NetAdaptV2: Efficient Neural Architecture Search with Fast Super-Network Training and Architecture Optimization

Tien-Ju Yang, Yi-Lun Liao, Vivienne Sze

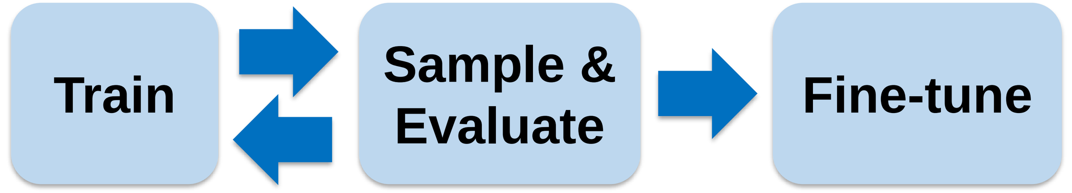
Massachusetts Institute of Technology

http://netadapt.mit.edu

CVPR VIRTUAL JUNE 19-25

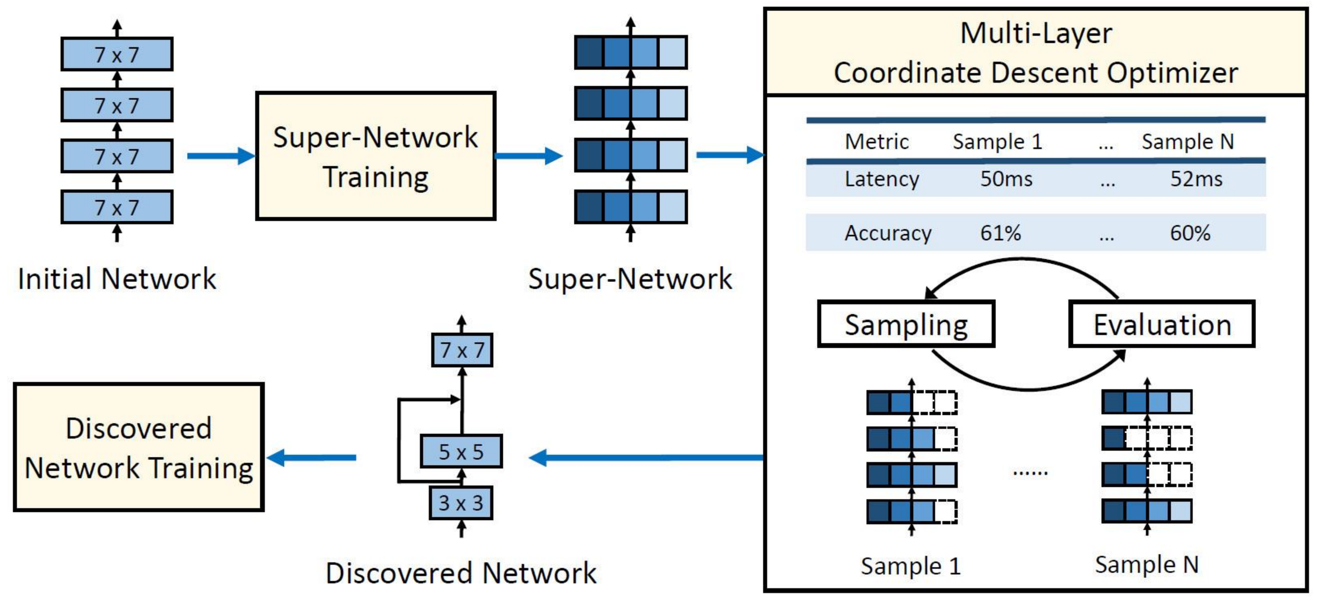
Neural Architecture Search (NAS)

- Two important metrics of NAS: network performance and search time.
- The search time mainly accounts for three steps: train, evaluate, fine-tune.
- NetAdaptV2 can discover **high-performance networks** in a **short time** by
  - Balancing and minimizing the time for each step (speed).
  - Supporting non-differentiable metrics (network performance).



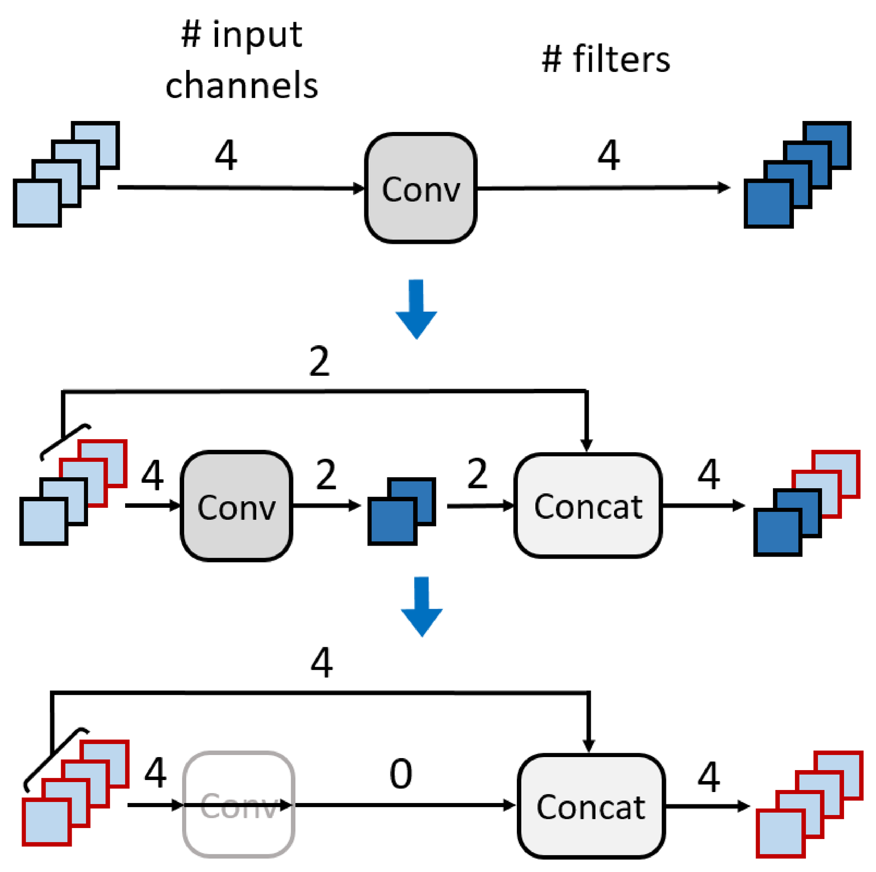
Algorithm Flow of NetAdaptV2

- Efficiently train the super-network using **ordered dropout**.
- Efficiently find high-performance networks using **channel-level bypass connections** and **multi-layer coordinate descent**.



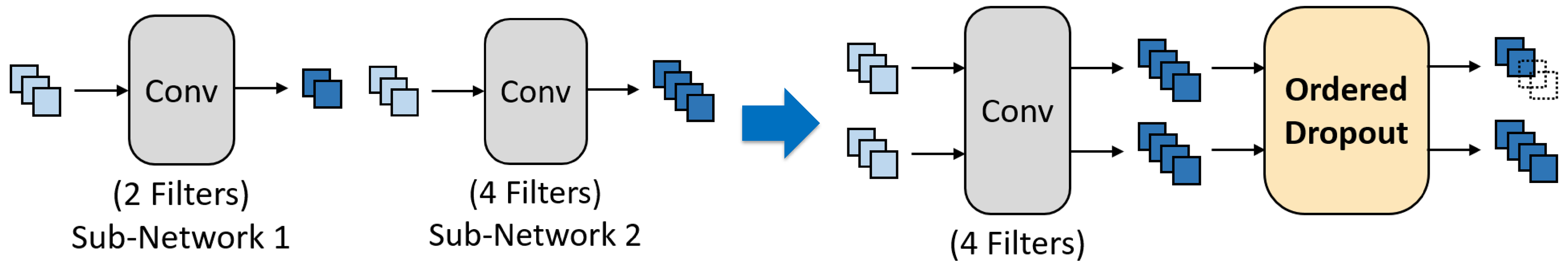
Technique: Channel-Level Bypass Connection (CBC)

- NetAdaptV2 searches layer width, network depth, and kernel size to improve network performance.
- **CBC reduces the time for evaluating networks**: merge layer width and network depth into a **single** search dimension and search only layer width.
  - Remove a layer when no filters inside.
- Idea: bypass an input channel when a filter is removed.
- Generalizing a network depth to a continuous value, e.g., 16.3 layers.



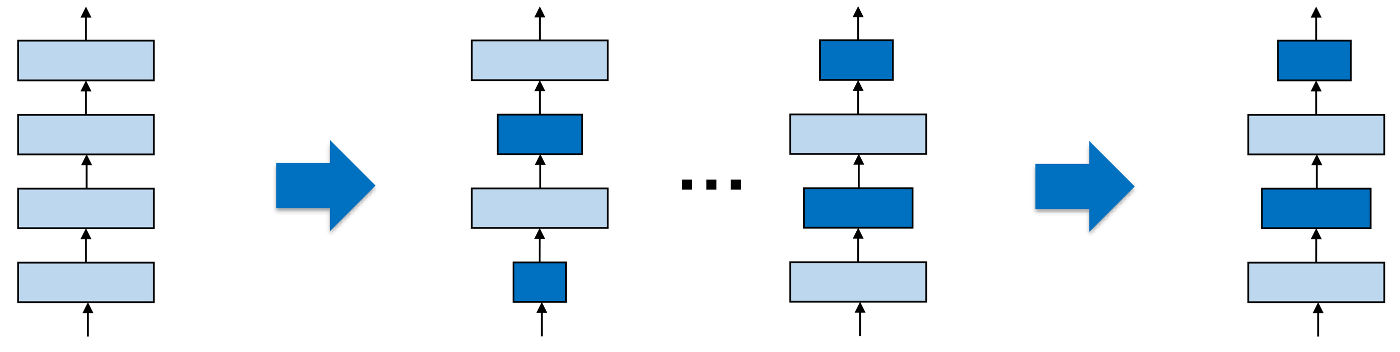
Technique: Ordered Dropout (OD)

- **OD reduces the time for training the super-network**: train **multiple** neural networks in a **single** forward-backward pass.
  - Architecture simulation: zero out different channels for different input images.
- To **avoid the training-evaluation mismatch**, OD always drops the last channels.



Technique: Multi-layer Coordinate Descent (MCD)

- **MCD 1) reduces the time for evaluating networks** and **2) supports non-differentiable metrics.**
- Idea: gradually and iteratively shrink a network until the constraints are satisfied.
- In each iteration, MCD generates *J* networks with similar metric values (e.g., 30ms) by randomly shrinking *L* layers in the network from the previous iteration and chooses the most accurate one for the next iteration.



Initial Network from Previous Iteration

***J* Networks with *L=2* Shrunk Layers and Similar Complexity**

Most Accurate Network for Next Iteration

Ablation Study of Proposed Techniques

- CBC and MCD improve the accuracy by 0.3% and 0.4%, respectively.
- Super-network + OD reduces the search time by **3.3x**.

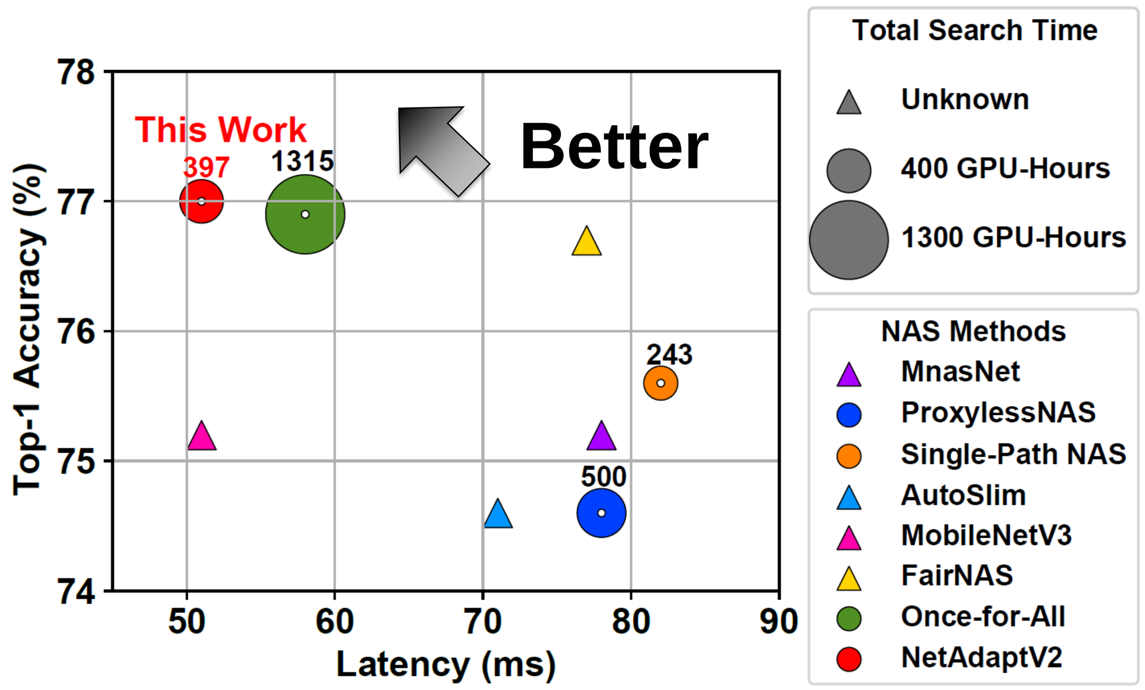
SN + OD	Top-1 Acc. (%)	Latency (ms)	Search Time (GPU-Hours)
	71.0 (+0)	43.9 (100%)	721 (100%)
✓	71.1 (+0.1)	44.4 (101%)	221 (31%)

Methods		Top-1 Acc. (%)
CBC	MCD	
		75.9 (+0)
✓		76.2 (+0.3)
✓	✓	76.6 (+0.7)

* Accuracy: ImageNet, Latency: Pixel 1 CPU, Search Time: V100.

Search Result – Image Classification

- Adopt a MobileNet-V3-based search space.
- **Latency-guided search result**
  - Up to **5.8x lower search time** with **better accuracy-latency/-MAC trade-off**.
  - NetAdaptV2 outperforms methods with hundreds of GPU-hours without sacrificing the **support of non-differentiable search metrics**.



Method	Top-1 Acc.	MAC (M)	Search Time[1]
FBNet	74.9%	375	-
BigNAS	76.5%	242	2304[2]
NetAdaptV2	77.0%	225	397

▲ **Networks with Unknown Latency**

- **MAC-guided search result**
  - Up to **2.6x lower search time** with comparable accuracy-MAC trade-off.
  - Up to **1.5% higher top-1 accuracy** with **fewer MACs**.

Method	Top-1 Acc.	MAC(M)	Search Time[1]
NSGANetV2-m	78.3%	312	1674
EfficientNet-B0	77.3%	390	-
MixNet-M	77.0%	360	-
NetAdaptV2	78.5%	314	656

1. The unit of the search time is GPU-hours on Nvidia V100s.
2. The search time on Google TPU V3s.
3. The latency is measured on a Pixel 1 CPU.

Search Result – Depth Estimation

- Adopt the sub-networks of the FastDepth* network as the search space.
- **2.4x lower search time on NYU Depth** with **better accuracy-latency trade-off**.

Method	RMSE (m)	Delta-1 Accuracy (%)	Latency (ms)	Search Time (GPU-Hours)	
				ImageNet	NYU Depth
NetAdaptV1	0.583	77.4	87.6	96	65
NetAdaptV2	**0.576**	**77.9**	**86.7**		**27**

* D. Wofk et al., "FastDepth: Fast Monocular Depth Estimation on Embedded Systems", ICRA 2019.