U-Boost NAS: Utilization-Boosted Differentiable Neural Architecture Search

Ahmet Caner Yüzügüler Nikolaos Dimitriadis♯ Pascal Frossard

EPFL

European Conference on Computer Vision 2022

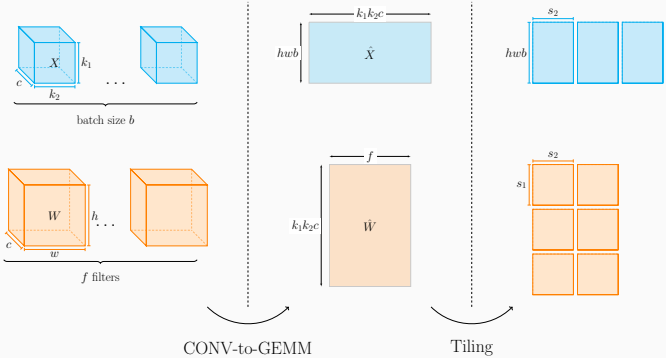
Tel Aviv, October 23-27, 2022

♯: presenter

Motivation

- inference latency is crucial in resource-constraint settings
- current DNN models are underutilizing resources
- no prior work optimizes for hardware utilization

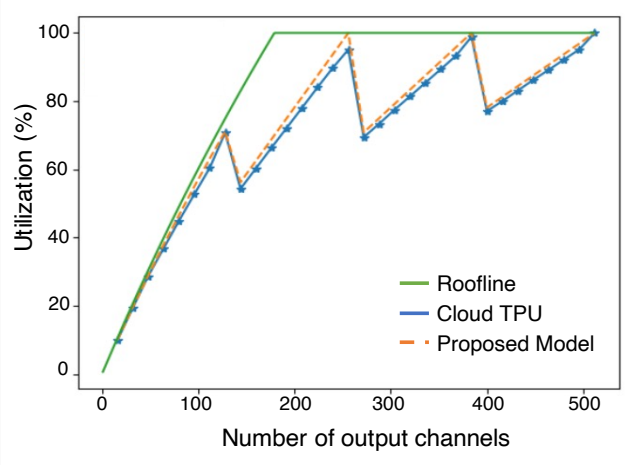
Modeling Resource Utilization in Inference Platforms



$$\texttt{RUNTIME} = \underbrace{\left\lceil \frac{k_1 k_2 c}{s_1} \right\rceil}_{\text{(\# vertical tiles)}} \times \underbrace{\left\lceil \frac{f}{s_2} \right\rceil}_{\text{(\# horizontal tiles)}} \times \underbrace{hwb}_{\text{(\# rows in } \hat{X})}$$

$$\texttt{UTIL} = \frac{\text{Avg throughput}}{\text{Peak Throughput}} = \frac{\frac{\#\text{MACs}}{\texttt{RUNTIME}}}{s_1 s_2} = \frac{k_1 k_2 c f}{s_1 s_2 \left\lceil \frac{k_1 k_2 c}{s_1} \right\rceil \left\lceil \frac{f}{s_2} \right\rceil}$$

Utilization in the real-world

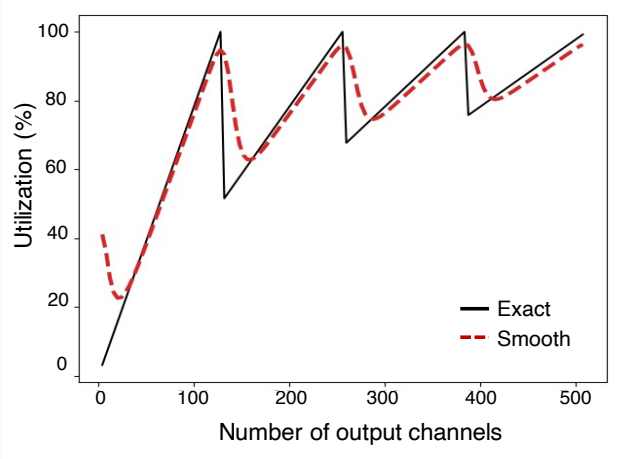


$$\texttt{UTIL} = \frac{k_1 k_2 c f}{s_1 s_2 \left\lceil \frac{k_1 k_2 c}{s_1} \right\rceil \left\lceil \frac{f}{s_2} \right\rceil} \xrightarrow{f = s_2} \frac{k_1 k_2 c}{s_1 \left\lceil \frac{k_1 k_2 c}{s_1} \right\rceil} = \begin{cases} 1, & k_1 k_2 c = s_1 \\ 0.5, & k_1 k_2 c = s_1 + 1 \end{cases}$$

Proposed Method

Smooth Approximation of ceiling function

$$\mathtt{CEIL}_{smooth}(x) = \sum_i \left[1 + \frac{\exp\left(-B(x - w_i)\right)}{C}\right]^{-1/v}$$

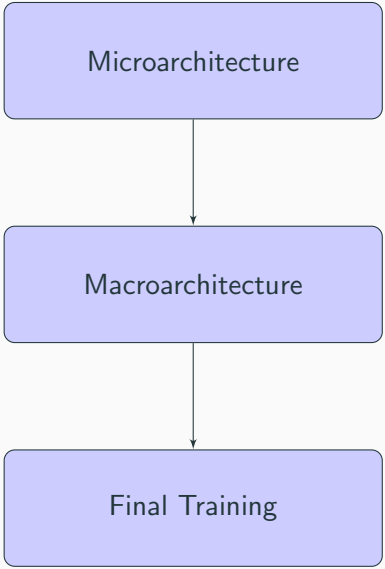


Multi-objective loss function

Let $\mathcal{F}$ be the hypothesis class of the search space and $\alpha \in \mathcal{F}$ the candidate architecture defining the function $f_\alpha : \mathcal{X} \to \mathcal{Y}$ for input and output domains $\mathcal{X}$ and $\mathcal{Y}$:

$$\mathcal{L}(\boldsymbol{x}, y, \alpha) = \mathcal{L}_{classification}(f_\alpha(\boldsymbol{x}), y) + \lambda \cdot \mathcal{L}_{latency}(\alpha) - \beta \cdot \mathcal{L}_{utilization}(\alpha)$$

Hierarchical three-stage Neural Architecture Search



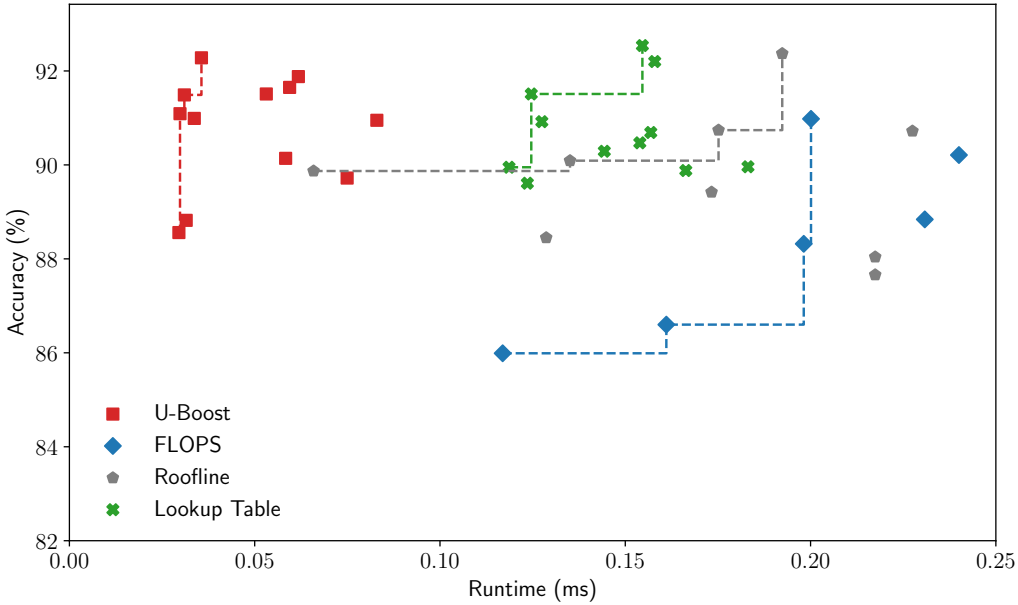
layer types and connections with single-cell model (fixed channel dims)

optimal channel dims search cell-wise for model with k sequential cells

train the selected architecture $\alpha \in \mathcal{F}$

Experiments

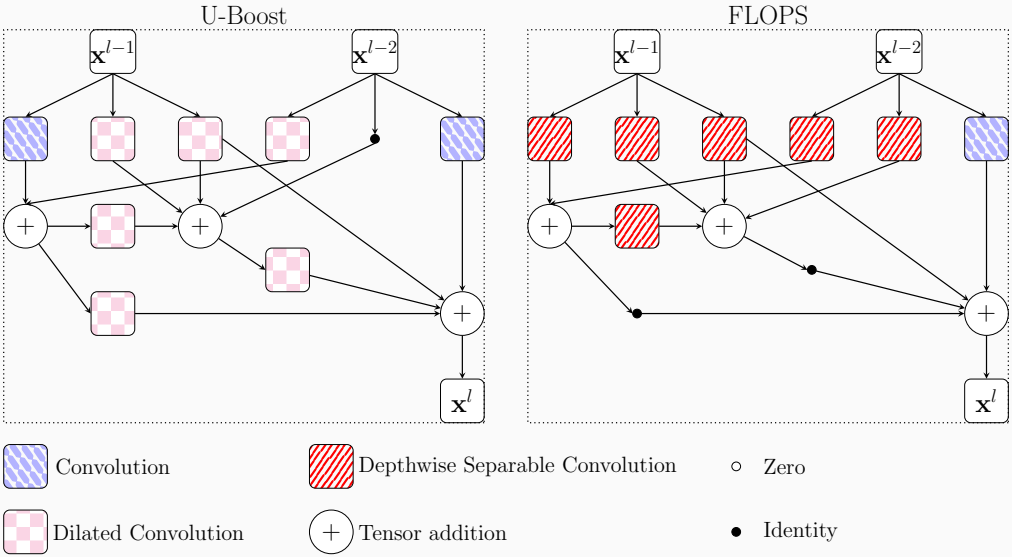
CIFAR10 experiments



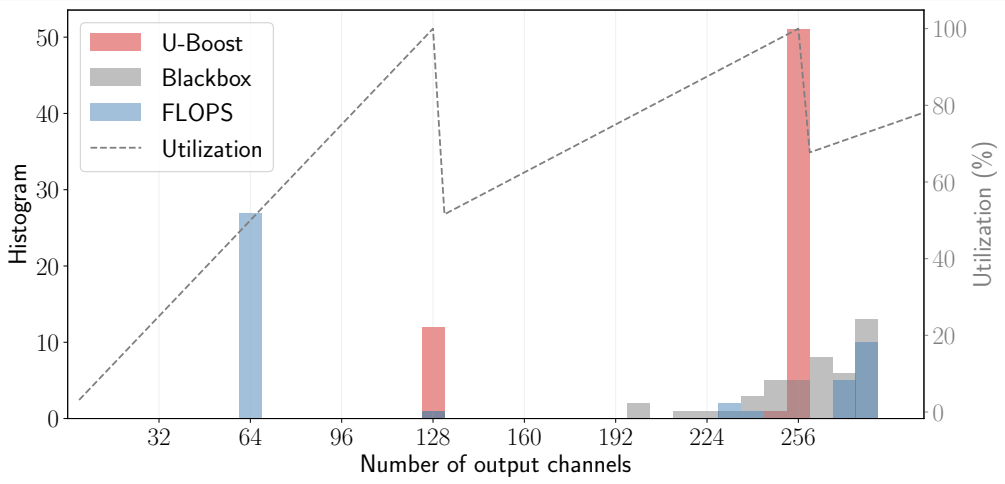
ImageNet100 experiments

	Acc. (%, ↑)		Runt. (ms, ↓)		Util. (%, ↑)		HV (↓)	# Params	
λ	0.1	1.0	0.1	1.0	0.1	1.0	(across λ)	0.1	1.0
Blackbox	87.5	87.8	4.8	4.05	69.3	68.5	49.4	70.5	55.5
Roofline	86.5	84.0	4.7	3.5	6.8	4.8	72.2	13.7	5.7
FLOPS	87.2	78.4	6.1	3.45	5.5	3.1	108	14.4	3.5
U-Boost	87.8	**87.9**	2.2	**1.05**	**91.1**	78.6	**12.7**	47.3	30.1

Cell microarchitecture: U-boost vs Baselines



Channel dimensions: U-boost vs Baselines



Thank you for your attention!

Acknowledgments
The work of Ahmet Caner Yüzügüler was supported by the Hasler Foundation (Switzerland) and Nikolaos Dimitriadis was supported by Swisscom (Switzerland) AG.