NEURAL INFORMATION PROCESSING SYSTEMS

Adaptive Depth Networks with Skippable Sub-Paths

38th Neural Information Processing Systems

Woochul Kang and Hyungsup Lee

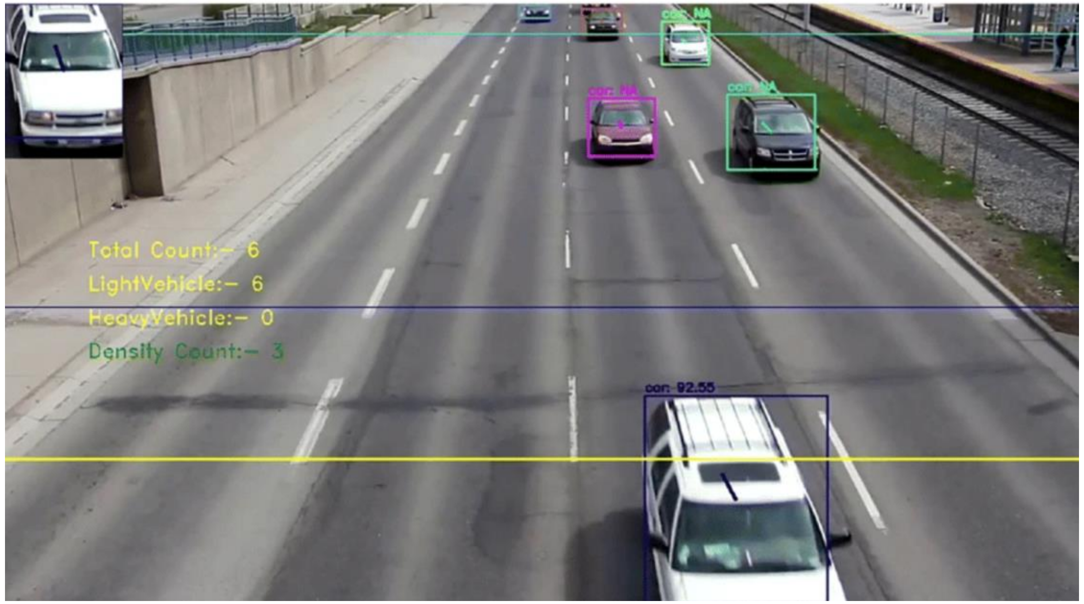
Embedded Systems Department
Incheon National University

INU 인천대학교 정보기술대학

"It is not the strongest of the species that survives.
It is the one that is most adaptable to change."

\- *Charles Darwin*

Neural networks also need adaptability



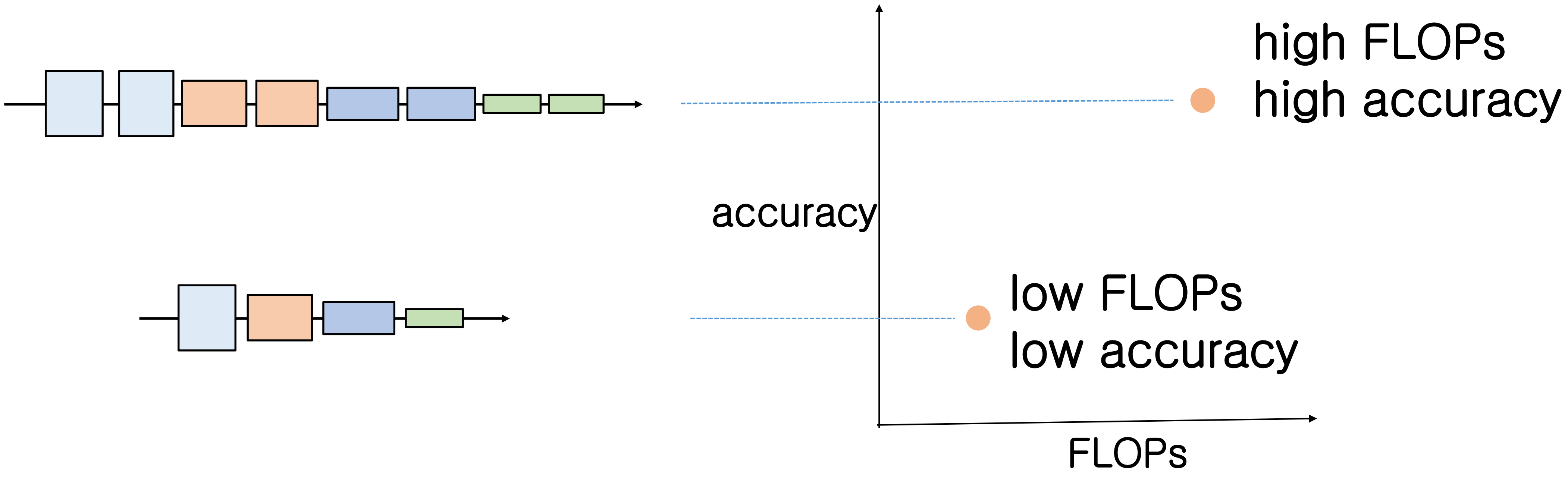
Accuracy : ★★★★

Latency : ★★★★

Accuracy : ★★★

Latency : ★★

Most deep learning networks are Not adaptable!



- Once trained, models' FLOPs and accuracy is fixed
- Many models need to be trained => huge cost!

Previous adaptive networks

- A single network can be trained to
  - adapt **widths**, or channels
  - adapt **depths**
  - adapt **input resolutions**
  - adapt active **tokens**
- **However,⋯**

Problem of previous adaptive networks

- High cost of training
- Lower accuracy than non-adaptive counterparts
- Low actual acceleration performance
- Not generally applicable to different network types

Low Practicality !

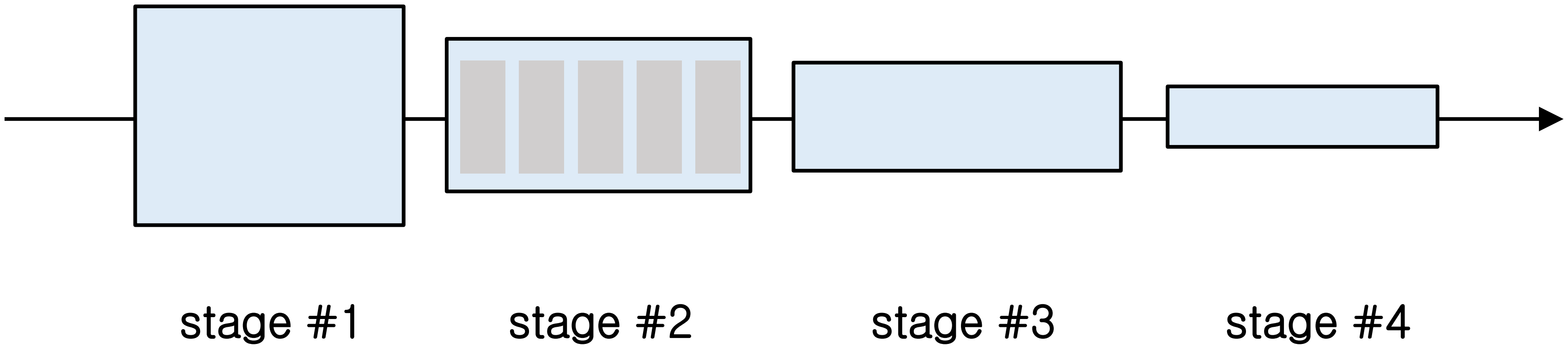
We propose Adaptive Depth Networks

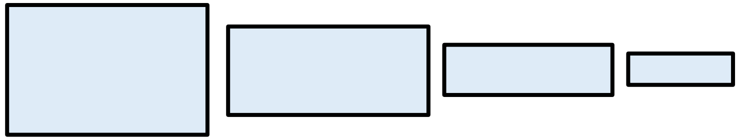
- Low cost of training
- Higher accuracy than non-adaptive counterparts
- High actual acceleration performance
- Generally applicable to CNNs and Transformers

High Practicality !

인천대학교 INCHEON NATIONAL UNIVERSITY

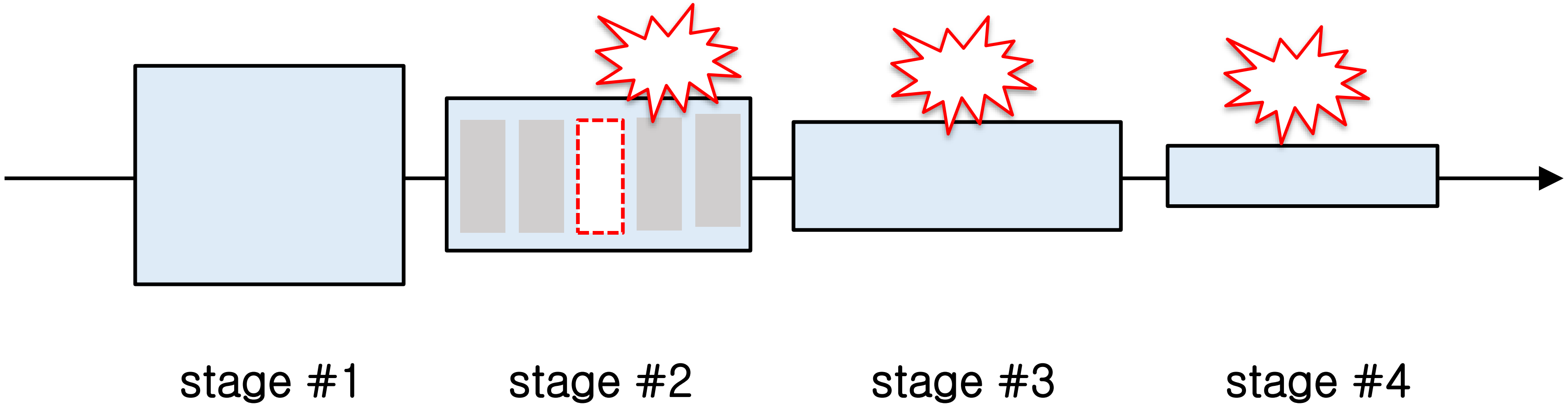
Adaptive Depth Networks with Skippable Sub-Paths



- Networks has 4~7 hierarchical stages 
- Each stage has 2~ 23 blocks
- Each block has 3~ 4 layers with identify shortcuts

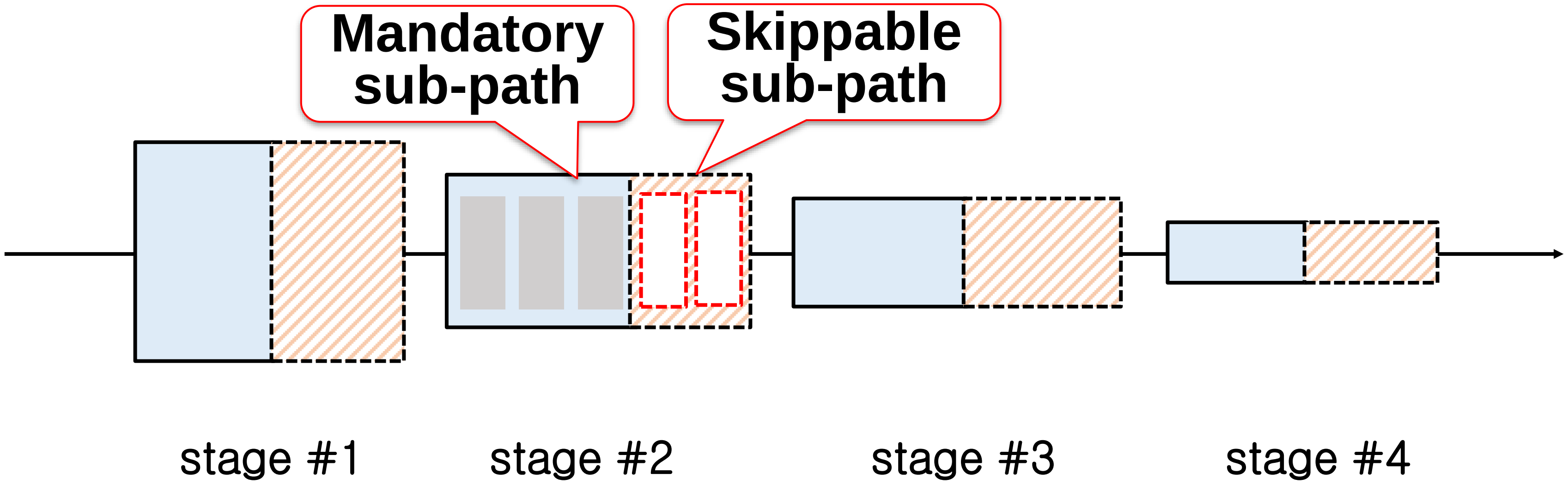
인천대학교
INCHEON NATIONAL UNIVERSITY

Adaptive Depth Networks with Skippable Sub-Paths



- Skipping blocks causes **severe loss of performance** because it changes the input feature distributions

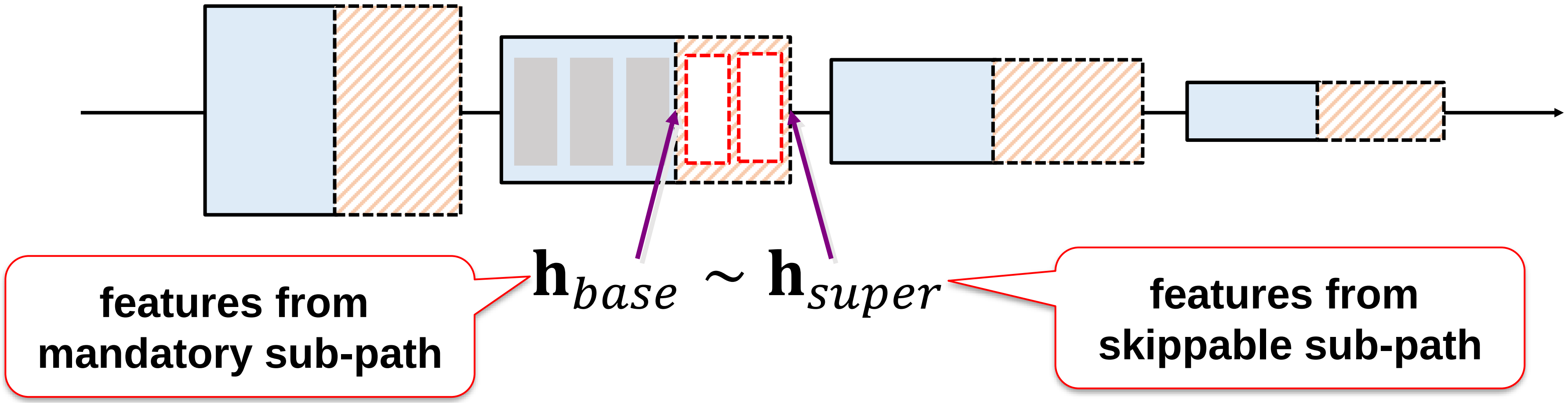
Adaptive Depth Networks with Skippable Sub-Paths



- Divide every stage into 2 groups, or 2 sub-paths
- Train every second sub-path to preserve input feature distributions and focus on refining the features
- Skipping second sub-paths has minor performance impact

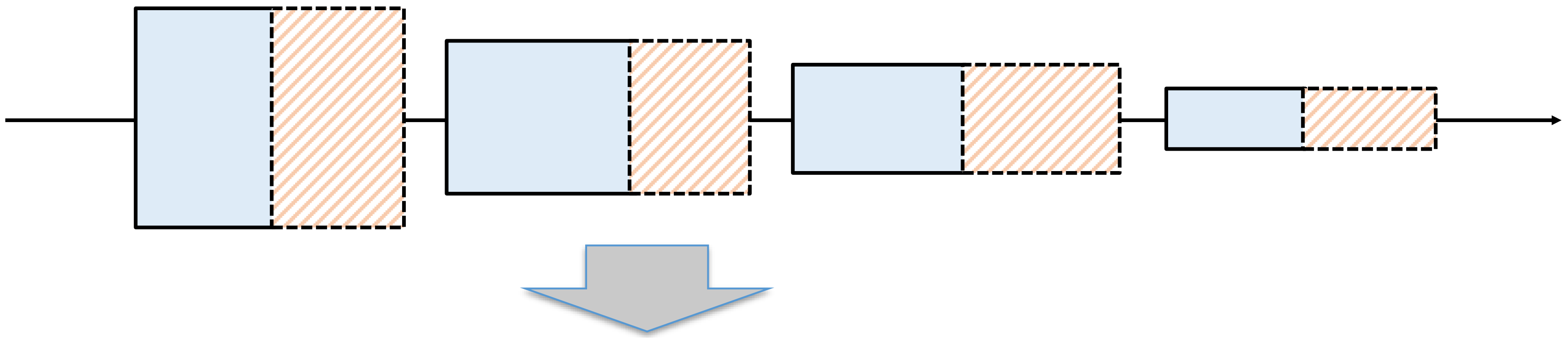
At training:

Encourage $\mathbf{h}_{base} \sim \mathbf{h}_{super}$ via self-distillation

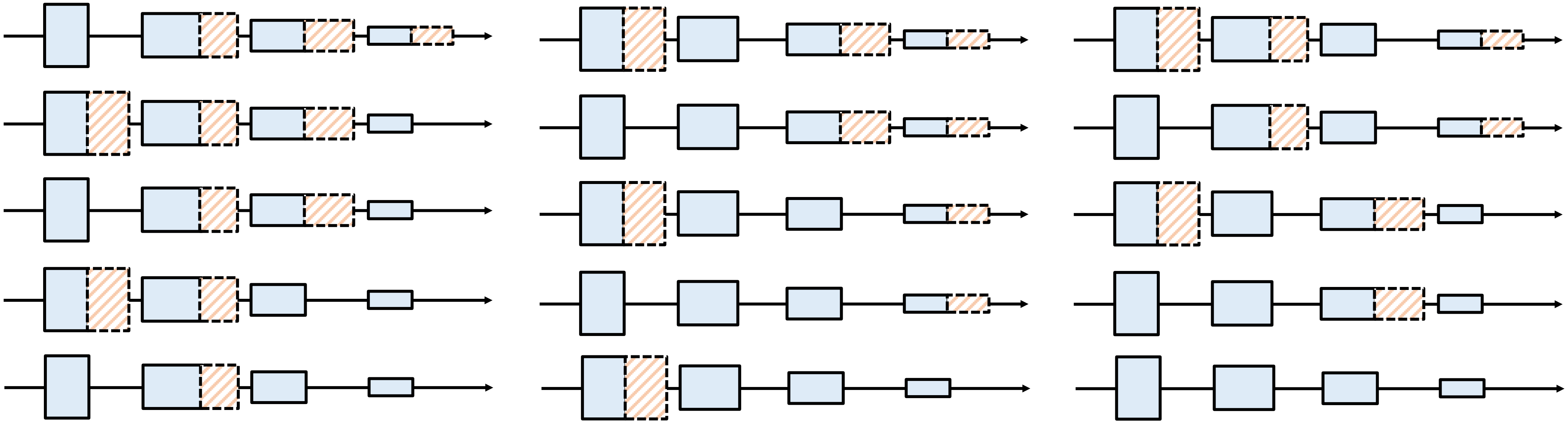


- Include $\mathrm{D}_{KL}(\mathbf{h}_{base}, \mathbf{h}_{super})$ in the loss function
- Only one additional forward/backward passes is required to the vanilla training loop

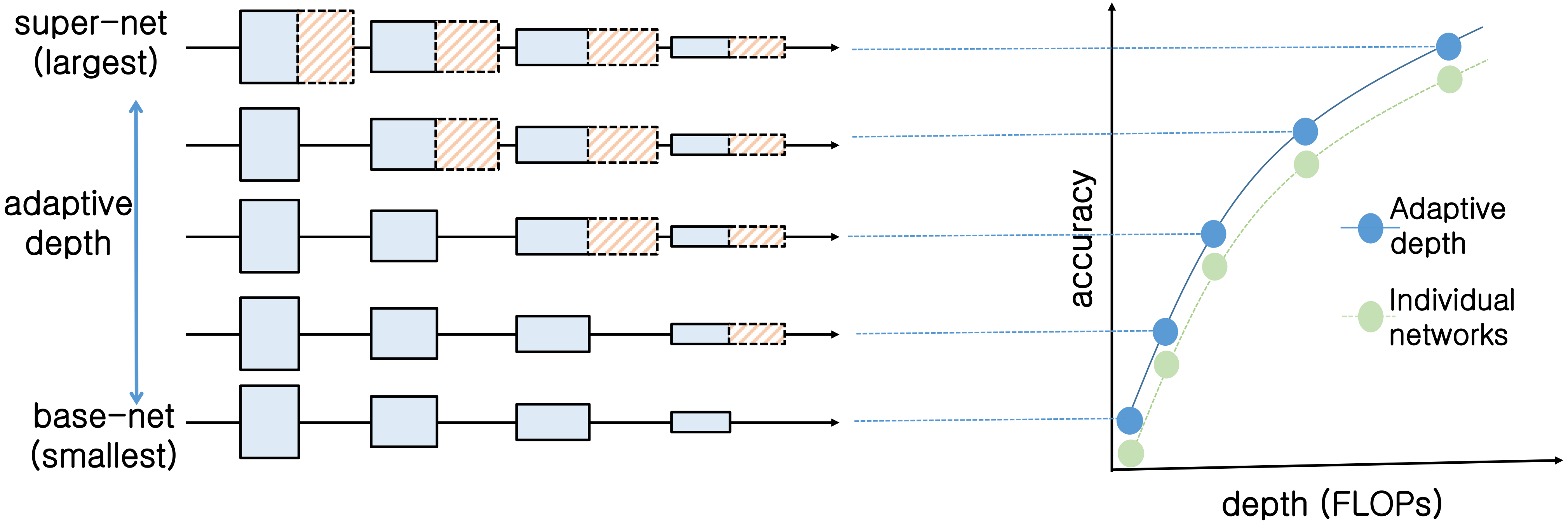
At test time



Many different sub-networks can be selected instantly at no cost



At test time

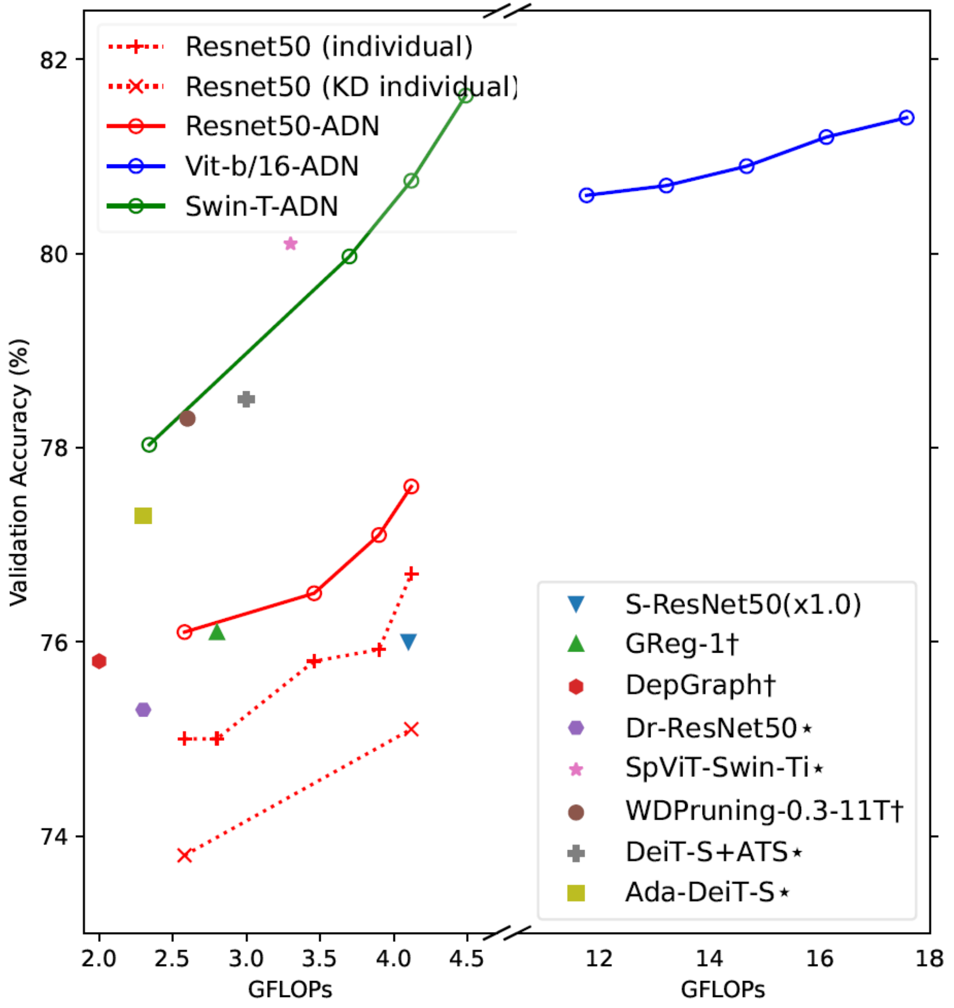


- Sub-networks outperform non-adaptive counterparts

Evaluation on CNNs and Vision Transformers

Model	Params (M)	FLOPs (G)	Acc (%)
ResNet50-ADN (FFFF)	25.58	4.11	77.6
ResNet50-ADN (TTTT)		2.58	76.1
ResNet50	25.56	4.11	76.7
ResNet50-Base	17.11	2.58	75.0
MbV2-ADN (FFFFF)	3.53	0.32	72.5
MbV2-ADN (TTTTT)		0.22	70.6
MbV2	3.50	0.32	72.1
MbV2-Base	2.98	0.22	70.2
ViT-b/16-ADN (FFFF)	86.59	17.58	81.4
ViT-b/16-ADN (TTTT)		11.76	80.6
ViT-b/16	86.57	17.58	81.1
ViT-b/16-Base	67.70	11.76	78.7
Swin-T-ADN (FFFF)	28.30	4.49	81.6
Swin-T-ADN (TTTT)		2.34	78.0
Swin-T	28.29	4.49	81.5
Swin-T-Base	15.34	2.34	77.4

(a) Results on ImageNet



(b) Pareto frontiers of our networks

인천대학교 INCHEON NATIONAL UNIVERSITY

See you at the poster session!

Source codes and weights are available at

https://github.com/wchkang/depth

인천대학교
INCHEON NATIONAL UNIVERSITY