Multicoated Supermasks Enhance Hidden Networks

Yasuyuki Okoshi*, Ángel López García-Arias*, Kazutoshi Hirose, Kota Ando,
Kazushi Kawamura, Thiem Van Chu, Masato Motomura, Jaehoon Yu*

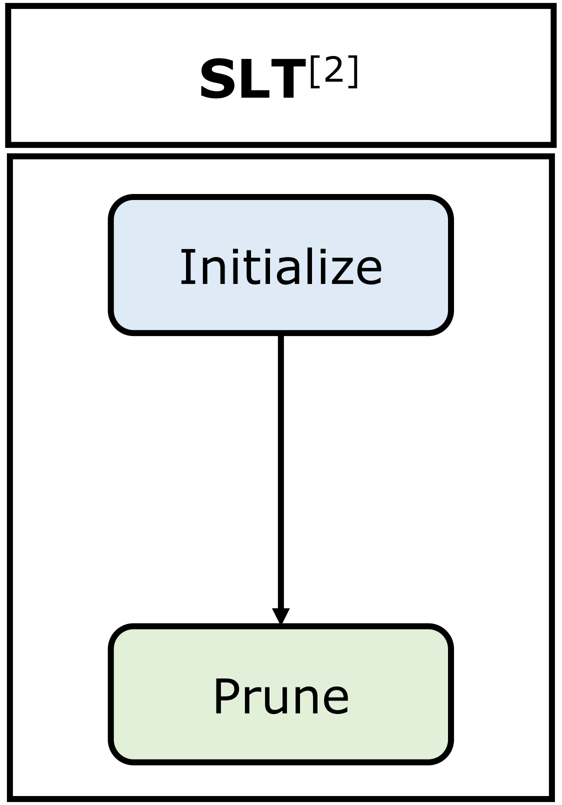
*Equally-Credited Authors

Code available

International Conference on Machine Learning

New Pruning Scheme : Strong Lottery Ticket (SLT)

SLT is a neural network obtained by learning only connections instead of weights

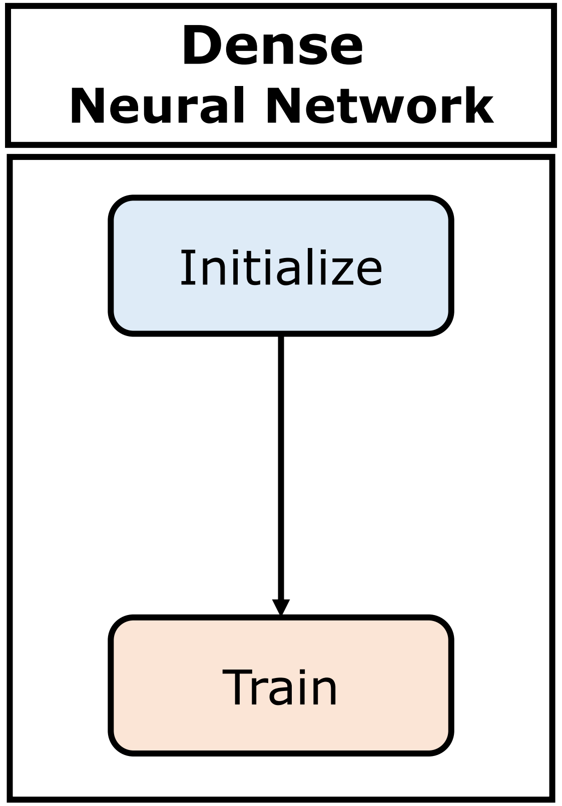


Find sparse NN without training

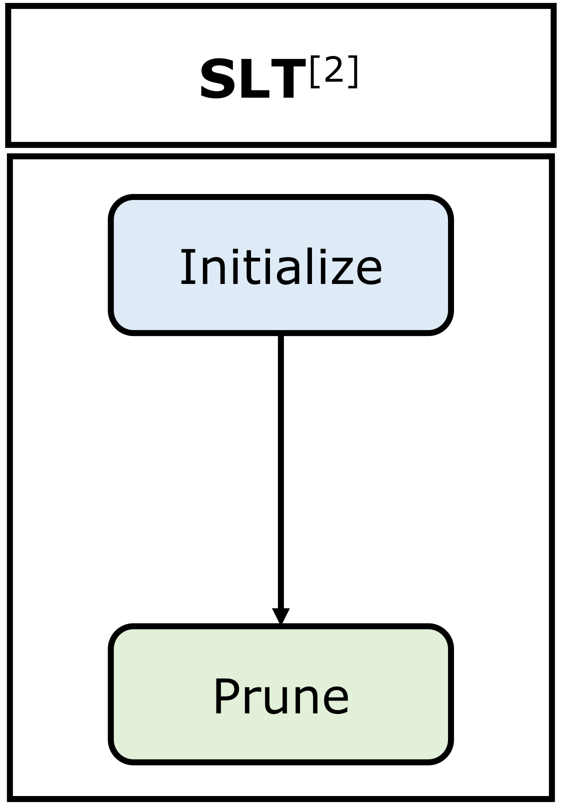
[1] J.Frankle, and C.Michael. "The Lottery Ticket Hypothesis: Finding Sparse, Trainable Neural Networks." ICLR. 2018.
[2] E.Malach, et al. "Proving the lottery ticket hypothesis: Pruning is all you need." ICML. 2020.

New Pruning Scheme : Strong Lottery Ticket (SLT)

SLT is a neural network obtained by learning only connections instead of weights



Overparametrized NN ⇨



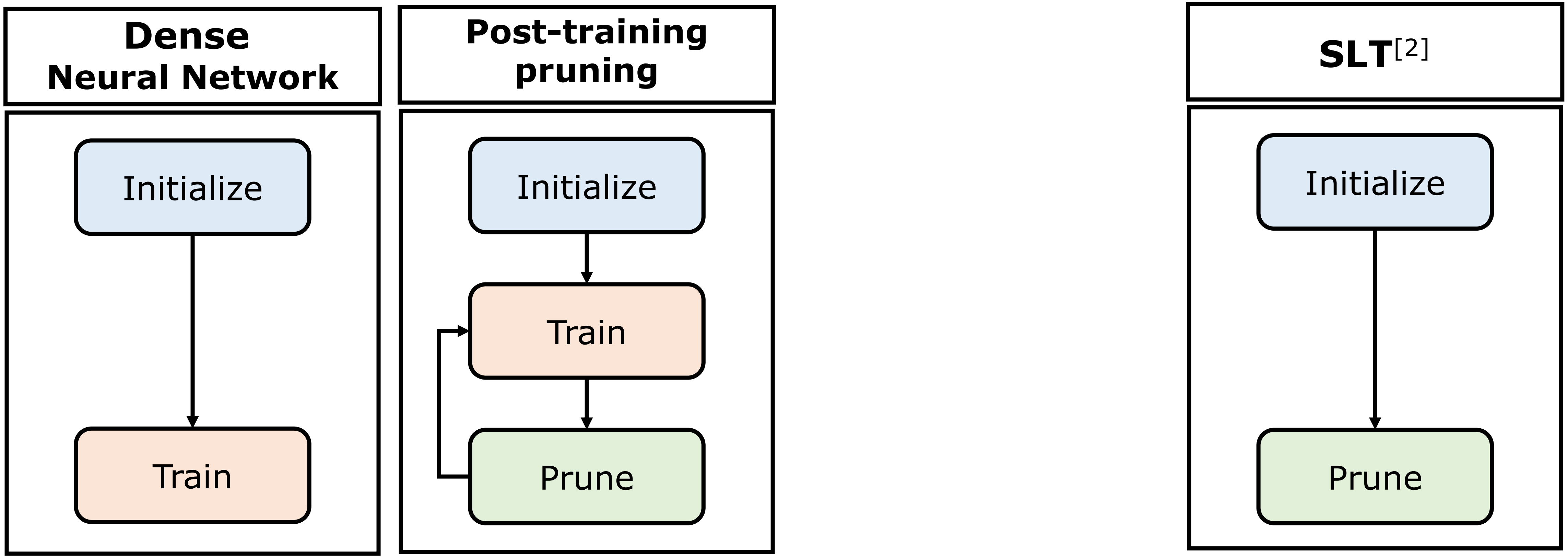
Find sparse NN without training

[1] J.Frankle, and C.Michael. "The Lottery Ticket Hypothesis: Finding Sparse, Trainable Neural Networks." ICLR. 2018.

[2] E.Malach, et al. "Proving the lottery ticket hypothesis: Pruning is all you need." ICML. 2020.

New Pruning Scheme : Strong Lottery Ticket (SLT)

SLT is a neural network obtained by learning only connections instead of weights



Dense Neural Network

Initialize

Train

Post-training pruning

Initialize

Train

Prune

SLT[2]

Initialize

Prune

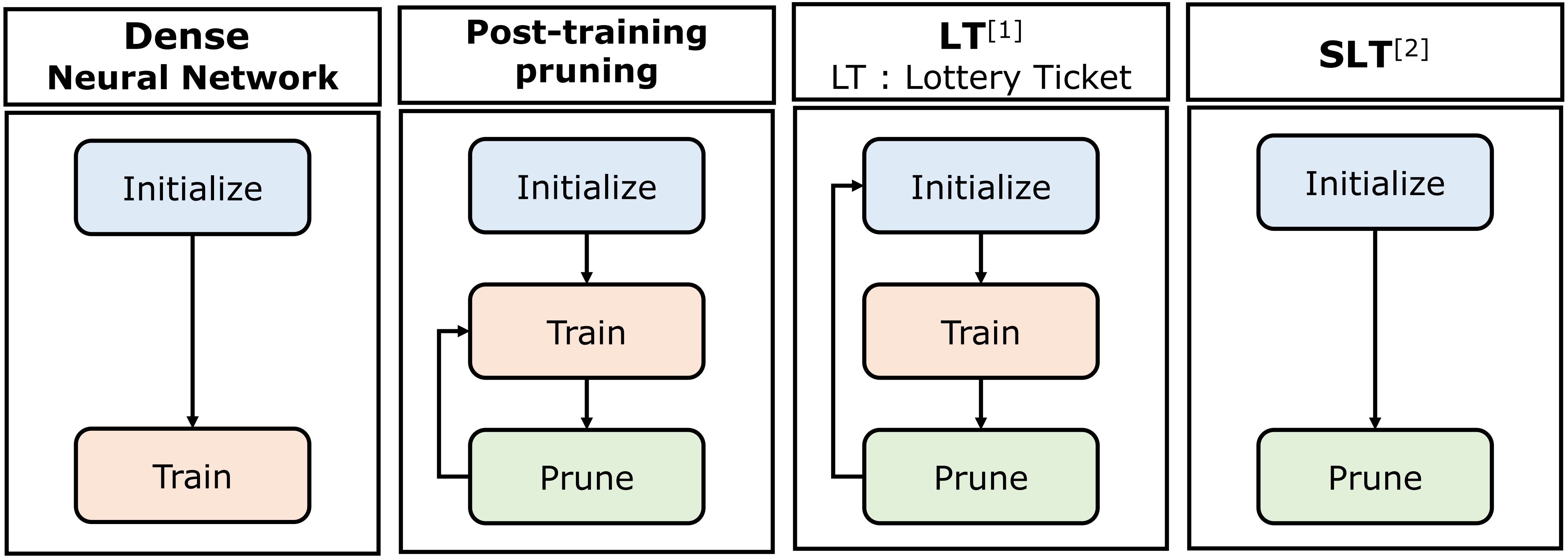
Overparametrized NN ⇨ Find sparse NN after training ⇨ Find sparse NN without training

[1] J.Frankle, and C.Michael. "The Lottery Ticket Hypothesis: Finding Sparse, Trainable Neural Networks." ICLR. 2018.

[2] E.Malach, et al. "Proving the lottery ticket hypothesis: Pruning is all you need." ICML. 2020.

New Pruning Scheme : Strong Lottery Ticket (SLT)

SLT is a neural network obtained by learning only connections instead of weights



Dense Neural Network

Initialize

Train

Post-training pruning

Initialize

Train

Prune

LT[1]

LT : Lottery Ticket

Initialize

Train

Prune

SLT[2]

Initialize

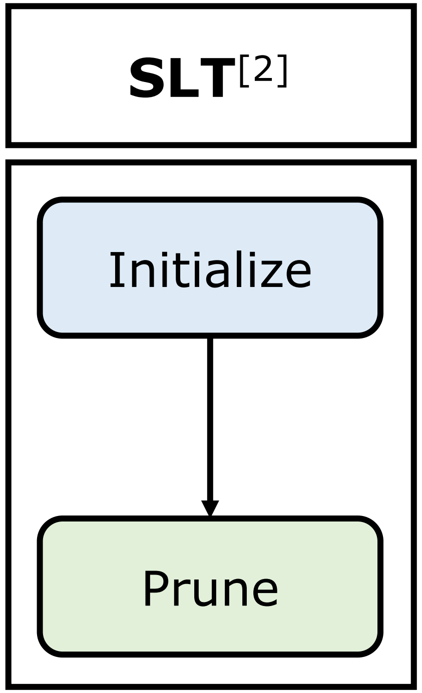
Prune

Overparametrized NN ⇨ Find sparse NN after training ⇨ Find trainable sparse NN ⇨ Find sparse NN without training

[1] J.Frankle, and C.Michael. "The Lottery Ticket Hypothesis: Finding Sparse, Trainable Neural Networks." ICLR. 2018.

[2] E.Malach, et al. "Proving the lottery ticket hypothesis: Pruning is all you need." ICML. 2020.

Comparison of SLT and Dense NN

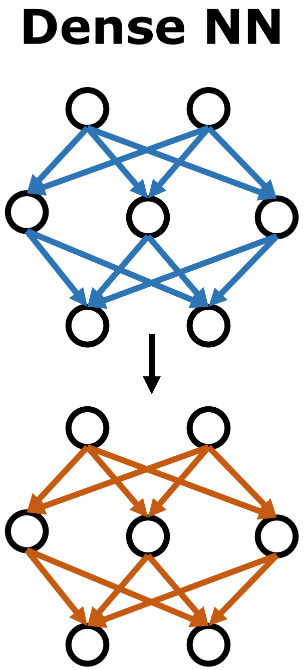


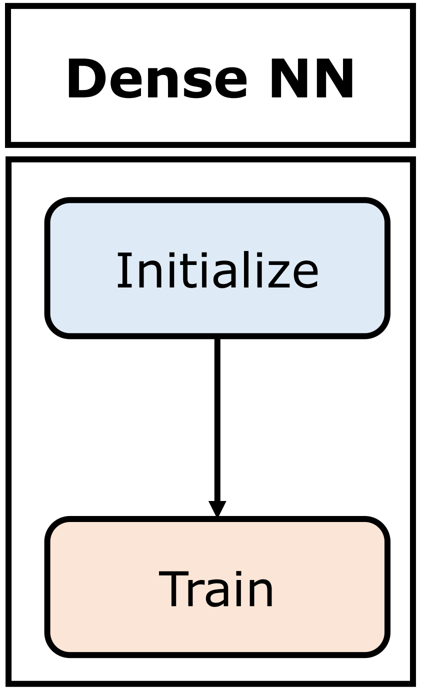
Pros of SLT

Small model size

Model Information

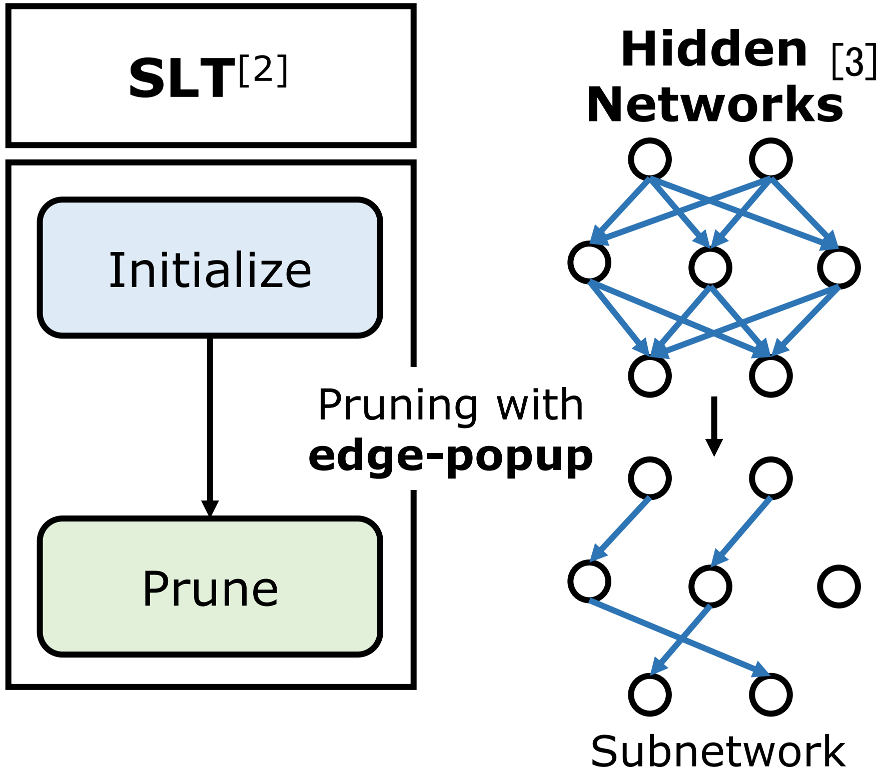
Connectivity (Supermask)	Learned weights





[3] V.Ramanujan, et al. "What's hidden in a randomly weighted neural network?." CVPR. 2020.

Comparison of SLT and Dense NN

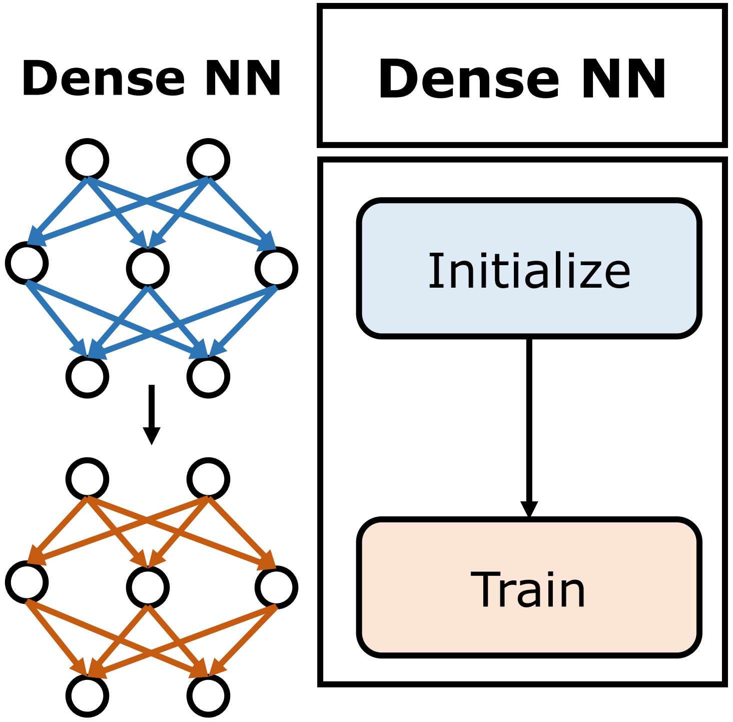


Pros of SLT

Small model size

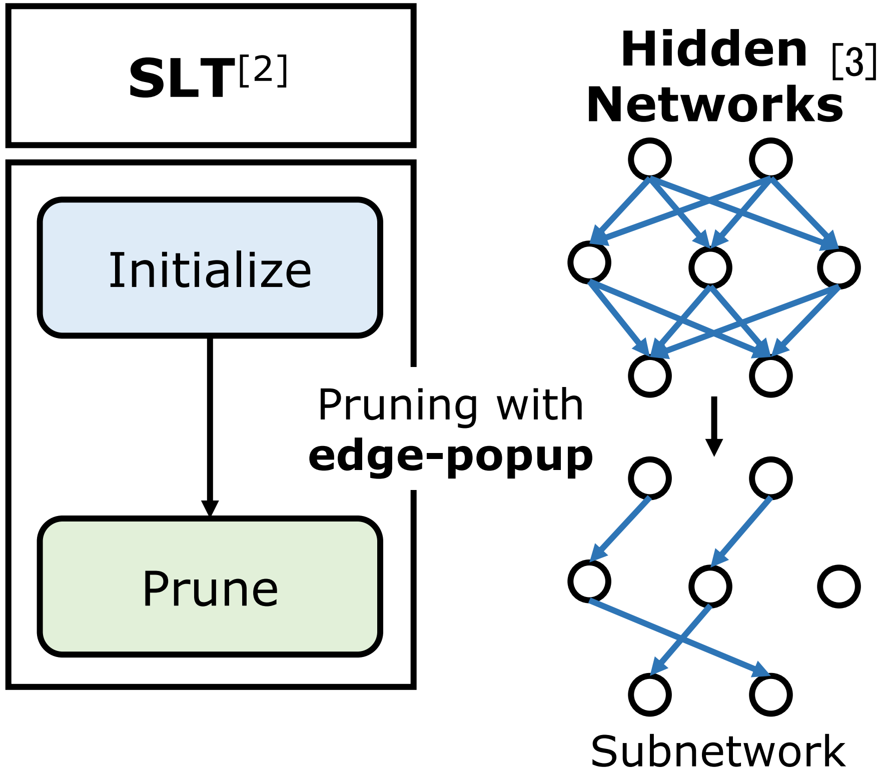
Model Information

Connectivity (Supermask)	Learned weights



[3] V.Ramanujan, et al. "What's hidden in a randomly weighted neural network?." CVPR. 2020.

Comparison of SLT and Dense NN



Pros of SLT

Small model size

Model Information

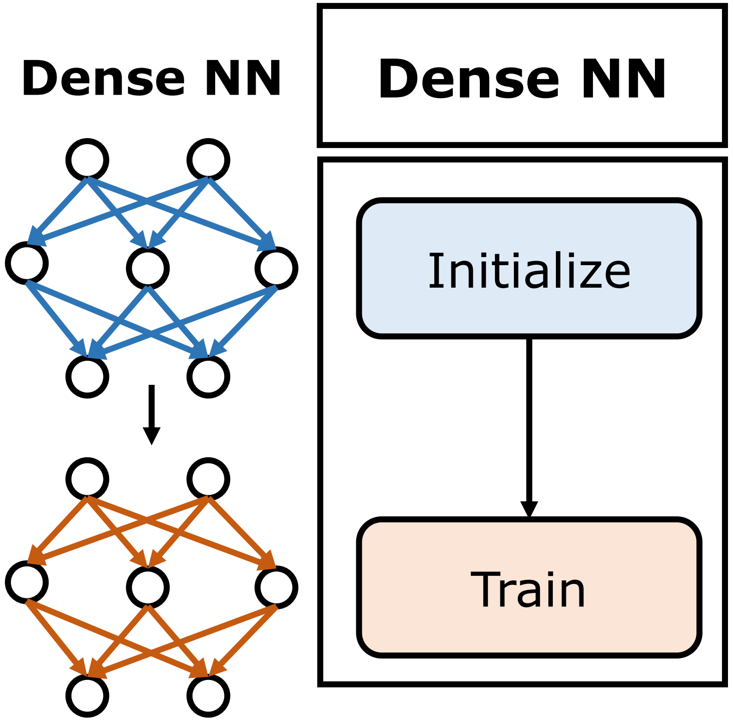
Connectivity (Supermask)	Learned weights

Cons of SLT

Lower accuracy

Search Space

Supermask	Weights



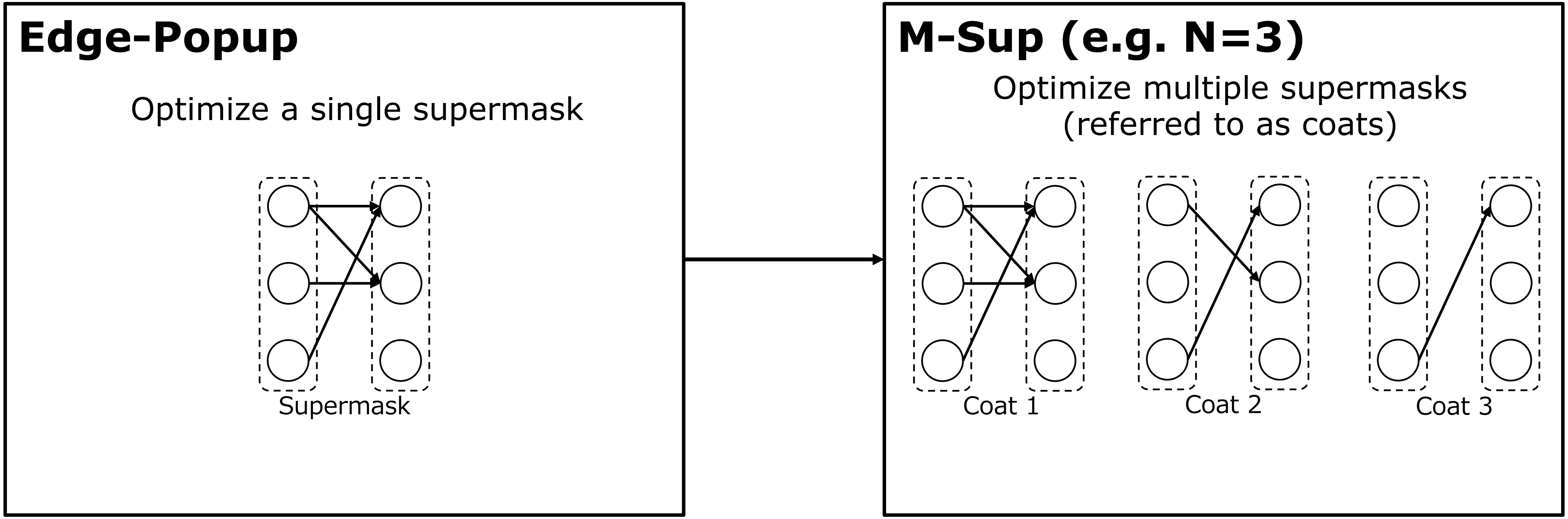
⇩

Multicoated Supermask **extend edge-popup to use multiple supermasks**

[3] V.Ramanujan, et al. "What's hidden in a randomly weighted neural network?." CVPR. 2020.

Multicoated Supermasks (M-Sup)

***Multicoated Supermasks* optimize multiple supermasks simultaneously**



Edge-Popup

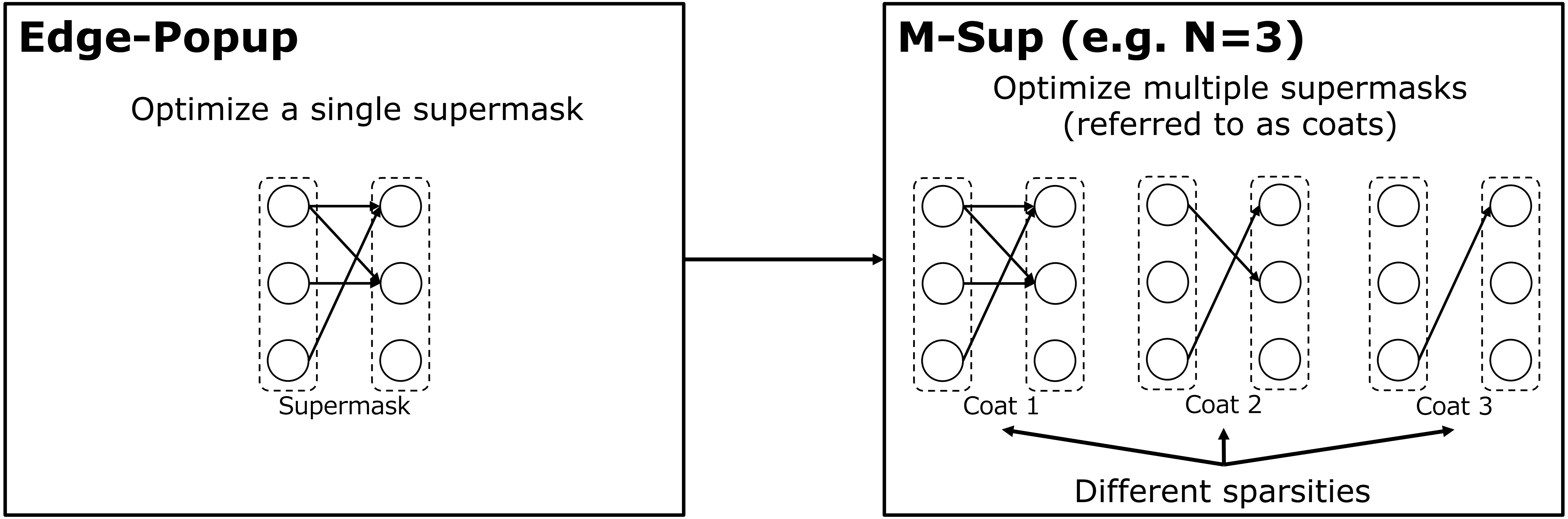
Optimize a single supermask

M-Sup (e.g. N=3)

Optimize multiple supermasks (referred to as coats)

Multicoated Supermasks (M-Sup)

Multicoated Supermasks **optimize multiple supermasks simultaneously**



Edge-Popup

Optimize a single supermask

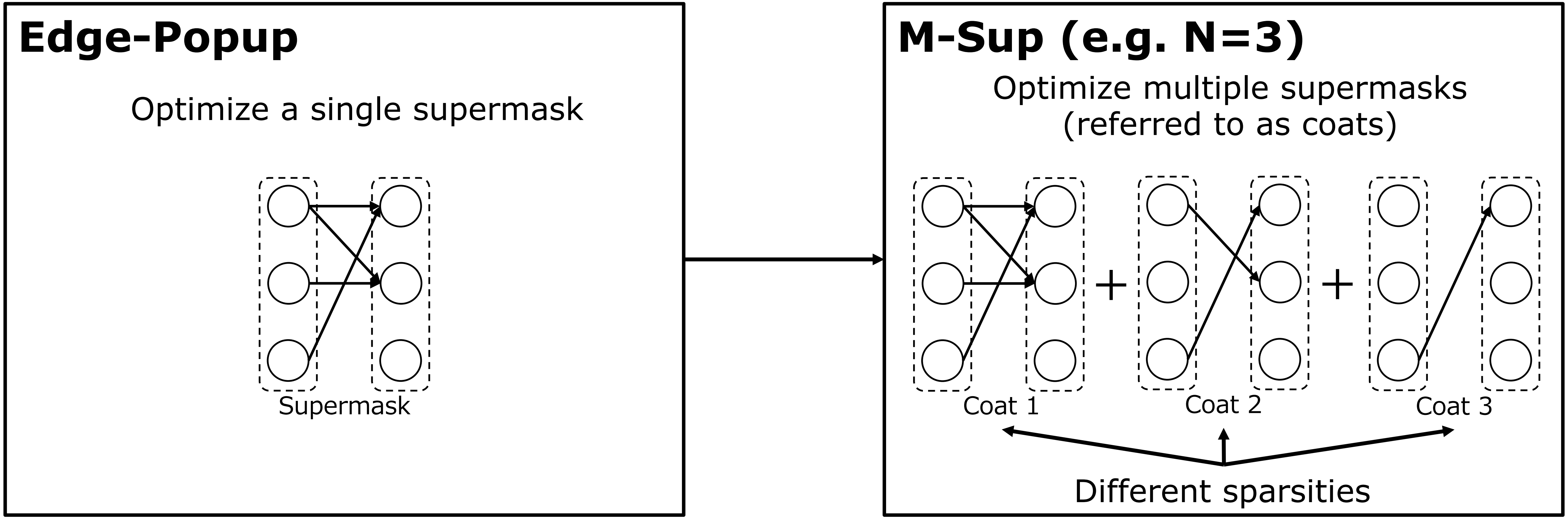
M-Sup (e.g. N=3)

Optimize multiple supermasks
(referred to as coats)

Different sparsities

Multicoated Supermasks (M-Sup)

Multicoated Supermasks **optimize multiple supermasks simultaneously**



Edge-Popup

Optimize a single supermask

M-Sup (e.g. N=3)

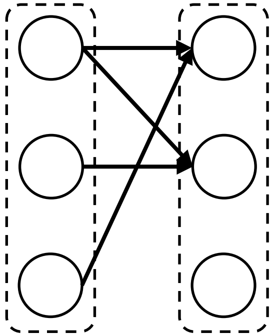
Optimize multiple supermasks (referred to as coats)

Multicoated Supermasks (M-Sup)

Compute subnetwork from random weights and supermasks

Edge-Popup

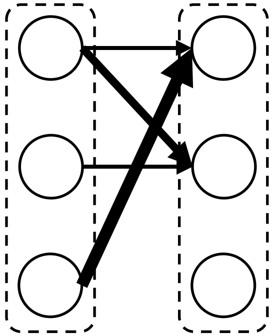
Optimize a single supermask



Supermask

M-Sup (e.g. N=3)

Optimize multiple supermasks
(referred to as coats)



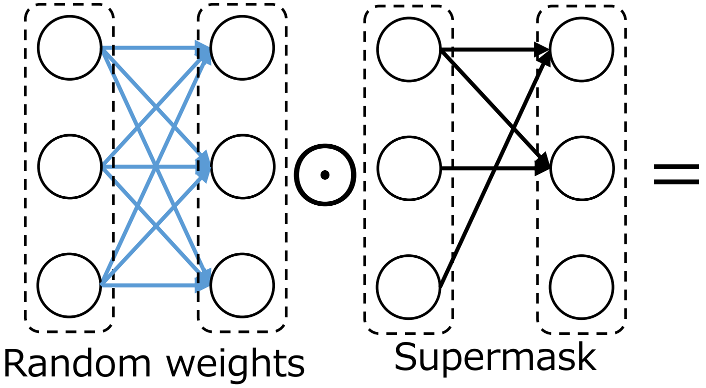
M-Sup

Multicoated Supermasks (M-Sup)

Compute subnetwork from random weights and supermasks

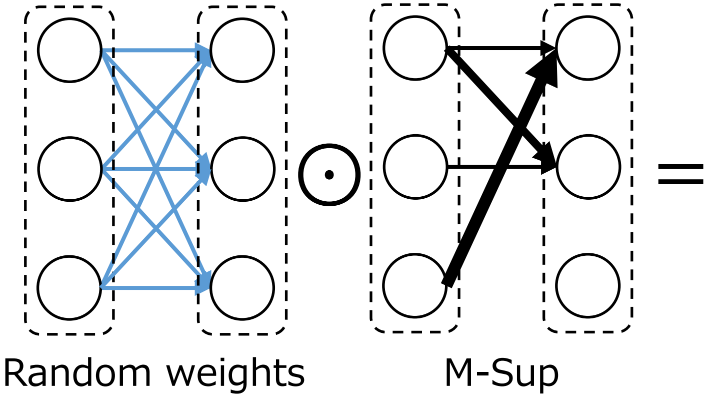
Edge-Popup

Optimize a single supermask



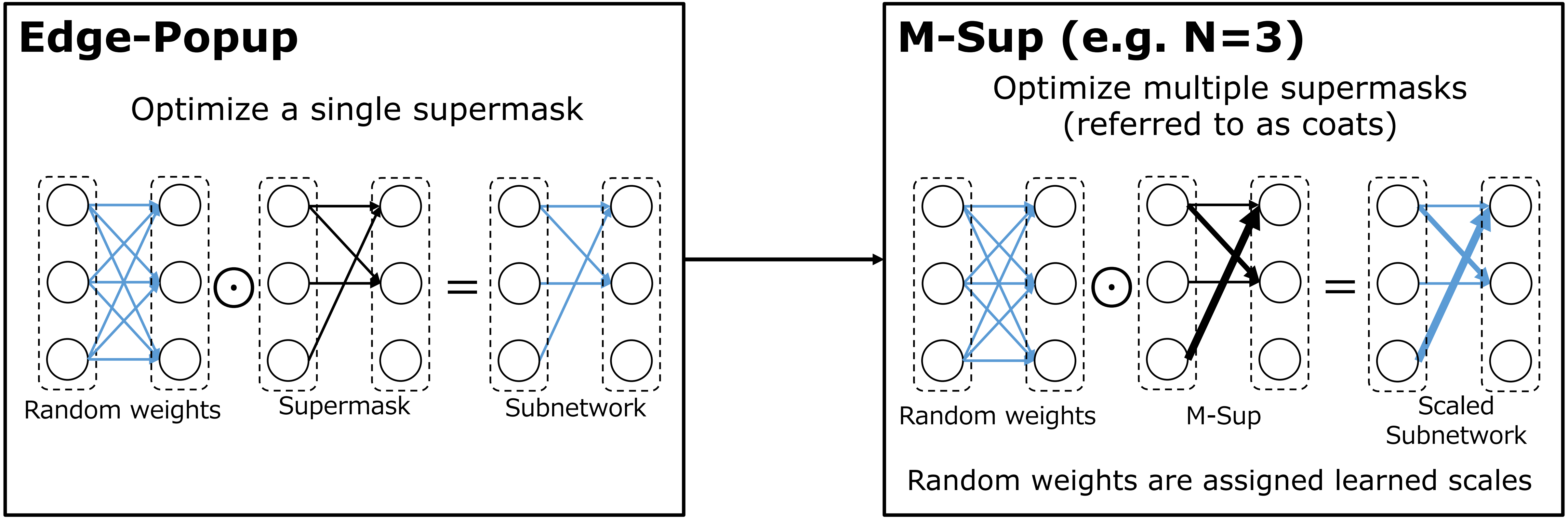
M-Sup (e.g. N=3)

Optimize multiple supermasks
(referred to as coats)



Multicoated Supermasks (M-Sup)

Compute subnetwork from random weights and supermasks



Edge-Popup

Optimize a single supermask

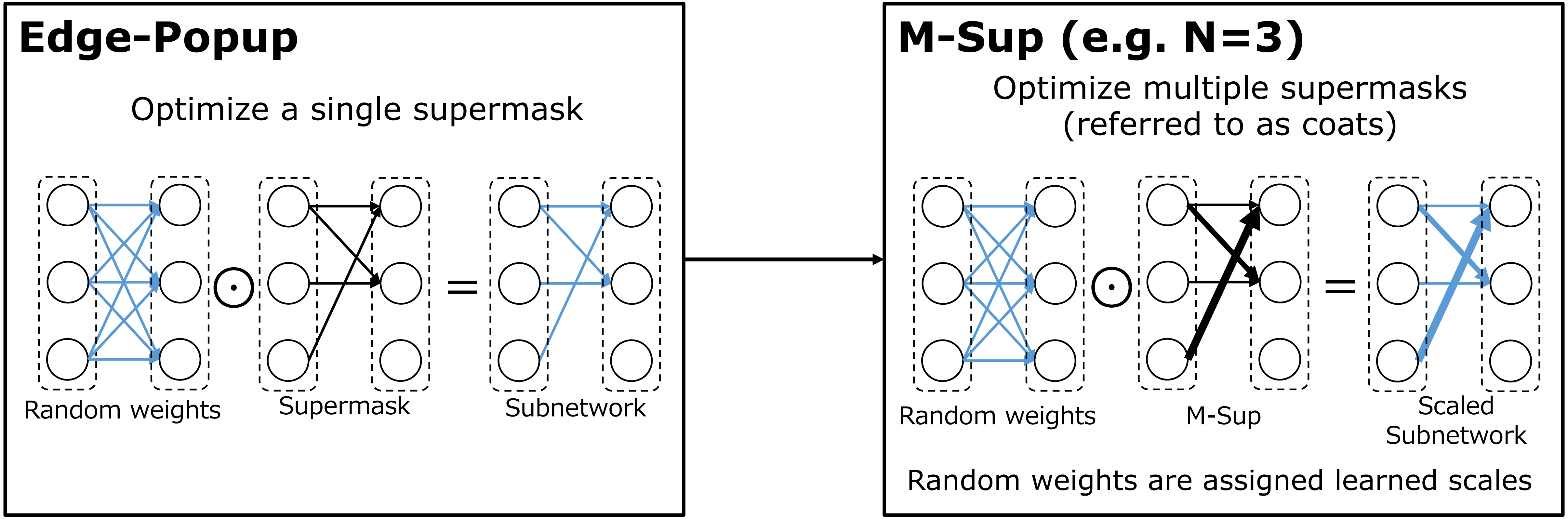
M-Sup (e.g. N=3)

Optimize multiple supermasks
(referred to as coats)

Random weights are assigned learned scales

Multicoated Supermasks (M-Sup)

Compute subnetwork from random weights and supermasks



Edge-Popup

Optimize a single supermask

M-Sup (e.g. N=3)

Optimize multiple supermasks
(referred to as coats)

Random weights are assigned learned scales

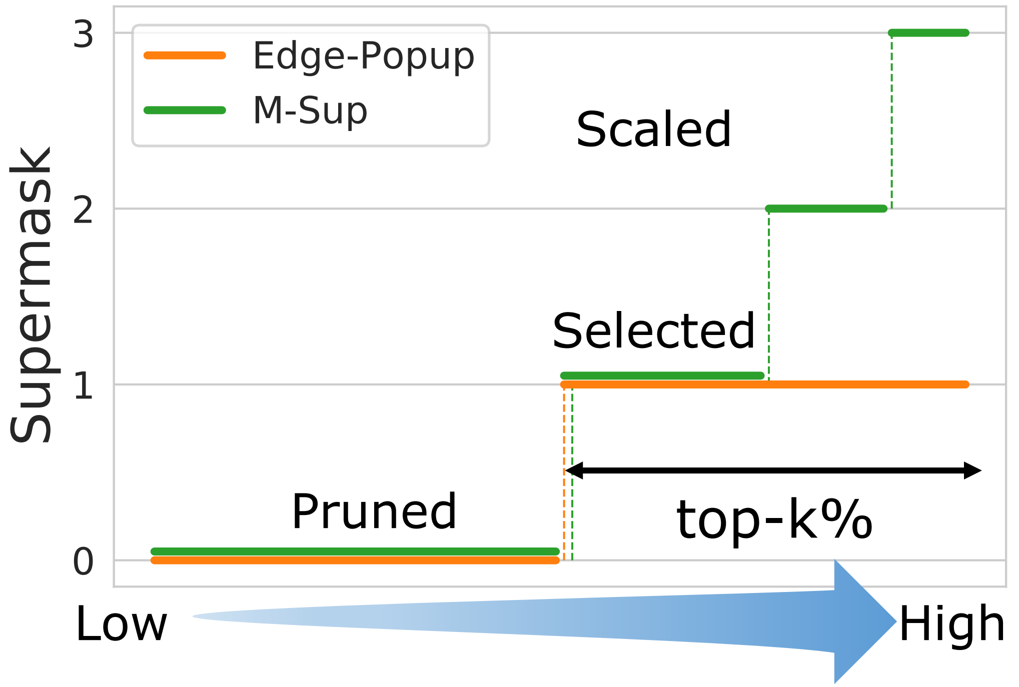
Multicoated Supermasks **expand search space with additional coats**

- Training connections of neural network
- Training scales of random weights

Edge-Popup vs M-Sup

Minimizing the additional cost and additional model size of our method

Score information
used more efficiently

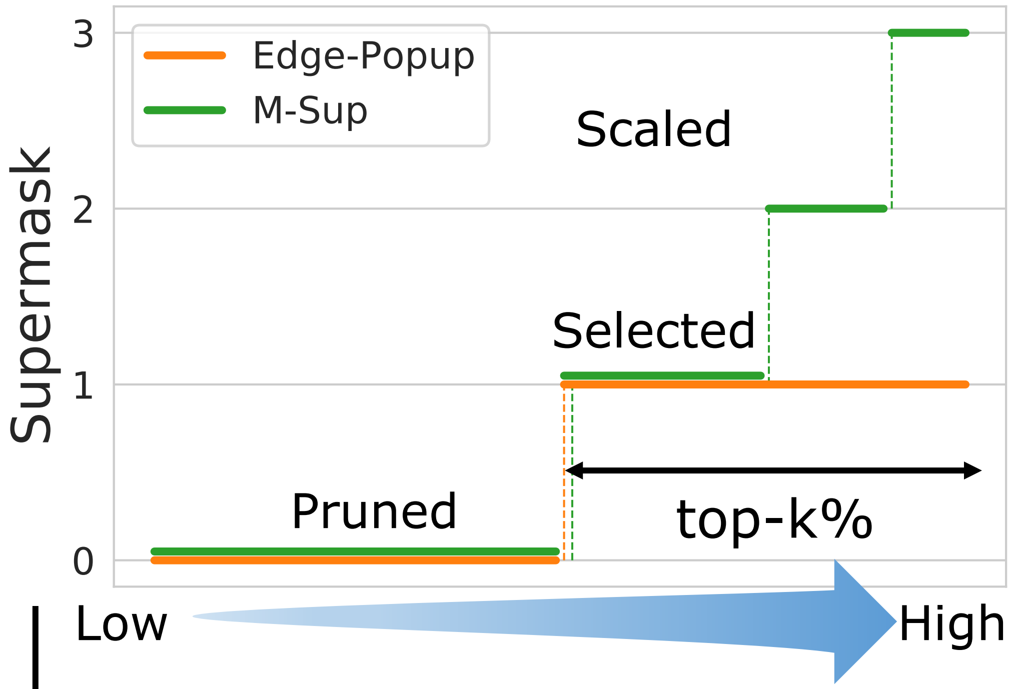


Edge-Popup vs M-Sup

Minimizing the additional cost and additional model size of our method

No additional pruning cost

Score information used more efficiently



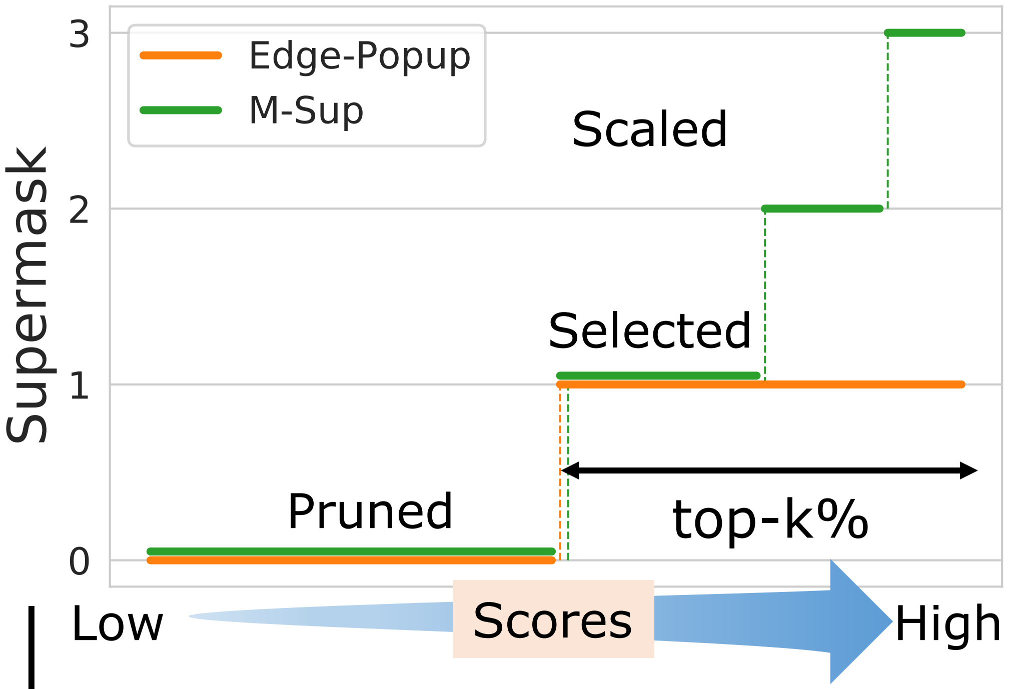
Edge-Popup vs M-Sup

Minimizing the additional cost and additional model size of our method

No additional pruning cost

Share pruning scores

Score information used more efficiently



Edge-Popup vs M-Sup

Minimizing the additional cost and additional model size of our method

No additional pruning cost

Share pruning scores

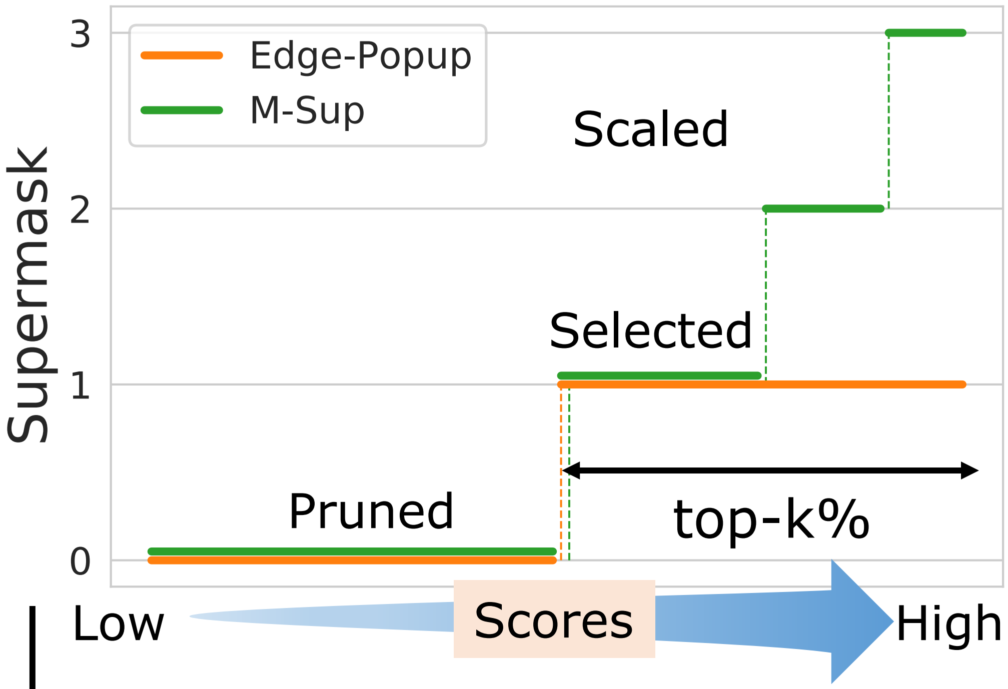
↓

No additional parameters

↓

No additional Training cost

Score information used more efficiently



Edge-Popup vs M-Sup

Minimizing the additional cost and additional model size of our method

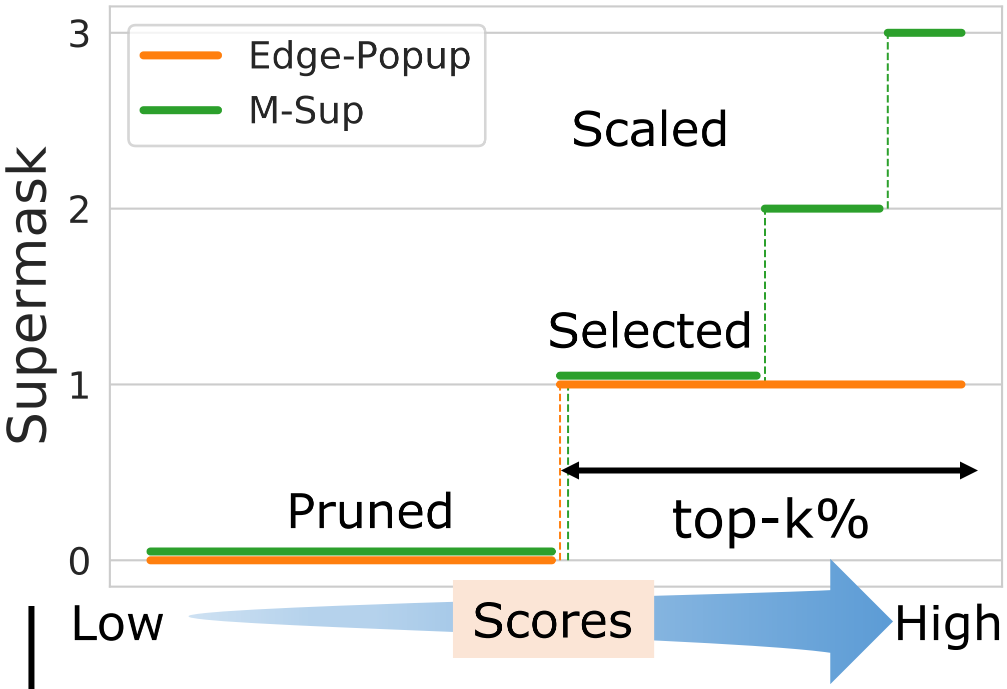
No additional pruning cost

Share pruning scores

No additional parameters

No additional Training cost

Score information used more efficiently



Small increase in model size

Edge-Popup vs M-Sup

Minimizing the additional cost and additional model size of our method

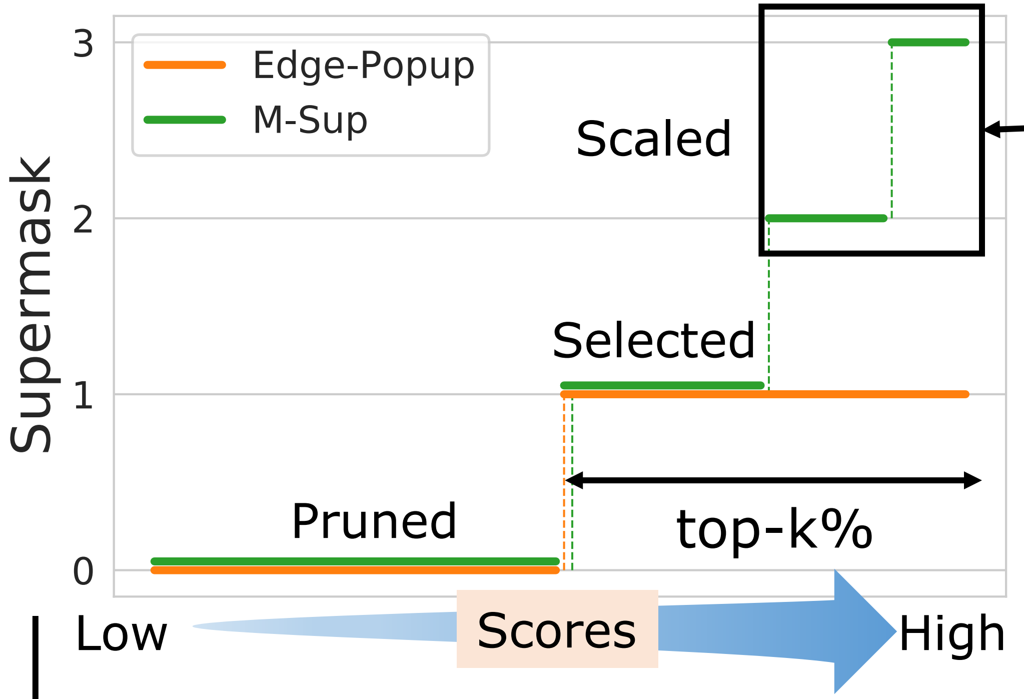
No additional pruning cost

Share pruning scores

No additional parameters

No additional Training cost

Score information used more efficiently



Small increase in model size

Additional masks are increasingly sparser

Edge-Popup vs M-Sup

Minimizing the additional cost and additional model size of our method

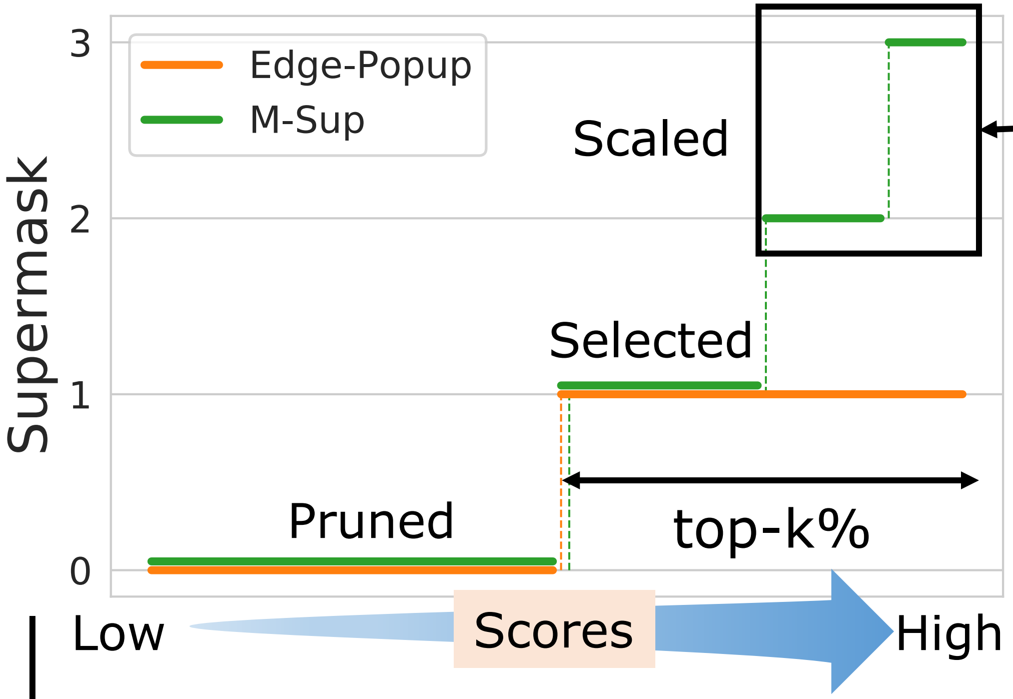
No additional pruning cost

Share pruning scores

No additional parameters

No additional Training cost

Score information used more efficiently



Small increase in model size

Additional masks are increasingly sparser

Entropy encoding

Larger mask

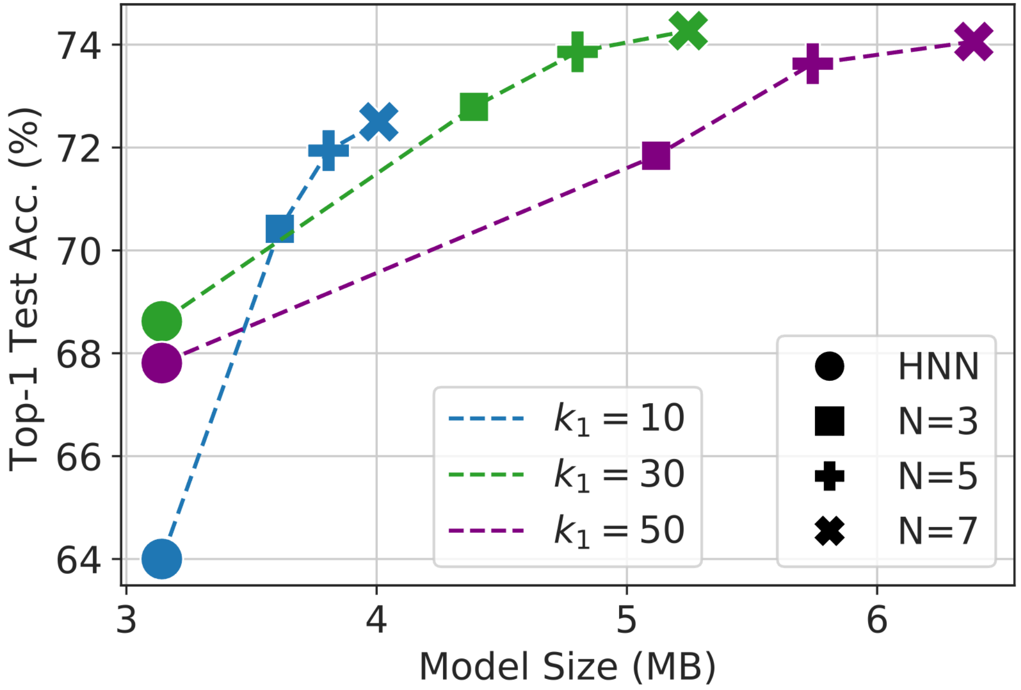
Mask	0	1	2	3
Code	0	10	110	111

Longer Length

Comparison – Model Size VS. Accuracy (ImageNet)

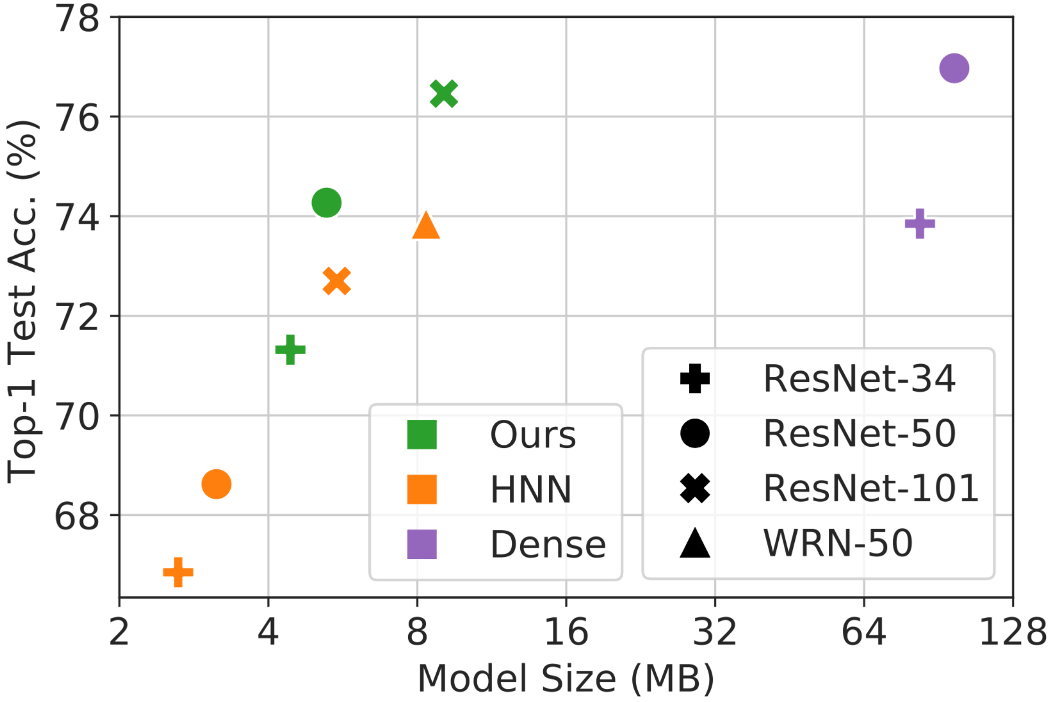
M-Sup achieve competitive results on ImageNet

- Comparison of #Coats
  - ResNet-50



- Comparison of Model Size

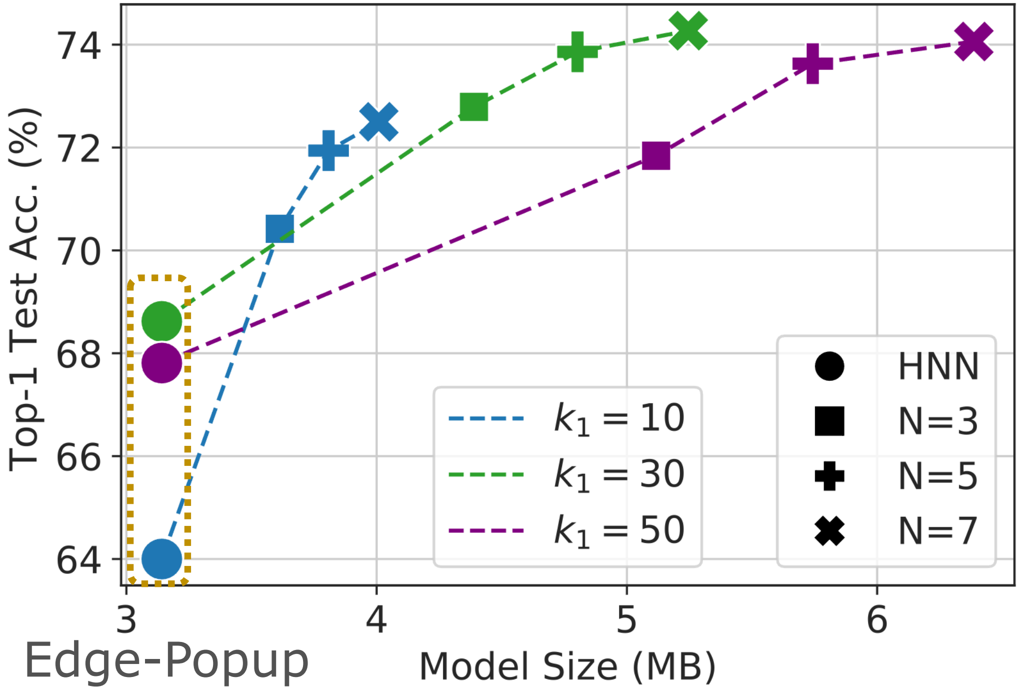
N=7,k_1=30



Comparison – Model Size VS. Accuracy (ImageNet)

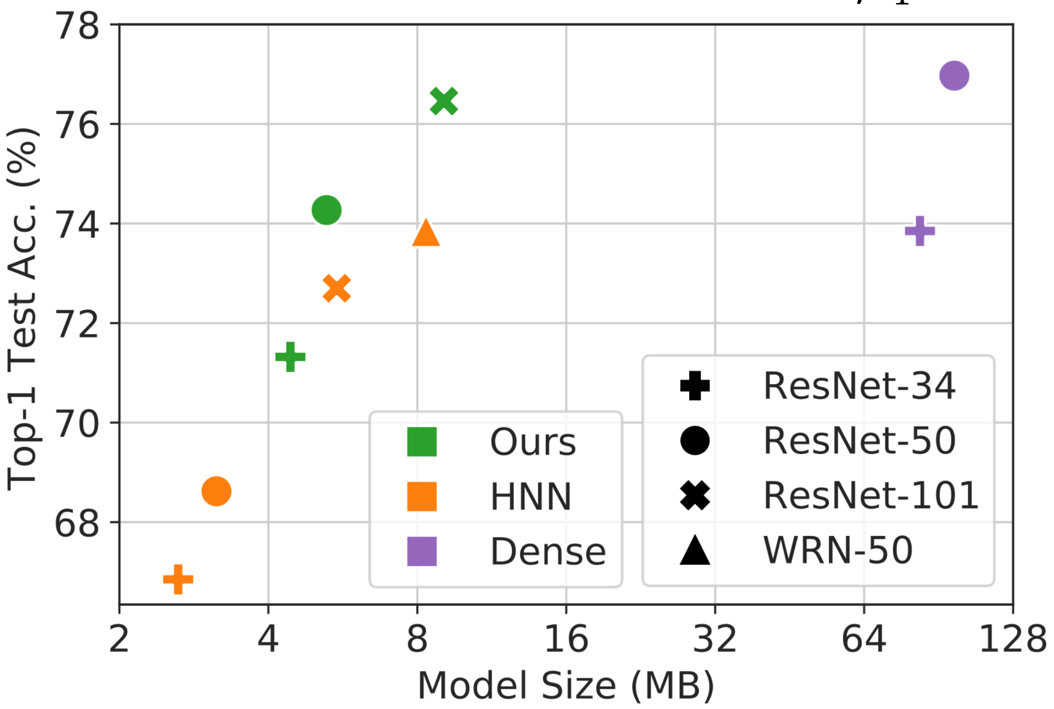
M-Sup achieve competitive results on ImageNet

- Comparison of #Coats
  - ResNet-50



- Comparison of Model Size

N=7,k_1=30

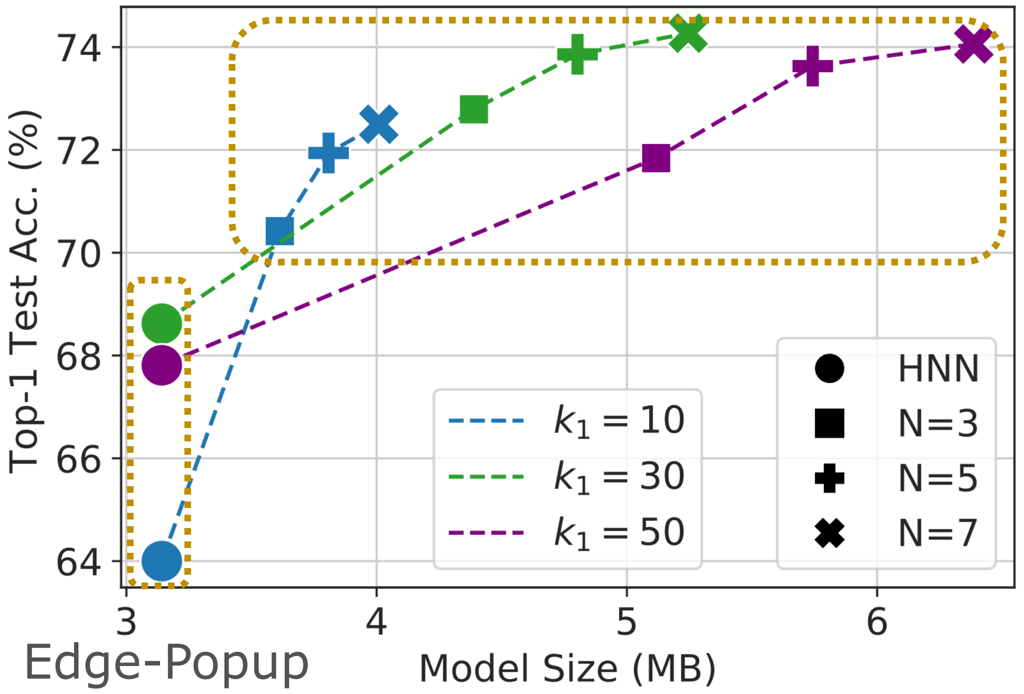


Comparison – Model Size VS. Accuracy (ImageNet)

M-Sup achieve competitive results on ImageNet

- Comparison of #Coats
  - ResNet-50

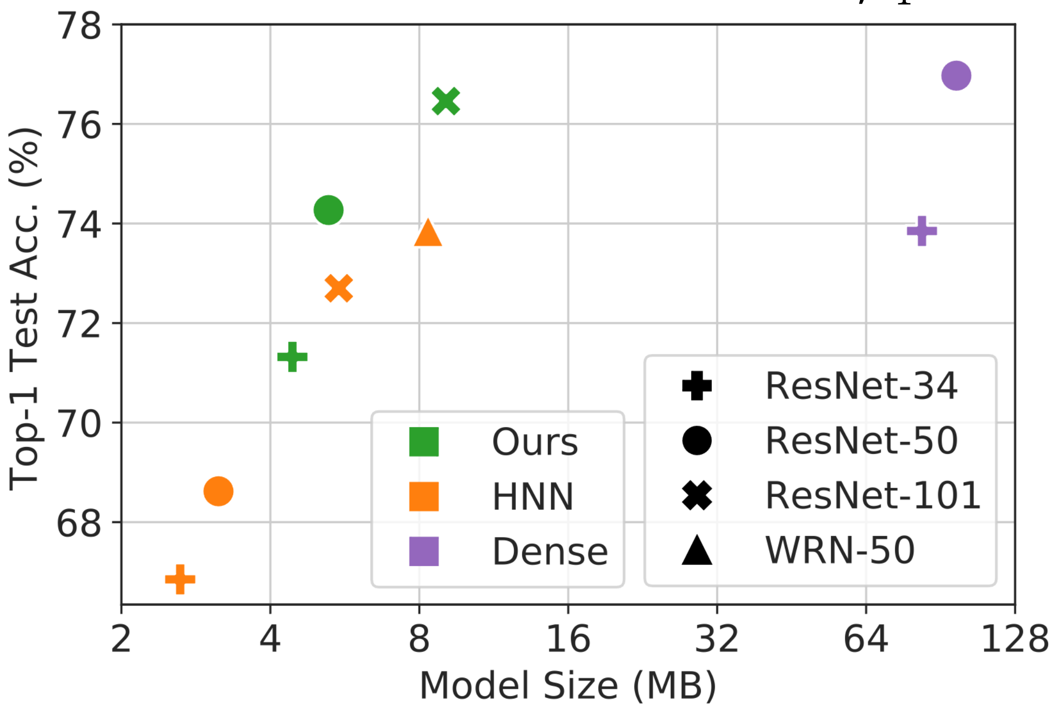
More supermask, higher accuracy



Edge-Popup

- Comparison of Model Size

N=7,k_1=30

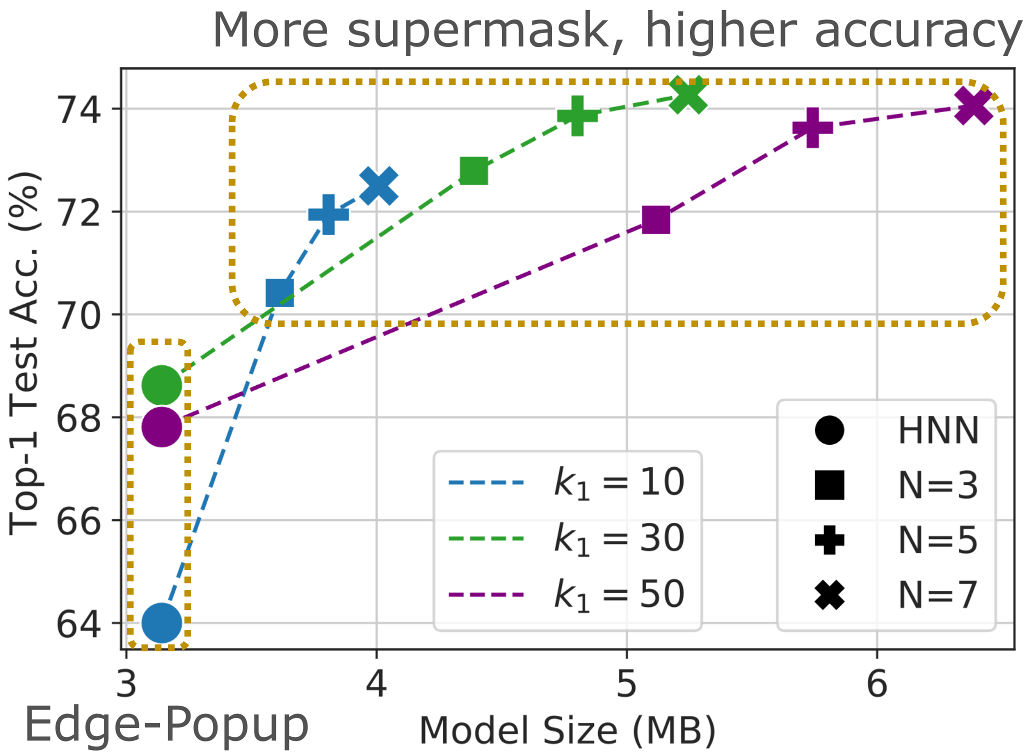


© 2022 by the authors.
International Conference on Machine Learning

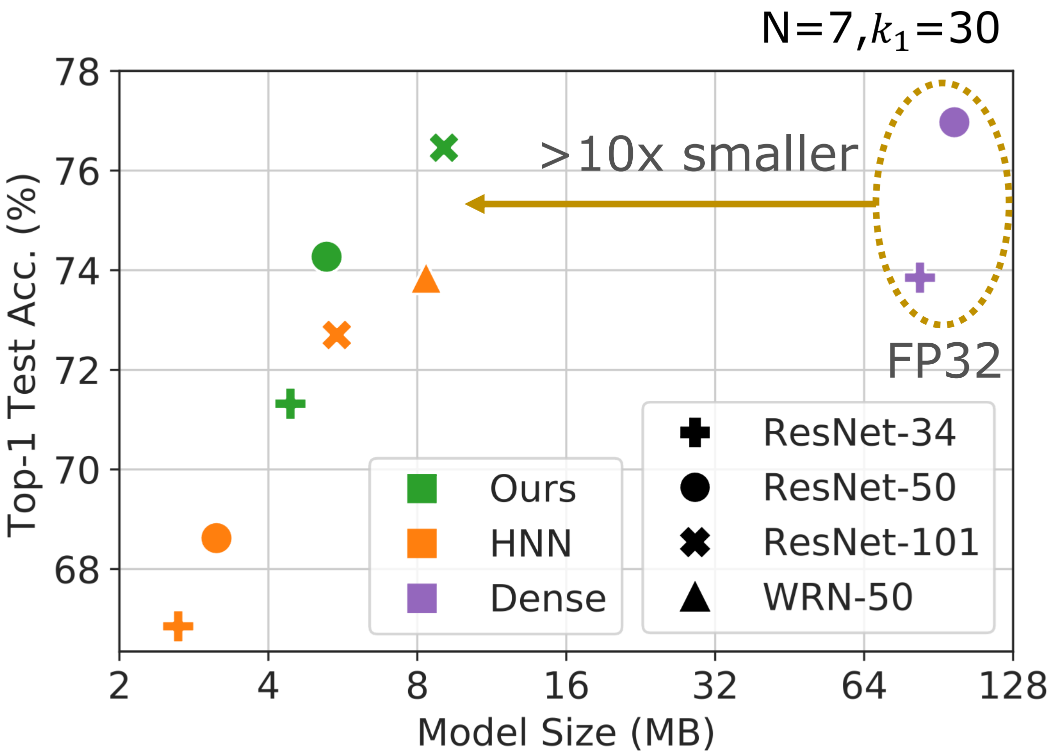
Comparison – Model Size VS. Accuracy (ImageNet)

M-Sup achieve competitive results on ImageNet

- Comparison of #Coats
  - ResNet-50



- Comparison of Model Size



Conclusion

- First work discussing multiple supermasks for trainable scaling
    - Expanded search space finds a network with high accuracy

Conclusion

- First work discussing multiple supermasks for trainable scaling
  - Expanded search space finds a network with high accuracy

- Multicoated supermasks achieve
  - **+5% Accuracy** on ImageNet w.r.t. Edge-Popup (ResNet-50)
  - **10x Smaller** size than dense model

Conclusion

- First work discussing multiple supermasks for trainable scaling
  - Expanded search space finds a network with high accuracy

- Multicoated supermasks achieve
  - **+5% Accuracy** on ImageNet w.r.t. Edge-Popup (ResNet-50)
  - **10x Smaller** size than dense model

- The Combination of pruning, quantization, and random weights achieves accurate, highly compressed models