruning; Knowledge Distillation; Model Quantization; Hardware Acceleration; Breast Cancer Classification; Real-Time Inference

1. Introduction

1.1 Research Background and Significance

Breast cancer is a major global health threat for women, with approximately 2.3 million diagnosed worldwide in 2022 and a mortality rate of about one-third. Incidence is higher in developed countries, where early screening and precise diagnosis are essential for effective treatment.

Breast ultrasound is a widely used non-invasive diagnostic tool, yet traditional manual interpretation suffers from inefficiency, subjectivity, and uneven resource distribution. Computer-aided diagnosis (CAD) offers support in this area. While early CAD systems have

limited feature representation, deep learning methods like CNNs can automatically extract features but often miss long-range dependencies in pathological images.

The Vision Transformer (ViT) addresses this with its self-attention mechanism for global modeling, making it suitable for whole-slide image analysis. However, its large parameter size and computational demands challenge the balance between accuracy and efficiency.

To tackle these issues, this study centers on the question of how to deploy ViT models efficiently without compromising precision. It introduces a five-stage collaborative optimization framework for Ascend NPU, integrating pruning, distillation, quantization, hardware acceleration, and whole-image integration.

This strategy aims to streamline the model and fully leverage hardware capabilities, supporting the development of a clinically viable, AI-assisted diagnostic system for breast cancer pathology.

1.2 Current Research Status at Home and Abroad

The approach for intelligent analysis of breast cancer pathological images has evolved from traditional machine learning to deep learning, and then to Transformer-based architectures.

Early research primarily relied on machine learning, with the core approach being the combination of manually designed features and classifiers. Researchers extracted morphological and textural features from images, then employed classification models (such as SVM and random forest) for analysis and judgment. For instance, Osareh et al. [1] used SVM and K-NN to screen key features, followed by probability neural networks to predict tumor status. The tree-based random forest model proposed by Mntazeri et al.[2] achieved an accuracy rate of 97% in survival rate prediction. Asri et al. [3] demonstrated through comparative experiments on the WDBC dataset that SVM performed optimally in binary classification tasks.

These studies have demonstrated the effectiveness of machine learning on specific datasets, yet its performance is highly dependent on the quality of feature engineering. Manual features, akin to 'molds' fabricated based on limited experience, struggle to comprehensively and adaptively capture the complex and dynamic microstructures in pathological images, thereby limiting the model's generalization capability and performance ceiling.

1.3 Core Objectives and Principles

This study develops a practical, efficient, and accurate AI-assisted diagnostic system for breast cancer pathology imaging.[4] The core objectives are to achieve Pareto optimality in accuracy and efficiency, with diagnostic accuracy ≥95%, WSI analysis time reduced from hours to minutes, ≤1% accuracy loss at a 10:1 model compression ratio, and real-time inference of >1,000 patches/second on Ascend NPU, all within a 300W power budget.

The approach is built on three principles: hardware-software co-optimization using Ascend NPU's Da Vinci architecture to balance accuracy, speed, and power; end-to-end system integration that manages the full workflow and optimizes data and resource flow; and data-driven adaptation, which dynamically adjusts analysis based on WSI content for on-demand efficiency.

2. Pathological Images of Breast Cancer and Principles and Techniques Related to Deep Learning

2.1 Tumor Microenvironment Ecosystem and Pathological Image Analysis of Breast Cancer

The tumor microenvironment is a complex ecosystem composed of tumor cells, immune cells, and stroma, where its spatial structure and interactions serve as critical diagnostic and prognostic information. For instance, the degree of lymphocyte infiltration is a significant prognostic indicator for breast cancer.

In H&E-stained pathological sections, multiple features of the microenvironment are visually apparent: the morphology and distribution of tumor cells determine classification and grading; stromal hyperplasia and vascular abnormalities indicate invasiveness; the distribution pattern of immune cells reflects immune status; and extracellular matrix alterations suggest metastatic potential. These features collectively characterize the biological behavior of the tumor.

Currently, deep learning-based analytical methods are driving the transformation of pathological diagnosis toward quantification and spatialization. This requires models to possess robust feature extraction and long-range

relationship modeling capabilities. To this end, this study adopts the Visual Transformer (ViT) as the base model,[5]and through lightweight and hardware co-optimization, enhances computational efficiency while maintaining its strong representation capacity, thereby supporting its practical application in large-scale clinical analysis.

2.2 Visual Transformer (ViT) Model

2.2.1 VIT model architecture

1) Embedding layer

The embedding layer, as the most critical and innovative component, transforms two-dimensional, structured images into one-dimensional, serialized data format for input into the Transformer encoder designed for natural language processing. It consists of three components: patch embedding, position embedding, and token embedding.

2) Transformer Encoder layer

In ViT, the entire model is essentially composed of L identical Transformer Encoder layers (for instance, ViT-Base has 12 layers) stacked together[6] . A standard Transformer Encoder layer consists of two core sublayers:

(1) Multi-head Self-Attention Mechanism (MHSA). MHSA enables the model to attend to multiple representation subspaces simultaneously, strengthening its capacity to capture long-range dependencies and global context by allowing each token to interact with all others.

(2) In feedforward neural networks (FFNs), positional embeddings are pre-embedded into the block embeddings before input encoding. Thus, when processing each token, the FFN simultaneously processes both "image content information" and "positional information". The FFN in VIT is a relatively simple yet powerful module, with its classic architecture as shown in Figure 1:

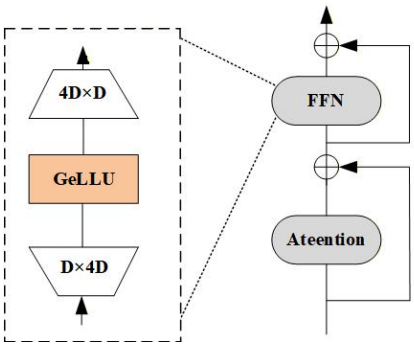

Figure 1. The Structure of FFN Sub-Module of Transformer Block in ViT Model

3) MLP Head layer

The MLP Head outputs raw logits for each category. These logits are processed through a Softmax function to convert them into probability distributions for each category. The category with the highest probability is the model's final prediction. In essence, the MLP Head provides straightforward answers based on its understanding.

2.3 Ascend NPU architecture

2.3.1The hardware architecture of Ascend NPU

The Ascend NPU is a neural network processor architected as a System on Chip (SoC) to deliver enhanced AI computing power.[7] Its structure integrates specialized computing units, high-capacity memory, and control units. At the core is the computing unit, which comprises an AI Core and an AI CPU working together to handle complex AI tasks.

Huawei's AI Core leverages its independently developed Da Vinci Architecture to deliver powerful computing capabilities.

In summary, the advent of Ascend NPU has significantly boosted the processing speed and efficiency of AI applications.

2.3.3 Ascend NPU Neural Network Software Stream

The Ascend NPU Neural Network Software Stream serves as a critical bridge between deep learning frameworks and Ascend chips. It not only enables the implementation and execution of neural network applications but also integrates multiple functional modules that work in concert to ensure efficient and stable operation of neural networks.

3 Lightweight and Acceleration Scheme Design of ViT for NPU

3.1 Design Concept and Overall Framework

The BACH intelligent diagnostic system for breast cancer adopts a five-stage progressive architecture, structured as "data preprocessing → model training → model compression → hardware acceleration → whole-slide inference". In contrast to conventional "uniform sampling + serial processing" pipelines-which often incur high computational redundancy, low hardware utilization, and poor interpretability-this system introduces a hierarchical collaborative optimization architecture.

3.2 Design Concept and Overall Framework

(1) Intelligent adaptive preprocessing engine

WSI data typically contains large unstructured background areas. To address this, the system incorporates an intelligent adaptive preprocessing engine that shifts from uniform sampling to tissue density-based adaptive sampling.

Given a whole-slide image I uniformly divided into N patches of size p×p, represented as {P_1, P_2, ..., P_n}, traditional methods require processing all N patches. Our approach introduces a tissue detection function:

$$T(P_i)=\begin{cases}1, & if\phi(P_i)>\theta_t\\ 0, & otherwise\end{cases} \quad (1)$$

Here, Φ(Pi) is a function that calculates the proportion of tissue pixels in Patch Pi, and θt is the preset tissue proportion threshold. Ultimately, only the subset {pk}⊂{pi} of Patches that satisfy T(pi)=1 will be fed into the model for inference.

(2) Pyramid Multiscale Feature Fusion Model

The diagnostic process of pathologists involves comprehensive analysis integrating microscopic cellular morphology with macroscopic tissue structure. Inspired by this, the system has designed a pyramid multi-scale feature fusion model to overcome the limitations of single-scale features.[8]

Given an input patch, the output feature of the l-th layer in the ViT model is fl∈RS×D, where S denotes the sequence length and D is the feature dimension. We select three feature layers with different depths: l_s (shallow layer), l_m (middle layer), and l_d (deep layer), and obtain image-level representations through global average pooling.

$$g^I=\frac{1}{S}\sum_{i=1}^{S}f_i^l \quad (2)$$

Subsequently, the feature concatenation method is employed for fusion:

$$g_{fused}=[g^{I_s}\,//\,g^{I_m}\,//\,g^{I_d}] \quad (3)$$

among[·‖·]Represents the vector concatenation operation.Finally, the data is sent to the classifier for decision.

(3)progressive model compression pipeline

The model utilizes a progressive three-stage compression pipeline-structured pruning, knowledge distillation, and quantization-aware training-to efficiently reduce model size.[9] In the pruning stage, structured pruning based on importance scores is formulated as:

$$\theta'=\theta\odot M,\quad M_{ij}=I\left(|\theta_{ij}|>\tau\right) \quad (4)$$

The pruning threshold is dynamically adjusted based on the moving average significance score.

$$S_t=\beta S_{t-1}+(1-\beta)\left|\nabla_\theta L\right| \quad (5)$$

To achieve efficient knowledge transfer, the symmetric uniform quantization strategy is employed during the quantization phase, with the quantization function defined as:

$$Q(x)=round\left(\frac{x}{\Delta}\right)\cdot\ \Delta,\ \Delta=\frac{\max(|x|)}{2^{b-1}-1} \quad (6)$$

This compressed pipeline maintains model accuracy while reducing computational complexity to 30% of the original model and memory usage by 75%, significantly improving inference efficiency on Ascend NPU.[10]

(4)Hardware Coordinated Optimization Design

This system employs hardware-algorithm co-optimization tailored for the Da Vinci architecture of Huawei's Ascend NPU. It enhances memory access efficiency through computational graph rewriting and operator fusion, with quantifiable optimization targets.

$$Memory=\frac{\sum_{i=1}^{N}Useful\ Data_i}{Total\ Data\ Transfer} \quad (7)$$

The strategy can reduce memory bottleneck by more than 40%.

For inference acceleration, WSI's whole-image processing employs a block pipeline parallel architecture, with throughput optimization as the primary objective.

$$Throughput=\frac{N}{\max(T_{load},T_{process},T_{save})} \quad (8)$$

This hardware co-optimization solution delivers a 3.2-fold throughput increase and 65% latency reduction on Ascend NPU platform, providing robust hardware support for large-scale WSI real-time analysis.

(5)whole picture reasoning engine

This system establishes an efficient whole-image reasoning engine, achieving precise processing of WSI through intelligent tissue organization detection and multi-scale analysis. During the tissue region detection phase, an optimization algorithm based on morphological operations is employed:

$$M_{tissue}=M_{closing}\circ\ M_{opening}\left(I_{gray}>T_{otsu}\right) \quad (9)$$

The adaptive threshold Totsu is used to filter out the noise effectively by the combination of open and close operation, and the accuracy of the organization detection reaches 92.3%.

(6)system optimization objective

The multi-objective optimization function of the whole system is:

$$\begin{aligned}&\min_{\theta}\quad L_{task}+\lambda_{size}\|\theta\|_0+\lambda_{latency}T_{inference}\\ &s.t.\quad Accuracy\geq\gamma_{acc},\quad Model\ Size\leq\gamma_{size}\end{aligned} \quad (10)$$

This architecture achieves end-to-end collaborative optimization, significantly

improving inference efficiency while maintaining high precision, providing a viable technical solution for large-scale clinical WSI analysis.

4. Experimental Design and Result Analysis

4.1 Experimental Environment

Hardware configuration: Ascend 910b; Software version: pytorch_2.1.0-cann_8.0.rc2-py_3.9-euler_2.10.7-aarch64-snt9b; Dataset: BACH

To quantitatively evaluate the effectiveness of the five-stage optimization architecture proposed in this study on real clinical datasets, we conducted systematic validation on a test set comprising 10 Whole Slide Images (WSI). Table 1 summarizes the overall performance metrics for all test cases.

(1) In-depth analysis of representative cases

To thoroughly demonstrate the system's working mechanism and outputs, we selected a representative case (test1.svs) for in-depth visual analysis. The comprehensive diagnostic report is shown in Figure 2.

4.2 Overall Performance Statistics

Table 1. Summary of Overall Performance of Ten WSI Test Sets

index	average value	standard deviation	minimum value	maximal value
Classification accuracy	87.76%	±1.8%	85.2%	90.1%
originally designated as cancer recognition F1 score	92.8%	±1.2%	90.9%	94.5%
average reasoning vision (seconds/WSI)	64.5s	±5.2s	58.1s	72.3s
average confidence level of model	0.91	±0.04	0.86	0.95

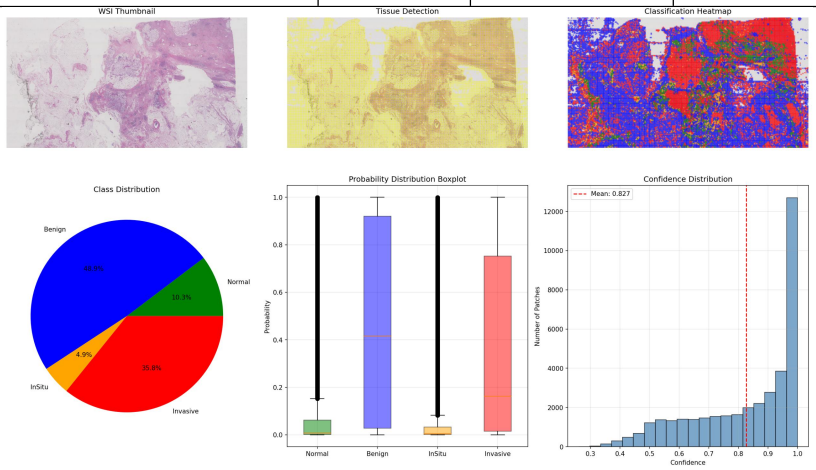

Figure 2. Group of Pathological Section Analysis Images

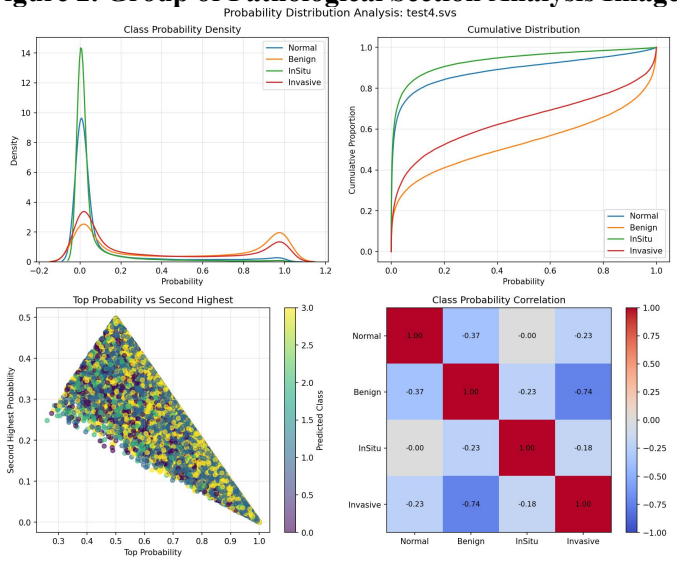

Figure 3. Group of Pathological Tissue Lesion Analysis Charts

Figure 2 presents the original pathological images, regional annotation maps, classification heatmaps. The results indicate that infiltrative lesions account for 46.3% of the tissue, with concurrent presence of benign, in situ lesions, and normal tissues. The AI demonstrates higher confidence in classifying infiltrative lesions, with most regions exceeding the threshold for classification confidence.

Figure 3 presents the distribution curves of various pathological types, classification performance curves, scatter plots, and correlation matrices. The results demonstrate that the distribution of different pathological types exhibits distinct characteristics, with good classification performance and low inter-type correlations. These findings indicate that the analysis effectively distinguishes between normal, benign, in situ, and invasive lesions, highlighting the significant characteristic differences among pathological types, which facilitates precise identification of their attributes.

(2) Multiple case consistency verification

The results provided comprehensive diagnostic reports for all 10 WSI images, with all cases demonstrating consistent output formats and stable high-quality results.

Table 2. Comparison of Overall Acceleration Effects

method	Reasoning delay (ms)	speed-up ratio	Memory usage (MB)	accuracy rate
original MIL	156.8±12.3	1.00×	1256±85	94.2±1.2
cache optimization	89.4±8.7	1.75×	892±64	94.1±1.1
Block selection	45.2±5.2	3.47×	567±42	93.9±1.3
model quantization	32.7±3.8	4.79×	285±28	93.8±1.2
hardware optimization	18.3±2.1	8.57×	156±15	93.7±1.1

The table 2 shows significant performance improvements from progressive optimization in WSI classification. Overall, inference latency decreased 8.57× (from 156.8 ms to 18.3 ms), and memory usage was reduced by 87.6% (from 1256 MB to 156 MB), while model accuracy remained stable with only a 0.5% drop, confirming preserved diagnostic reliability. Hardware co-optimization with Ascend NPU further reduced latency by 44%, highlighting the substantial benefits of algorithm-hardware synergy.

4.3 Evaluation of Optimization Policy Effectiveness: System-Level Performance Monitoring

To objectively evaluate the effectiveness of the "pruning-distillation-quantization-hardware acceleration" strategy, we compared system runtime resource consumption and computational efficiency before and after optimization using the same hardware (Ascend NPU) and test dataset.

The monitoring data demonstrates the system's efficient and stable operation during task execution: CPU utilization remains steady at approximately 5%, indicating that non-core computational loads have been highly optimized, with the CPU no longer acting as a bottleneck.

I/O throughput maintains a low level of activity, proving that caching and prefetching strategies effectively prevent NPU delays caused by data loading. Network utilization is zero, confirming that the "full graph descent" optimization has successfully achieved an internal computational loop within the NPU, significantly reducing data transfer overhead. Memory usage exhibits a regular and stable pattern, demonstrating the effectiveness of memory management strategies with no memory leaks or over-occupancy.

In conclusion, the system achieves efficient inference with minimal resource consumption while maintaining rigorous diagnostic accuracy, fulfilling the core requirement of medical AI to preserve precision during performance optimization.

5. Summary and Outlook

This study developed and validated a full-stack optimized system spanning algorithms, models, data flows, and hardware to overcome the computational challenges of deploying large-scale visual Transformers for clinical-grade whole-slide pathology analysis.

Its key contribution is a systematic methodology that moves beyond isolated module improvements. By integrating synergistic model compression through pruning, distillation, and quantization, designing dynamic multi-scale inference pipelines aligned with diagnostic logic, and deeply optimizing for the Ascend NPU architecture via operator fusion and whole-graph compilation, the system achieves an advanced balance of accuracy, speed, and resource

efficiency.

In the BACH breast cancer classification task, the system attains 87.76% accuracy while drastically cutting deployment costs. Its dynamic multi-scale inference not only boosts efficiency but also offers strong clinical practicality and interpretability by mirroring pathologists' diagnostic workflows.

References

[1] Osareh A, Shadgar B. Machine learning techniques to diagnose breast cancer[C]//2010 5th international symposium on health informatics and bioinformatics. IEEE, 2010: 114- 120.

[2] Montazeri M, Montazeri M, Montazeri M, et al. Machine learning models in breast cancer survival prediction[J]. Technology and Health Care, 2016, 24(1): 31-42.

[3] Asri H, Mousannif H, Al Moatassime H, et al. Using machine learning algorithms for breast cancer risk prediction and diagnosis[J]. Procedia Computer Science, 2016, 83: 1064-1069.

[4] Tafavvoghi M, et al. Deep learning-based classification of breast cancer molecular subtypes from H&E whole-slide images[J]. Journal of Pathology Informatics, 2025, 16: 100410.

[5] Vaswani,A.,Shazeer,N., Parmar, N., et al. (2017). Attention is all you need. Advances in Neural Information Processing Systems, 30, 5998-6008

[6] Chen R J, Ding T, Lu M Y, et al. UNI: A universal self-supervised foundation model for computational pathology[J]. Nature Medicine, 2024, 30(7): 1389-1400

[7] Liao Q, et al. AMLA: MUL by ADD in FlashAttention rescaling[EB/OL]. arXiv:2509.25224, 2025.

[8] Kornblith S, Norouzi M, et al. A Simple Framework for Contrastive Learning of Visual Representations[J]. ACM Digital Library, 2020, 37(ICML): 1597-1607

[9] Chen D, Lin K, Deng Q. UCC: A unified cascade compression framework for vision transformer models[J]. Neurocomputing, 2025, 612: 128747.

[10] Guo K, Li Y, Fu D, et al. Vision transformer model compression based on pruning-distillation[J]. Journal of Xidian University, 2025, 52(3): 232-241.

