Hardware-Aware Neural Architecture Search (NAS) Methods

MOUGENOT, **LY-MANSON**, and **MORLIER**

Gustave MOUGENOT
Student IMT-Atlantique
gustave.mougenot@imt-atlantique.net

Timotée Ly-Manson
PhD student IMT-Atlantique
timotee.ly-manson@imt-atlantique.fr

Jérémy MORLIER
PhD student IMT-Atlantique
jeremy.morlier@imt-atlantique.fr

Abstract

Neural Architecture Search (NAS) seeks to automate the design of neural network architectures, a process traditionally reliant on human expertise. However, most existing approaches, though effective, primarily focus on maximizing accuracy while neglecting hardware constraints. This work proposes integrating hardware-aware metrics such as latency, energy consumption, and memory footprint into the architecture evaluation criteria. The approach builds upon state-of-the-art NAS methods and frameworks, such as the *einspace* [1] search space. We seek to demonstrate that it is possible to obtain a strategy that performs hardware-aware optimization in the einspace search space.

Keywords: NAS, Hardware-aware, *einspace*, evaluation criteria, evolutionary algorithms, regularized evolution.

Background

The increase in computational power and the availability of large-scale datasets have recently made neural network architectures effective. However, the design of high-performing architectures has long relied on human expertise, a process guided by intuition, experience, and extensive empirical experimentation. While this approach has proven effective, it remains costly in terms of time, computational resources, and required expertise.

In this context, Neural Architecture Search (NAS) has emerged as a research field aiming to automate the design of neural network architectures. NAS seeks to automatically explore a predefined search space of architectures to identify the most promising structures. Nevertheless, early approaches, although successful in improving accuracy, largely ignored hardware-related constraints such as inference latency, energy consumption, or memory footprint.

Building upon existing work, we investigate how hardware-aware criteria can be integrated into the model search process.

Aim

This work is part of an effort to explore and extend Neural Architecture Search (NAS) approaches that explicitly incorporate hardware constraints. The main objective is to demonstrate the feasibility of hardware-aware constraints within a single framework.

we are focusing on adapting the search strategy proposed in einspace [1] to explicitly account for hardware constraints. We first reproduce selected published results before exploring how hardware considerations can be effectively incorporated into the search process.

We explore different strategies for incorporating hardware awareness, across several metrics such as the number of parameters, FLOPs, latency, and energy cost.

Finally, the results are compared both quantitatively (computational results) and conceptually (methodological limitations), highlighting the contributions and limitations of our hardware-aware NAS approach.

Methods

As previously indicated, our work primarily builds on the einspace experimental framework [1]. This paper presents a NAS method based on regularized evolution of architectures sampled from base blocks in 4 families: **branching**, **aggregation**, **routing**, and **computation**.

The sampling is based on: choosing some architectures in the population, comparing them based on a criterion, mutating the best one to create a new individual, and updating the population by removing the oldest individuals as represented in Figure 1.

As criteria, we try 3 policies that use different parameters. For simplicity, we use here only **accuracy** and the **number of parameters** of the architecture (we could imagine similar usage of FLOPS, latency, or energy cost). Figure 2 shows our policy, some of these policies are inspired from [2].

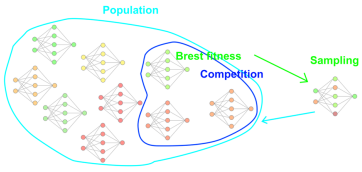

Figure 1: Regularised evolution illustration

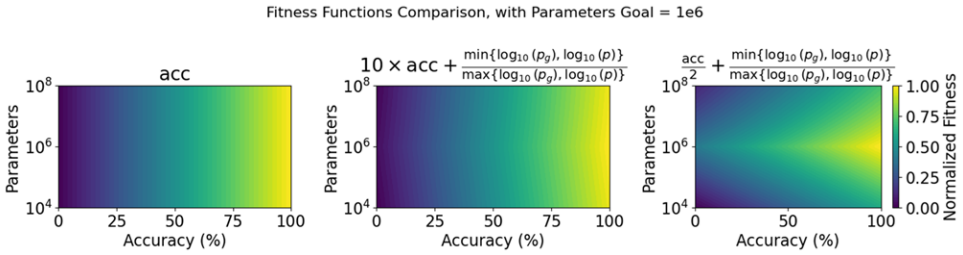

Figure 2: Fitness functions comparisons

As datasets, we use the **CIFARTile** dataset [3], which is a 4-class image dataset with 60,000 images.

Expected results and preliminary results

First of all, we expect results that are reproducible and openly shared on a free platform (e.g., GitHub) to facilitate verification and reuse of experiments. These results should include:

- the partial reproduction of existing benchmarks;
- the integration and measurement of new hardware metrics;
- a quantitative and qualitative comparison of the tested optimization strategies.

At this stage, we have obtained results presented in Figure 3, which indicate that we can successfully optimize accuracy while constraining the number of parameters (target: 10^6 parameters). **We achieved an accuracy of 53.64% with 10^6 parameters, whereas the initial architecture reached 47.13% with 10^7 parameters**.

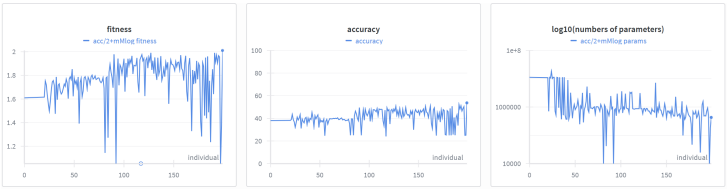

Figure 3: Our result, hardware-aware(blue)

Conclusion

Our experiments show that it is possible to optimize neural network architectures with respect to both accuracy and parameter count. However, these results require further validation, particularly by extending the search duration and running experiments with multiple random seeds to ensure statistical significance.

References

[1] Linus Ericsson, Miguel Espinosa, Chenhongyi Yang, Antreas Antoniou, Amos Storkey, Shay B. Cohen, Steven McDonagh, and Elliot J. Crowley. einspace: Searching for Neural Architectures from Fundamental Operations, October 2024. arXiv:2405.20838 [cs].

[2] Mingxing Tan, Bo Chen, Ruoming Pang, Vijay Vasudevan, Mark Sandler, Andrew Howard, and Quoc V. Le. MnasNet: Platform-Aware Neural Architecture Search for Mobile, May 2019. arXiv:1807.11626 [cs].

[3] David Towers, Rob Geada, Andrew Stephen McGough, and Amir Atapour-Abarghouei. CIFARTile Dataset. https://data.ncl.ac.uk/articles/dataset/CIFARTile_Dataset/24551539, 11 2023. Dataset.