The Heterogeneity Hypothesis: Finding Layer-Wise Differentiated Network Architectures

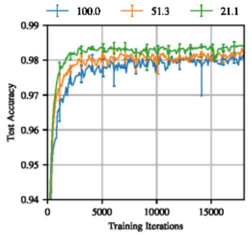
[1]ETH Zurich, Switzerland; [2]UESTC, China; [3]The University of Sydney, Australia; [4]KU Leuven, Belgium

Yawei Li[1] W. Li[2], M. Danelljan[1], K. Zhang[1] S. Gu[3], Luc van Gool[1,4], R. Timofte[1]

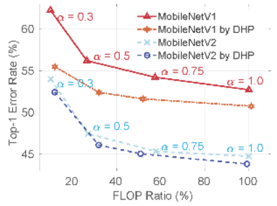
VIRTUAL JUNE 19-25

Introduction

Aim: Cost-free fine-grained architecture optimization



Lottery Ticket Hypothesis



Channel Pruning (DHP)

Observation 1: Pruned network performs better than the original network.
Observation 2: Channel pruned network outperforms the original network under different model complexities.

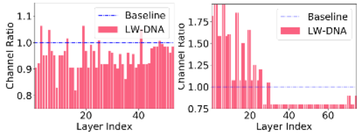
Limitation of previous works:

- Lottery ticket hypothesis targets unstructured pruning.
- Longer training epochs for channel pruning.
- Small datasets.

Question 1: The Heterogeneity Hypothesis

With the same training protocol, there exists a layer-wise differentiated network architecture (LW-DNA) that can outperform the original network with regular channel configurations but with a lower level of model complexity.

- The same training protocol.
- The existence of LW-DNA models.
- Lower level of model complexity (parameters, computation).



(a) ResNet50, ImageNet. (b) RegNet-4GF, ImageNet.

Network	Method	Top-1 Error (%)	FLOPs [G] / Ratio (%)	Params [M] / Ratio (%)
ResNet50	Baseline	23.28	4.1177 / 100.0	25.557 / 100.0
	LW-DNA	23.00	3.7307 / 90.60	23.741 / 92.90
RegNet-4GF	Baseline	23.05	4.0005 / 100.0	22.118 / 100.0
	LW-DNA	22.74	3.8199 / 95.49	15.285 / 69.10

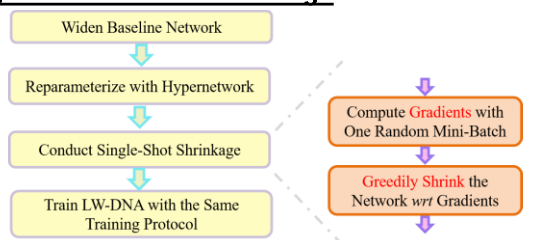
Question 2: How to identify?

Starting from a baseline architecture

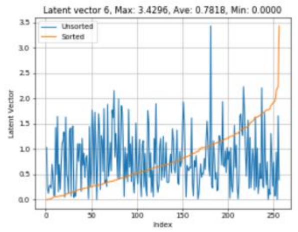
Cost-free architecture optimization

- Fair comparison. Identify the origin of the improvement

Single-shot network shrinkage

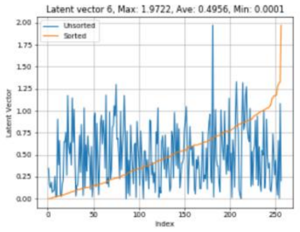


- 1) Initialize; 2) Single-shot shrinkage; 3) Train from scratch



Layer 6, Epoch 1

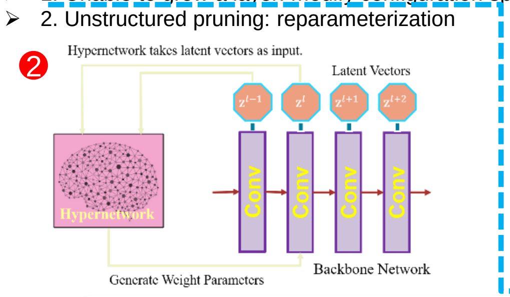
Why single-shot?

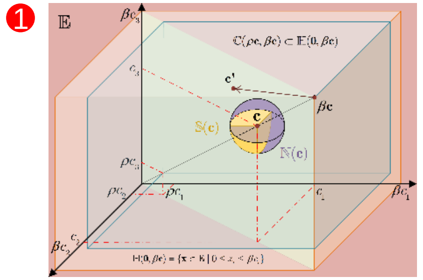


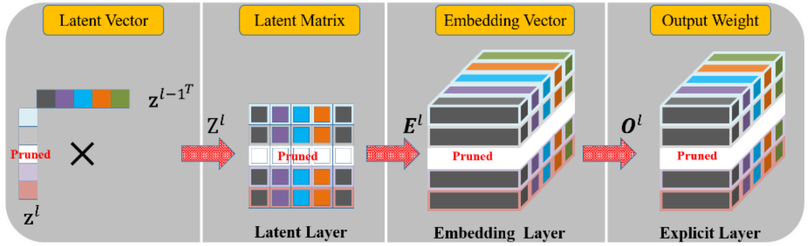
Layer 6, Epoch 4

Two problems of single-shot pruning:

- 1. Unable to grow a layer: modify configuration space
- 2. Unstructured pruning: reparameterization







Question 3: Why the benefit?

1. CNNs are redundant. It is possible to find a layer-wise specific channel configuration comparable with the baseline under lower model complexity.

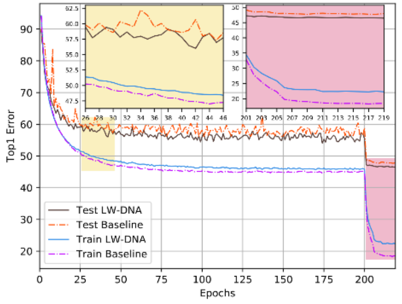
2. The redistribution of computational budget could help to improve the performance.

3. Maybe related to overfitting

- Higher training error while lower test error.
- Easier to find an LW-DNA model for larger networks.
- Improvement on smaller dataset is more significant.

Dataset	Network	Method	Top-1 Error (%)	FLOPs [G] / Ratio (%)
ImageNet [6]	ResNet50 [14]	Baseline	23.28	4.1177 / 100.0
		LW-DNA	23.00	3.7307 / 90.60
	RegNet [39] X-4.0GF	Baseline	23.05	4.0005 / 100.0
		LW-DNA	22.74	3.8199 / 95.49
	MobileNetV3 small [16]	Baseline	34.91	0.0612 / 100.0
		LW-DNA	34.84	0.0605 / 98.86
Tiny-ImageNet	MobileNetV1 [17]	Baseline	51.87	0.0478 / 100.0
		Baseline KD	48.00	0.0478 / 100.0
		LW-DNA	46.44	0.0460 / 96.23
	MobileNetV2 [44]	Baseline	44.38	0.0930 / 100.0
		Baseline KD	41.25	0.0930 / 100.0
		LW-DNA	40.74	0.0872 / 93.76

Table 6: Image classification results.



Training and test log. MobileNetV1, Tiny-ImageNet

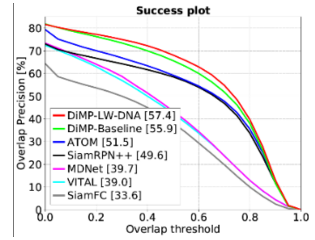
Extensive Results on Image Super-Resolution and Visual Tracking

Network	Method	PSNR [dB] Set5 [2]	Set14 [52]	B100 [35]	Urban100 [19]	DIV2K [1]	FLOPs [G] / Ratio (%)	Params [M] / Ratio (%)
SRResNet [23]	Baseline	32.02	28.50	27.52	25.88	28.84	32.81 / 100.0	1.53 / 100.0
	LW-DNA	32.07	28.51	27.52	25.88	28.85	28.79 / 87.75	1.36 / 88.43
EDSR [29]	Baseline	32.10	28.55	27.55	26.02	28.93	90.37 / 100.0	3.70 / 100.0
	LW-DNA	32.13	28.61	27.59	26.09	28.99	55.44 / 61.34	2.84 / 76.94

Table 2: Results on single image super-resolution networks. The upscaling factor is ×4.

Metric	DiMP-Baseline	DiMP-LW-DNA
TrackingNet [36]		
Precision	68.06	68.27
Norm. Prec. (%)	79.70	79.64
Success (AUC) (%)	73.77	73.83
LaSOT [8]		
Precision	54.97	57.30
Norm. Prec. (%)	63.70	65.82
Success (AUC) (%)	55.87	57.43

Table 3: Tracking test results. DiMP-LW-DNA and DiMP-Baseline use the identified LW-DNA and baseline version of ResNet50, respectively.



Success plot on the LaSOT dataset for visual tracking.

References

[9] Jonathan Frankle and Michael Carbin. The lottery ticket hypothesis: Finding sparse, trainable neural networks. arXiv preprint arXiv:1803.03635, 2018.

[25] Namhoon Lee, Thalaiyasingam Ajanthan, and Philip HS Torr. SNIP: Single-shot network pruning based on connection sensitivity. arXiv preprint arXiv:1810.02340, 2018.

[28]Yawei Li, Shuhang Gu, Kai Zhang, Luc Van Gool, and Radu Timofte. DHP: Differentiable meta pruning via hypernetworks. arXiv preprint arXiv:2003.13683, 2020.