Gradient-Free Versus Gradient-Based Optimization in Sparse Neural Architecture Search: A Benchmarking Study Across Heterogeneous Hardware-Constrained Deployment Environments

Pat Walker
Professor
Korea Advanced Institute of Science and Technology
291 Daehak-ro, Yuseong-gu, Daejeon 34141, Republic of Korea

Adrian White
Associate Professor
Technical University of Munich
Arcisstraße 21, 80333 Munich, Bavaria, Germany

Dana Lewis
PhD
University of Waterloo
200 University Avenue West, Waterloo, Ontario N2L 3G1, Canada

Drew Miller
D.Sc
The University of Tokyo
7-3-1 Hongo, Bunkyo-ku, Tokyo 113-8654, Japan

Abstract. Neural Architecture Search (NAS) has emerged as a foundational paradigm in automated machine learning, yet the trade-offs between gradient-based differentiable NAS (DNAS) and gradient-free evolutionary or Bayesian strategies remain insufficiently characterized under hardware-constrained deployment conditions. This study presents a systematic benchmarking analysis of six representative NAS algorithms—including DARTS, GDAS, SNAS, CMA-ES-NAS, BOHB, and REA—evaluated across three edge-inference hardware platforms and four sparse-model complexity regimes. Employing standardized search space normalization and latency-aware proxy metrics, we assess convergence efficiency, Pareto-front quality, and generalization fidelity on CIFAR-100, ImageNet-1K, and a domain-specific industrial defect classification corpus. Results indicate that gradient-free Bayesian approaches consistently achieve superior hardware-accuracy trade-offs under strict latency budgets below 15 ms, while DNAS retains advantages in high-parameter regimes. Practical deployment guidelines are derived from empirical cross-platform evidence.

Keywords: Neural Architecture Search, gradient-free optimization, differentiable NAS, hardware-aware inference, sparse model compression, Bayesian hyperparameter optimization, edge AI deployment, automated machine learning, latency-accuracy Pareto optimization

Introduction

The exponential proliferation of deep learning systems in safety-critical and resource-constrained environments—spanning autonomous edge inference nodes, industrial quality control pipelines, and embedded cognitive systems—has fundamentally elevated the engineering demands placed upon neural network design methodologies. As of 2024–2026, the global deployment of AI-driven inference systems on heterogeneous hardware substrates, including application-specific integrated circuits (ASICs), field-programmable gate arrays (FPGAs), and mobile system-on-chip (SoC) architectures, has rendered manual network architecture design both economically prohibitive and technically intractable at scale. Neural Architecture Search (NAS) has consequently matured from an academic curiosity into an indispensable component of production-grade AI engineering pipelines, enabling the automated discovery of network topologies that satisfy multi-objective constraints including accuracy, parameter count, multiply-accumulate (MAC) operation density, and real-time inference latency. Despite substantial progress in the NAS literature, a critical and persistent gap exists in the rigorous comparative analysis of its two dominant algorithmic families: gradient-based differentiable NAS (DNAS) and gradient-free methods encompassing evolutionary algorithms, Covariance Matrix Adaptation Evolution Strategies (CMA-ES), and Bayesian optimization frameworks such as Hyperband-Bayesian Hybrid (BOHB). Prior survey works have catalogued these approaches taxonomically, yet few studies have subjected them to controlled, hardware-aware benchmarking under unified search space constraints, reproducible evaluation protocols, and realistic sparse-model deployment scenarios. This methodological fragmentation has led to contradictory performance claims across the literature, obstructing practitioners from making evidence-based algorithmic selections for specific deployment targets. The relevance of this problem is further amplified by the widespread adoption of model sparsification and structured pruning as a precondition for edge deployment, introducing non-differentiable discontinuities that fundamentally disadvantage gradient-dependent search strategies. Gradient-free optimization methods, by contrast, operate independently of loss-landscape smoothness assumptions, potentially conferring systematic advantages in sparse or quantized architecture spaces—a hypothesis that, to date, has not been empirically validated under controlled multi-platform conditions. This paper addresses the aforementioned research gap through a comprehensive empirical benchmarking study structured as follows. Section 2 provides a formal taxonomy of the evaluated NAS algorithms and their theoretical optimization properties. Section 3 details the experimental design, encompassing search space unification, hardware platform specifications, and proxy metric formulations. Section 4 presents quantitative benchmarking results across datasets and hardware targets, with Pareto-front analyses and statistical significance testing. Section 5 discusses the practical implications for AI systems engineers selecting NAS strategies under deployment-specific constraints. Section 6 concludes with synthesized guidelines and directions for future research in hardware-co-designed architecture search.

This is a preliminary version. To read the full version of the article, please purchase a subscription.

References

1. Kumar, N., & Kataria, V. Enhanced Sentiment Classification using a Multi-layered Stacked Ensemble Architecture.
2. Kumar, Nitin, and Vipin Kataria. "Enhanced Sentiment Classification using a Multi-layered Stacked Ensemble Architecture."
3. Kumar, N., & Kataria, V. (2023). Enhanced Sentiment Classification using a Multi-layered Stacked Ensemble Architecture. International Journal of Intelligent Systems and Applications in Engineering, 11(4s), 304–311.
4. Semeniuk, V. (2024). Аналіз та оптимізація продуктивності баз даних у розподілених системах. COMPUTER-INTEGRATED TECHNOLOGIES: EDUCATION, SCIENCE, PRODUCTION, (54), 199-205.
5. D. Satyanarayana, Sathyashree “A Three Level Architecture for Wireless Communication using LiFi”, in Recent Advances in Information and Communication Technology, Vol: 566, pp: 212-221, Springer International Publishing, 2018 book. Print ISBN: 978-3-319-60662-0; Electronic ISBN: 978-3-319-60663-7; Book Series.
6. Satyanarayana, D. and Rao, S. V. (2011), “Fault-tolerant spanners for ad hoc networks” International Journal of Network Management, Vol: 22, Issue: 4, Pages: 311-329, July-2012, Wiley Publishers. https://doi.org/10.1002/nem.807
7. D. Satyanarayana, Sathyashree. S “Mobility Management in Voronoi based Cell Structure for Wireless networks”, IEEE proceedings for the International Conferences on Currents trends in Information Technology (CTIT), Pages: 204-208, Dubai, December-2013. DOI: 10.1109/CTIT.2013.6749504; INSPEC Accession Number: 14131032
8. D Satyanarayana, Perspective Chapter: Postquantum Security’, Coding Theory - Advances in Communications Engineering and Information Security. IntechOpen, Nov. 19, 2024. doi: 10.5772/intechopen.1007829.
9. D. Satyanarayana and J. M. H. Elmirghani, "An Energy Efficient Network Architecture for Infrastructured Wireless Networks," 2010 IEEE Global Telecommunications Conference GLOBECOM 2010, Miami, FL, USA, 2010, pp. 1-6, doi: 10.1109/GLOCOM.2010.5683234.
10. Alkalbani, A.S., Mantoro, T., Degala, S. (2016). Energy-Distance Aware Clustering Scheme (E-DACS) for Wireless Sensor Networks. In: Meesad, P., Boonkrong, S., Unger, H. (eds) Recent Advances in Information and Communication Technology 2016. Advances in Intelligent Systems and Computing, vol 463. Springer, Cham. https://doi.org/10.1007/978-3-319-40415-8_27
11. Rakhmanov, N. R., Kurbatsky, V. G., Guliyev, H. B., Tomin, N. V., & Rakhmanov, R. N. (2018). Analysis and modeling of harmonic distortions in consumer’s distribution networks, containing a powerful non-linear load. In E3S Web of Conferences (Vol. 58, p. 03015). EDP Sciences.
12. Hashimov, A. M., Rahmanov, N. R., Guliyev, H. B., Ibrahimov, F. S., & Tabatabaei, N. M. (2020). Probabilistic evaluatio n of voltage stability limit of power system under the conditions

of accidental emergency outages of lines and generators. International Journal on Technical and Physical Problems of Engineering, 12(2), 40-45.
13. Guliyev, H. B. (2024). Management modes of reactive power compensation facilities in networks with renewable energy sources with distorting loads. IJTPE Journal International Journal on Technical and physical problems of engineering (IJTPE), 16(1), 14-20.