The Heterogeneity Hypothesis: Finding Layer-Wise Differentiated Network Architectures

Yawei Li

Wen Li

Martin Danelljan

Kai Zhang

Shuhang Gu

Luc Van Gool

Radu Timofte


Contents

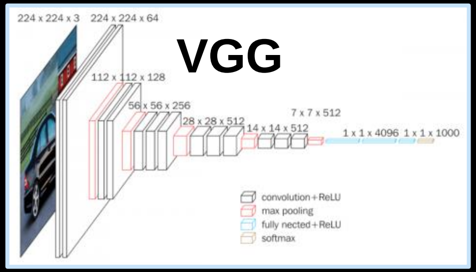
Introduction

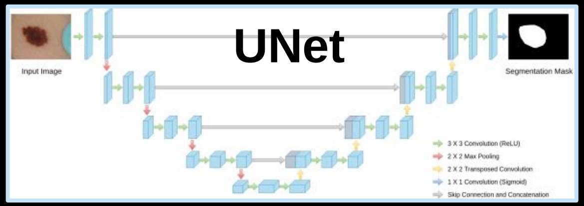
Cost-free
Fine-grained
Architecture Optimization

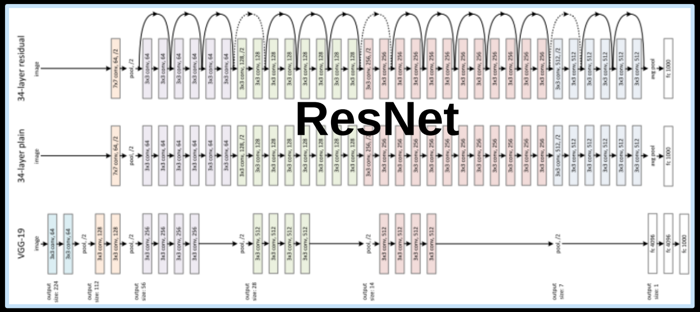
Introduction

Neural Architecture Design

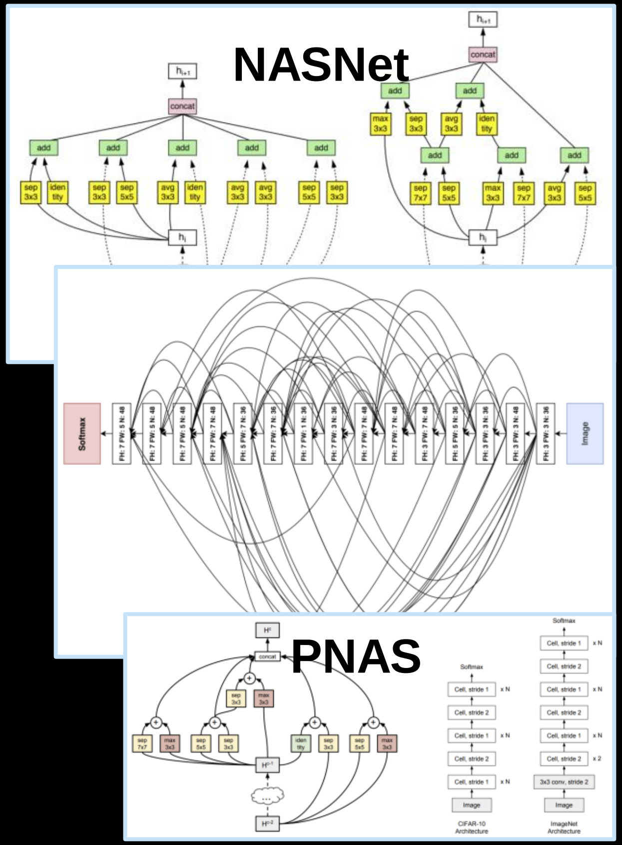
Manual Design



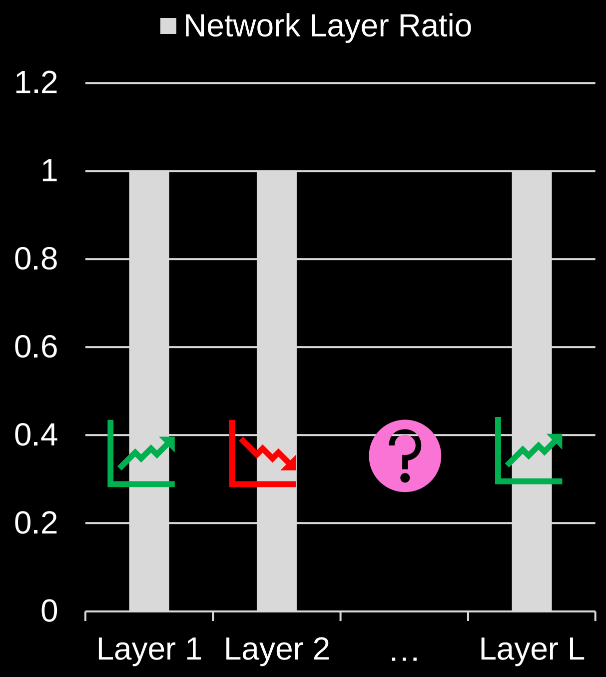




Architecture Search

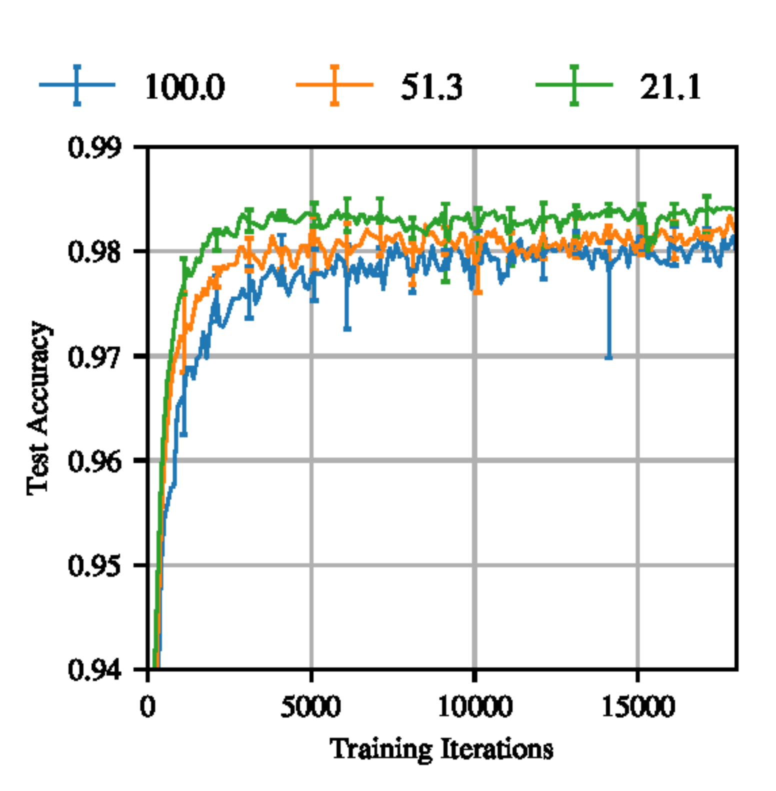


Architecture Optimization



Hints: Lottery Ticket Hypothesis (Unstructured)

- MNIST

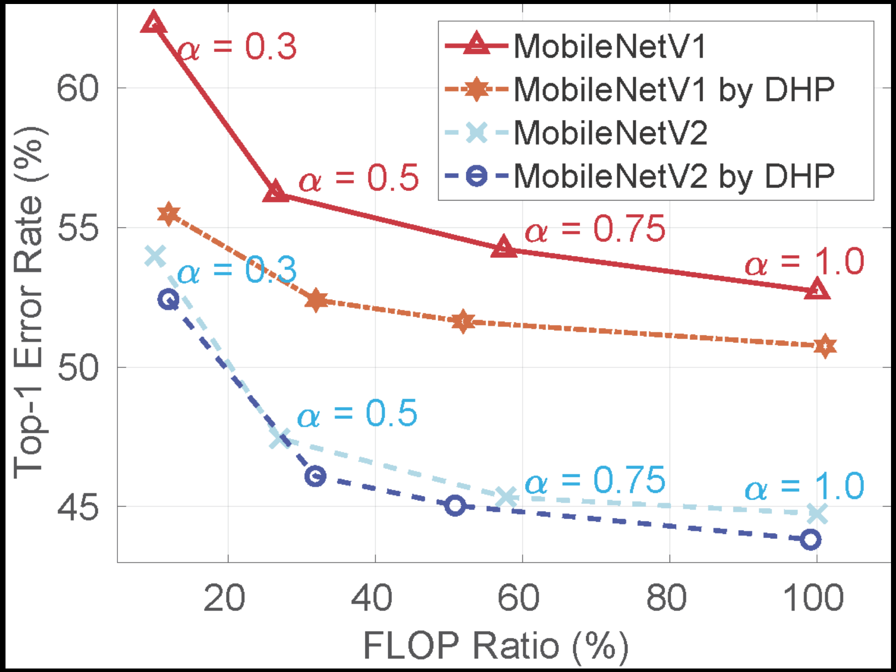


Observation 1: Pruned network performs better than the original network.

[1] Jonathan Frankle, Michael Carbin. **The Lottery Ticket Hypothesis: Finding Sparse, Trainable Neural Networks. ICLR 2019.**

Hints: Channel Pruning (Structured)

- Tiny-ImageNet
- α: width multiplier



Observation 2: Channel pruned network outperforms the original network under different model complexities.

[1] Yawei Li, Shuhang Gu, Kai Zhang, Luc Van Gool, Radu Timofte. **DHP: Differentiable Meta Pruning via HyperNetworks. ECCV 2020.**

Limitation of Previous Work

- The lottery ticket hypothesis is only valid under the setting of weight removal.
  - Extension to architecture optimization in terms of channel reconfiguration is not studied.
- The optimized network architectures are derived under different training protocols (epoch).
  - Where the improvement comes from.
- Small dataset (MNIST, Tiny-ImageNet).

The Heterogeneity Hypothesis:

The Existence of LW-DNA models

The Heterogeneity Hypothesis

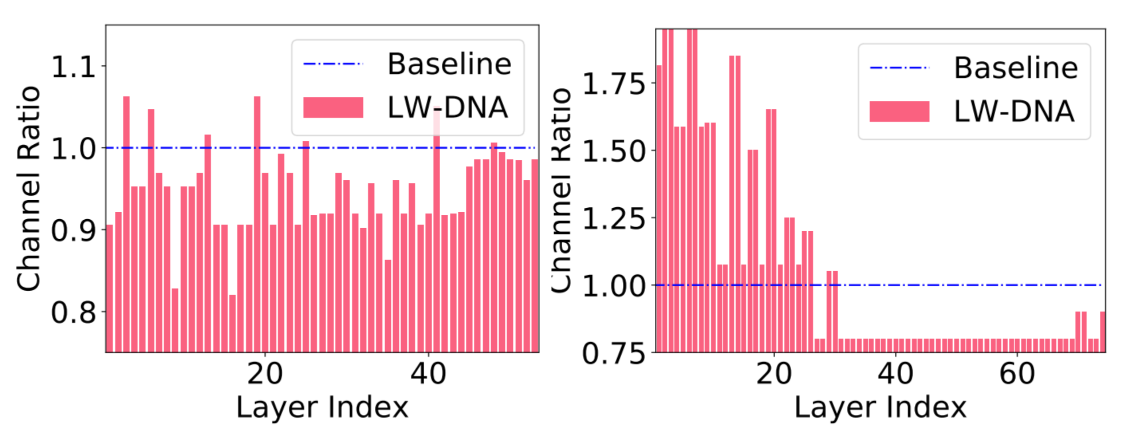
Question 1: The existence LW-DNA models

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
- ✓ LW-DNA
- ✓ Lower level of model complexity
  - Parameters
  - Computation

The Heterogeneity Hypothesis

Question 1: The existence LW-DNA models



(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

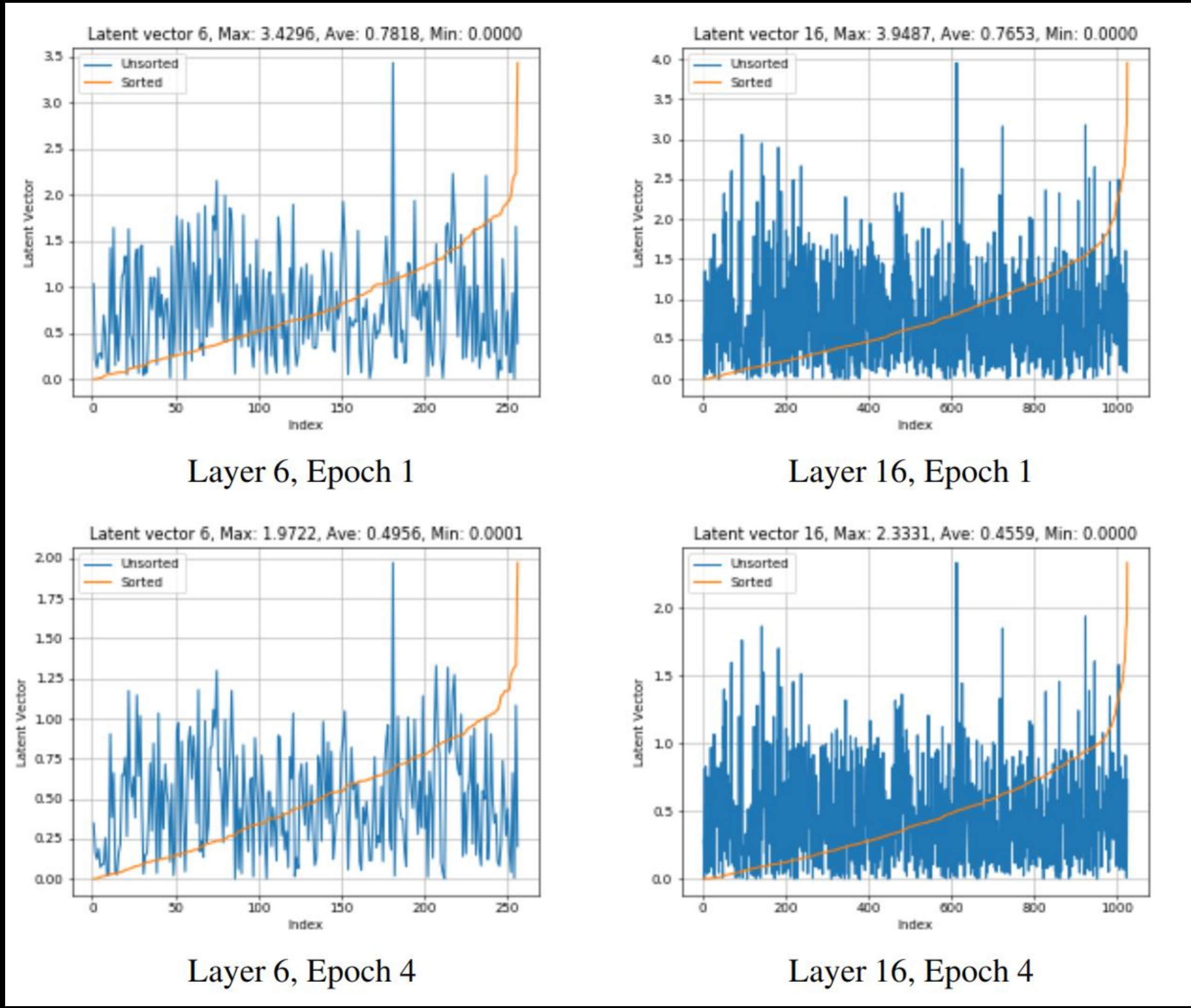
Methodology

How to identify LW-DNA models

Methodology

Question 2: How to identify an LW-DNA model efficiently?

- Starting from a baseline architecture
- Cost-free architecture optimization
  - Fair comparison
- Single-shot network shrinkage
  - Initialize a network
  - Prune the initialized network
  - Train the pruned the network
- Why single-shot?
- Two problems:
  - 1. Unable to grow a layer
  - 2. Unstructured pruning

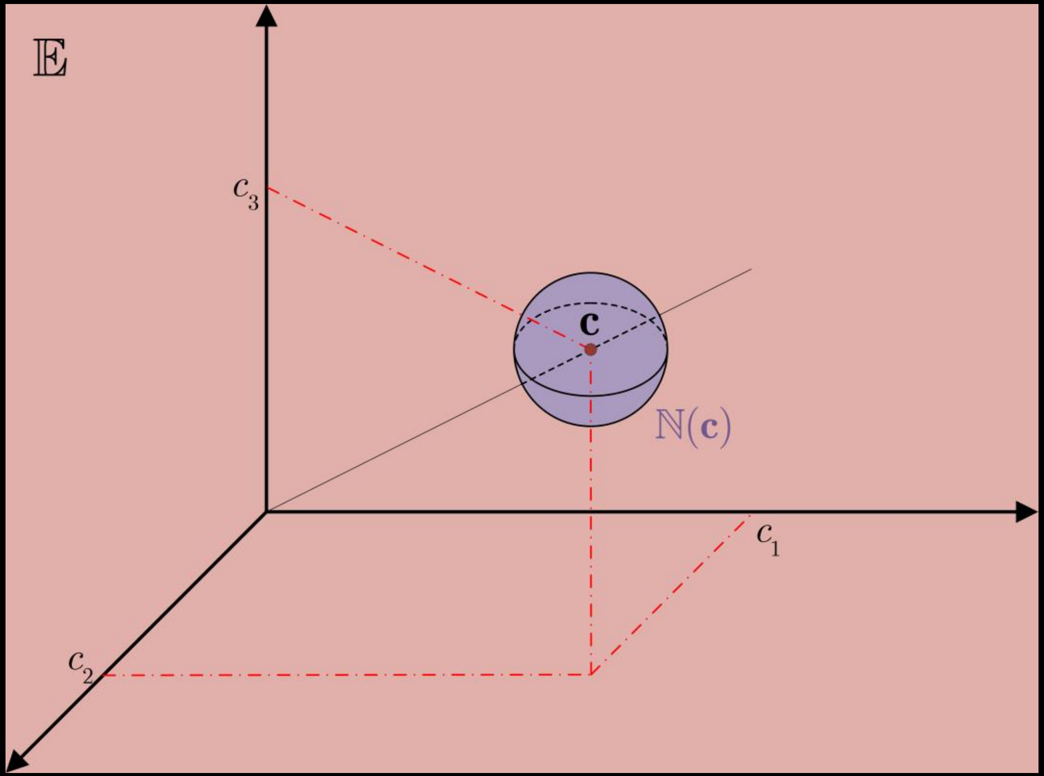


Layer 6, Epoch 1 — Layer 16, Epoch 1

Layer 6, Epoch 4 — Layer 16, Epoch 4

Question 2: How to identify an LW-DNA model efficiently?

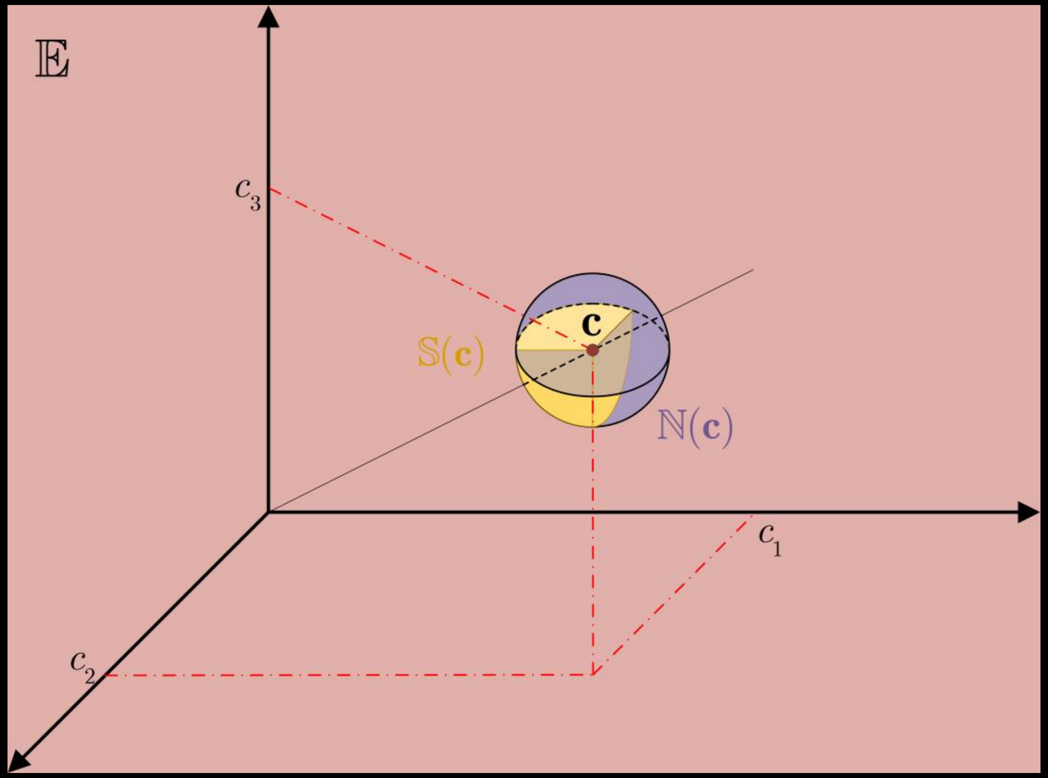
- Problem One: Unable to grow a layer
- Channel configuration vector in the configuration space
  - Assembly of channel number into a vector.



Methodology

Question 2: How to identify an LW-DNA model efficiently?

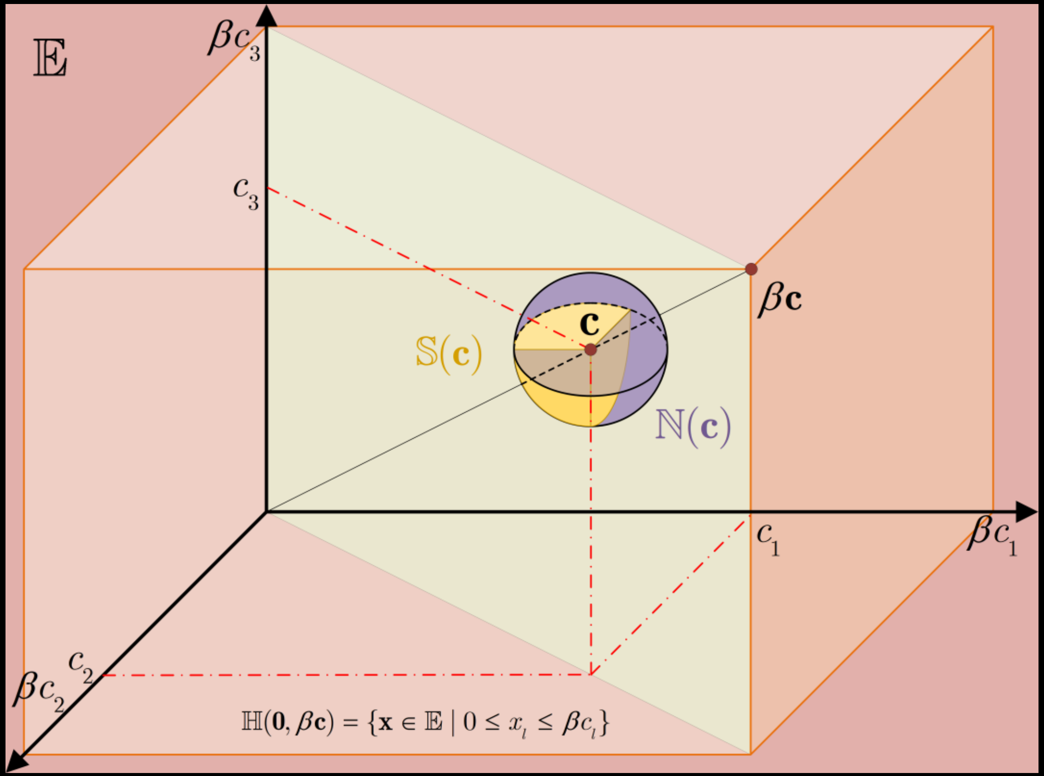
- **Problem One: Unable to grow a layer**



Methodology

Question 2: How to identify an LW-DNA model efficiently?

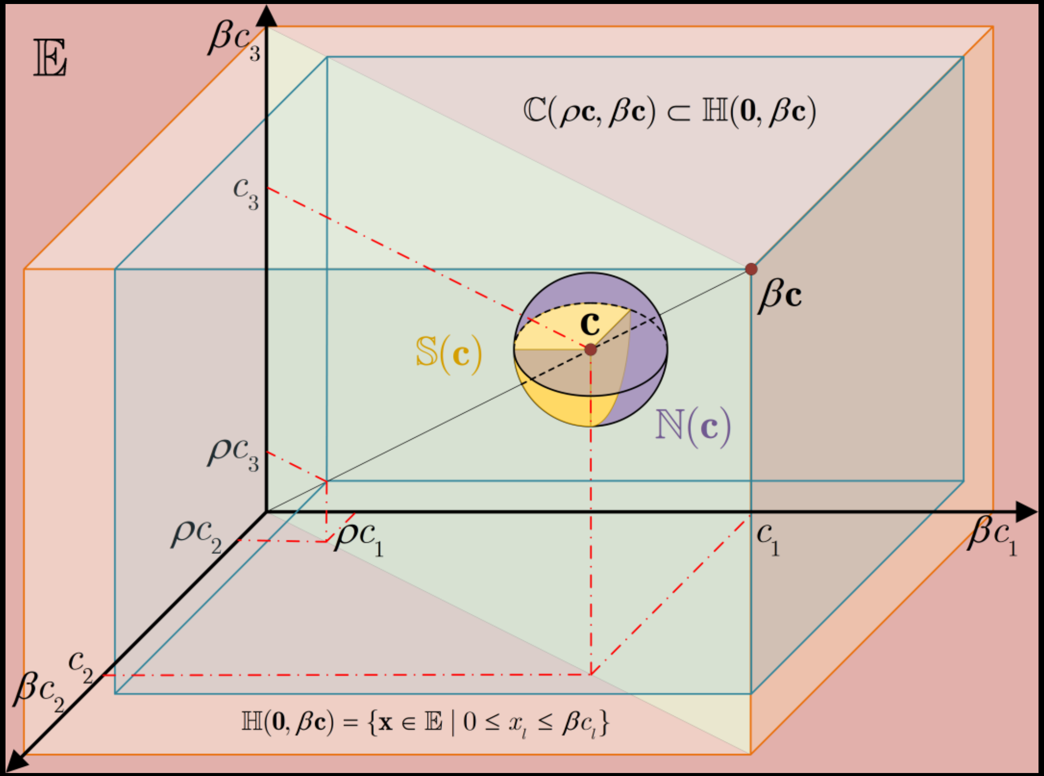
- **Problem One: Unable to grow a layer**
  - **Solution: Expand the network by an upscaling factor**



Methodology

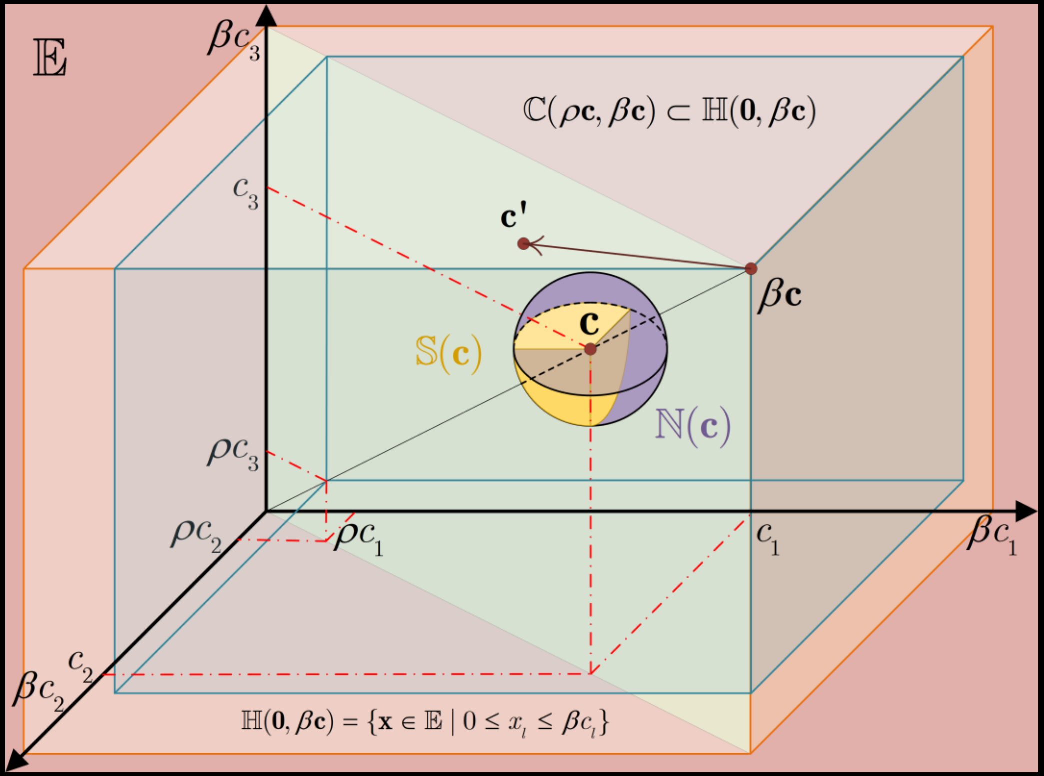
Question 2: How to identify an LW-DNA model efficiently?

- **Problem One: Unable to grow a layer**
  - **Solution: Expand the network by an upscaling factor**
  - **Constrain the minimum channel width by a factor ρ**



Question 2: How to identify an LW-DNA model efficiently?

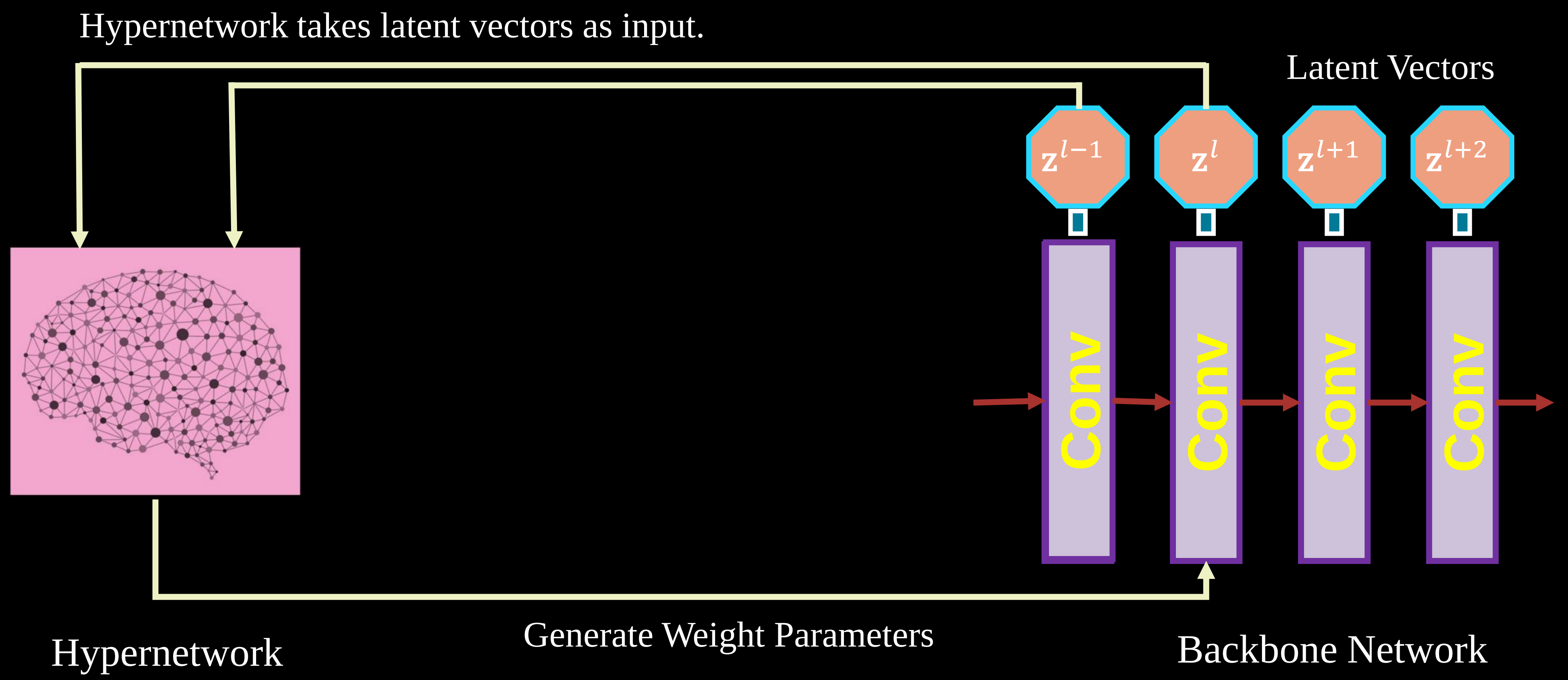
- Problem One: Unable to grow a layer
  - Shrink to the optimal solution $\mathbf{c}'$



Methodology

Question 2: How to identify an LW-DNA model efficiently?

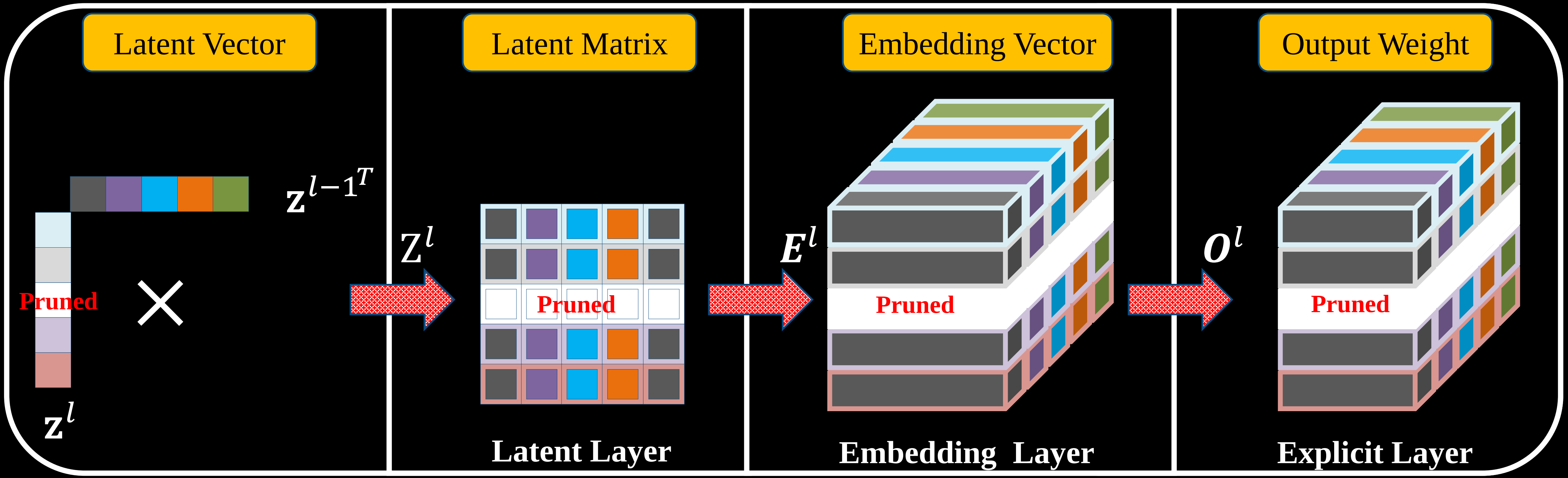
- **Problem 2: Unstructured Pruning**
  - **Reparameterization of the network**



[1] Yawei Li, Shuhang Gu, Kai Zhang, Luc Van Gool, Radu Timofte. **DHP: Differentiable Meta Pruning via HyperNetworks. ECCV 2020.**

Question 2: How to identify an LW-DNA model efficiently?

- **Problem 2: Unstructured Pruning**
  - **Reparameterization of the network**



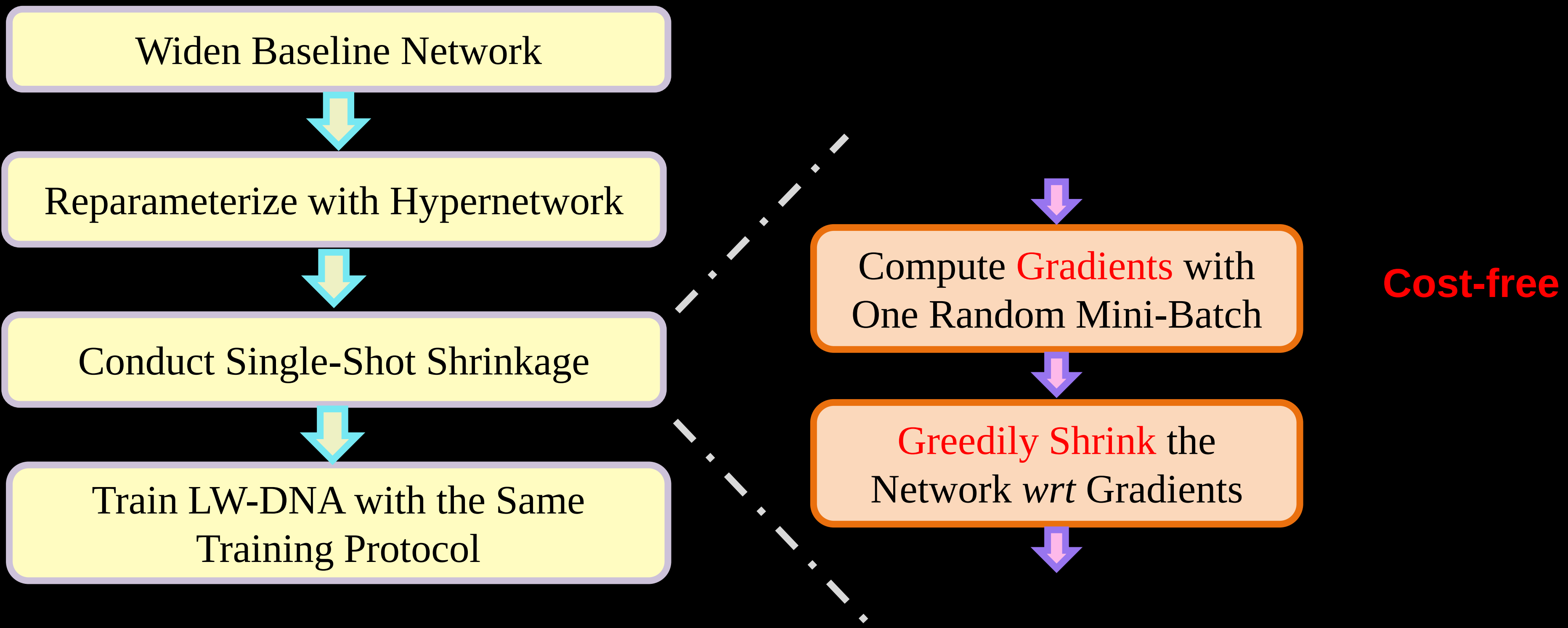
Hypernetwork.

[1] Yawei Li, Shuhang Gu, Kai Zhang, Luc Van Gool, Radu Timofte. **DHP: Differentiable Meta Pruning via HyperNetworks. ECCV 2020.**

Methodology

Question 2: How to identify an LW-DNA model efficiently?

- **Steps of the architecture optimization method**



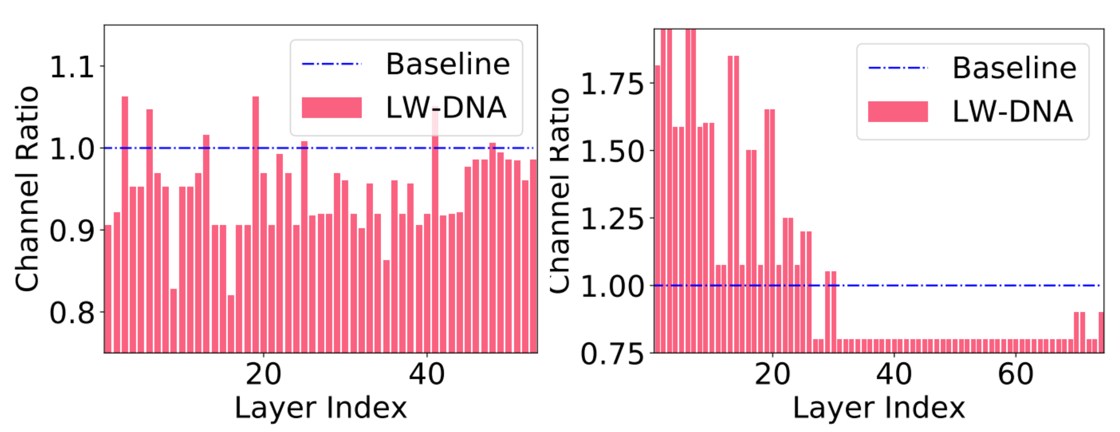
Explanation:

Why LW-DNA models performs better?

Explaination

Question 3: How to explain the benefits of LW-DNA?

- **CNNs are redundant.**
  - It is possible to find a layer-wise specific channel configuration comparable with the baseline under lower model complexity.
- **The redistribution of computational budget could help to improve the performance.**

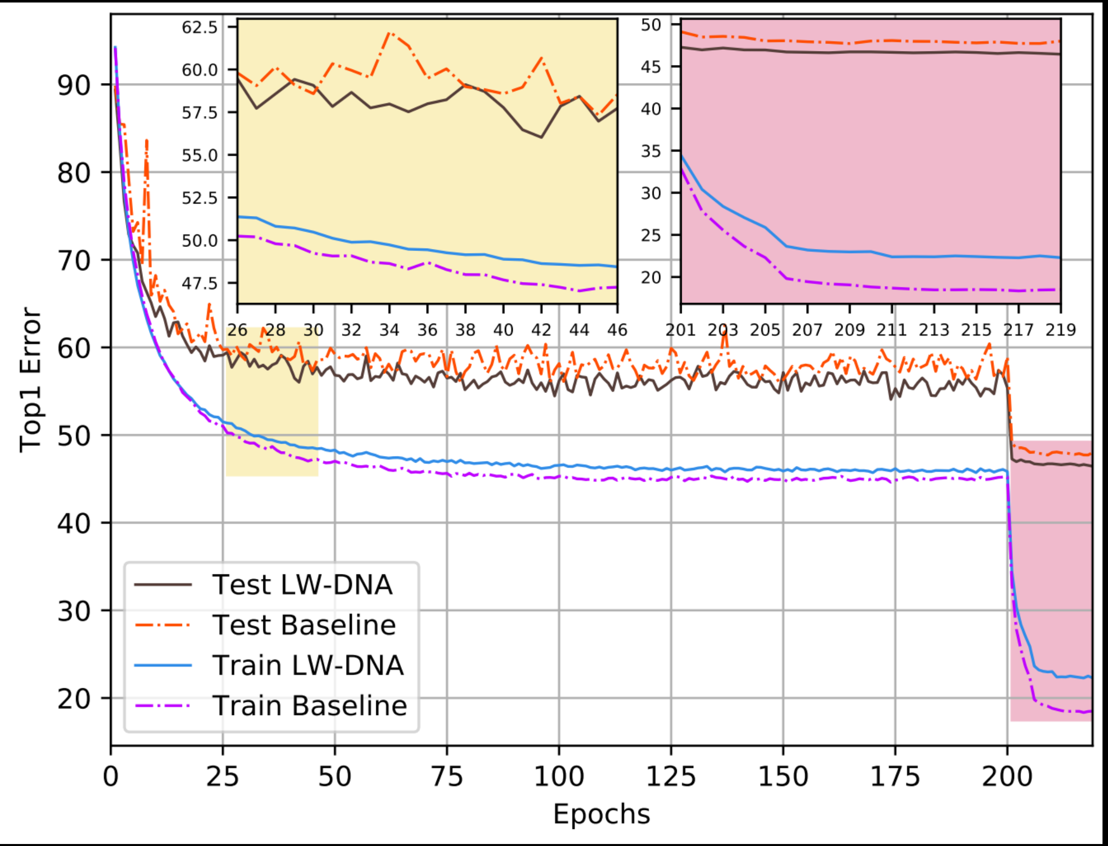


(a) ResNet50, ImageNet. (b) RegNet-4GF, ImageNet.

Explanation

Question 3: How to explain the benefits of LW-DNA?

- **Maybe related to overfitting**
  - **Evidence one: training and test log.**



MobileNetV1

Explanation

Question 3: How to explain the benefits of LW-DNA?

- **Maybe related to overfitting**
  - **Evidence three:** On the same dataset, it is easier to identify an LW-DNA model version for larger networks than for smaller networks.

Dataset	Network	Method	Top-1 Error (%)	FLOPs [G] / Ratio (%)	Params [M] / Ratio (%)
ImageNet [6]	ResNet50 [14]	Baseline	23.28	4.1177 / 100.0	25.557 / 100.0
		LW-DNA	23.00	3.7307 / 90.60	23.741 / 92.90
	RegNet [39] X-4.0GF	Baseline	23.05	4.0005 / 100.0	22.118 / 100.0
		LW-DNA	22.74	3.8199 / 95.49	15.285 / 69.10
	MobileNetV3 small [16]	Baseline	34.91	0.0612 / 100.0	3.108 / 100.0
		LW-DNA	34.84	0.0605 / 98.86	3.049 / 98.11

Table 4: Image classification results.

Explanation

Question 3: How to explain the benefits of LW-DNA?

- **Maybe related to overfitting**
  - **Evidence two:** The accuracy gain of an LW-DNA model is larger for smaller datasets (Tiny-ImageNet) compared with larger datasets (ImageNet).

Dataset	Network	Method	Top-1 Error (%)	FLOPs [G] / Ratio (%)	Params [M] / Ratio (%)
ImageNet [6]	MobileNetV3 small [16]	Baseline	34.91	0.0612 / 100.0	3.108 / 100.0
		LW-DNA	34.84	0.0605 / 98.86	3.049 / 98.11
Tiny-ImageNet	MobileNetV1 [17]	Baseline	51.87	0.0478 / 100.0	3.412 / 100.0
		Baseline KD	48.00	0.0478 / 100.0	3.412 / 100.0
		LW-DNA	46.44	0.0460 / 96.23	1.265 / 37.08
	MobileNetV2 [44]	Baseline	44.38	0.0930 / 100.0	2.480 / 100.0
		Baseline KD	41.25	0.0930 / 100.0	2.480 / 100.0
		LW-DNA	40.74	0.0872 / 93.76	2.230 / 89.90

Table 5: Image classification results.

Explanation

Extension to other vision tasks

- **Visual Tracking**

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.

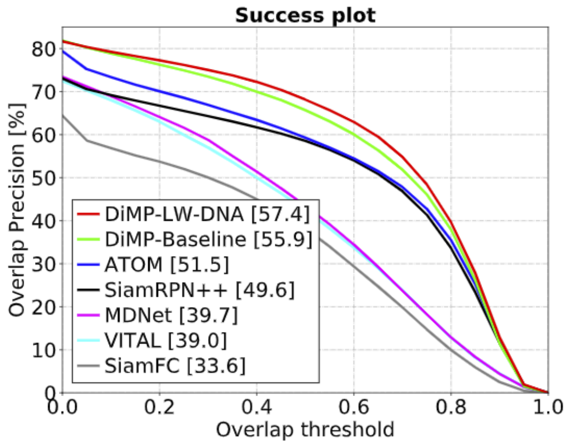


Figure 6: Success plot on the LaSOT dataset for visual tracking.

Explanation

Extension to other vision tasks

- **Single image super-resolution**

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 $\times 4$.

Conclusion

Conclusion

- We empirically validate the heterogeneity hypothesis proposed in this paper.
  - It's possible to identify an LW-DNA model.
  - This could be used as a post-searching mechanism complementary to semi- or fully automated neural architecture search.
- Secondly, an almost cost-free fine-grained architecture optimization method is proposed.
  - This method only needs the computation of one random batch.
- Thirdly, the possible reason for the improved performance of an LW-DNA is explained by observing the experimental results.

Thanks for your attention!
Q&A