A Pre-Search Evaluation Framework for Assessing Search Space Complexity

Anonymous authors
Paper under double-blind review

Abstract

The search space represents the most important component of Neural Architecture Search. It defines the ranges of performance, encapsulating the potential for discovering optimal architectures. This paper presents a framework for evaluating these spaces based on size, performance diversity, architecture diversity, and multi-task ability. Our comparative analysis across seven established benchmarks highlights their complexity and adaptability to target multiple tasks, offering a comprehensive tool for NAS strategy assessment [1].

1 Introduction & Related Works

Neural Architecture Search (NAS) (Cai et al., 2019; Liu et al., 2019) is pivotal in driving automated neural network design. Yet, assessing NAS methods is challenged by the complexity and uncharted nature of their search spaces. Current NAS methodologies oscillate between proposing novel search spaces and relying on pre-defined benchmarks, each exhibiting distinct characteristics. We evaluate these spaces based on size, performance diversity, architecture diversity, and multi-task ability. Despite large search spaces, our analysis indicates that limited diversity can simplify exploration.

Existing studies recognize the importance of search space complexity, particularly operator importance (Lopes et al., 2023). This paper introduces a framework to evaluate NAS search spaces without training, featuring a scoring mechanism that gauges strategy efficiency across complexity levels, addressing the lack of systematic search space assessment in current literature.

2 Search Space Complexity Scoring

In this section, we detail how each property is measured.

- **Search Space Size**: *The size is quantified simply as the total number of unique architectures within the search space.*

- **Performance Diversity** refers to the variance in the effectiveness of different architectures within the space. A large performance diversity is desirable. It indicates a wide range of effectiveness among different architectures, offering greater opportunities to discover highly efficient models.

We utilize the variance of $jacob\_cov$ *estimation strategy (Abdelfattah et al., 2021).* $jacob\_cov$ measures the correlation of activations within a network when subject to different inputs within a mini-batch of data – the lower the correlation, the better the network is expected to perform as it can differentiate between different inputs well. Figure 1 compares this variance to the variance of the actual performance post-training, to validate the correlation between predicted and realized performance diversities in different benchmark. We define performance diversity score, denoted as *PDS*, using Cohen's d measure (Kotz et al., 2005) to compare the variance of the jacobian covariance metric and the actual accuracy in a search space. The jacob_cov measure performs badly at evaluating the ranking of the architectures, as reflected by the low values of the kendal Tau Correlation (red), in figure 1. However, comparing the variances reflected in the low PDS (blue), indicates its efficacy as a reliable estimator of the overall distribution of performance within different search spaces.

[1] Code will be available in the Camera ready.

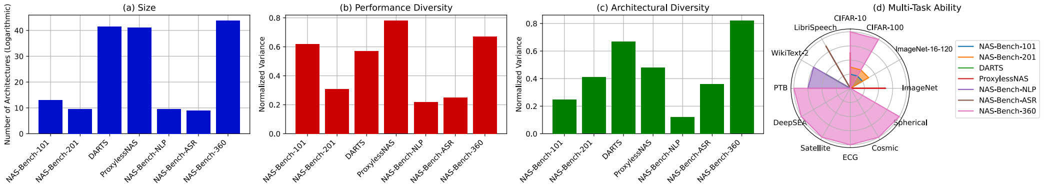

Figure 2: Property analysis for common NAS benchmarks.

We detail PDS computation as follows:

$$PDS = \frac{|SD_{\text{jacob\_cov}} - SD_{\text{accuracy}}|}{SD_{\text{pooled}}} \tag{1}$$

where $SD_{\text{jacob\_cov}}$ and SD_{accuracy} denote the standard deviation for both the Jacobian covariance and the actual accuracy metrics respectively. The pooled standard deviation (SD_{pooled}), is the average of these two standard deviations: $SD_{\text{pooled}} = \frac{SD_{\text{jacob\_cov}} + SD_{\text{accuracy}}}{2}$

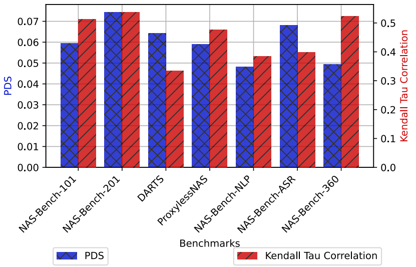

Figure 1: Performance diversity Comparison.

- **Architecture Diversity** is critical as it expands the scope for innovation by encompassing a broad spectrum of structural design possibilities. *We construct a vector representing each architecture's features, derived from an analysis of models in a model zoo. This vector includes elements like convolutional layers with their kernel sizes, attention layers, feedforward networks (ffn) layers, PvT blocks, residual blocks, MBConv blocks, etc. We then calculate the variance in these vectors across the search space, providing a quantitative measure of architectural diversity.*

- The final property, **Multi-task Ability**, evaluates the capacity of architectures within the space to efficiently handle multiple tasks. This allows multi-task search for backbone architectures able to adapt to different tasks, making the search more complex. *This is measured by the number of distinct tasks that the architectures within a benchmark or search space are designed to target.*

3 Comparison of SOTA search spaces

Figure 2 showcases a comparison across multiple NAS benchmarks based on our first properties. NAS-Bench-101 (Ying et al., 2019), with its considerable number of architectures, offers a wide range of designs, yet its performance and architectural diversity are not as extensive. NAS-Bench-201 (Dong & Yang, 2020) shows greater architectural uniformity, which might limit its adaptability across different tasks and renders the search simple. We consider these two benchmarks as not complex.

DARTS (Liu et al., 2019) and ProxylessNAS (Cai et al., 2019), both known for continuous search spaces, show moderate performance diversity, suggesting a reasonable spread in model efficacy without extreme outliers. This may reflect a balance between exploration and exploitation in their search strategies. While building a supernetwork allows for an extensive search space, the continuous space restricts architecture diversity.

When it comes to NAS-Bench-NLP (Klyuchnikov et al., 2022) and NAS-Bench-ASR (Mehrotra et al., 2021), despite having fewer architectures, they exhibit significant performance diversity. This suggests a low architecture modification may reflect on significant performance improvement for NLP and Speech tasks.

Finally, NAS-Bench-360 (Tu et al., 2022) stands out in its multi-task ability, as indicated by the spider chart. This benchmark appears to be the most complex and versatile, likely due to its design that integrates diverse datasets and tasks, making it a comprehensive choice for evaluating NAS methods.

4 Conclusion

In this work, we introduced a systematic framework to evaluate the complexity of NAS search spaces, underscoring the significance of architecture diversity as a catalyst for innovation. This work complements NAS best practices (Lindauer & Hutter, 2020) with a dedicated search space analysis. Future work will extend our framework to include hardware efficiency metrics, examining the impact of search spaces on latency and energy consumption, crucial factors for real-world deployment.

URM STATEMENT

The authors acknowledge that at least one key author of this work meets the URM criteria of ICLR 2024 Tiny Papers Track.

REFERENCES

Mohamed S. Abdelfattah, Abhinav Mehrotra, Lukasz Dudziak, and Nicholas Donald Lane. Zero-cost proxies for lightweight NAS. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL `https://openreview.net/forum?id=0cmMMy8J5q`.

Han Cai, Ligeng Zhu, and Song Han. Proxylessnas: Direct neural architecture search on target task and hardware. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019.

Xuanyi Dong and Yi Yang. Nas-bench-201: Extending the scope of reproducible neural architecture search. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020.

Nikita Klyuchnikov, Ilya Trofimov, Ekaterina Artemova, Mikhail Salnikov, Maxim V. Fedorov, Alexander Filippov, and Evgeny Burnaev. Nas-bench-nlp: Neural architecture search benchmark for natural language processing. *IEEE Access*, 10:45736–45747, 2022.

Samuel Kotz, Narayanaswamy Balakrishnan, Campbell B Read, and Brani Vidakovic. *Encyclopedia of Statistical Sciences, Volume 1*. John Wiley & Sons, 2005.

Marius Lindauer and Frank Hutter. Best practices for scientific research on neural architecture search. *The Journal of Machine Learning Research*, 21(1):9820–9837, 2020.

Hanxiao Liu, Karen Simonyan, and Yiming Yang. DARTS: differentiable architecture search. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019.

Vasco Lopes, Bruno Degardin, and Luís A. Alexandre. Are neural architecture search benchmarks well designed? A deeper look into operation importance. *Inf. Sci.*, 650:119695, 2023. doi: 10.1016/J.INS.2023.119695. URL `https://doi.org/10.1016/j.ins.2023.119695`.

Abhinav Mehrotra, Alberto Gil C. P. Ramos, Sourav Bhattacharya, Lukasz Dudziak, Ravichander Vipperla, Thomas Chau, Mohamed S. Abdelfattah, Samin Ishtiaq, and Nicholas Donald Lane. Nas-bench-asr: Reproducible neural architecture search for speech recognition. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021.

Renbo Tu, Nicholas Roberts, Mikhail Khodak, Junhong Shen, Frederic Sala, and Ameet Talwalkar. Nas-bench-360: Benchmarking neural architecture search on diverse tasks. In *NeurIPS*, 2022.

Chris Ying, Aaron Klein, Eric Christiansen, Esteban Real, Kevin Murphy, and Frank Hutter. Nas-bench-101: Towards reproducible neural architecture search. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pp. 7105–7114. PMLR, 2019.

A NAS Benchmarks

In this section, we describe each benchmark evaluated in this paper.

- **NAS-Bench-101:** (Ying et al., 2019) The first architecture dataset for NAS research, containing over 423,000 unique convolutional architectures evaluated on CIFAR-10, providing performance metrics and training times.
- **NAS-Bench-201:** (Dong & Yang, 2020) A tabular benchmark for NAS that extends upon NAS-Bench-101, covering three datasets (CIFAR-10, CIFAR-100, and ImageNet-16-120) and offering a fixed search space of 15,625 neural cell-based architectures.
- **DARTS:** (Liu et al., 2019) Differentiable Architecture Search introduces a gradient-based optimization method for NAS, focusing on a continuous search space that allows efficient architecture selection through gradient descent. The search space is extensive and targets ImageNet task.
- **ProxylessNAS:** (Cai et al., 2019) A method that directly learns architectures for large-scale target tasks and hardware platforms, eliminating the need for proxy tasks, thus named for its ability to bypass proxy stages common in other NAS methods. They proposed a search space of over 6.64x10^{17} composed of convolutional blocks.
- **NAS-Bench-NLP:** (Klyuchnikov et al., 2022) A benchmark specifically designed for Natural Language Processing tasks, containing a collection of architectures for evaluating recurrent neural networks (RNNs) on language modeling.
- **NAS-Bench-ASR:** (Mehrotra et al., 2021) Tailored for Automatic Speech Recognition, this benchmark provides a platform to study the architectures of RNNs and their performance in speech-related tasks. It provides 8242 different architectures with their performance metrics.
- **NAS-Bench-360:** (Tu et al., 2022) A recent benchmark aimed at generalizing NAS studies across diverse data modalities and tasks, offering a 360-degree view of NAS performance across various domains beyond vision and language. It is heavily based on DARTS.

B Experimental Setup

Given these established NAS benchmarks, we conduct a comparison and analysis of each one. The sizes of each benchmark are available in the corresponding papers and verified either using the available code or by confirming the search space definition. For the performance diversity, the PDS metric was computed using all the architectures in the search spaces. For DARTS and proxylessNAS, because of the huge size, we average the PDS over 5 randomly sampled 10,000 architectures. The architectural diversity, required the definition of a list of different architectural blocks. This list was extracted from 168 architectures from Model Zoo [2], targeting multiple tasks. Similarly the the performance diversity, the variance to measure architectural diversity was measured using the full size of the search spaces, except for DARTS and proxylessNAS.

[2] https://modelzoo.co/framework/pytorch