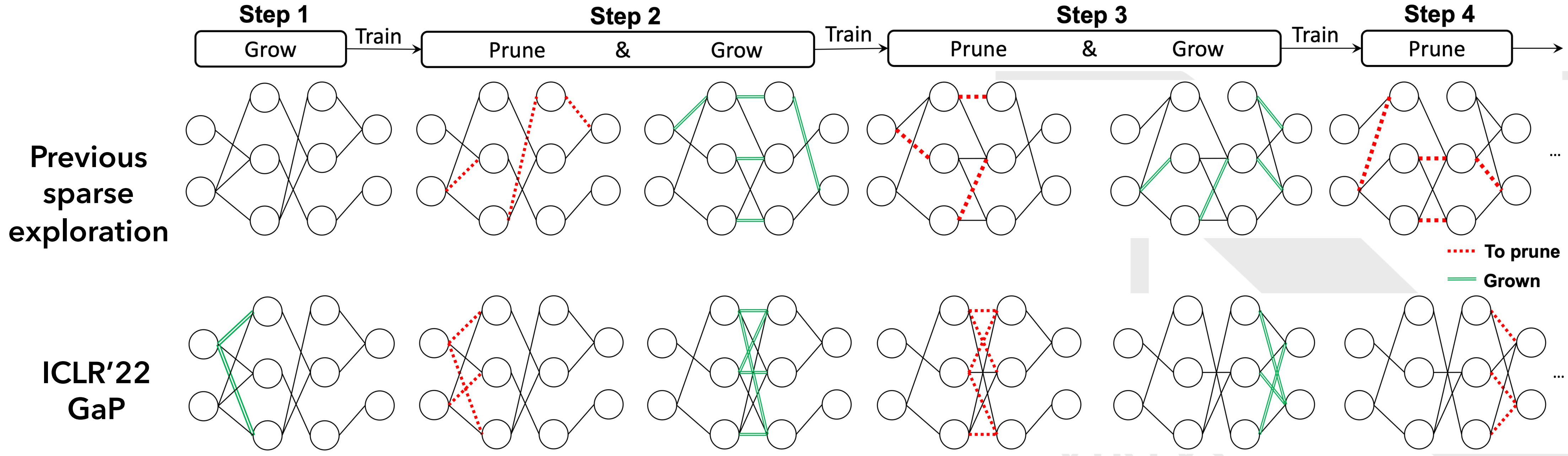
Effective Model Sparsification by Scheduled Grow-and-Prune Methods

Xiaolong Ma
Department of Electrical and Computer Engineering

Motivation

Two different schedules



Previous: Greedy and random

- No guarantee for full exploration
- Mutation in model-level – less flexibility

GaP: Scheduled Grow-and-Prune

- Guaranteed full exploration for sparse mask – approximately **3x** efficiency.
- Mutation in layer-level – more flexibility and design space.
- Mask parallelism

Alibaba DAMO Academy 达摩院 Northeastern University

Effective Grow-and-Prune Methods

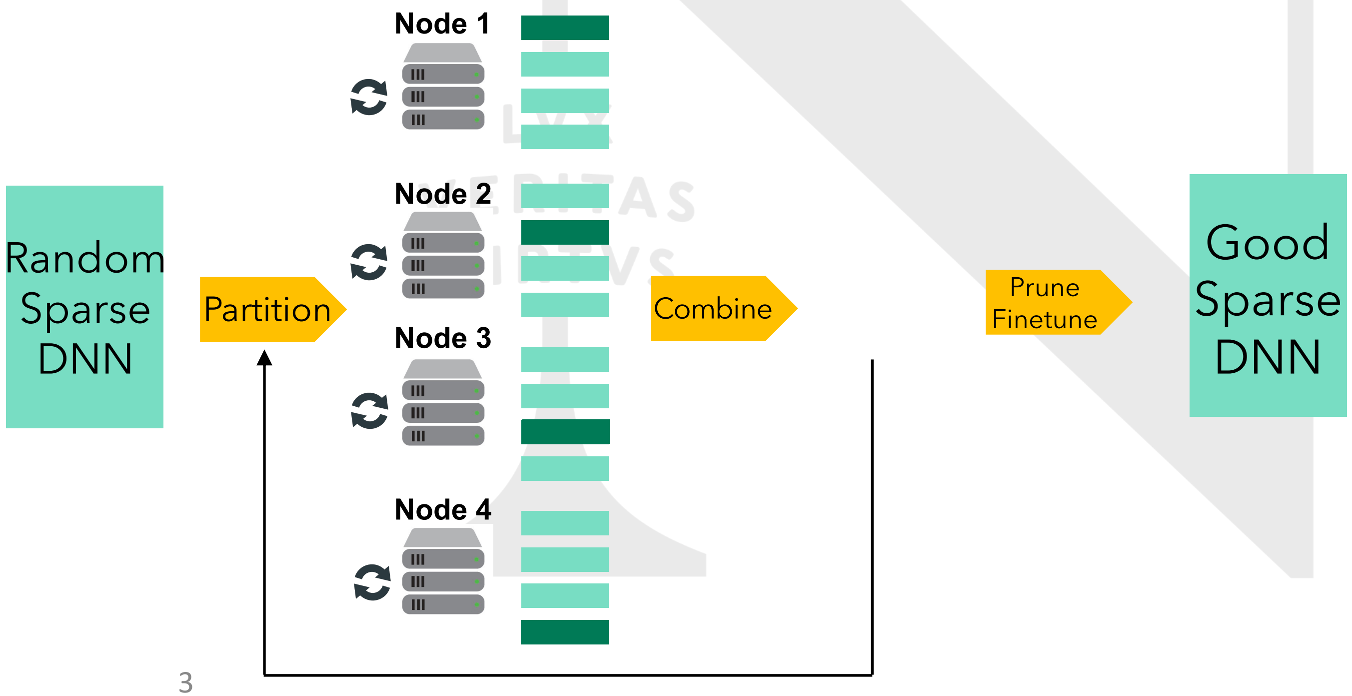
ICLR

Cyclic GaP (C-GaP)

- Train a sparse model whole time
- Training on one machine.
- Grow and prune DNN partition in a sequential order.

Parallel GaP (P-GaP)

- Train a sparse model whole time
- Training each partition at the same time on multiple machine.
- Combine explored partitions and discard others.



Alibaba DAMO Academy 达摩院 Northeastern University

Effective Grow-and-Prune Methods

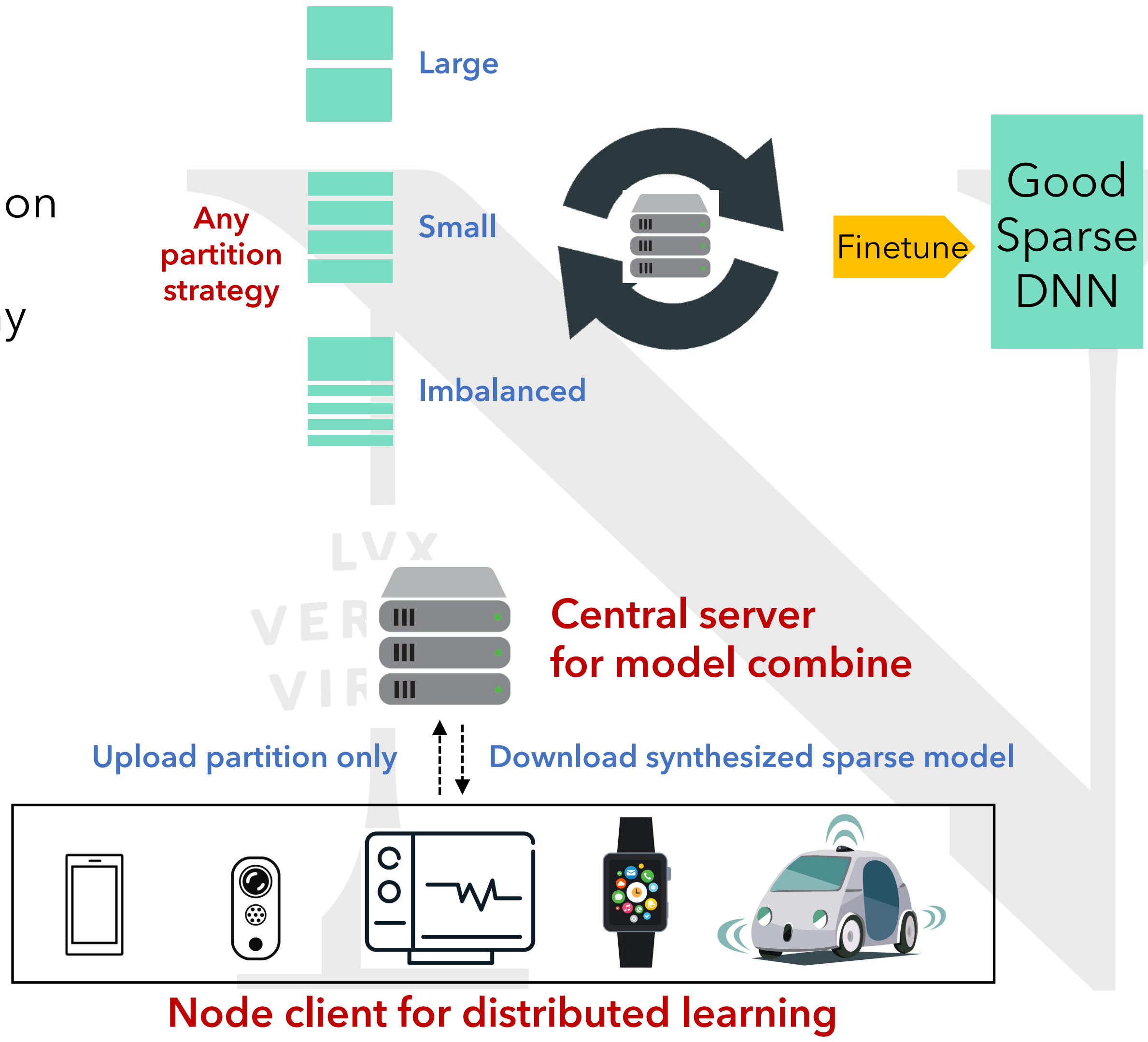
Why is the GaP method important

High flexibility

- The partition strategy – partition number, partition size, partition sequence order.
- Export and finetune a good sparse model at any time.

P-GaP distributed training

- Does not need to use large batch size to fully utilize computing resource.
- Data communication between training nodes are kept in minimal (low-bandwidth friendly).
- Masks in different partitions are less correlated.



Experimental Results: Accuracy

GaP is proved to be effective on multiple ML tasks

1. Image classification

- ResNet-50 on ImageNet.
- Fair comparison with extended baselines.
- We show the best accuracy.

Method	Distribution	Epochs	80% sparse Acc	90% sparse Acc
Dense	-	250	Dense acc: 78.2%	
Prune from dense	uniform	750/1250	77.1	75.8
RigL	uniform	100	74.6	72.0
$RigL_{5x}$		500	76.6	75.7
$RigL_{12x}$ (SOTA)		1200	77.1	76.0
C-GaP (Ours)		990	**77.9**	**76.3**
P-GaP		1110	77.5	76.1
RigL	Non-uniform (ERK)	100	75.1	73.0
$RigL_{5x}$		500	77.1	76.4
$RigL_{12x}$ (SOTA)		1200	77.4	76.8
C-GaP (Ours)	Non-uniform	990	**78.1**	**77.9**

Experimental Results: Flexibility

Partition number and partition strategy

1. Partition number

- From 1 partition (DSD) to multiple.

Method	Sparsity	Task	# Part	Acc (%)
C-GaP	0.9	ResNet-50	1	75.9
			4	**76.3**
		Transformer	1	26.8
			3	**27.7**
			6	27.1

2. Partition strategy

- Cyclic vs. random.

Method	Sparsity	Task	# Part	Strategy	Acc (%)
C-GaP	0.9	ResNet-50	4	Cyclic	**77.9**
			4	Random	77.8
		Transformer	3	Cyclic	**27.7**
			3	Random	27.0