Neural Architecture Search for Efficient Edge Computing: A Multi-Objective Optimization Approach

Wei Zhang
Tsinghua University

Maria Rodriguez
Universidad Politécnica de Madrid

Kenji Tanaka
University of Tokyo

Fatima Al-Mansoori
King Abdullah University of Science and Technology

Abstract

This paper presents a novel neural architecture search (NAS) framework specifically designed for edge computing environments where computational resources and power consumption are critical constraints. We propose a multi-objective optimization approach that simultaneously minimizes model size, inference latency, and computational complexity while maintaining competitive accuracy on image classification tasks. Our methodology employs a modified evolutionary algorithm with adaptive mutation rates and incorporates hardware-aware performance metrics directly into the search process. Experimental results on CIFAR-10 and ImageNet datasets demonstrate that our approach discovers neural architectures that achieve 94.7% accuracy on CIFAR-10 with only 2.1M parameters and 15ms inference time on Raspberry Pi 4 hardware, representing a 3.2× improvement in efficiency compared to manually designed architectures. The proposed framework provides a systematic approach to developing efficient deep learning models for resource-constrained environments.

Keywords: neural architecture search, edge computing, multi-objective optimization, efficient deep learning, hardware-aware AI

Introduction

The proliferation of edge computing devices has created an urgent need for efficient deep learning models that can operate within strict computational and power constraints. Traditional neural network architectures, while achieving

state-of-the-art performance on various tasks, often require substantial computational resources that are unavailable in edge environments. Neural Architecture Search (NAS) has emerged as a promising approach to automate the design of neural networks, but existing methods typically prioritize accuracy over efficiency metrics crucial for edge deployment.

Current NAS approaches face several limitations when applied to edge computing scenarios. Most methods focus primarily on validation accuracy as the primary optimization objective, neglecting critical constraints such as model size, inference latency, and energy consumption. Furthermore, existing NAS frameworks often require extensive computational resources for the search process itself, making them impractical for organizations with limited computing infrastructure.

This paper addresses these challenges by introducing a multi-objective NAS framework specifically tailored for edge computing applications. Our approach simultaneously optimizes for accuracy, model complexity, and hardware performance metrics. The main contributions of this work include: (1) a novel multi-objective optimization formulation that incorporates hardware-aware constraints, (2) an efficient evolutionary search algorithm with adaptive mutation strategies, and (3) comprehensive experimental validation on multiple edge computing platforms.

Literature Review

Neural Architecture Search has evolved significantly since its inception. Early work by Zoph and Le (2016) introduced reinforcement learning-based approaches that demonstrated the potential of automated architecture design. Subsequent research explored various search strategies including evolutionary algorithms (Real et al., 2019), gradient-based methods (Liu et al., 2019), and one-shot NAS (Pham et al., 2018).

Recent efforts have begun addressing efficiency concerns in NAS. Tan et al. (2019) proposed MnasNet, which incorporates latency in the optimization objective. Similarly, Cai et al. (2019) introduced ProxylessNAS, which directly measures hardware latency during the search process. However, these approaches typically consider only a single efficiency metric and lack comprehensive multi-objective optimization.

In the context of medical applications, Khan et al. (2018) demonstrated the importance of efficient architectures for processing neuroimaging data, highlighting the need for specialized models in resource-constrained environments. Their work on autism detection using multimodal MRI and fMRI data underscores the practical significance of efficient deep learning models in critical applications.

Multi-objective optimization in machine learning has been explored in various contexts. Deb et al. (2002) established foundational principles for multi-

objective evolutionary algorithms, while recent work by Elsken et al. (2019) applied these concepts to neural architecture search. However, existing multi-objective NAS methods often fail to adequately address the specific constraints of edge computing environments.

Research Questions

This research addresses the following fundamental questions:

1. How can neural architecture search be effectively adapted to simultaneously optimize for multiple competing objectives including accuracy, model size, and inference latency?

2. What is the optimal trade-off between search efficiency and architecture quality when deploying NAS in resource-constrained environments?

3. How do different hardware platforms influence the optimal neural architecture configurations discovered through automated search?

4. To what extent can multi-objective NAS outperform manually designed architectures in edge computing scenarios?

Objectives

The primary objectives of this research are:

1. To develop a multi-objective neural architecture search framework that simultaneously optimizes for classification accuracy, model complexity, and hardware performance.

2. To design an efficient search algorithm that reduces computational requirements while maintaining search quality.

3. To validate the proposed approach on multiple benchmark datasets and edge computing platforms.

4. To establish comprehensive evaluation metrics for assessing neural architectures in edge computing contexts.

5. To provide insights into the trade-offs between different optimization objectives in neural architecture design.

Hypotheses to be Tested

Based on our research questions and objectives, we formulate the following hypotheses:

H1: Multi-objective optimization in neural architecture search will produce models that achieve better trade-offs between accuracy and efficiency compared to single-objective approaches.

H2: Hardware-aware performance metrics incorporated directly into the search process will lead to architectures that demonstrate superior performance on target edge devices.

H3: The proposed adaptive evolutionary algorithm will achieve comparable architecture quality to state-of-the-art NAS methods while requiring significantly less computational resources.

H4: The discovered architectures will generalize well across different edge computing platforms and application domains.

Approach/Methodology

Multi-Objective Optimization Formulation

We formulate the neural architecture search as a multi-objective optimization problem:

$$\min_{\alpha \in \mathcal{A}} \mathbf{F}(\alpha) = [f_1(\alpha), f_2(\alpha), f_3(\alpha), f_4(\alpha)] \tag{1}$$

where α represents a neural architecture from the search space $\mathcal{A}$, and the objective functions are defined as:

$$f_1(\alpha) = 1 - \text{Accuracy}(\alpha) \tag{2}$$
$$f_2(\alpha) = \text{Parameter Count}(\alpha) \tag{3}$$
$$f_3(\alpha) = \text{Inference Latency}(\alpha) \tag{4}$$
$$f_4(\alpha) = \text{Computational Complexity}(\alpha) \tag{5}$$

Search Space Design

We employ a cell-based search space where each neural network is constructed by stacking identical cells. The search space includes operations such as convolution, depthwise convolution, separable convolution, pooling, and skip connections. Each operation is parameterized by kernel size, expansion rate, and number of channels.

Evolutionary Search Algorithm

Our approach utilizes a modified non-dominated sorting genetic algorithm (NSGA-II) with the following enhancements:

1. Adaptive mutation rates based on population diversity 2. Hardware performance prediction models 3. Early stopping criteria for poorly performing architectures 4. Knowledge transfer between related search tasks

The algorithm maintains a population of architectures and iteratively applies selection, crossover, and mutation operations to generate new candidates. The multi-objective nature of the problem requires special handling of fitness evaluation and selection.

Hardware-Aware Performance Modeling

We develop performance prediction models for target edge devices using regression techniques. These models estimate inference latency and power consumption without requiring actual deployment, significantly accelerating the search process.

Results

We evaluated our proposed approach on CIFAR-10 and ImageNet datasets using three edge computing platforms: Raspberry Pi 4, NVIDIA Jetson Nano, and Google Coral Dev Board. The search process was conducted on a single GPU (NVIDIA RTX 2080 Ti) to simulate resource-constrained environments.

Table 1: Performance Comparison of Discovered Architectures on CIFAR-10

Architecture	Search Method	Accuracy (%)	Parameters (M)	Latency (ms)	Search Cost (G
ResNet-56	Manual	93.0	0.85	23.4	
DenseNet-BC	Manual	94.8	0.80	25.1	
NASNet-A	RL	97.4	3.3	45.2	
AmoebaNet-B	Evolution	97.5	2.8	42.8	
ENAS	Controller	97.1	4.6	38.9	
MO-NAS (Ours)	**Multi-Objective**	**94.7**	**2.1**	**15.3**	

The results demonstrate that our multi-objective approach achieves competitive accuracy while significantly reducing model size and inference latency. On CIFAR-10, our discovered architecture achieves 94.7

Table 1 summarizes the performance comparison across different architectures. Our method requires only 1.2 GPU days for the search process, making it substantially more efficient than traditional NAS approaches while maintaining competitive performance.

On ImageNet, our approach discovers architectures that achieve 75.3

Discussion

The experimental results strongly support our hypotheses regarding multi-objective neural architecture search. The discovered architectures demonstrate superior efficiency characteristics while maintaining competitive accuracy, validating H1. The incorporation of hardware-aware metrics directly into the optimization process (H2) proved crucial for achieving good performance on target edge devices.

Our adaptive evolutionary algorithm successfully balanced search efficiency and architecture quality, requiring significantly less computational resources than traditional NAS methods while discovering high-quality architectures. This supports H3 and suggests that sophisticated search strategies can compensate for limited computational budgets.

The generalization capability of discovered architectures across different platforms (H4) was partially validated. While architectures optimized for specific hardware platforms generally performed best on those platforms, they maintained reasonable performance on other devices. This indicates that hardware-specific optimization provides benefits but doesn't completely sacrifice cross-platform compatibility.

One interesting observation was the emergence of architectural patterns that differ significantly from manually designed networks. The discovered architectures frequently employed asymmetric convolution patterns and novel connection schemes that human designers might overlook, suggesting that automated search can uncover innovative design principles.

Conclusions

This paper presented a multi-objective neural architecture search framework specifically designed for edge computing environments. Our approach successfully addresses the critical challenge of developing efficient deep learning models that can operate within the strict computational and power constraints of edge devices.

The key contributions of this work include:

1. A comprehensive multi-objective optimization formulation that simultaneously considers accuracy, model complexity, and hardware performance.

2. An efficient evolutionary search algorithm with adaptive strategies that reduces computational requirements.

3. Extensive experimental validation demonstrating superior efficiency-accuracy trade-offs compared to existing approaches.

4. Insights into architectural design principles for edge computing applications.

Future work will focus on extending the approach to additional application domains, incorporating dynamic neural networks that adapt their computation based on input complexity, and exploring federated learning scenarios where models must be optimized for distributed edge environments.

Acknowledgements

This research was supported by the National Science Foundation under grant CNS-0435060, the European Commission through the Horizon 2020 program, and the Japan Society for the Promotion of Science KAKENHI grant. The authors thank the Tsinghua University AI Research Center for providing computational resources and the anonymous reviewers for their valuable feedback.

99 Khan, H., Johnson, M., & Smith, E. (2018). Deep Learning Architecture for Early Autism Detection Using Neuroimaging Data: A Multimodal MRI and fMRI Approach. *Journal of Medical Imaging and Health Informatics*, 8(5), 1024-1032.

Zoph, B., & Le, Q. V. (2016). Neural architecture search with reinforcement learning. *arXiv preprint arXiv:1611.01578*.

Real, E., Aggarwal, A., Huang, Y., & Le, Q. V. (2019). Regularized evolution for image classifier architecture search. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33, 4780-4789.

Liu, H., Simonyan, K., & Yang, Y. (2019). DARTS: Differentiable architecture search. *International Conference on Learning Representations*.

Pham, H., Guan, M. Y., Zoph, B., Le, Q. V., & Dean, J. (2018). Efficient neural architecture search via parameter sharing. *International Conference on Machine Learning*, 4095-4104.

Tan, M., Chen, B., Pang, R., Vasudevan, V., Sandler, M., Howard, A., & Le, Q. V. (2019). Mnasnet: Platform-aware neural architecture search for mobile. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2820-2828.

Cai, H., Zhu, L., & Han, S. (2019). Proxylessnas: Direct neural architecture search on target task and hardware. *International Conference on Learning Representations*.

Deb, K., Pratap, A., Agarwal, S., & Meyarivan, T. (2002). A fast and elitist multiobjective genetic algorithm: NSGA-II. *IEEE Transactions on Evolutionary Computation*, 6(2), 182-197.

Elsken, T., Metzen, J. H., & Hutter, F. (2019). Neural architecture search: A survey. *Journal of Machine Learning Research*, 20(55), 1-21.