

AutoHLS: Learning to Accelerate Design Space Exploration for HLS Designs

Ahmed, Md Rubel; Koike-Akino, Toshiaki; Parsons, Kieran; Wang, Ye

TR2023-097 August 08, 2023

Abstract

High-level synthesis (HLS) is a design flow that leverages modern language features and flexibility, such as complex data structures, inheritance, templates, etc., to prototype hardware designs rapidly. However, exploring various design space parameters can take much time and effort for hardware engineers to meet specific design specifications. This paper proposes a novel framework called AutoHLS, which integrates a deep neural network (DNN) with Bayesian optimization (BO) to accelerate HLS hardware design optimization. Our tool focuses on HLS pragma exploration and operation transformation. It utilizes integrated DNNs to predict synthesizability within a given FPGA resource budget. We also investigate the potential of emerging quantum neural networks (QNNs) instead of classical DNNs for the AutoHLS pipeline. Our experimental results demonstrate up to a 70-fold speedup in exploration time.

International Midwest Symposium on Circuits and Systems (MWSCAS) 2023

© 2023 MERL. This work may not be copied or reproduced in whole or in part for any commercial purpose. Permission to copy in whole or in part without payment of fee is granted for nonprofit educational and research purposes provided that all such whole or partial copies include the following: a notice that such copying is by permission of Mitsubishi Electric Research Laboratories, Inc.; an acknowledgment of the authors and individual contributions to the work; and all applicable portions of the copyright notice. Copying, reproduction, or republishing for any other purpose shall require a license with payment of fee to Mitsubishi Electric Research Laboratories, Inc. All rights reserved.

AutoHLS: Learning to Accelerate Design Space Exploration for HLS Designs

Md Rubel Ahmed, Toshiaki Koike-Akino, Kieran Parsons, Ye Wang
Mitsubishi Electric Research Laboratories (MERL), 201 Broadway, Cambridge, MA 02139, USA.
{mdahmed, koike, parsons, yewang}@merl.com

Abstract—High-level synthesis (HLS) is a design flow that leverages modern language features and flexibility, such as complex data structures, inheritance, templates, etc., to prototype hardware designs rapidly. However, exploring various design space parameters can take much time and effort for hardware engineers to meet specific design specifications. This paper proposes a novel framework called AutoHLS, which integrates a deep neural network (DNN) with Bayesian optimization (BO) to accelerate HLS hardware design optimization. Our tool focuses on HLS pragma exploration and operation transformation. It utilizes integrated DNNs to predict synthesizability within a given FPGA resource budget. We also investigate the potential of emerging quantum neural networks (QNNs) instead of classical DNNs for the AutoHLS pipeline. Our experimental results demonstrate up to a 70-fold speedup in exploration time.

Index Terms—HLS acceleration, design space exploration, optimization, design automation, FPGA

I. INTRODUCTION

HLS is a widely used rapid design and prototyping method in industry and academia. Still, it poses several challenges for source code optimization due to the rich features of modern programming languages such as C/C++. Careless optimization can result in inefficient and resource-hungry designs with high latency or, in some cases, loss of synthesizability under a reasonable FPGA resource budget. HLS compilers such as Vitis [1] offer optimization tactics such as pragma directives and timing/closure analysis to tackle these issues which have spurred active research areas in design-space exploration (DSE) for HLS. Accelerated DSE is required since downstream tools used for RTL generation, such as Vitis [1], can take significant time to compile and report synthesis results. This limits the number of designs evaluated during DSE, resulting in sub-optimal solutions. Besides, the time required for RTL generation can increase the DSE time from hours to days, depending on the complexity of the design. The quest for faster and more efficient DSE in HLS has led to the development of machine learning (ML) and analytical methods. In this context, ScaleHLS [2] presents an analytical approach that leverages a Quality-of-Results (QoR) estimator to accelerate the DSE process. By statically analyzing code blocks and modeling latency and resource utilization, the QoR estimator enables ScaleHLS’s DSE engine to explore the design space efficiently and converge to the Pareto front faster. Other methods [2]–[7] use statistical, heuristic, ML, or meta-learning approaches to accelerate DSE. For instance, using an ML model, Pyramid [8] estimates the maximum achievable throughput. At

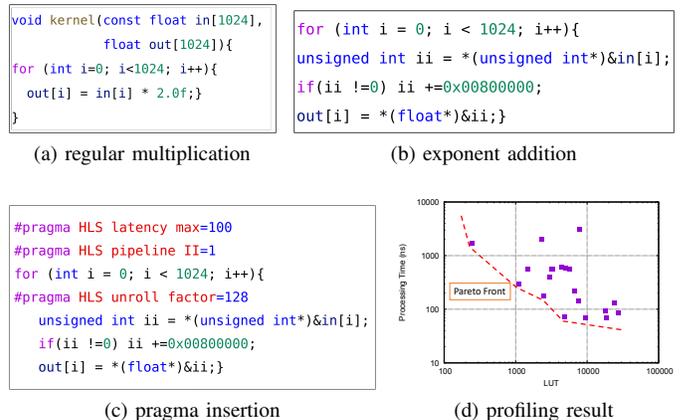


Fig. 1: Array doubling kernels having functionally equivalent operations but different hardware profiles.

the same time, a recent work [9] predicts resource usage for synthesizing convolutional neural networks. Sherlock [10], another DSE tool, uses active learning with a surrogate model to find Pareto front, highlighting the challenges in handling conflicting objectives in parameter optimization. We consider Optuna [11], a Bayesian Optimization (BO) framework, as a baseline multi-objective optimization tool. BO is generally slow to find the Pareto front as the downstream HLS flow takes much time to generate QoR for each sample design point. Therefore, we add an early failure prediction network with the BO to accelerate the DSE. To the best of our knowledge, no current works focus on reducing the search space based on synthesizability constraints, such as FPGA footprints (DSP, FF, LUT) or synthesis time budget.

Our proposed method, AutoHLS, optimizes the design by considering synthesizability constraints as a multi-objective optimization problem. AutoHLS efficiently determine loop unrolling factor, pipeline depth, array partition, etc., for pragma installments in order to optimize HLS designs considering signal processor (DSP), flip-flop (FF), look-up table (LUT), power consumption, and latency. Furthermore, AutoHLS also includes kernel operations transformation to further optimize the designs. The **contributions** of this work are as follows.

- We reveal that existing multi-objective optimization tools can fail to meet the budget-centric design approach.
- We propose AutoHLS framework to accelerate the DSE using ML models.
- A novel QNN model is employed to predict synthesis

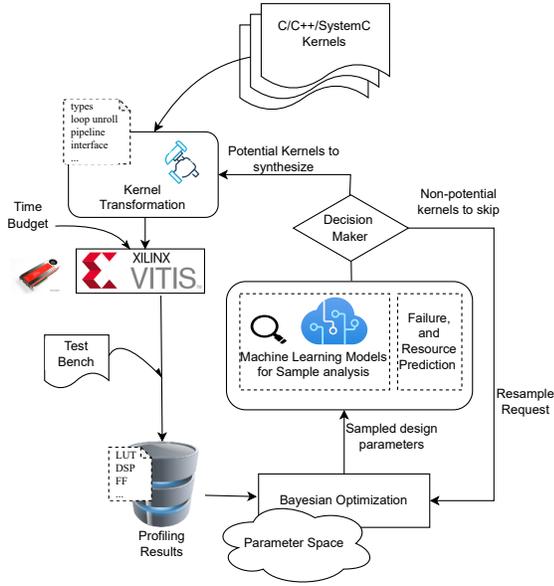


Fig. 2: Overview of AutoHLS.

failure and resource usage accurately.

II. AUTOHLS FOR EFFICIENT HARDWARE DESIGN

In designing hardware in HLS, several critical factors must be considered, such as the target device, available resources, required precision level, simulation, synthesis, co-simulation time, etc. In this regard, Fig. 1 (a) presents a regular array doubling kernel in C++ as an example. However, HLS provides several alternative implementations, such as Fig. 1 (b), where a functionally equivalent exponent addition replaces the multiplication operation. Additionally, Fig. 1 (c) shows a more optimized implementation that utilizes pragma insertion. The synthesis profiling results over different pragma factors and kernel operation transforms, shown in Fig. 1 (d), demonstrate a tradeoff behavior in multi-objective optimization, where the reduction in LUT resources and runtime would compete. Nevertheless, due to the high degree of flexibility in pragma installment and kernel operation transforms, finding an optimal Pareto front in constrained development time remains challenging. AutoHLS, as depicted in Fig. 2, takes an unoptimized kernel and efficiently explores different design alternatives to meet design objectives such as runtime, precision level, DSP, FF, LUT, etc., usage. It can discover an optimal set of pragma and kernel operation transforms with the help of an ML-based synthesizability prediction mechanism.

A. Scope and Definition

1) *Pragma Selection*: Pragma and their parameters guide the HLS compiler toward optimal designs. For example, AutoHLS uses a categorical sampling of BO to decide the set of HLS pragma insertions $P_K \subseteq \mathbf{P}$, where \mathbf{P} includes pipeline, unroll, etc.

2) *Pragma Parameter Selection*: Each HLS pragma P can have a set of parameters \mathbf{A}_P . Given a kernel K , AutoHLS decides a parameter set $A_K \subseteq \{\mathbf{A}_P\}$ for each HLS pragma

$P \in P_K$ in the selection, using BO sampling. For example, Fig. 1 (c) uses the parameters set of $A_K = \{100, 1, 128\}$ for the pragma set P_K .

3) *Kernel/Operation Transformation*: HLS synthesis tools often utilize high-cost resources, such as DSP blocks, to meet high throughput requirements, which may not be available for resource-constrained applications like edge/embedded devices. Therefore, considering alternative operations that can save resources at a potential cost of throughput or precision. For example, a regular multiplication kernel in Fig.1 (a) can be functionally equivalent to an exponent addition kernel in Fig.1 (b) for a floating-point operation when the multiplicand is a power-of-two (PoT) value. Furthermore, simplifications can be achieved by reducing bit-width precision and using fixed-point operations with bit-shifting. Given an HLS kernel K , *kernel/operation transformation* produces another kernel K_T such that the outputs from both kernels are almost equivalent or exactly equivalent within a specified tolerance range. In addition, recent green ML models have also demonstrated that quantized DNNs, such as DeepShift [12], can outperform floating-point DNNs. Therefore, we explore PoT and additive-Power-of-Two (APoT) quantization for further optimization.

B. AutoHLS Flow

AutoHLS explores both kernel and parameter space. Given a set of kernels K , an objective function, and an HLS design constraint, AutoHLS analyzes the kernels and returns a set of optimal synthesizable kernels for the given objectives that meet the design constraint.

1) *Kernel Transformation*: AutoHLS first parses the input C/C++ kernels and constructs pragmas using the selected set \mathbf{P} , which includes the pipeline, unroll, latency, array partition, etc. These kernels are then checked for feasibility before being synthesized.

2) *Kernel Synthesis*: The transformed kernel is synthesized using standard HLS tools (e.g., Vitis) with pre-set device-specific parameters for FPGA. The synthesis process involves functional correctness checking with `csim` and feasibility checking with `synth`.

3) *Kernel Profiling*: After the synthesis step, the Quality of Results (QoR), kernel type, and pragma parameters are collected. The synthesis can be complete or fail for the given constraint. These data are utilized directly or indirectly in the objective function.

4) *Bayesian Optimization*: AutoHLS adopts the BO method based on a tree-structured Parzen estimator (TPE) [11] for DSE, which can handle multi-objective optimization. The TPE-based optimizer suggests a set of optimized design parameters from the parameter space based on an acquisition function for efficient Pareto optimization.

5) *Decision Maker*: AutoHLS tool incorporates machine learning techniques to predict the synthesis failure and estimate the resource utilization of the designed kernel. Specifically, DNN and QNN provide the failure prediction scores on each sample set generated by the BO. Based on the prediction results, the tool decides whether to synthesize or discard the

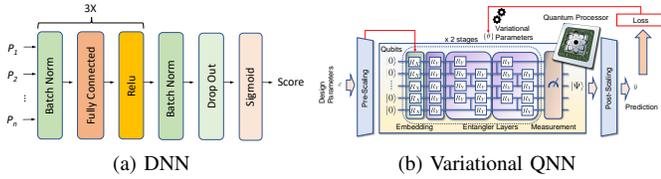


Fig. 3: Failure/resource prediction models.

kernel and move to the next one. This approach enables accelerated design space exploration and reduces the overall design time.

C. ML models for Sample Analysis

AutoHLS employs ML models to predict synthesis failure and estimate the resource profile of a design. These models, including classifiers and regression models, are trained on the already explored samples and assign a score to a new sample generated by BO. A decision is then made based on a threshold τ . Finally, the sample is sent for synthesis only if it passes the decision-maker.

1) *DNN*: We propose a DNN model shown in Fig. 3 (a), for predicting a design’s synthesizability score and resource usage. The model takes design parameters as input and consists of three batch normalization layers, a fully-connected layer, and a ReLU activation. A batch normalization layer, a dropout, and a sigmoid are applied at the end. The model has 3243 trainable parameters and is designed to learn with limited training samples, which is essential in DSE due to the long synthesis time of HLS tools.

2) *QNN*: The recent advancements in quantum technology have led to the availability of high-qubit processors, such as the 433-qubit processors released by IBM in 2022. This has given rise to a new paradigm of ML models known as QNNs, which have the universal approximation property [13] and are more compact than modern DNNs. We propose a proof-of-concept evaluation of QNNs for HLS acceleration. First, we present a QNN architecture shown in Fig. 3 (b) with only 54 trainable parameters. It has five quantum bits.

3) *Classical ML Algorithms*: We evaluate various classical ML models, including SVM and LR for failure prediction and linear regression, lasso, KRR, and Bayesian ridge regression for hardware profile prediction.

III. AUTOHLS VALIDATION

Our experiments are performed on a machine with an Intel® Core™ i7-8700K CPU @ 3.70GHz and 64GB of main memory, running on Ubuntu 20.04.5 LTS. The Xilinx ZCU104 board is used as a target FPGA, and Vitis HLS 2022.1 is used for kernel synthesis.

A. Problem Setup

We investigate the effectiveness of AutoHLS for the DSE of a CNN block. We consider synthesis time t as a design resource budget or constraint. The CNN block comprises a window size L , an input channel C_{in} , and an output channel C_{out} , where the convolution operation involves element-wise

TABLE I: Resource usage of convolution kernels

Kernel	FF	LUT	DSP	Latency	MSE
MAC	40922	17761	5	3072	-
MAC<16, 6>	24784	8650	1	1352	5.09e-06
PoT	42396	18067	4	4533	3.78
PoT<16, 6>	24893	6947	0	925	3.78
APoT	45207	18593	4	4952	0.019
APoT<16, 6>	26061	7090	0	1039	0.019

multiplication and accumulation of the window and input channel elements. Table I provides the area utilization of the conventional multiplier-based implementation. MAC stands for multiplication and accumulation-based convolution. In Table I, we present the QoR results for $C_{in} = 100$, $L = 7$, $C_{out} = 106$, and `float32` as the datatype. The table shows different types of kernels, such as PoT<16, 6> and APoT<16, 6>, which are arbitrary precision (ap) fixed-point data types. The results reveal kernel transformation significantly impacts the hardware footprint. However, kernel transformation may cause some loss in precision, which MSE indicates. Additionally, the ap-type PoT has the lowest resource usage but higher MSE than APoT, which has better MSE but consumes more area than PoT. Finding the optimal kernel requires DSE to determine the pragma and appropriate pragma parameters. We conduct a case study on two kernels, PoT and APoT that entirely eliminates the multipliers. We then evaluate the performance of the conventional BO method and subsequently employ AutoHLS for further optimization.

1) *Quantizations*: We use PoT and APoT quantizations as kernel transformation schemes to create hardware-friendly designs of green ML models [12]. A regular MAC with W as a weight, b as the bias: $y = Wx + b$; the PoT quantization of weight, W , $u \in \mathbb{Z}$: $W = \pm 2^u$; and APoT quantization of weight, W : $W = \pm 2^u \pm 2^v$, where $u, v \in \mathbb{Z}$ and $v < u$.

2) *Bayesian Optimization for DSE*: To investigate the performance of BO on parameter optimization, the kernels are instrumented with four pragmas: unroll factor, pipeline instantiation interval, latency max and min.

Table II presents the performance evaluation of BO on the exploration process. The column ‘Time’ indicates the synthesis time budget in minutes. The columns ‘Comp.’ and ‘Fail’ denote the number of samples for which the kernel synthesis succeeded and failed, respectively. The exploration involves 3302 designs, taking 4 to 6 minutes to complete, regardless of the synthesis status. However, most of the parameters suggested by BO failed to synthesize, resulting in a vain attempt to synthesize the wrong design. To address this problem, AutoHLS leverages an early failure prediction mechanism.

B. Training and Validation

We generate 3302 convolution design points with BO, and 961 of them are synthesizable within the given time budget. Each sample has five independent variables, one dependent variable, and a kernel identifier. We use all samples to train classification models and synthesizable samples to train regression models. Classification models predict the sample outcome

TABLE II: Kernel synthesis using BO with a given time budget

Kernel	Time (min.)	Comp.	Fail	Comp. + Fail	%Fail	% Comp.
APoT	2.00	14	194	208	93.26	6.73
	2.20	12	289	301	96.01	3.98
	2.50	20	480	500	96.00	4.00
	2.75	11	389	400	97.25	2.75
	3.00	42	551	593	92.76	7.07
PoT	1.50	19	381	400	95.25	4.75
	1.75	364	36	400	9.00	94.00
	2.00	291	9	300	3.00	97.00
	2.20	188	12	200	6.00	94.00
Total		961	2341	3302	70.90	29.10

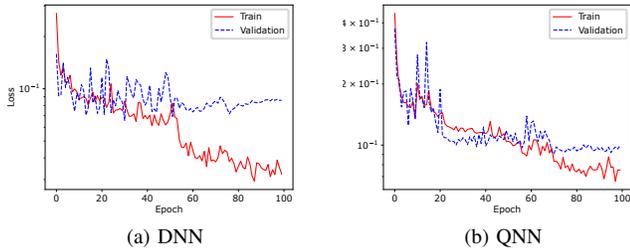


Fig. 4: ML model training: Cross-entropy loss over epoch.

TABLE III: AutoHLS parameter search with failure prediction.

τ	samples	TP	FP	%TP	BO hrs.	AutoHLS hrs.	Speedup
0.95	2000	48	4	2.5	~333	~8	~38
0.85	2000	23	4	1.15	~333	~4.5	~74
0.75	200	14	1	7.0	~33	~2.5	~14

and are used as early failure prediction models. Regression models predict FPGA resource usage. We show the model training process over 100 epochs and the corresponding loss in Fig. 4. Our models converge quickly on the training data. Fig. 5 (b) demonstrates that our models can learn from a small number of training samples and achieve high accuracy on the test data.

Proposed models are validated under various conditions. The results show high true positive rates in the ROC curve, as demonstrated in Fig. 5 (a). The models’ robustness and generalization capabilities are also confirmed. They still achieve high true positive rates even when trained on only 5% of the samples. Proposed DNN and QNN models outperform classical regression methods, as shown in Fig. 6 (a). Regarding Pareto fronts, AutoHLS outperforms BO as highlighted in the blue dotted line in Fig. 6 (b).

We evaluate the effectiveness of the proposed early failure prediction model by running a pragma parameter exploration for the APoT kernel. The estimated time for each design point synthesis is ten minutes. Table III shows the results for different threshold values τ , demonstrating a speedup in synthesizable design exploration time ranging from 15 to 74 times faster when using the failure prediction model.

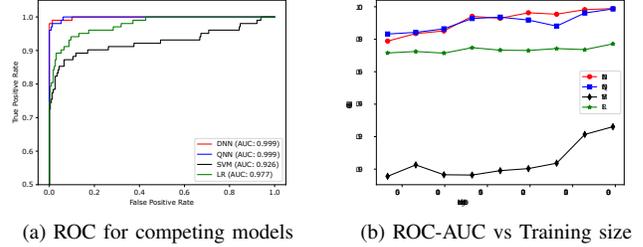


Fig. 5: ML model Accuracy of Training data size

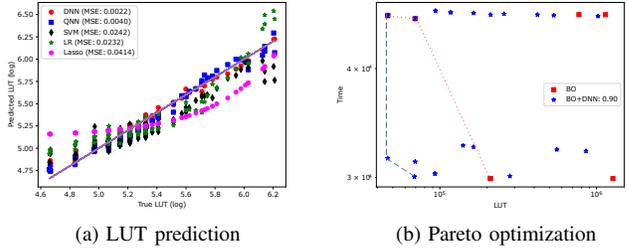


Fig. 6: FPGA synthesis results using different models.

C. Discussion

An important concern regarding AutoHLS is its generality. Future research can experiment with unseen designs to evaluate the generalizability of this framework. Improvements can also be made by leveraging the vast amount of open-source FPGA synthesis data available in DB4HLS [14], which contains more than 100,000 design points. Our experiments with the CNN kernel demonstrate AutoHLS’s efficacy, even with an imbalanced training set. The low false positive rate achieved by AutoHLS indicates that the machine learning models can learn effectively. We consider synthesizability within a given time budget and note that early failure prediction could be possible for other metrics, such as DSP and clock cycle numbers. Due to the nature of HLS synthesis data, AutoHLS can learn from a small number of training data. Finally, we suggest exploring multi-objective reinforcement learning methods to enhance the robustness of this framework.

IV. CONCLUSION

This paper presents AutoHLS, a framework for accelerating DSE for HLS using DNN/QNN-enabled multi-objective BO. It addresses the shortcomings of BO in HLS optimization. Furthermore, it provides resource prediction mechanisms and faster exploration of the Pareto front. It demonstrates the effectiveness of this framework in achieving specific design goals through accelerated DSE and kernel operation transformation. Our experiments significantly speed up finding optimal FPGA design parameters for the CNN kernel.

REFERENCES

[1] Xilinx. Vitis high-level synthesis user guide (ug1399). <https://docs.xilinx.com/r/en-US/ug1399-vitis-hls/HLS-Pragmas>, February 2022.

- [2] Hanchen Ye, HyeGang Jun, Hyunmin Jeong, Stephen Neuendorffer, and Deming Chen. Scalehls: A scalable high-level synthesis framework with multi-level transformations and optimizations: Invited. In *Proceedings of the 59th ACM/IEEE Design Automation Conference, DAC '22*, page 1355–1358, New York, NY, USA, 2022. Association for Computing Machinery.
- [3] Guyue Huang, Jingbo Hu, Yifan He, Jialong Liu, Mingyuan Ma, Zhaoyang Shen, Juejian Wu, Yuanfan Xu, Hengrui Zhang, Kai Zhong, Xuefei Ning, Yuzhe Ma, Haoyu Yang, Bei Yu, Huazhong Yang, and Yu Wang. Machine learning for electronic design automation: A survey. *ACM Transactions on Design Automation of Electronic Systems*, 26(5):1–46, Jun 2021.
- [4] Atefeh Sohrabizadeh, Yunsheng Bai, Yizhou Sun, and Jason Cong. Automated accelerator optimization aided by graph neural networks. In *Proceedings of the 59th ACM/IEEE Design Automation Conference, DAC '22*, page 55–60, New York, NY, USA, 2022. Association for Computing Machinery.
- [5] HyeGang Jun, Hanchen Ye, Hyunmin Jeong, and Deming Chen. Autoscaledse: A scalable design space exploration engine for high-level synthesis. *ACM Trans. Reconfigurable Technol. Syst.*, feb 2023. Just Accepted.
- [6] Brooks Olney, Shakil Mahmud, Md Adnan Zaman, and Robert Karam. An eda framework for design space exploration of on-chip ai in bioimplantable applications. In *2022 IEEE 65th International Midwest Symposium on Circuits and Systems (MWSCAS)*, pages 1–4, 2022.
- [7] Farah Fahim, Benjamin Hawks, Christian Herwig, James Hirschauer, Sergio Jindariani, Nhan Tran, Luca P. Carloni, Giuseppe Di Guglielmo, Philip Harris, Jeffrey Krupa, Dylan Rankin, Manuel Blanco Valentin, Josiah Hester, Yingyi Luo, John Mamish, Seda Orgreneci-Memik, Thea Aarrestad, Hamza Javed, Vladimir Loncar, Maurizio Pierini, Adrian Alan Pol, Sioni Summers, Javier Duarte, Scott Hauck, Shih-Chieh Hsu, Jennifer Ngadiuba, Mia Liu, Duc Hoang, Edward Kreinar, and Zhenbin Wu. hls4ml: An open-source codesign workflow to empower scientific low-power machine learning devices, 2021.
- [8] H. Mohammadi Makrani, F. Farahmand, H. Sayadi, S. Bondi, S. Pudukotai Dinakarrao, H. Homayoun, and S. Rafatirad. Pyramid: Machine learning framework to estimate the optimal timing and resource usage of a high-level synthesis design. In *2019 29th International Conference on Field Programmable Logic and Applications (FPL)*, pages 397–403, Los Alamitos, CA, USA, sep 2019. IEEE Computer Society.
- [9] Pingakshya Goswami, Masoud Shahshahani, and Dinesh Bhatia. Robust estimation of fpga resources and performance from cnn models. In *2022 35th International Conference on VLSI Design and 2022 21st International Conference on Embedded Systems (VLSID)*, pages 144–149, 2022.
- [10] Quentin Gautier, Alric Althoff, Christopher L. Crutchfield, and Ryan Kastner. Sherlock: A multi-objective design space exploration framework. *ACM Trans. Des. Autom. Electron. Syst.*, 27(4), mar 2022.
- [11] Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2019.
- [12] Toshiaki Koike-Akino, Ye Wang, Keisuke Kojima, Kieran Parsons, and Tsuyoshi Yoshida. Zero-multiplier sparse dnn equalization for fiber-optic qam systems with probabilistic amplitude shaping. In *2021 European Conference on Optical Communication (ECOC)*, pages 1–4, 2021.
- [13] Adrián Pérez-Salinas, Alba Cervera-Lierta, Elies Gil-Fuster, and José I. Latorre. Data re-uploading for a universal quantum classifier. *Quantum*, 4:226, February 2020.
- [14] Lorenzo Ferretti, Jihye Kwon, Giovanni Ansaloni, Giuseppe Di Guglielmo, Luca Carloni, and Laura Pozzi. Db4hls: A database of high-level synthesis design space explorations. *IEEE Embedded Systems Letters*, 13(4):194–197, 2021.