CVPR

Building Optimal Neural Architectures using Interpretable Knowledge

Keith G. Mills[1,2], Fred X. Han[2], Mohammad Salameh[2], Shengyao Lu[1], Chunhua Zhou[3], Jiao He[3] Fengyu Sun[3] and Di Niu[1]

[1]Dept. ECE, University of Alberta [2]Huawei Technologies Canada Co., Ltd [3]Huawei Kirin Solution, Shanghai, China

UNIVERSITY OF ALBERTA

HUAWEI

ALBERTA INNOVATES

You want an *efficient* model
for real-world task deployment.
What does it look like?

#Blocks? Kernel size? Attn heads? Hidden dimensions? DW-Sep Conv? Mixed Query Attention? …

"…a surprisingly large number of subnetworks (>10^{19}) that can fit different hardware platforms…"

Cai et al., 2020
Once for All, ICLR'20

Search Space	Cardinality
ProxylessNAS	$\sim 10^{21}$
ProxylessNAS-Enlarged	$\sim 10^{28}$
MobileNetV3-Like	$\sim 10^{43}$

Bender and Liu et al., 2020
TuNAS, CVPR'20

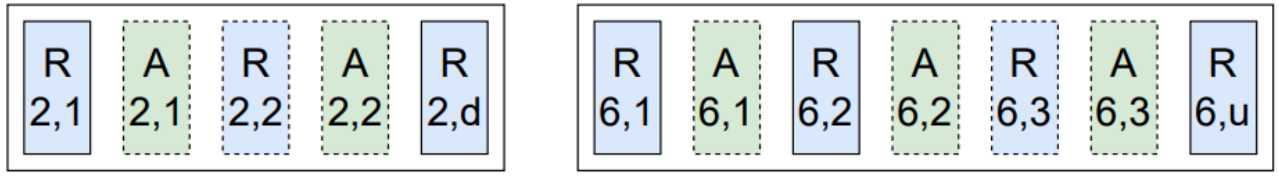
Hands off approach: AutoML NAS

Issue:
Still too many architectures.

Is there a better way to approach?

We answer in the affirmative with AutoBuild

Broad Idea: Consider modules, not architectures.

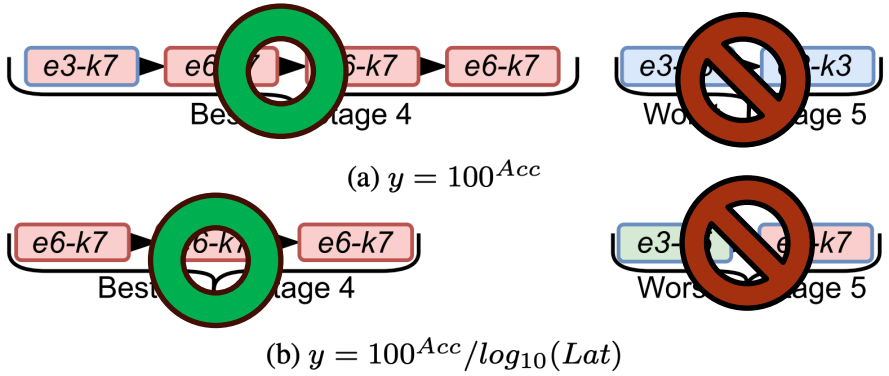


Example Stages from SDv1.5
A – Attention; R – ResNet Blk

Why do this?

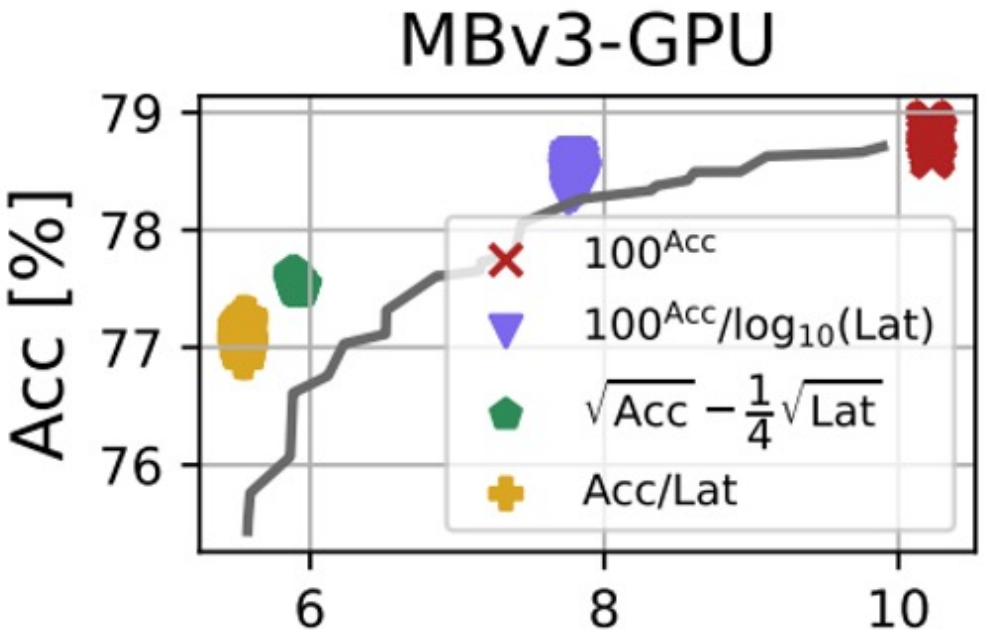
- Combinatorial

What do we do?



(a) $y = 100^{Acc}$

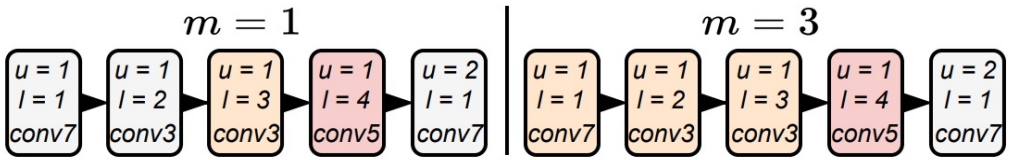
(b) $y = 100^{Acc}/log_{10}(Lat)$



Challenge: How to do if with end-to-end metrics?

Preliminary: Graphs and GNNs

- $(arch, perf) = (G_1, y_1)$
- Learn $y'_1 = GNN(G_1)$

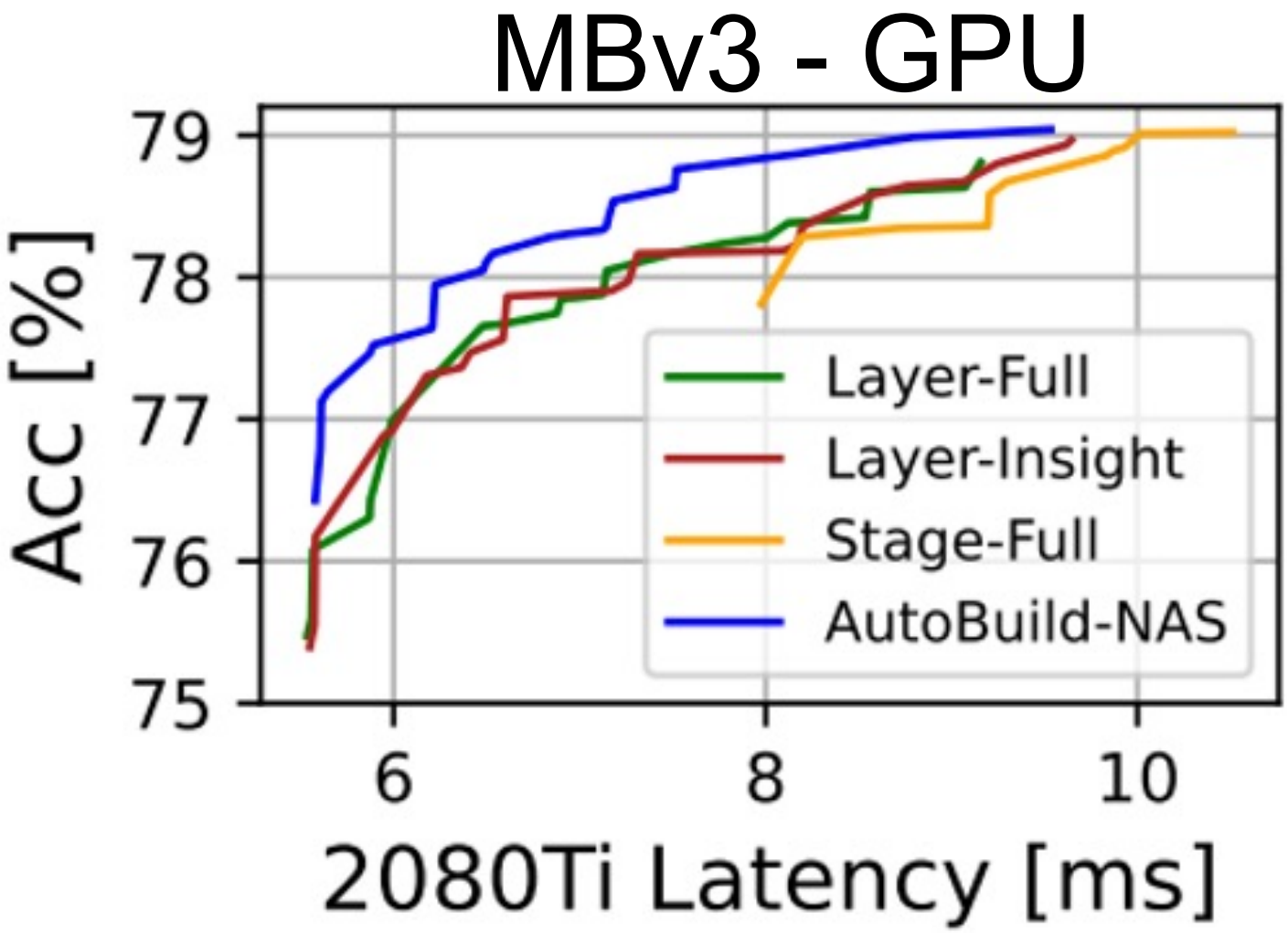


Intermediate workings: Node and Graph Embeddings

- $GNN(G) = MLP(h_G^m);\ h_G^m = \frac{1}{|V_G|}\sum_{v \in V_G} h_v^m$
- m is hop-level => h_v^m represents an *entire* subgraph/module!

Key learning constraint: if $y_1 > y_2$, then $\|h_{G_1}\|_1 > \|h_{G_2}\|_1$

Experimental Results:
Enhancing NAS with Restricted Search Space



Experimental Results: Stable Diffusion v1.4 Inpainting

Randomly sample 68/800k archs to learn from.
Aim to *minimize* FID

Arch Set	Eval Archs (68)	Exhaustive Search (4)	AutoBuild (4)
Ave. FID	22.13	10.82	**10.13**
Best FID	10.54	10.29	**9.96**

(a) Original (b) ES (c) AutoBuild

CVPR

Building Optimal Neural Architectures using Interpretable Knowledge

Thank you for watching ‘till the end!
See you in Seattle!

UNIVERSITY OF ALBERTA

HUAWEI

ALBERTA INNOVATES