Surrogate Neural Architecture Codesign Package (SNAC-Pack)

Jason Weitz[1], Dmitri Demler[1], Benjamin Hawks[2], Nhan Tran[2], Javier Duarte[1]

[1]University of California San Diego, [2]Fermi National Accelerator Laboratory

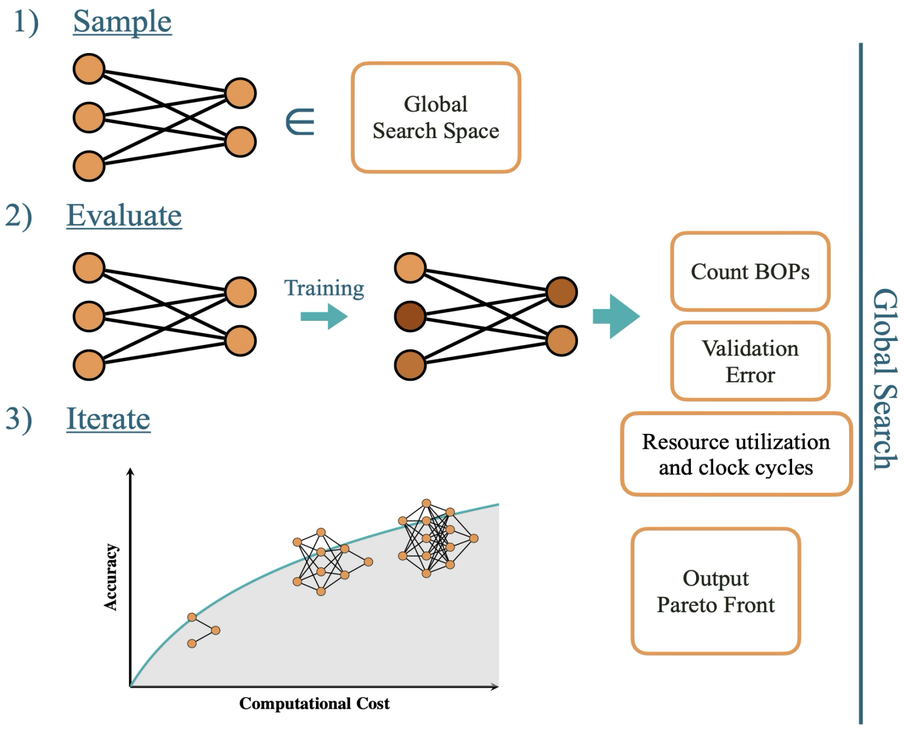
FERMILAB-POSTER-25-0265-CSAID

Introduction

- Neural Architecture Search is a powerful approach for automating model design, but existing methods struggle to accurately optimize for real hardware performance, often relying on proxy metrics such as bit operations.
- We present Surrogate Neural Architecture Codesign Package (SNAC-Pack), an integrated framework that automates the discovery and optimization of neural networks focusing on FPGA deployment.

Method

•The SNAC-Pack tool builds upon NAC by introducing the additional objectives that can be estimated with rule4ml.

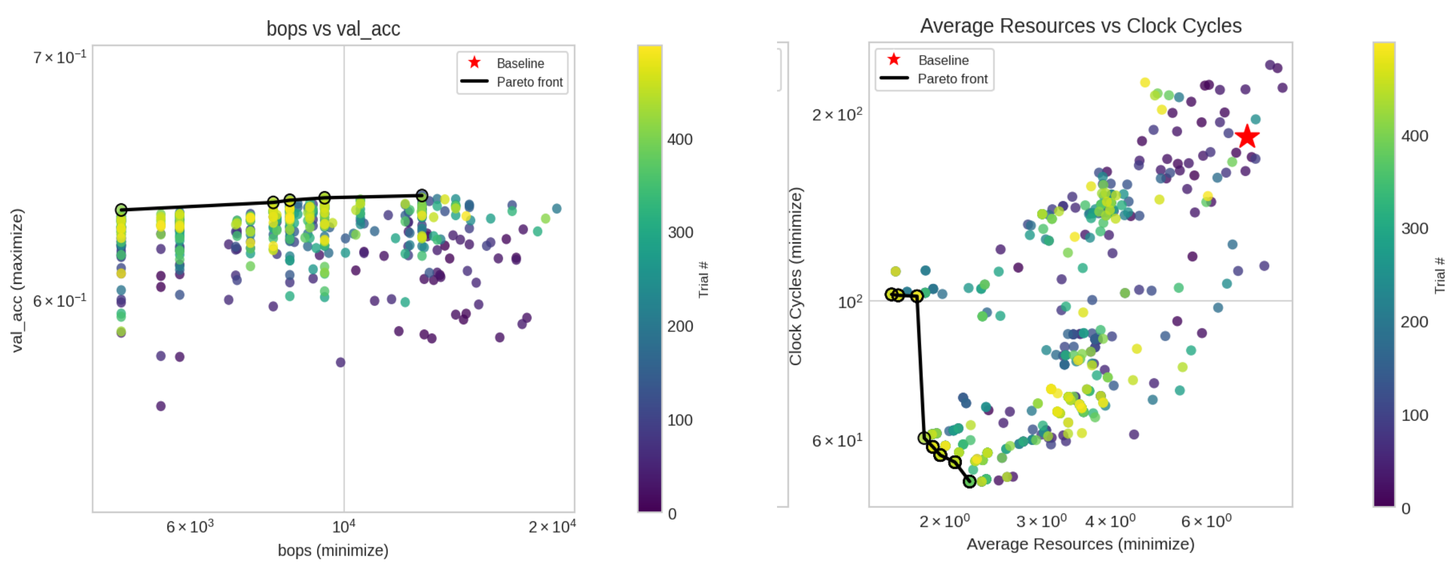


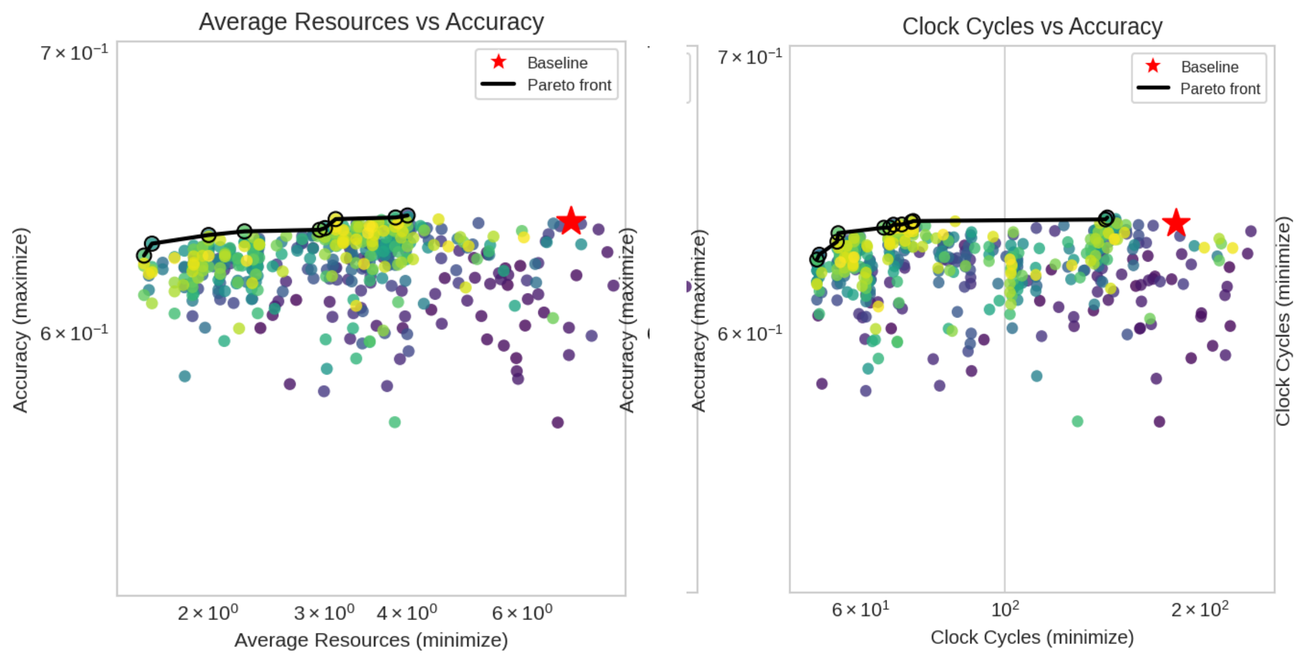
- With NAC integration, the global search stage begins with a user-defined search space that specifies the range of possible architectures, such as the types of layers, number of neurons, activations, and other hyperparameters.
- A multi-objective search algorithm then explores this space, samples candidate architectures, and performs evaluation.
- All metrics estimated by rule4ml, BOPs, and accuracy can be used as objectives in the search.

Parameter	Space
Number of layers	{4, 5, 6, 7, 8}
Hidden units per layer	
Layer 1	{64, 120, 128}
Layer 2	{32, 60, 64}
Layer 3	{16, 32}
Layer 4	{32, 64}
Layer 5	{32, 64}
Layer 6	{32, 64}
Layer 7	{16, 32}
Layer 8	{32, 44, 64}
Activation function	{ReLU, Tanh, Sigmoid}
Batch normalization	{True, False}
Learning rate	{0.0010, 0.0015, 0.0020}
L1 regularization	{0.0, 10^{-6}, 10^{-5}, 10^{-4}}
Dropout rate	{0.0, 0.05, 0.1}

Jet Classification Implementation

• To show the effectiveness of SNAC-Pack, we apply it to jet classification, a common and challenging task in high energy physics at the Large Hadron Collider (LHC). The goal is to accurately classify collision-created jets into one of five categories (light quark, gluon, W boson, Z boson, top quark) based on their kinematic properties. This is showcased with the hls4ml LHC dataset.





Model	Accuracy [%]	BOPs	Est. average resources	Est. clock cycles
Baseline [12]	63.77	25,916	7.10	183.74
Optimal NAC [1]	63.81	**7,904**	3.60	**62.69**
Optimal SNAC-Pack	**63.84**	8,352	**3.12**	72.24

Model	Lat. [ns] (cc)	II [ns] (cc)	DSP	LUT	FF	BRAM
Baseline [12]	**105 (21)**	**5** (1)	262 (2.1%)	155080 (9.0%)	25714 (0.7%)	4 (0.1%)
Optimal NAC [1]	125 (25)	60 (12)	**0**	**54075 (3.13%)**	**12016 (0.35%)**	8 (0.3%)
Optimal SNAC-Pack	140 (24)	60 (12)	**0**	57728 (3.34%)	12605 (0.36%)	**0**

Conclusion

•This work introduced the Surrogate Neural Architecture Codesign Package (SNAC-Pack), a framework that extends NAC by incorporating resource-aware objectives estimated with rule4ml.

• Applied to the jet classification task, the optimal model produced by SNAC-Pack performed comparably to the NAC method and baseline reference.

Acknowledgements

This manuscript has been authored by Fermi Forward Discovery Group, LLC under Contract No. 89243024CSC000002 with the U.S. Department of Energy, Office of Science, Office of High Energy Physics. BH and NT are supported by Fermi Research Alliance, LLC under Contract No. DE-AC02-07CH11359 with the United States Department of Energy (DOE), Office of Science, Office of High Energy Physics. BH and NT are also supported under the DOE Early Career Research program under Award No. DE-0000247070. KT is also supported by DOE Grant KA2401045. BH, JD, and NT are supported by the U.S. Department of Energy (DOE), Office of Science, Office of Advanced Scientific Computing Research under the ``Real-time Data Reduction Codesign at the Extreme Edge for Science" Project(DE-FOA-0002501). JD is also supported by the DOE, Office of Science, Office of High Energy Physics Early Career Research program under Grant No. DE-SC0021187, and the U.S. National Science Foundation (NSF) Harnessing the Data Revolution (HDR) Institute for Accelerating AI Algorithms for Data Driven Discovery (A3D3) under Cooperative Agreement No. PHY-2117997. JW is supported by a WATCHEP fellowship sponsored by the DOE, Office of High-Energy Physics under Award No. DE-SC-0023527.

Surrogate Neural Architecture Codesign Package (SNAC-Pack)

Jason Weitz[1], Dmitri Demler[1], Benjamin Hawks[2], Nhan Tran[2], Javier Duarte[1]

[1]University of California San Diego, [2]Fermi National Accelerator Laboratory

UC San Diego

Fermilab

Introduction

Introduction

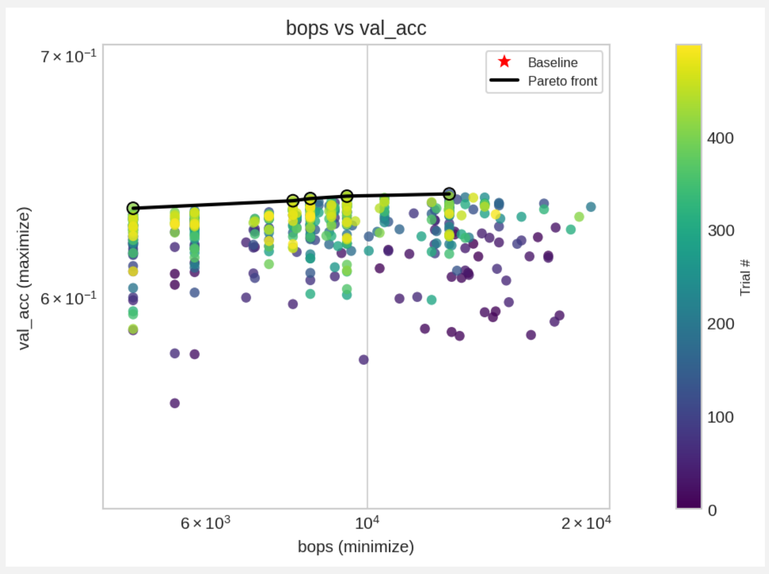


Table 1: comprehensive parameter space for multilayer perceptron (mlp) search

Parameter	Space
Number of layers	{4, 5, 6, 7, 8}
Hidden units per layer	
Layer 1	{64, 120, 128}
Layer 2	{32, 60, 64}
Layer 3	{16, 32}
Layer 4	{32, 64}
Layer 5	{32, 64}
Layer 6	{32, 64}
Layer 7	{16, 32}
Layer 8	{32, 44, 64}
Activation function	{ReLU, Tanh, Sigmoid}
Batch normalization	{True, False}
Learning rate	{0.0010, 0.0015, 0.0020}
L1 regularization	$\{0.0, 10^{-6}, 10^{-5}, 10^{-4}\}$
Dropout rate	{0.0, 0.05, 0.1}

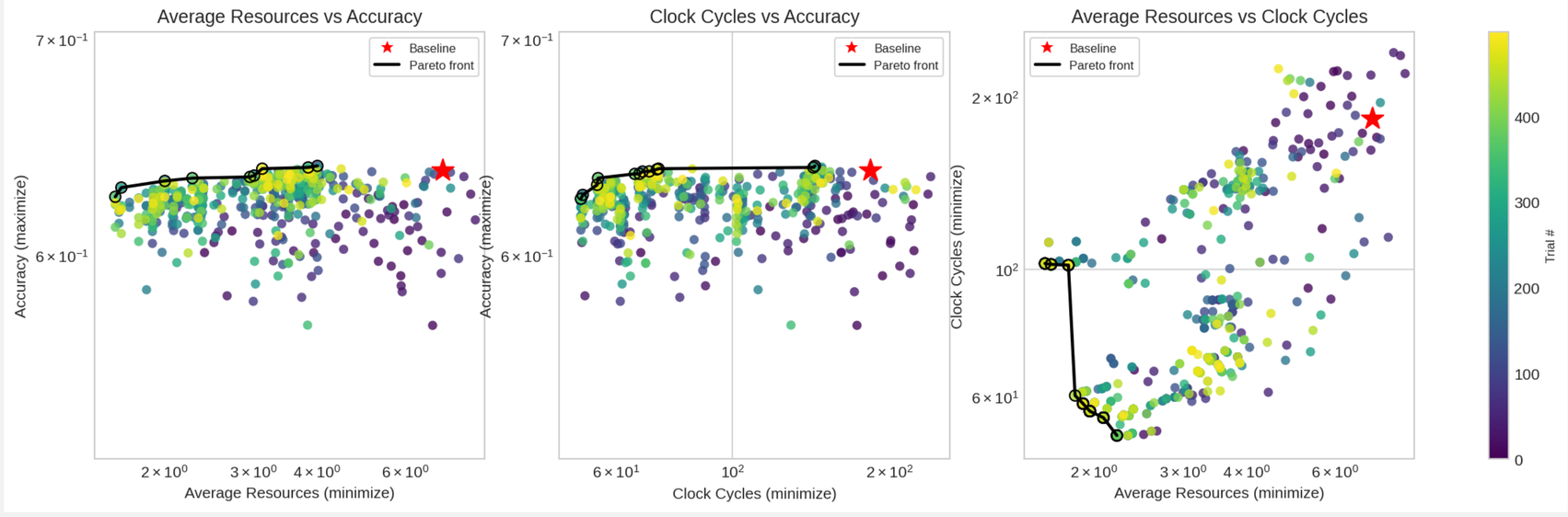


Table 2: Comparison of model accuracy, BOPs, and estimated hardware metrics from global search. Note that while all metrics are reported here for consistency, the Baseline was optimized for accuracy, NAC for accuracy and BOPs, and SNAC-Pack for accuracy, estimated average resources and clock cycles. The best values are reported in bold.

Model	Accuracy [%]	BOPs	Est. average resources	Est. clock cycles
Baseline [12]	63.77	25,916	7.10	183.74
Optimal NAC [1]	63.81	**7,904**	3.60	**62.69**
Optimal SNAC-Pack	**63.84**	8,352	**3.12**	72.24

Table 3: Hardware resource utilization and latency estimates for the selected models. cc is the number of clock cycles. The best values are reported in bold.

Model	Lat. [ns] (cc)	II [ns] (cc)	DSP	LUT	FF	BRAM
Baseline [12]	**105 (21)**	**5** (1)	262 (2.1%)	155080 (9.0%)	25714 (0.7%)	4 (0.1%)
Optimal NAC [1]	125 (25)	60 (12)	**0**	**54075 (3.13%)**	**12016 (0.35%)**	8 (0.3%)
Optimal SNAC-Pack	140 (24)	60 (12)	**0**	57728 (3.34%)	12605 (0.36%)	**0**