NebulOS

PRINCIPAL ACADEMIC TUTOR

Barbara Caputo, DAUIN, Politecnico di Torino

ACADEMIC TUTOR

Barbara Maja Broadbent, Dipartimento di Design, Politecnico di Milano

EXTERNAL INSTITUTIONS

Nebuly

EXTERNAL TUTOR

Julien Roux, Nebuly

TEAM MEMBERS

Francesco Capuano, Data Science and Engineering, Politecnico di Torino

Matteo Matteotti, Data Science and Engineering, Politecnico di Torino

Marco Alberto Pellicanò, Materials Engineering and Nanotechnology, Politecnico di Milano

Davide Rinaldoni, Mathematical Engineering, Politecnico di Milano

Federica Rosellini, Biomedical Engineering, Politecnico di Milano

Executive summary

The unprecedented explosion of Artificial Intelligence (AI) and its applications in everyday personal and professional life is posing a major concern for the environment. Among the various reasons that make AI development extremely energy-demanding is the fact that there are generally no all-purpose guidelines when searching for the best model for a given task. Usually, many different model configurations are tested, and only the one which yields the best results is retained. Not only is this *trial-and-error* approach a significant waste of energy and resources, but it also leads to a centralization of innovation capabilities in a handful of big players, who can speed up the process by relying on unmatched computational firepower. To tackle the issue, we are proud to introduce NebulOS. Unlike traditional methods that search for the best model only considering the performance on the downstream task, our novel approach also takes into the energy consumption when training a model on a specific hardware, producing tailored designs on a variety of different devices. This ensures that the final model aligns with the end-user's unique needs, resources, and preferences.

Automating the procedure of architecture design and tailoring it to be as energy-efficient as possible holds the promise of enabling a new generation of genuinely green AI models, and to stimulate a much broader involvement of underprivileged players - such as public research institutes - in the AI scene.

Key Words

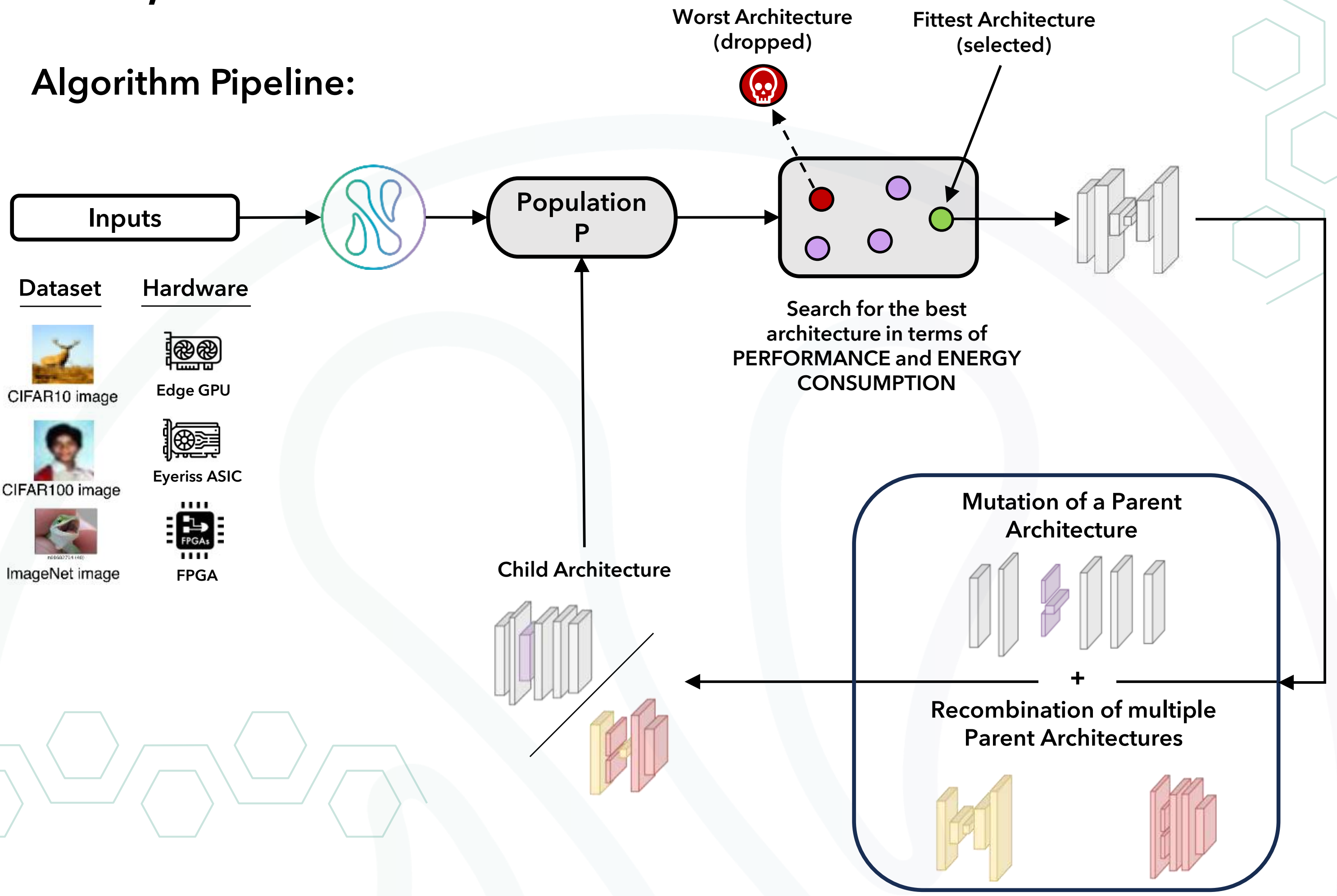
Green AI, democratic AI, Energy saving

NebulOS

ALTA SCUOLA POLITECNICA
POLITECNICO DI MILANO | POLITECNICO DI TORINO

FAIR, GREEN AI

Algorithm Pipeline:



Inputs

Dataset

CIFAR10 image

CIFAR100 image

ImageNet image

Hardware

Edge GPU

Eyeriss ASIC

FPGA

Population P

Worst Architecture (dropped)

Fittest Architecture (selected)

Search for the best architecture in terms of PERFORMANCE and ENERGY CONSUMPTION

Mutation of a Parent Architecture

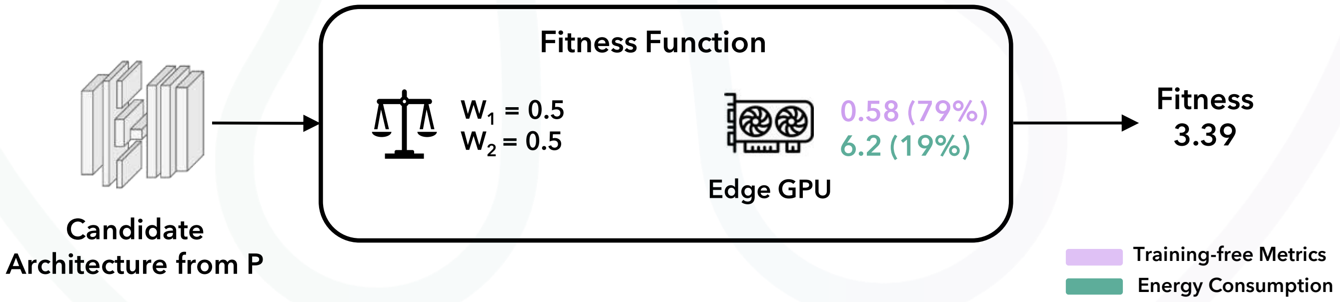
+

Recombination of multiple Parent Architectures

Child Architecture

Novelty:
NebulOS integrates hardware-specific energy consumption in the definition of the fitness function.

Fitness Function $= w_1 * TF\ Metric(Architecture) + w_2 * Energy\ Consumption(Architecture, Device)$



Fitness Function

$W_1 = 0.5$
$W_2 = 0.5$

0.58 (79%)
6.2 (19%)

Edge GPU

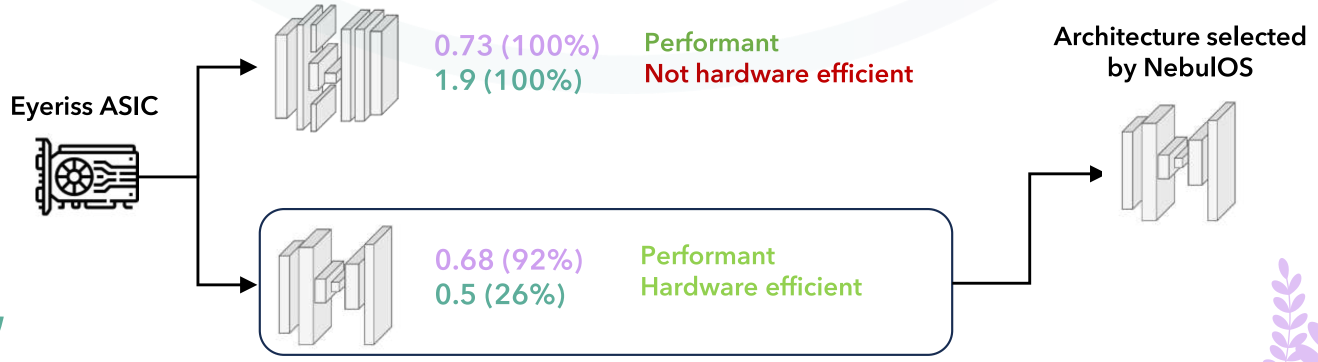
Fitness
3.39

Candidate Architecture from P

Training-free Metrics

Energy Consumption

Given a specific hardware, NebulOS selects the architecture with the best trade-off between performance and hardware efficiency.



Eyeriss ASIC

0.73 (100%)
1.9 (100%)

Performant
Not hardware efficient

0.68 (92%)
0.5 (26%)

Performant
Hardware efficient

Architecture selected by NebulOS

Project description written by the Principal Academic Tutor

AI is now everywhere. Fraud detection algorithms that help us not receive scam emails, vocal assistants that control alarms and memos for us, web mapping services that calculate in a fraction of time the shortest way from any point A to any other point B on the planet, the most recent large language models that allow users to experience a pseudo-human conversation with a chatbot. Whilst AI has undoubtedly improved some aspects of our lives, its associated carbon footprint is a major reason for concern.

In fact, Deep Learning is intrinsically extremely computationally intensive. Due to its *black-box* nature, the most suitable model to tackle a given problem is not known a priori. This results in an exhaustive *trial-and-error* approach, and in a significant waste of energy and resources. Even though some techniques have been introduced in the literature to automatize the search for the best model, these approaches are mostly hardware-agnostic. In other words, the best model can still be intricately complex, and its training can span over several weeks and demand multiple GPUs, often not available to casual users. Moreover, the state of the art in the field of AI research is always defined in terms of accuracy, and energy consumption is always overlooked.

By analyzing the needs of the different stakeholders involved, the students need to come up with a hardware-aware search algorithm, which is tailored to the user's unique resources. In so doing, the hardware will no longer be a constraint, but it will become a feature.

Team description by skill

Our team is a mix of expertise spanning various fields of engineering. Each member brings a unique skill set and perspective to the table, and this allowed us to tackle the complex problem we were facing with a multi-disciplinary approach.

- Francesco & Matteo: Specializing in Data Science and Engineering, both Francesco and Matteo have a strong foundation in Data Science, programming, and algorithm design. Their academic background provided them with the tools necessary to craft a system capable of performing AI optimization. They are the technical backbone of the project.

- Davide: With a background in Mathematical Engineering, Davide exploited the power of mathematical tools and programming skills to support the technical development of NebulOS, together with Francesco and Matteo.

- Federica and Marco: Coming from domains not directly linked to AI - specifically Biomedical Engineering and Material Engineering and Nanotechnology - their goal was to focus on the community side of the project. First, they tried to identify the more relevant AI applications in their fields. Then, they researched the best launch strategy for NebulOS. Through Reddit's Machine Learning channels, they conducted interviews with developers of open-source projects to learn more about the best platforms to launch our project, the communication strategy to adopt, and how to attract new users to our open-source repository.

Abstract

Neural Architecture Search (NAS) has emerged as a powerful technique to automatically discover optimal neural network architectures for specific tasks. Traditional NAS approaches require extensive training of multiple candidate architectures, resulting in significant time and energy consumption. To address this limitation, Training-Free NAS (TF-NAS) methods have been proposed, wherein the search is guided by hardware-agnostic metrics that correlate with validation accuracy. However, while these metrics are informative, they may overlook crucial hardware constraints that can impact the deployment of the discovered architecture. In this work, we propose a hardware-aware Training-Free NAS approach that considers both training-free metrics and hardware constraints, aiming to find a Pareto-optimal solution that strikes a balance between validation accuracy and energy consumption. We present a framework for incorporating hardware metrics into the search process, allowing users to customize the trade-offs based on their specific requirements.

Understanding the problem

Rather than relying on human intuition or prior experience, NAS employs search algorithms to explore a vast space of possible architectures and determine the optimal configuration. The process is guided by using the model's performance on a given task as an objective function. Whilst NAS has proven to outperform human designers on multiple occasions, it typically requires the training of every single candidate architecture it evaluates. Whilst the research for the best architecture becomes less trial-and-error, its exhaustive nature still makes the traditional NAS approach inaccessible or impractical for many researchers and practitioners, as well as extremely energy-demanding.

Recent works have been able to rank models via the computation of metrics that do not require any training and have been proven to correlate well with the performances of the said models (Training-Free NAS). Whilst this approach significantly reduces the computational cost of traditional NAS, it does not solve all its issues. The typically-used metrics usually correlate well with the model's performances, but really little is known about the corresponding required energy consumption.

Our solution and its opportunities

Our proposed solution, NebulOS, drives the research in terms of performance (by using training-free metrics) but also considers the energy consumed when training a model for a specific task on the hardware at the user's disposal. At its core, NebulOS utilizes a genetic algorithm to search through a population of candidate architectures. The solution has been fully tested for image classification on three datasets (ImageNet16-120, CIFAR-10, and CIFAR-100), across three different devices (NVIDIA Edge GPU Jetson TX2, ASIC-Eyeriss, and FPGA).

Our strategy paves the way for a significant reduction in computational effort, achieved through hardware optimization. Such a framework makes it possible to democratize the use of advanced AI solutions, traditionally the purview of large tech entities, by extending their applicability to public research institutes and to enterprises with limited computational resources. Ultimately, NebulOS goes someway in making AI greener and fairer.

An open-source solution

Our solution is completely open-source on GitHub (https://www.github.com/fracapuano/NebulOS). A well-documented open-source AI solution offers multiple benefits. It fosters collaboration, ensures higher code readability, and increases the likelihood of its adoption across various companies. Having a community around our product would help establish a reciprocal relationship of trust with the users and gain credit in the eyes of the general public, ultimately resulting in an increase in reputation. Whilst building a highly engaged community from scratch falls off the scope of this project, we conducted a thorough analysis to learn how to launch an open-source product and to understand the main mechanisms behind online communities.

Main bibliographic references

Patterson, D., Gonzalez, J., Hölzle, U., Le, Q., Liang, C., Rothchild, D., Texier, M., and Dean, J. The carbon footprint of machine learning training plateau, then shrink

Cavagnero, N., Robbiano, L., Caputo, B., and Averta, G. FreeREA: Training-free -based architecture search. *In 2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV).* IEEE, jan 2023.

Dong, X. and Yang, Y. Nas-bench-201: Extending the scope of reproducible neural search, 2020.

Li, C., Yu, Z., Fu, Y., Zhang, Y., Zhao, Y., You, H., Yu, Q., Wang, Y. HW-nas-bench: Hardware-aware neural architecture search benchmark. In *International Conference on Learning Representations*, 2021