CLOSE: Curriculum Learning On the Sharing Extent Towards Better One-shot NAS

Zixuan Zhou*, Xuefei Ning*, Yi Cai, Jiashu Han,
Huazhong Yang, Yu Wang*

NICS-EFC lab, Tsinghua University
*zhouzx17@gmail.com *foxdoraame@gmail.com
*yu-wang@tsinghua.edu.cn

2022.08.07

Menu

Neural Architecture Search (NAS)

Traditional NAS

[Zoph et al., ICLR 2017] explore 13k architectures, each trained from scratch for 50 epochs.

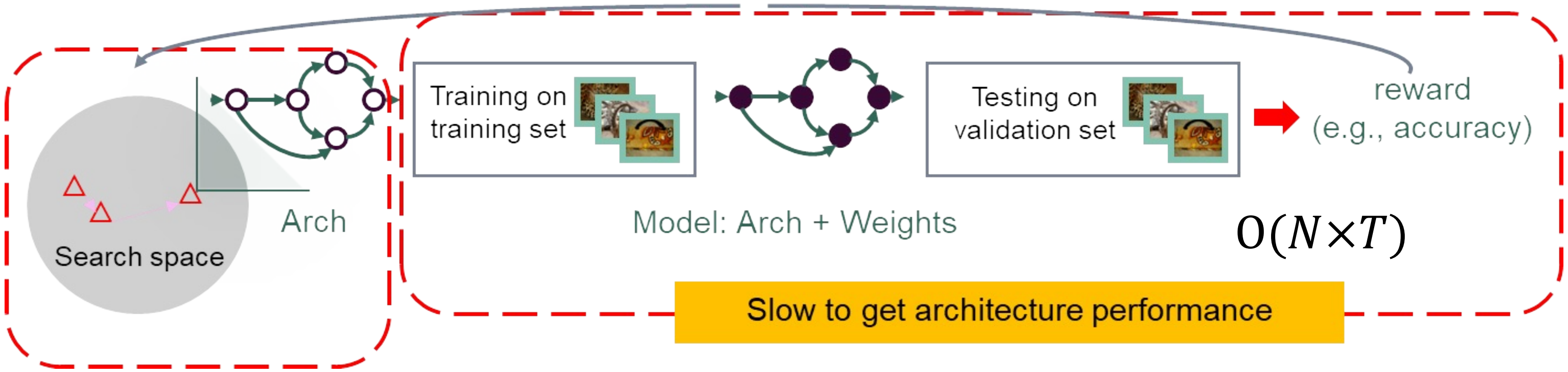
~48k GPU hours! Very expensive!

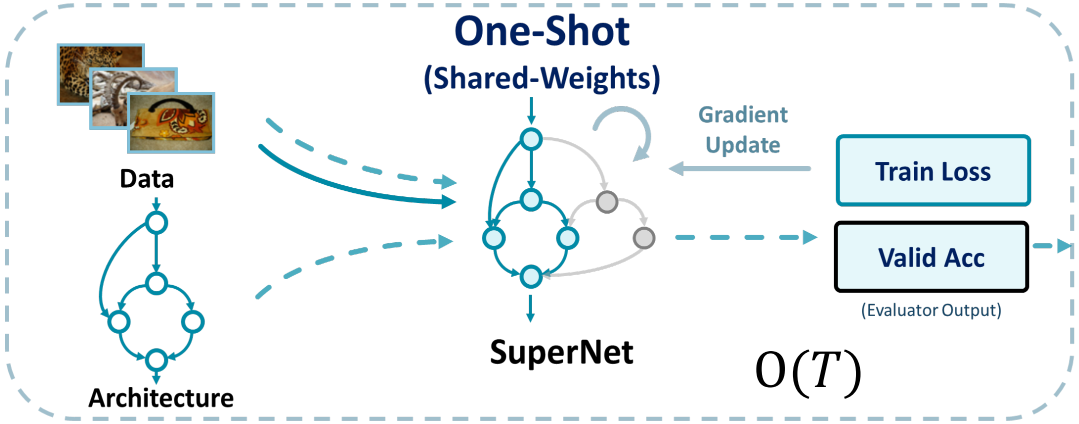
More Efficient

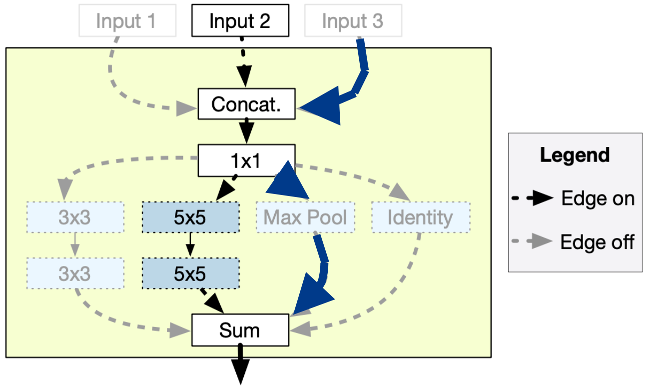
One-shot NAS

[Pham et al., ICML 2018] adopt parameter sharing technique to search the optimal architectures.

~10.8 GPU hours! 5000x faster!







Conv 1x1 shared between different architectures

- **Input 2 -> Concat -> 1x1 -> 5x5 -> 5x5**
- **Input 3 -> Concat -> 1x1 -> Max Pool**

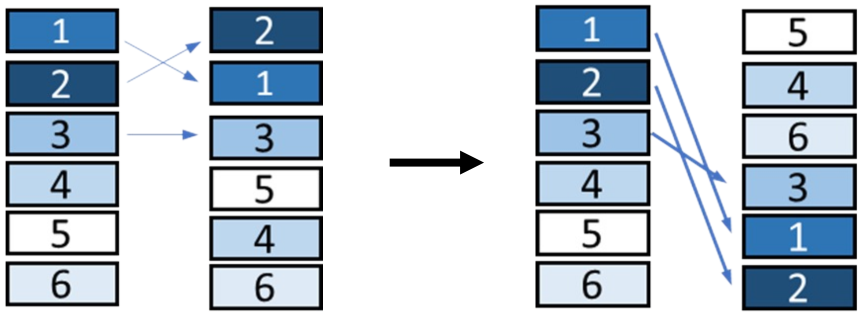
Thomas Elsken, et al., Neural Architecture Search: A Survey, JMLR 2019.

Weakness & Improvement of One-shot NAS

- One-shot NAS suffers from the poor ranking correlation between the one-shot performances and stand-alone performances of architectures.

The performance ranking is more important than the performance itself.

Kendall's Taus of the SuperNet trained for 1000 epochs are poor.
NB101: 0.369 NB201: 0.766 NB301: 0.515 [1]

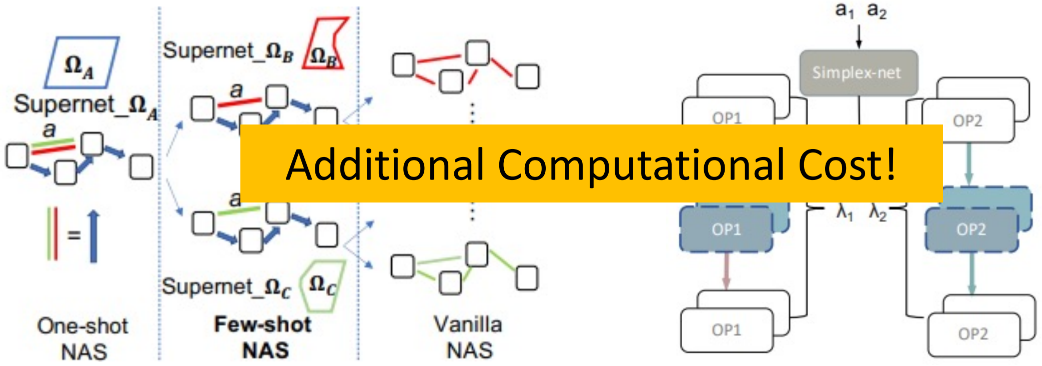


Traditional NAS
high ranking correlation

One-shot NAS
poor ranking correlation

A main factor causes to the poor correlation is the **unsuitable sharing extent** [1].

Few-shot NAS [2] K-shot NAS [3]



[1]. Ning et al., Evaluating Efficient Performance Estimators of Neural Architectures, NeurIPS 2021.
[2] Zhao et al. Few-shot Neural Architecture Search, ICML 2021.
[3] Su et al. K-shot NAS: Learnable Weight-Sharing for NAS with K-shot Supernets, ICML 2021.

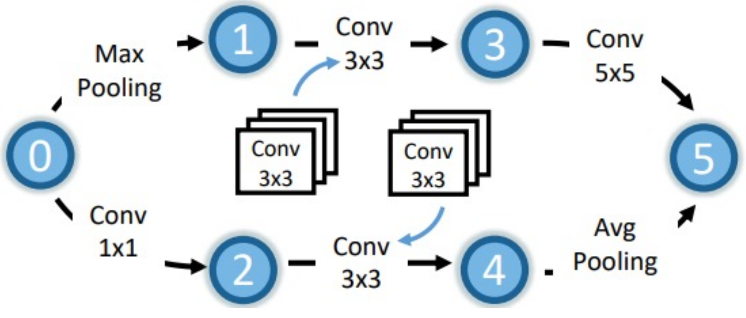
Menu

1. Background
2. **Methodology**
3. Experiment
4. Conclusion

Motivation

- Observation 1: Using a larger sharing extent can accelerate the training speed, but cannot achieve a high saturating performance.

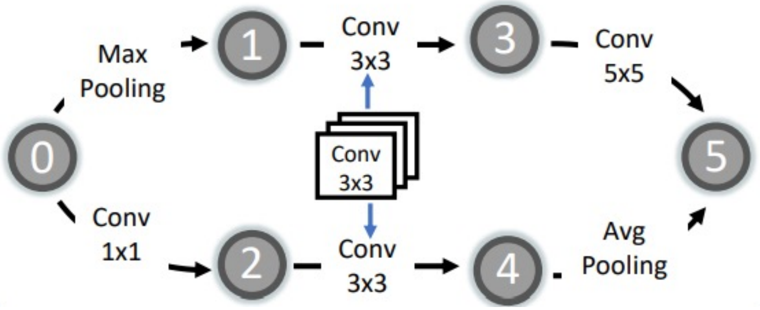
Supernet-1: Supernet with vanilla sharing extent



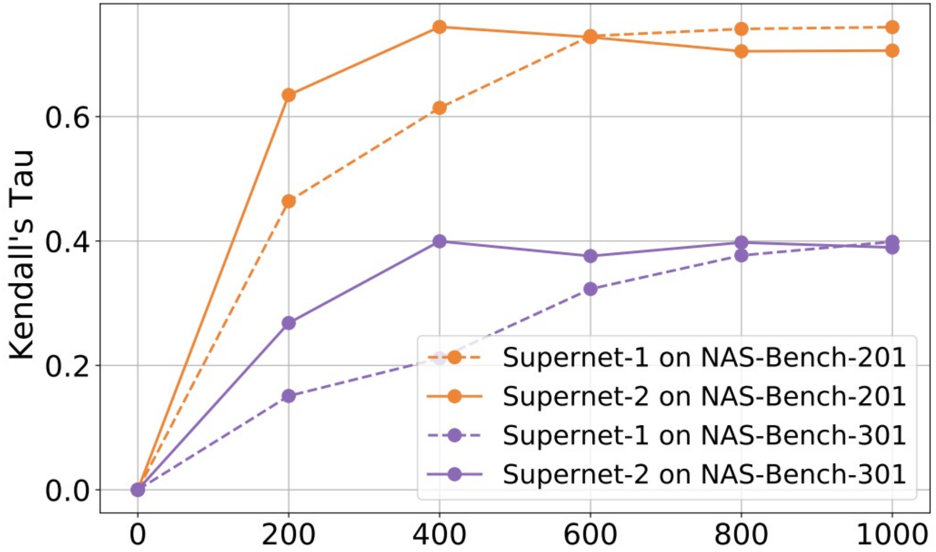
Ops with same type but in different **position** use **different** parameters.

Sharing Extent Increasing

Supernet-2: Supernet with larger sharing extent



Ops with same type but in different **position** share the **same** parameters.

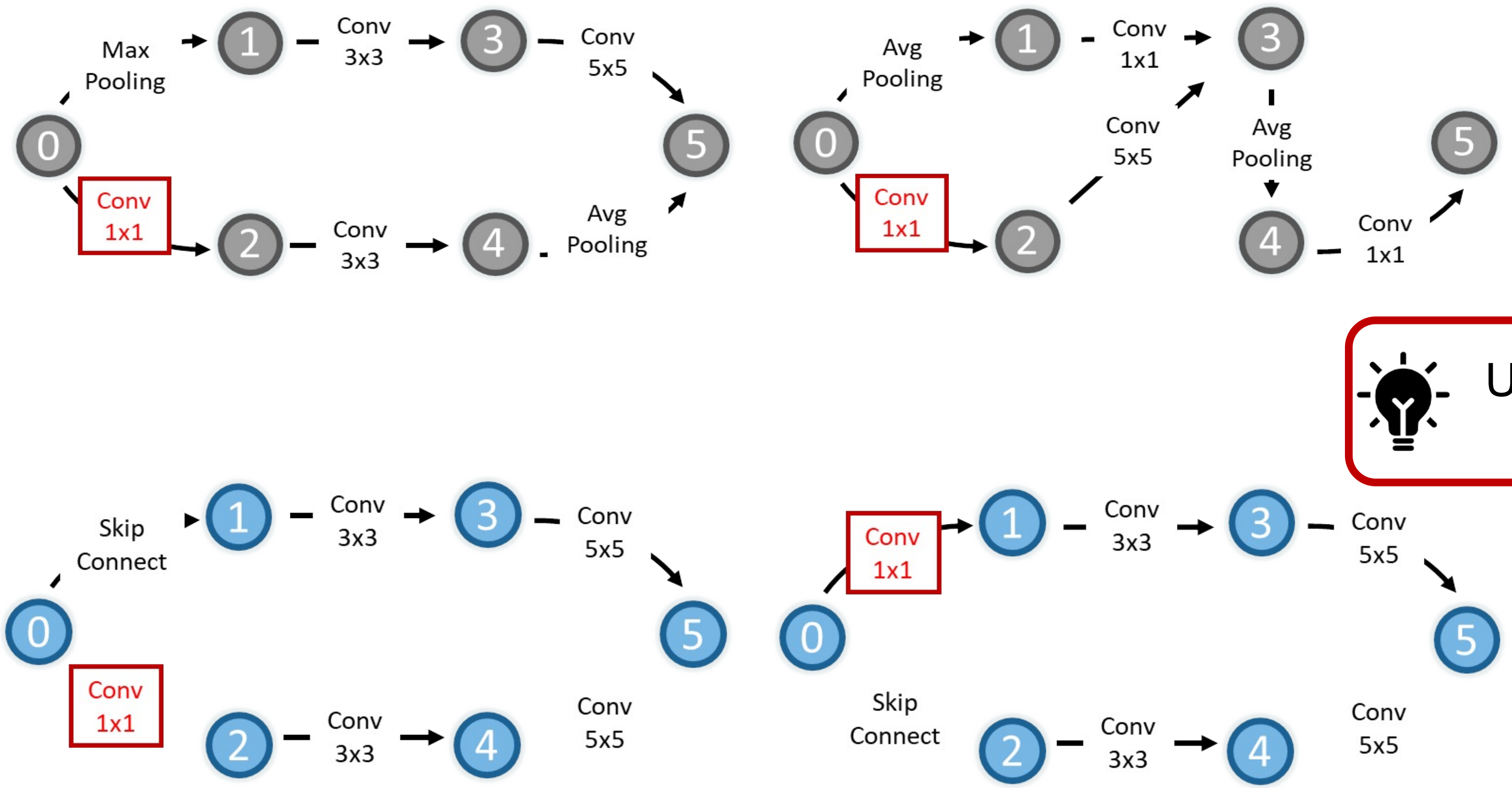


1. Higher KD in the early training stage. (0~600)
2. Lower KD when trained convergence. (800~1000)

Use large sharing extent in the early stage, and then gradually reduce it.

Motivation

- Observation 2: Using operations' positions to decide the sharing scheme is inappropriate.



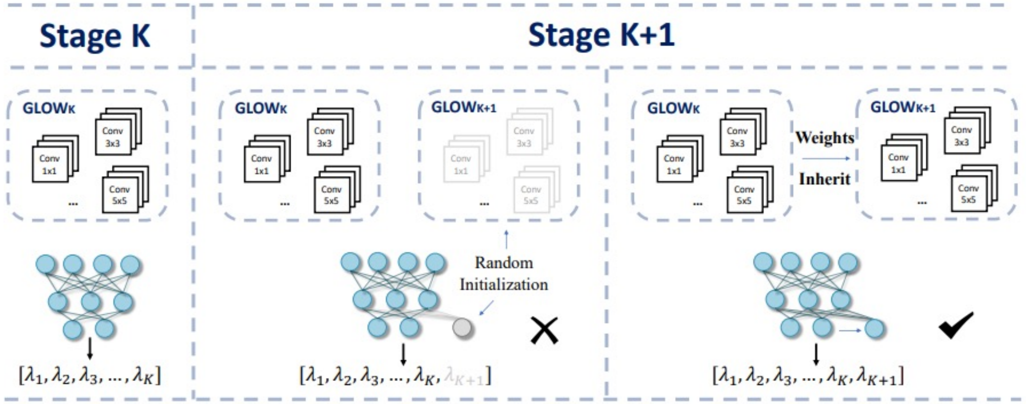
Should have the flexibility to **assign different Params** to Ops with vastly **different** functionalities (thus different optimal Params)

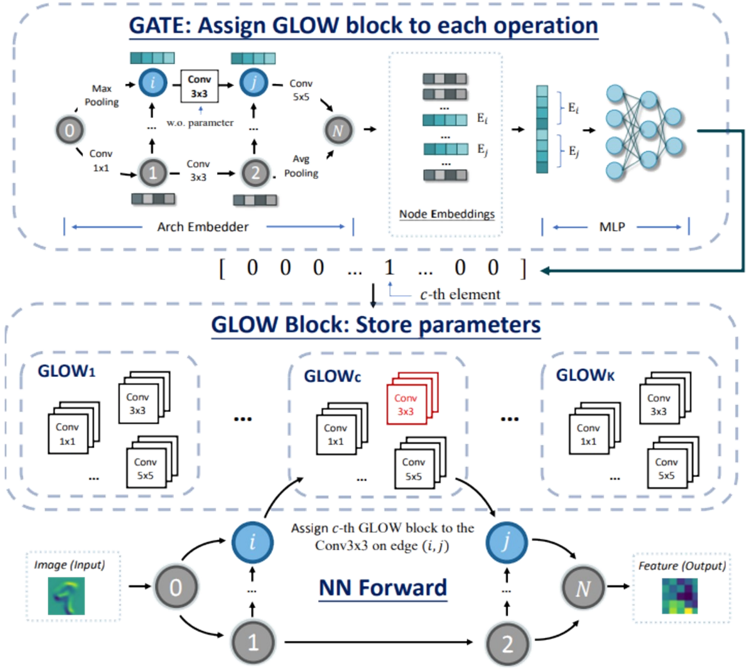
Use operations' functionality to decide the sharing scheme.

Should **share Params** for Ops with **similar or equivalent** functionalities across architectures

Curriculum Learning On Sharing Extent (CLOSE)

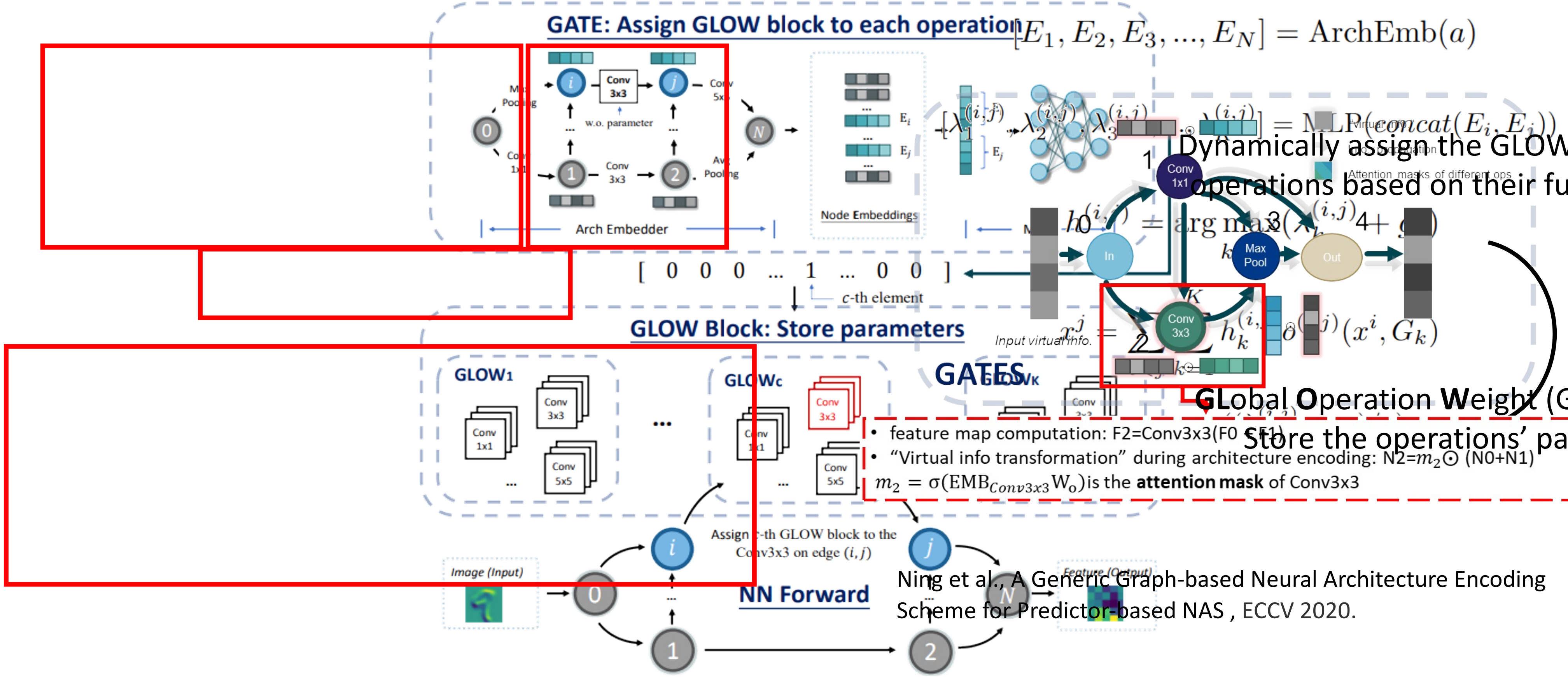
- **Dynamically** Adjust **Sharing Extent and Scheme** in the supernet
  - A curriculum learning-like supernet training strategy
    - Use larger sharing extent in the early training stage to accelerate the training process.
    - Gradually reduce the sharing extent in the later stage to boost the saturating performance.
  - A novel supernet with adjustable sharing extent and scheme
    - Decouple the operations and parameters to simply support the sharing extent adjustment.
    - Adopt a control module to flexibly and more properly decide the sharing scheme.





Curriculum Learning On Sharing Extent (CLOSE)

- CLOSENet: A novel and flexible supernet
  - Enable flexible sharing scheme (pick the shared parameters based on the functionality by the GATE module)
  - Enable adjustable sharing extent (change the extent by simply adding the GLOW block)



Dynamically assign the GLOW blocks to the operations based on their functionalities.

GLobal **O**peration **W**eight (GLOW) block: Store the operations' parameters

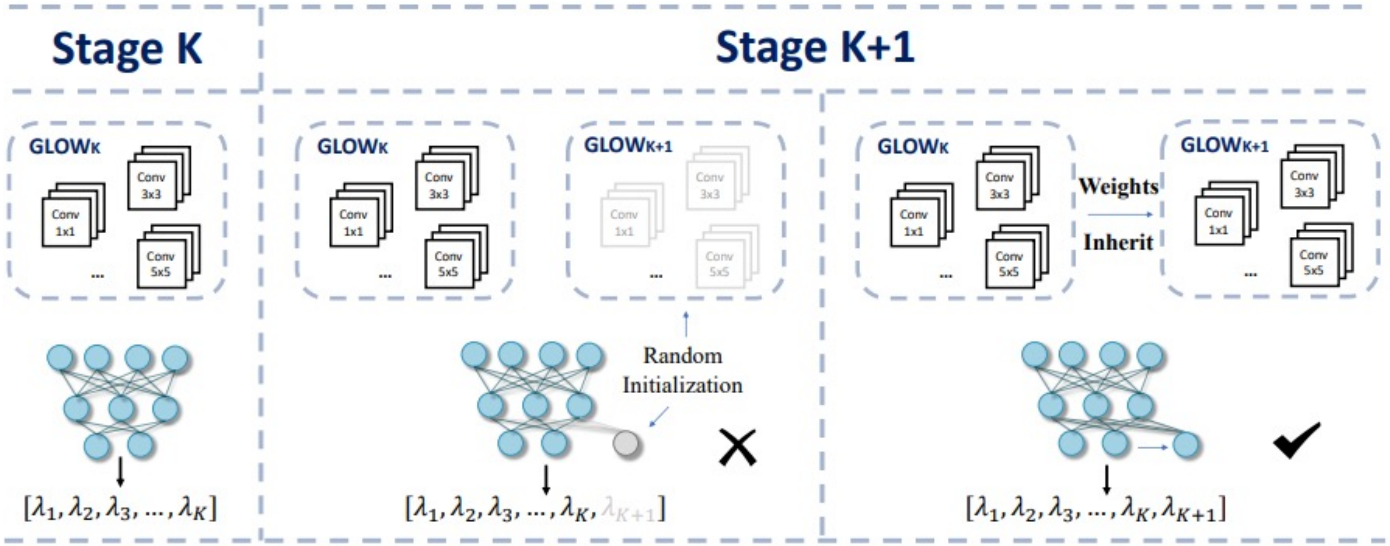
- feature map computation: F2=Conv3x3(F0 + F1)
- "Virtual info transformation" during architecture encoding: N2=m_2⊙ (N0+N1)

$m_2 = \sigma(\mathrm{EMB}_{Conv3x3}\mathrm{W_o})$is the **attention mask** of Conv3x3

Ning et al., A Generic Graph-based Neural Architecture Encoding Scheme for Predictor-based NAS , ECCV 2020.

Curriculum **L**earning **O**n **S**haring **E**xtent (CLOSE)

- CLOSE: A curriculum learning-like supernet training strategy
  - Using large sharing extent is “easy” for supernet to train.
  - Reducing the sharing extent increases the “difficulty”, but can push its limits.



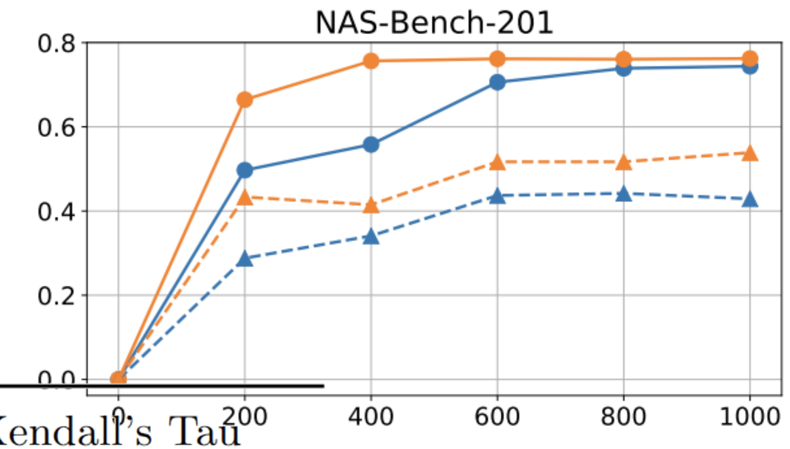
Challenge 1: Performance drop after adding a new randomly-initialized GLOW block.
Technique 1: **W**eight **I**nherit **T**echnique (WIT). Make the new GLOW block to inherit the parameters from the previous one.

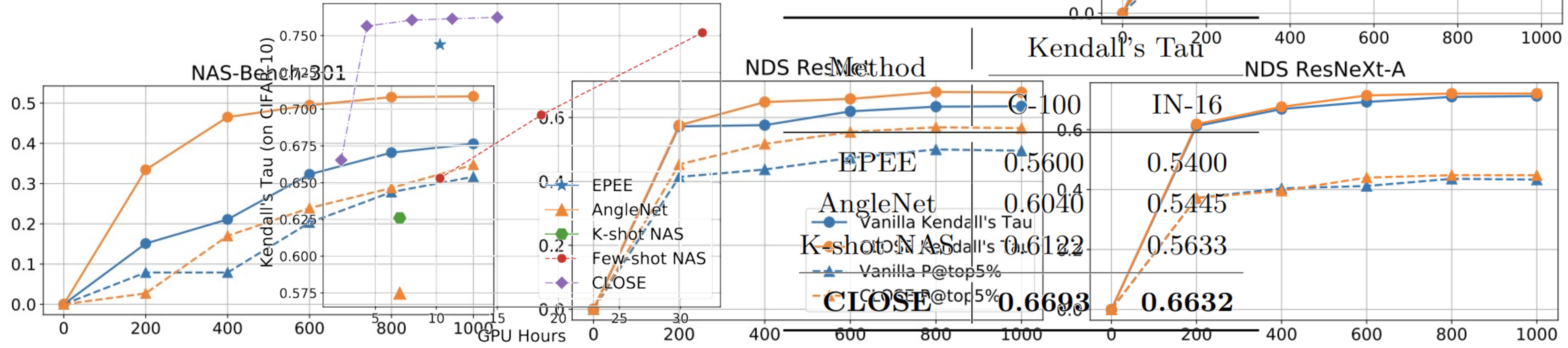
Challenge 2: Smaller learning rate makes the training hard to jump out of the local optimal solution.
Technique 2: **S**chedule **R**estart **T**echnique (SRT). Reset the learning rate and its schedule at some preset epochs.

Menu

Ranking Quality on Four NAS Benchmarks

- Evaluation Criteria
  - **Kendall`s Tau (KD)**: The relative difference of the number of concordant pairs and discordant pairs
  - **P@top5%**: The proportion of true top-5% architectures in the top-5% architectures according to the one-shot estimations
- NAS Benchmark
  - NAS-Bench-201 / NAS-Bench-301: Topological search space
  - NDS-ResNet / ResNeXt-A: Non-topological search space

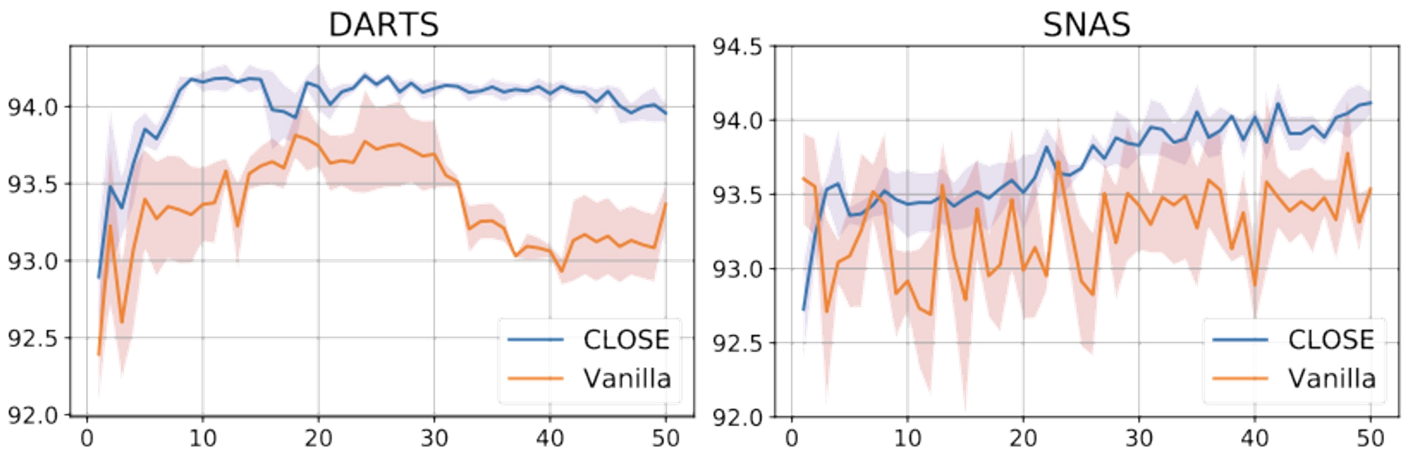




Method	Kendall's Tau C-100	IN-16
EPEE	0.5600	0.5400
AngleNet	0.6040	0.5445
K-shot NAS	0.6122	0.5633
CLOSE	**0.6693**	**0.6632**

Search Performance on DARTS Search Space

- DARTS Search Space
  - A generic topological search space that contains 10^{18} architectures
  - The architectures' performances are provided by NAS-Bench-301
- Search Strategy
  - DARTS , SNAS , CARS



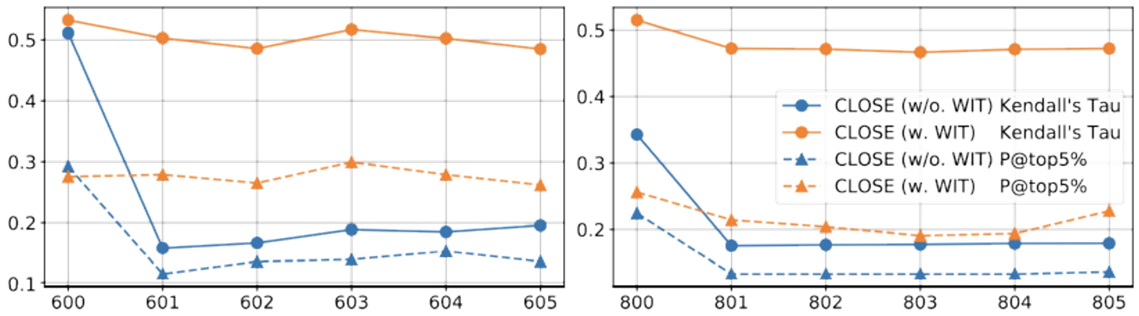
Method	CIFAR-10			ImageNet	
	Top-1 Error (%)	Param (M)	Search Cost (GPU days)	Top-1 Error (%)	Param (M)
NASNet-A [41]	2.65	3.3	2000	26.0	5.3
AmoebaNet-B [26]	2.55	2.8	3150	26.0	5.3
PNAS [17]	3.41	5.1	225	25.8	5.1
ENAS [23]	2.89	4.6	0.5	-	-
DARTS [18]	2.76	3.3	1.5	26.9	4.9
SNAS [33]	2.85	2.8	1.5	27.3	4.3
BayesNAS [39]	2.81	3.4	0.2	26.5	3.9
GDAS [5]	2.82	2.5	0.17	27.5	4.4
CLOSE (Ours)	2.72 ± 0.04	4.1	0.6	24.7	4.8

Liu et al., DARTS: Differentiable Architecture Search, ICLR 2019.
Xie et al., SNAS: Stochastic Neural Architecture Search, ICLR 2019.
Yang et al., CARS: Continuous Evolution for Efficient Neural Architecture Search, CVPR 2020.

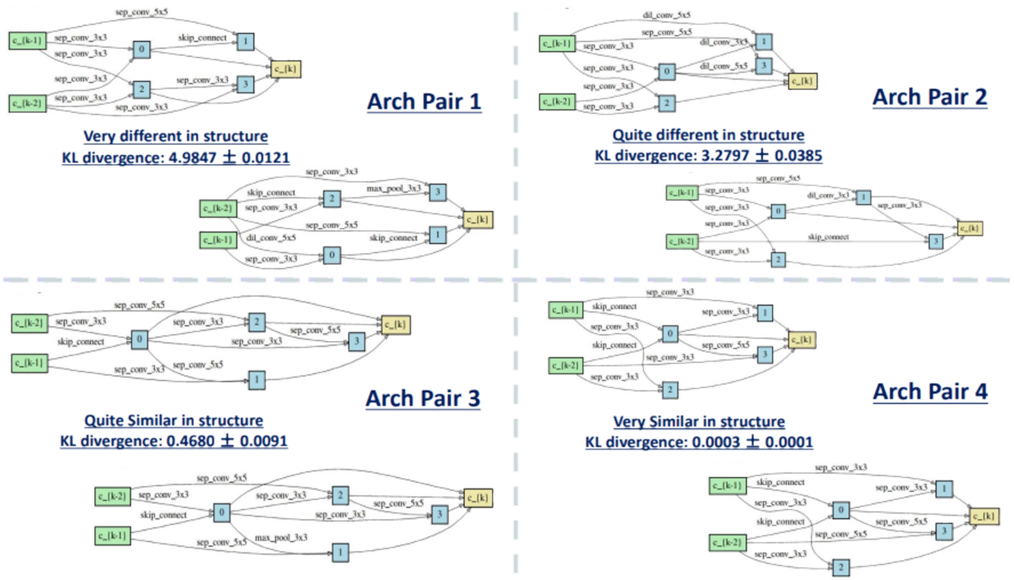
Ablation Studies

Effect of the proposed techniques WIT and SRT

WIT	SRT	NAS-Bench-301		NDS ResNet	
		KD	P@top5%	KD	P@top5%
		0.1104	0.1145	0.6339	0.5387
✓		0.1047	0.1122	0.6550	0.5520
	✓	0.2004	0.1610	0.6448	0.5280
✓	✓	**0.5168**	**0.3470**	**0.6786**	**0.5667**



Effect of the GATE module



GATE	NAS-Bench-201		NAS-Bench-301	
	KD	P@top5%	KD	P@top5%
w/o.	0.3627	0.2014	0.2236	0.1924
w.	**0.7622**	**0.5387**	**0.5168**	**0.3470**

Effect of gradually adding the GLOW blocks

Benchmark	Fixed number of blocks				CLOSE
	2	3	4	5	
NB201	0.7320	0.7247	0.7073	-	**0.7622**
NB301	0.4533	0.3427	0.3301	0.3106	**0.5168**

Menu

Conclusion

- Knowledge
  - Large sharing extent also has some positive effects on one-shot supernet training, which means that improving both the efficiency and efficacy is a promising direction.
- CLOSE: A curriculum learning-like supernet training strategy
  - An intuitive training approach based on the observations that different sharing extents have different effects on different training stage.
  - Design effective techniques to help switch the curriculum appropriately.
- CLOSENet: A novel and flexible supernet
  - Decouple the operations and parameters to simply support the sharing extent adjustment.
  - Adopt a control module to flexibly and more properly decide the sharing scheme.

Thanks for listening!

Contact us at:

Zixuan Zhou zhouzx17@gmail.com, Xuefei Ning foxdoraame@gmail.com, Prof. Yu Wang yu-wang@tsinghua.edu.cn

Paper

https://arxiv.org/abs/2207.07868

Code

https://github.com/walkerning/aw_nas

Contributions, suggestions and discussions are all welcome!