Neural Architecture Search for Efficient Convolutional Networks: 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 that optimizes convolutional neural networks for both accuracy and computational efficiency. Traditional NAS methods often focus solely on accuracy metrics, leading to computationally expensive models that are impractical for resource-constrained environments. Our approach employs a multi-objective optimization strategy that simultaneously considers classification accuracy, model size, and inference speed. We introduce a modified evolutionary algorithm with specialized mutation and crossover operations tailored for neural architecture exploration. Experimental results on CIFAR-10 and ImageNet datasets demonstrate that our method discovers architectures that achieve competitive accuracy while reducing computational requirements by 35-60

Keywords: neural architecture search, multi-objective optimization, convolutional networks, computational efficiency, evolutionary algorithms

Introduction

The rapid advancement of deep learning has revolutionized numerous fields, from computer vision to natural language processing. However, the increasing complexity of neural network architectures poses significant challenges for practical deployment, particularly in resource-constrained environments such as mobile devices, embedded systems, and edge computing platforms. Traditional approaches to neural network design rely heavily on human expertise and

extensive trial-and-error experimentation, which is both time-consuming and suboptimal.

Neural Architecture Search (NAS) has emerged as a promising alternative, automating the process of discovering optimal network architectures. While early NAS methods demonstrated impressive results, they often prioritized accuracy at the expense of computational efficiency, resulting in models that are impractical for real-world applications. This limitation highlights the need for multi-objective optimization approaches that balance competing performance metrics.

This paper introduces a novel NAS framework that addresses these challenges through a comprehensive multi-objective optimization strategy. Our approach simultaneously optimizes for classification accuracy, model complexity, and inference speed, enabling the discovery of architectures that are both accurate and computationally efficient. By incorporating specialized evolutionary operators and a sophisticated fitness evaluation mechanism, our method navigates the complex search space of neural architectures more effectively than existing approaches.

The contributions of this work are threefold: First, we propose a modified evolutionary algorithm specifically designed for neural architecture search. Second, we introduce a multi-objective fitness function that balances accuracy with computational constraints. Third, we demonstrate through extensive experiments that our approach discovers architectures that significantly outperform both hand-designed networks and single-objective NAS methods in terms of computational efficiency while maintaining competitive accuracy.

Literature Review

The field of neural architecture search has evolved rapidly since its inception. Early work by Zoph and Le (2016) introduced reinforcement learning-based approaches for NAS, demonstrating that automated methods could discover architectures competitive with human-designed networks. However, these methods required enormous computational resources, making them impractical for widespread adoption.

Subsequent research focused on improving the efficiency of NAS through various techniques. ENAS (Pham et al., 2018) introduced weight sharing across architectures to reduce computational requirements. DARTS (Liu et al., 2018) proposed a differentiable architecture search method that significantly accelerated the search process. While these approaches improved efficiency, they still primarily focused on accuracy as the primary optimization objective.

The importance of multi-objective optimization in NAS has gained increasing recognition. Tan et al. (2019) introduced MnasNet, which incorporated latency as an optimization objective using reinforcement learning. Similarly, Cai et al.

(2018) proposed ProxylessNAS, which directly optimized for hardware-specific metrics. These works demonstrated the potential of multi-objective approaches but were limited in their ability to handle multiple competing objectives simultaneously.

Evolutionary algorithms have shown particular promise for multi-objective NAS. Real et al. (2019) demonstrated that evolutionary approaches could discover high-performing architectures, while Elsken et al. (2018) provided a comprehensive survey of evolutionary NAS methods. However, existing evolutionary approaches often struggle with the high-dimensional search space of neural architectures and the computational cost of fitness evaluation.

Our work builds upon these foundations by introducing a specialized evolutionary algorithm that incorporates domain knowledge about neural architecture design. Unlike previous approaches, our method explicitly models the trade-offs between accuracy, model size, and inference speed, enabling more effective navigation of the Pareto front in multi-objective optimization.

Research Questions

This research addresses the following fundamental questions:

1. How can neural architecture search be effectively formulated as a multi-objective optimization problem that simultaneously considers accuracy, computational efficiency, and model complexity?

2. What evolutionary operators and search strategies are most effective for exploring the high-dimensional space of neural architectures while balancing competing objectives?

3. To what extent can automated architecture search discover networks that outperform both human-designed architectures and single-objective NAS approaches in terms of the accuracy-efficiency trade-off?

4. How do the discovered architectures generalize across different datasets and computational constraints?

Objectives

The primary objectives of this research are:

1. To develop a multi-objective neural architecture search framework that optimizes for classification accuracy, model size, and inference speed simultaneously.

2. To design and implement specialized evolutionary operators for neural architecture exploration, including mutation and crossover operations tailored to convolutional network components.

3. To establish a comprehensive evaluation methodology for assessing the performance of discovered architectures across multiple metrics and datasets.

4. To validate the proposed approach through extensive experiments on standard benchmarks and compare its performance against state-of-the-art methods.

5. To analyze the characteristics of discovered architectures and derive insights about efficient neural network design principles.

Hypotheses to be Tested

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

H1: Multi-objective neural architecture search will discover architectures that achieve better accuracy-efficiency trade-offs compared to single-objective approaches.

H2: The proposed specialized evolutionary operators will enable more effective exploration of the neural architecture search space compared to standard genetic algorithms.

H3: Architectures discovered through our method will demonstrate consistent performance advantages across different datasets and computational constraints.

H4: The multi-objective optimization approach will lead to architectures with distinctive structural patterns that differ systematically from those designed through human intuition or single-objective optimization.

Approach/Methodology

Search Space Design

We define a flexible search space that encompasses various convolutional operations, including standard convolutions, depthwise separable convolutions, dilated convolutions, and various activation functions. The search space also includes skip connections, branching patterns, and attention mechanisms. Each architecture is represented as a directed acyclic graph where nodes represent feature maps and edges represent operations.

Multi-Objective Optimization Formulation

Our optimization problem is formulated as:

$$\min_{\theta \in \Theta} [-A(\theta), S(\theta), T(\theta)] \tag{1}$$

where θ represents an architecture in the search space Θ, $A(\theta)$ denotes classification accuracy, $S(\theta)$ represents model size (number of parameters), and $T(\theta)$ indicates inference time. We employ the NSGA-II algorithm for multi-objective optimization due to its effectiveness in handling non-dominated sorting and crowding distance computation.

Evolutionary Algorithm Design

Our modified evolutionary algorithm incorporates several key innovations:

1. **Specialized Mutation Operators**: We design mutation operators that specifically target different architectural components, including layer type changes, filter size adjustments, and connectivity modifications.

2. **Knowledge-Guided Crossover**: The crossover operation combines architectural components from parent networks while preserving functional modules that contribute to performance.

3. **Adaptive Search Strategy**: The algorithm dynamically adjusts exploration and exploitation based on population diversity and convergence metrics.

Fitness Evaluation

Each candidate architecture undergoes training on a subset of the target dataset for a fixed number of epochs. The fitness evaluation considers:

$$F(\theta) = \alpha \cdot A(\theta) - \beta \cdot \frac{S(\theta)}{S_{max}} - \gamma \cdot \frac{T(\theta)}{T_{max}} \tag{2}$$

where α, β, and γ are weighting coefficients that balance the importance of accuracy, model size, and inference time, respectively. S_{max} and T_{max} represent maximum acceptable values for model size and inference time.

Results

We evaluated our proposed method on two standard benchmarks: CIFAR-10 and ImageNet. The experiments were conducted using a cluster of NVIDIA Tesla V100 GPUs, with each architecture search requiring approximately 200 GPU-days.

On CIFAR-10, our method discovered architectures that achieved 94.8

On ImageNet, the results were equally promising. Our best discovered architecture achieved 75.3

Table 1: Performance Comparison on CIFAR-10 Dataset

Architecture	Accuracy (%)	Parameters (M)	Inference Time (ms)	Search Cost (GPU-days)
ResNet-56	93.0	0.85	12.3	-
DenseNet-BC	94.8	0.80	15.1	-
NASNet-A	97.4	3.3	23.5	2000
ENAS	97.1	4.6	19.8	0.5
Our Method (Pareto-1)	94.8	1.2	6.8	200
Our Method (Pareto-2)	96.2	2.1	8.9	200
Our Method (Pareto-3)	97.0	3.8	12.3	200

The table above demonstrates the effectiveness of our multi-objective approach. While our method may not achieve the absolute highest accuracy of some single-objective NAS approaches, it discovers architectures that provide superior trade-offs between accuracy and computational efficiency. The Pareto-optimal solutions cover a range of operating points, allowing practitioners to select architectures based on specific deployment constraints.

Discussion

The results validate our primary hypothesis that multi-objective neural architecture search can discover networks with superior accuracy-efficiency trade-offs. The discovered architectures exhibit several interesting characteristics that differentiate them from both human-designed networks and single-objective NAS results.

First, we observed a preference for heterogeneous layer compositions, with different parts of the network employing different types of convolutional operations optimized for their specific roles. This contrasts with the homogeneous structures often found in hand-designed networks.

Second, the architectures frequently incorporated efficient operations like depthwise separable convolutions in early layers where spatial information is more important, while using standard convolutions in later layers where channel interactions become more critical. This pattern suggests that the search process learned to allocate computational resources strategically.

Third, we found that skip connections were used more selectively than in architectures like ResNet, appearing primarily where they provided significant performance benefits rather than as a universal design pattern.

The computational cost of our approach, while substantial, represents a significant improvement over early NAS methods and provides good value given the quality of discovered architectures. The ability to discover multiple Pareto-optimal solutions in a single search run is particularly valuable for practical applications where deployment constraints may vary.

Conclusions

This paper has presented a novel multi-objective neural architecture search framework that effectively balances classification accuracy with computational efficiency. Our approach demonstrates that explicitly considering multiple competing objectives during architecture search leads to networks that are better suited for practical deployment in resource-constrained environments.

The key contributions of this work include: (1) a comprehensive formulation of NAS as a multi-objective optimization problem, (2) specialized evolutionary operators tailored for neural architecture exploration, and (3) extensive experimental validation demonstrating significant improvements in computational efficiency without compromising accuracy.

Future work will focus on reducing the computational cost of the search process through more efficient fitness evaluation strategies and incorporating additional objectives such as energy consumption and memory bandwidth requirements. We also plan to extend the approach to other domains beyond computer vision, including natural language processing and speech recognition.

Acknowledgements

This research was supported by the National Science Foundation under grant CNS-0435060 and the European Research Council under the Horizon 2020 program. The authors thank the Tsinghua University AI Research Center for providing computational resources and the anonymous reviewers for their valuable feedback. We also acknowledge the contributions of the open-source deep learning community, whose tools and libraries made this research possible.

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 Artificial Intelligence*, 2(1), 45-62.

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

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

Liu, H., Simonyan, K., & Yang, Y. (2018). DARTS: Differentiable architecture search. *arXiv preprint arXiv:1806.09055*.

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 Conference on Computer Vision and Pattern Recognition*.

Cai, H., Zhu, L., & Han, S. (2018). Proxylessnas: Direct neural architecture search on target task and hardware. *arXiv preprint arXiv:1812.00332*.

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*.

Elsken, T., Metzen, J. H., & Hutter, F. (2018). Neural architecture search: A survey. *Journal of Machine Learning Research*.