Embedded Technoloy and Intelligent Systems 嵌入式技术与智能系统, 2025, 2(4), 255-260
Published Online August 2025 in Hans. https://www.hanspub.org/journal/etis
https://doi.org/10.12677/etis.2025.24023

异构智能嵌入式系统AI模型推理与部署优化

——从模型轻量化到系统级加速的综述与展望

庞为光

齐鲁工业大学(山东省科学院)，山东省计算中心(国家超级计算济南中心)，山东 济南

收稿日期：2025年10月28日；录用日期：2025年11月20日；发布日期：2025年12月1日

摘 要

随着人工智能技术与嵌入式硬件的快速发展，嵌入式人工智能系统(如移动机器人、自动驾驶汽车和星载无人机)在工业自动化、交通运输和航空航天等关键领域变得越来越重要。作为集成CPU、GPU、NPU等多种异构处理器单元的智能实时系统，其核心任务是通过计算密集型的深度神经网络(DNN)实现环境感知、决策控制等复杂功能，同时面临严格的时间约束与资源瓶颈。文章从网络模型在嵌入式系统加速推理优化的角度，将围绕DNN模型轻量化、推理加速优化与动态任务调度三个方面，详细分析嵌入式智能系统的国内外研究现状。

关键词

异构嵌入式系统，深度神经网络，推理加速，实时调度

AI Model Inference and Deployment Optimization for Heterogeneous Intelligent Embedded Systems

—A Survey and Perspective from Model Lightweighting to System-Level Acceleration

Weiguang Pang

Shandong Computer Science Center (National Supercomputing Center in Jinan), Qilu University of Technology (Shandong Academy of Sciences), Jinan Shandong

Received: October 28, 2025; accepted: November 20, 2025; published: December 1, 2025

文章引用: 庞为光. 异构智能嵌入式系统 AI 模型推理与部署优化[J]. 嵌入式技术与智能系统, 2025, 2(4): 255-260.
DOI: 10.12677/etis.2025.24023

Abstract

With the rapid advancement of artificial intelligence (AI) technologies and embedded hardware, embedded AI systems—such as mobile robots, autonomous vehicles, and spaceborne unmanned aerial vehicles—are becoming increasingly significant in key domains including industrial automation, transportation, and aerospace. As intelligent real-time systems integrating heterogeneous processing units such as CPUs, GPUs, and NPUs, their core mission is to execute computationally intensive deep neural networks (DNNs) to achieve complex functions such as environmental perception and decision control, all under stringent temporal constraints and resource limitations. From the perspective of accelerating and optimizing neural network inference on embedded systems, this paper provides a comprehensive analysis of the current research progress at home and abroad, focusing on three major aspects: DNN model lightweighting, inference acceleration optimization, and dynamic task scheduling.

Keywords

Heterogeneous Embedded Systems, Deep Neural Networks, Inference Acceleration, Real-Time Scheduling

1. 引言

深度神经网络模型轻量化技术是突破人工智能应用在嵌入式系统部署瓶颈的关键路径，其通过算法重构与参数压缩的双重优化，在可接受网络精度损失范围内构建高效推理模型。在算法层面，知识蒸馏技术实现复杂模型向轻量化架构的能力迁移，配合模块化网络设计降低结构冗余；参数剪枝(结构化/非结构化)、量化(二值化/混合精度)及低秩分解等方法系统性地减少模型计算量[1]。硬件适配层面则通过稀疏矩阵加速器、多分支网络架构等定制化设计，提升轻量化模型在嵌入式异构平台的能效表现，形成算法–硬件协同优化方法[2]。

面向大语言模型的嵌入式部署需求，轻量化技术呈现细粒度创新趋势：一方面，根据量化所应用的不同阶段，可以将量化方法分为三类：量化感知训练(QAT, Quantization-Aware Training)、量化感知微调(QAF, Quantization-Aware Fine-tuning)及训练后量化(PTQ, Post-Training Quantization) [3]。QAT 在模型的训练过程中采用量化，QAF 在预训练模型的微调阶段应用量化，PTQ 在模型完成训练后对其进行量化，并结合硬件特性开发出极限低比特的整型(如 INT4、INT8)压缩方案；另一方面，混合专家模型等异构架构革新了模型部署范式，通过大小模型动态协作实现推理效率的阶跃式提升[4]。此类技术使百亿参数级模型在嵌入式设备端的实时推理成为可能，推动嵌入式系统向智能认知层级跨越。

2. 嵌入式智能系统推理优化加速技术发展现状

当前嵌入式智能系统的网络模型部署主要集中在推理加速优化，其技术策略在保持模型精度的前提下提升运行效率。核心优化方向包括网络模型编译优化、异构资源调度以及存储计算优化。英伟达的 TensorRT 推理框架通过算子融合与内存优化技术有效提升了推理速度[5]。关于网络模型推理任务在异构计算单元上的分配方法，当前研究工作采用模型并行、数据并行和流水线并行等模型的推理加速方法，

进一步提升了嵌入式系统上的模型推理性能[6]。

在大语言模型在嵌入式系统上优化部署方面，伊利诺伊大学针对大模型输出长度不确定导致的端到端推理时间不可预测问题，提出了一种推测性最短作业优先调度器。该方案利用轻量级代理模型预测大模型输出序列长度，有效解决了传统先到先服务调度的队首阻塞问题[7]。英伟达开发了动态内存压缩技术，通过在推理过程中在线压缩键值缓存，成功缓解因输入序列长度与批处理规模线性增长引发的缓存膨胀问题[8]。首尔大学提出的细粒度调度机制实现了迭代级连续批处理，可通过动态整合多个大模型请求显著提升推理效率[9]。针对 Transformer 架构的计算特性，学界提出了 KV 缓存复用、FlashAttention 以及 PageAttention 等加速方法[10]，并结合投机采样与混合专家模型技术，在保证模型精度的前提下实现推理效率突破。

国内研究团队在模型推理加速领域取得显著进展。北京邮电大学在片上神经处理单元实现高效设备端大模型预填充加速的系统，该系统通过在提示供工程、张量和模型三个层次上优化了大模型在端侧设备上的推理，从而显著减少了推理延迟[11]。东北大学在边端系统推理加速方面积累了较多的系统部署优化基础，其中 GPU 并行加速方面研究了 GPU 内部异构计算核心的并行策略，提升了系统整理利用率和任务吞吐量[12]。国内人工智能团队 DeepSeek 通过创新的多头隐式注意力(MLA)设计，突破了现有优化方案的瓶颈，使得模型在存储和计算效率上达到了新的高度[13]。

3. 动态智能任务实时调度方法发展现状

动态神经网络通过运行时自适应调整模型结构或参数，成为实时系统应对计算资源约束的关键技术。其核心优势在于能够根据输入特征(如图像尺寸、批处理规模)及系统约束(如截止期限、资源限制)，如图 1 所示，动态神经网络通过灵活调整网络压缩率、分支路径或输出节点，实现负载的动态适配[14]-[16]。例如，通过动态调节输入图像分辨率或网络分支选择，模型可在保证模型精度的同时显著降低推理延迟，满足工业物联网、自动驾驶等场景的实时性需求。

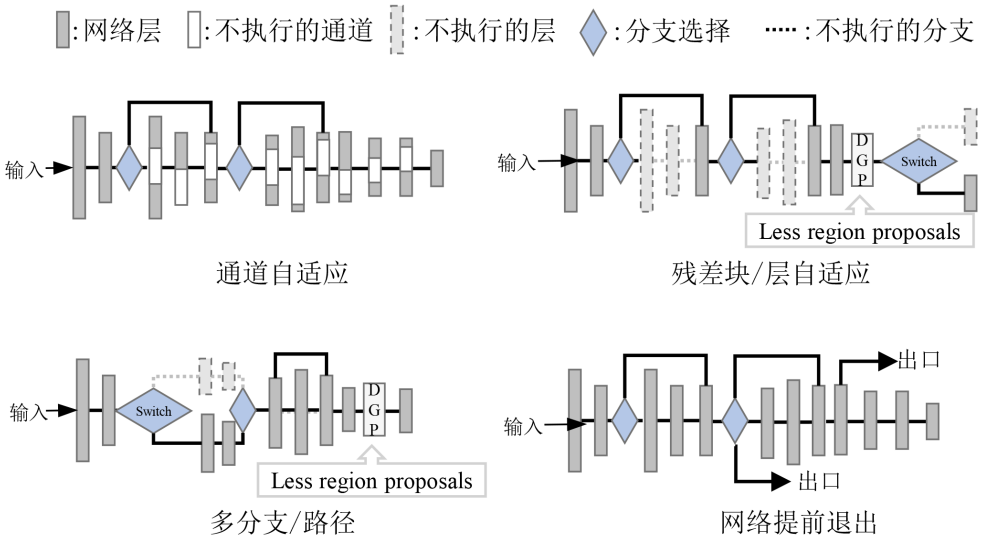

Figure 1. Dynamic neural network-based workload adjustment method
图 1. 动态神经网络调节计算负载方法

在动态推理 DNN 任务方面，学术界提出了多维度的系统调度方法。美国得克萨斯大学所提出的近似网络，量化了计算负载缩减与精度/时延的关联模型，支持运行时动态负载调整[17] [18]。韩国庆熙大学研

究者结合 GPU 最坏执行时间分析与自适应图像缩放技术，设计了动态路径切换机制，在任务截止期约束下将精度损失降至最低[19] [20]。工业界则聚焦轻量化动态架构创新，如三星公司提出的分支条件神经网络(BPNet)实现了系统化的时间与精度权衡[21]。苹果公司开发的 UPSCALE 通道剪枝策略通过权重重排序技术，实现了无显著时延代价的动态网络裁剪[22]。微软提出基于全局的大批量 LLM 推理优化前缀共享和面向吞吐量的令牌批处理方法，通过全局前缀识别与请求调度重组、内存中心的分批处理及水平融合注意力核优化，实现共享前缀的 KV 上下文高效复用、预填充与解码阶段的 GPU 负载均衡，显著提升工业场景下大批量 LLM 推理效率[23]。北卡罗莱纳大学提出的 SubFlow 框架从模型结构层面出发，利用动态诱导子图策略在运行时根据任务截止期自适应选择子网络路径，实现了可变时间预算下的低时延高精度推理，为网络任务动态推理提供了新思路[24]。韩国汉阳大学提出的 Exegpt 系统则从系统层面出发，引入约束感知资源调度机制，通过联合优化批量大小与 GPU 分配，在延迟约束下实现高吞吐并发推理，体现了动态推理在资源调度与 QoS 保障方面的潜力[25]。

国内学者在动态自适应负载建模与部署优化方面取得显著进展。清华大学团队系统阐述了动态神经网络的理论框架[16]。上海交通大学通过扩展深度学习编译器实现了动态网络的高效推理支持[26]。上海科技大学进一步提出带时间约束的自适应任务模型，构建了兼顾服务质量与实时性的调度优化框架[1]。西北工业大学则聚焦环境自适应技术，通过动态调整模型参数降低资源消耗，为智能物联网系统提供高效解决方案[27]。香港中文大学利用深度学习编译技术在 GPU 上实现多 DNN 推理任务调度，在不损失网络精度的情况下，通过神经网络图和内核优化，提高 GPU 并行性，减少多任务之间的资源争用[28]。东北大学在异构 CPU-GPU 平台上的多 DNN 调度方面[29]，采用有效的 CUDA 流优先级管理方法实现了不同优先级多 DNN 任务在共享 GPU 上的实时调度策略。

4. 发展趋势与展望

随着大模型逐步渗透至边缘端，主流技术的发展推动了模型轻量化和压缩技术的突破。通过模型压缩、量化和知识蒸馏等手段，使得模型在资源受限的嵌入式设备(如手机和机器人)上实现高效推理和实时响应，同时配合实时调度技术，确保动态任务处理能力。2025 年被视为“具身智能元年”，嵌入式系统借助轻量化和压缩技术，助力人形机器人在工业、医疗、家庭和自动驾驶等场景中完成复杂操作与实时决策，体现了主流技术在物理交互领域的应用优势和调度能力。原生多模态大模型整合视觉、音频、文本及 3D 数据，通过端到端训练实现数据对齐，并借助低功耗 AI 芯片和边缘计算平台降低推理延迟。此过程中，模型轻量化与实时调度技术是实现综合感知与实时处理的关键支撑。未来嵌入式智能系统将向垂直领域定制化发展，例如医疗诊断、农业机器人和消费电子。主流技术的发展促使模型更轻量、压缩更高效，同时借助实时调度实现自主智能体的动态任务管理，推动“All-in-One”超级应用的崛起，实现多场景智能服务。总之，嵌入式智能系统的发展正依托主流技术的模型轻量化、压缩技术及实时调度能力，实现高效推理、多模态融合和精细化物理交互。未来，这些技术将在垂类应用与自主智能体领域发挥核心作用。

基金项目

本文受山东省自然科学基金资助项目 ZR2024QF052。

参考文献

[1] Wang, W., Chen, W., Luo, Y., Long, Y., Lin, Z., Zhang, L., *et al.* (2024) Model Compression and Efficient Inference for Large Language Models: A Survey. arXiv: 2402.09748.

[2] Liu, D., Kong, H., Luo, X., Liu, W. and Subramaniam, R. (2022) Bringing AI to Edge: From Deep Learning’s Perspective.

Neurocomputing, **485**, 297-320. https://doi.org/10.1016/j.neucom.2021.04.141

[3] Zhou, Z., Ning, X., Hong, K., *et al.* (2024) A Survey on Efficient Inference for Large Language Models. https://arxiv.org/abs/2404.14294

[4] Dai, D., Deng, C., Zhao, C., Xu, R.X., Gao, H., Chen, D., *et al.* (2024) DeepSeekMoE: Towards Ultimate Expert Specialization in Mixture-Of-Experts Language Models. *Proceedings of the* 62*nd Annual Meeting of the Association for Computational Linguistics* (*Volume* 1: *Long Papers*), Bangkok, 11-16 August 2024, 1280-1297. https://doi.org/10.18653/v1/2024.acl-long.70

[5] NVIDIA. (2024). TensorRT-LLM [Computer Software]. GitHub. https://github.com/NVIDIA/TensorRT-LLM

[6] Ascend. (2024). AscendSpeed [Computer Software]. GitHub. https://github.com/Ascend/AscendSpeed

[7] Qiu, H., Mao, W., Patke, A., *et al.* (2024) Efficient Interactive LLM Serving with Proxy Model-Based Sequence Length Prediction. arXiv: 2404.08509.

[8] Nawrot, P., Łańcucki, A., Chochowski, M., *et al.* (2024) Dynamic Memory Compression: Retrofitting LLMs for Accelerated Inference. arXiv: 2403.09636.

[9] Yu, G.I., Jeong, J.S., Kim, G.W., *et al.* (2022) Orca: A Distributed Serving System for {Transformer-Based} Generative Models. 16*th USENIX Symposium on Operating Systems Design and Implementation* (*OSDI* 22), 521-538.

[10] Kwon, W., Li, Z., Zhuang, S., Sheng, Y., Zheng, L., Yu, C.H., *et al.* (2023) Efficient Memory Management for Large Language Model Serving with PagedAttention. *Proceedings of the* 29*th Symposium on Operating Systems Principles*, Koblenz, 23-26 October 2023, 611-626. https://doi.org/10.1145/3600006.3613165

[11] Xu, D., Zhang, H., Yang, L., *et al.* (2024) Empowering 1000 Tokens/Second On-Device LLM Prefilling with MLLM-NPU. arXiv: 2407.05858v1.

[12] Pang, W., Jiang, X., Liu, S., Qiao, L., Fu, K., Gao, L., *et al.* (2024) Control Flow Divergence Optimization by Exploiting Tensor Cores. *Proceedings of the* 61*st ACM/IEEE Design Automation Conference*, San Francisco, 23-27 June 2024, 1-6. https://doi.org/10.1145/3649329.3658462

[13] Meng, F., Yao, Z. and Zhang, M. (2025) TransMLA: Multi-Head Latent Attention Is All You Need. arXiv: 2502.07864.

[14] 王子曦, 邵培南, 邓畅. 异构并行平台的 Caffe 推理速度提升方法[J]. 计算机系统应用, 2022, 31(2): 220-226.

[15] 尚绍法, 蒋林, 李远成, 等. 异构平台下卷积神经网络推理模型自适应划分和调度方法[J]. 计算机应用, 2023, 43(9): 2828-2835.

[16] Han, Y., Huang, G., Song, S., Yang, L., Wang, H. and Wang, Y. (2022) Dynamic Neural Networks: A Survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **44**, 7436-7456. https://doi.org/10.1109/tpami.2021.3117837

[17] Bo, Z., Guo, C., Leng, C., Qiao, Y. and Wang, H. (2024) RTDeepEnsemble: Real-Time DNN Ensemble Method for Machine Perception Systems. 2024 *IEEE* 42*nd International Conference on Computer Design* (*ICCD*), Milan, 18-20 November 2024, 191-198. https://doi.org/10.1109/iccd63220.2024.00037

[18] Han, Y., Liu, Z., Yuan, Z., Pu, Y., Wang, C., Song, S., *et al.* (2024) Latency-Aware Unified Dynamic Networks for Efficient Image Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **46**, 7760-7774. https://doi.org/10.1109/tpami.2024.3393530

[19] Heo, S., Jeong, S. and Kim, H. (2022) RTScale: Sensitivity-Aware Adaptive Image Scaling for Real-Time Object Detection. 34th *Euro-Micro Conference on Real-Time Systems*, Modena, 5-8 July 2022, 1-22.

[20] Heo, S., Cho, S., Kim, Y. and Kim, H. (2020) Real-Time Object Detection System with Multi-Path Neural Networks. 2020 *IEEE Real-Time and Embedded Technology and Applications Symposium* (*RTAS*), Sydney, 21-24 April 2020, 174-187. https://doi.org/10.1109/rtas48715.2020.000-8

[21] Park, K., Oh, C. and Yi, Y. (2020) BPNet: Branch-Pruned Conditional Neural Network for Systematic Time-Accuracy Tradeoff. 2020 57*th ACM/IEEE Design Automation Conference* (*DAC*), San Francisco, 20-24 July 2020, 1-6. https://doi.org/10.1109/dac18072.2020.9218545

[22] Wan, A., Hao, H., Patnaik, K., *et al.* (2023) UPSCALE: Unconstrained Channel Pruning. arXiv: 2307.08771.

[23] Zheng, Z., Ji, X., Fang, T., Zhou, F., Liu, C. and Peng, G. (2024) BatchLLM: Optimizing Large Batched LLM Inference with Global Prefix Sharing and Throughput-Oriented Token Batching. arXiv: 2412.03594.

[24] Lee, S. and Nirjon, S. (2020) SubFlow: A Dynamic Induced-Subgraph Strategy toward Real-Time DNN Inference and Training. 2020 *IEEE Real-Time and Embedded Technology and Applications Symposium* (*RTAS*), Sydney, 21-24 April 2020, 15-29. https://doi.org/10.1109/rtas48715.2020.00-20

[25] Oh, H., Kim, K., Kim, J., Kim, S., Lee, J., Chang, D., *et al.* (2024) ExeGPT: Constraint-Aware Resource Scheduling for LLM Inference. *Proceedings of the* 29*th ACM International Conference on Architectural Support for Programming Languages and Operating Systems*, *Volume* 2, La Jolla, 27 April-1 May 2024, 369-384. https://doi.org/10.1145/3620665.3640383

[26] Cui, W., Han, Z., Ouyang, L., *et al.* (2023) Optimizing Dynamic Neural Networks with Brainstorm. 17*th USENIX Symposium on Operating Systems Design and Implementation* (*OSDI* 23), Boston,10-12 July 2023, 797-815.

[27] Wang, H., Zhou, X., Yu, Z., Liu, S., Guo, B., Wu, Y., *et al*. (2020) Context-aware Adaptation of Deep Learning Models for IoT Devices. *Scientia Sinica Informationis*, **50**, 1629-1644. https://doi.org/10.1360/ssi-2020-0067

[28] Zhao, Z., Ling, N., Guan, N. and Xing, G. (2022) Aaron: Compile-Time Kernel Adaptation for Multi-DNN Inference Acceleration on Edge GPU. *Proceedings of the* 20*th ACM Conference on Embedded Networked Sensor Systems*, Boston, 6-9 November 2022, 802-803. https://doi.org/10.1145/3560905.3568050

[29] Pang, W., Luo, X., Chen, K., Ji, D., Qiao, L. and Yi, W. (2023) Efficient CUDA Stream Management for Multi-DNN Real-Time Inference on Embedded GPUs. *Journal of Systems Architecture*, **139**, Article ID: 102888. https://doi.org/10.1016/j.sysarc.2023.102888