SGAS: Sequential Greedy Architecture Search – Supplementary Material –

https://www.deepgcns.org/auto/sgas

Guohao Li*[1], Guocheng Qian*[1], Itzel C. Delgadillo*[1], Matthias Müller[2], Ali Thabet[1], Bernard Ghanem[1]
[1]King Abdullah University of Science and Technology (KAUST), Saudi Arabia
[2]Intelligent Systems Lab, Intel Labs, Germany

1. Discussion

1.1. Greedy Methods in NAS

The idea of incorporating greedy algorithms into NAS has been explored in several works. PNAS [7] proposes a sequential model-based optimization (SMBO) approach to accelerate the search for CNN architectures. They start from a simple search space and a learn a predictor function. Then they greedily grow the search space by predicting scores of candidates cells using the learned predictor function. GNAS [4] learns a global tree-like architecture for multi-attribute learning by iteratively updating layer-wise local architectures in a greedy manner. P-DARTS [1] can also be regarded as a greedy approach, in which they bridge the depth gap between search and evaluation by gradually increasing the depth of the search networks while shrinking the number of candidate operations.

1.2. Selection Criteria and Hyper-parameters

Edge Importance and Selection Certainty. Edge importance and selection certainty are combined into a single criterion, since the algorithm will be agnostic to the selection distribution of an edge, if we only consider edge importance. In this case, an edge may be selected with a sub-optimal operation at early epochs. On the other hand, we need to select 8 out of 14 edges in a DAG with 4 intermediate nodes for a fair comparison with DARTS. Only considering selection certainty may fail to select the optimal 8 edges, since an edge with a high selection certainty may have a high weight on *Zero* operation (low edge importance).

Choices of Hyper-parameters. Three extra parameters are introduced in *SGAS*: **(1)** *length of warm-up phase* **(2)** *interval of greedy decisions* **(3)** *history window size for Cri.2*. We provide a discussion on the default choices of them:
(1) Since the softmax weights of operations are initialized under a uniform distribution, choosing an operation for an edge after a short period of warming up leads to stable results. We simply set the *length of the warm-up phase* to 9 epochs so that the first greedy decision will be made at the 10th epoch. **(2)** For CNN experiments, the *interval between greedy decisions* is chosen to be 5. Since designing a normal cell with 4 intermediate nodes needs to select 8 out of 14 edges (8 decisions to be made). For a fair comparison to our baseline DARTS, we want the search phase to last up to 50 epochs, which is the length of search epochs in DARTS. For GCN architectures, in order to learn a compact network, we search a normal cell with 3 intermediate nodes. Thus, we have 6 decisions to make (6 out of 9 edges). Similarly, to keep the length of the search phase less than 50 epochs, we set the *interval between greedy decisions* to be 7. **(3)** The *history window size for Cri.2* is always set as 4, which is simply chosen to be slightly smaller than the *interval between greedy decisions*.

Ablation Study on Hyper-parameters. In order to better understand the effects of the choices of hyper-parameters, we conduct ablation studies on *interval of greedy decisions* T and *history window size* K *for Cri.2* on CIFAR-10 in Table 1. The default values of T and K are 5 and 4 respectively. We find that larger T and K stabilize the search and produce standard deviations in the test error. The test error only has a standard deviation as 0.08 when $T = 7$. When $T = 3$, the average test error increases significantly from 2.67% to 2.86%. We also find K is less sensitive than T.

2. Experimental Details

2.1. GCN Experiments

GCN operators. Similar as the search for CNN, SGAS selects one operation from a candidate operation search space for each edge in the DAG. We choose the following 10 operations as our candidate operations: *conv-1×1*, *MRConv* [6], *EdgeConv* [10], *GAT* [9], *SemiGCN* [5], *GIN* [11], *SAGE* [3], *RelSAGE*, *skip-connect*, and *zero* operation. *conv-1×1* is a basic convolution operation without aggregating information from neighbors, which is similar to PointNet [8].

*equal contribution

T	K	Avg. Params (M)	Avg. Test Err.(%)	Best Params (M)	Best Test Err.(%)
5	4	3.91±0.22	2.67±0.21	4.09	2.44
3	4	4.09±0.24	2.86±0.12	4.39	2.69
7	4	3.66±0.16	2.65±0.08	3.68	2.54
5	2	3.87±0.20	2.73±0.16	3.94	2.51
5	6	3.93±0.26	2.67±0.17	3.70	2.47

Table 1. **Ablation study on *interval of greedy decisions* T and *history window size* K *for Cri.2* on CIFAR-10.** We use SGAS (Cri.2) as our search method. We report the average and best performance of searched architectures.

MRConv [6], *EdgeConv* [10], *GAT* [9], *SemiGCN* [5], *GIN* [11] and *SAGE* [3] are widely used GCN operators in the graph learning domain and the 3D computer vision domain. RelSAGE is a modified GraphSAGE (SAGE) [3] which combines the ideas from *MRConv* [6] and GraphSAGE [3]. Instead of aggregating the node features with its neighbor features directly, we aggregate the node features with the difference between the node features and its neighbor features:

$$\mathbf{h}_v^{(k)} = \sigma\left(\mathbf{W}^{(k)} \cdot f_k\left(\mathbf{h}_v^{(k-1)}, \left\{\mathbf{h}_u^{(k-1)} - \mathbf{h}_v^{(k-1)}, \forall u \in \mathcal{N}(v)\right\}\right)\right)$$

where $\mathbf{h}_v^{(k)}$ is the feature of the center node v in k-th layer. $\mathcal{N}(v)$ denotes the neighbors of node v. f_k is a max aggregation function and σ is a ReLU activation function as GraphSAGE [3]. The GCNs operators are implemented using Pytorch Geometric [2]. We also add *skip-connect* (similar as residual graph connections [6]) and *zero* operation in our search space.

Ablation Study on GCN Cells. We conduct an ablation study on the parameter size of the best cell searched on PPI by SGAS (Cri.1 best). Table 2 shows the trade-off between the parameter size and the final performance. To derive a compact model, we can use a smaller number of cells or less channels in the architecture searched by SGAS.

Number of Cells	Channel Size	Params (M)	micro-F1(%)
5	64	0.40	98.894
5	128	1.52	99.369
5	256	5.89	99.429
5	512	23.18	99.462
1	256	1.22	99.157
3	256	3.52	99.418
5	256	5.89	99.429
7	256	8.25	99.433

Table 2. **Ablation study on channel size and number of cells on node classification on PPI.** We use SGAS (Cri.1 best), the best architecture we discovered by using *Criterion 1* to conduct experiments.

2.2. More Details

Cell Visualizations. We visualize the best cells discovered by SGAS with different criteria (*Criterion 1* and *Criterion 2*) mentioned in the experiment section. Figure 1 shows the best cells for CNNs on CIFAR-10 and ImageNet. Figure 2 shows the best cells for GCNs on ModelNet-40 and PPI.

Detailed results. Here we provide the detailed results mentioned in the experimental section of the paper. In the CNN experiments, we compare SGAS with DARTS on CIFAR-10 and ImageNet. We execute the search phase 10 times for both SGAS (Cri.1 and Cri.2) and DARTS (1st and 2nd order) to obtain 10 different architectures per method. For each resulting architecture, we run the evaluation phase and assign a ranking based on the evaluation accuracy. To measure the discrepancy between the search and evaluation, we calculate the Kendall τ correlation between the ranking of the search phase and the evaluation phase. We show these results in Table 3 and Table 4 for SGAS, and Table 5 and Table 6 for DARTS. For ImageNet, we evaluate the top three architectures found on CIFAR-10. We show the results in Table 7 and Table 8 for both *Criterion 1* and *Criterion 2*.

In the GCN experiments, we compare SGAS (Cri.1 and Cri.2) with a random search baseline on ModelNet and PPI. Similar as in experiments for CNNs, we conduct the search phase 10 times for each method. For experiments on ModelNet, we search cells on ModelNet10 and then evaluate the searched cells on ModelNet40. The results are shown for *Criterion 1*, *Criterion 2* and random search in Table 9, Table 10 and Table 11 respectively. The results on PPI are presented in Table 12 and Table 13 for each *Criterion* and in Table 14 for random search.

References

[1] Xin Chen, Lingxi Xie, Jun Wu, and Qi Tian. Progressive differentiable architecture search: Bridging the depth gap between search and evaluation. *arXiv preprint arXiv:1904.12760*, 2019. 1

[2] Matthias Fey and Jan Eric Lenssen. Fast graph representation learning with pytorch geometric. *arXiv preprint arXiv:1903.02428*, 2019. 2

[3] Will Hamilton, Zhitao Ying, and Jure Leskovec. Inductive representation learning on large graphs. In *Advances in Neural Information Processing Systems*, pages 1024–1034, 2017. 1, 2

[4] Siyu Huang, Xi Li, Zhi-Qi Cheng, Zhongfei Zhang, and Alexander Hauptmann. Gnas: A greedy neural architecture search method for multi-attribute learning. In *2018 ACM Multimedia Conference on Multimedia Conference*, pages 2049–2057. ACM, 2018. 1

[5] Thomas Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *ArXiv*, abs/1609.02907, 2016. 1, 2

[6] Guohao Li, Matthias Müller, Ali Thabet, and Bernard Ghanem. Deepgcns: Can gcns go as deep as cnns? In *The IEEE International Conference on Computer Vision (ICCV)*, 2019. 1, 2

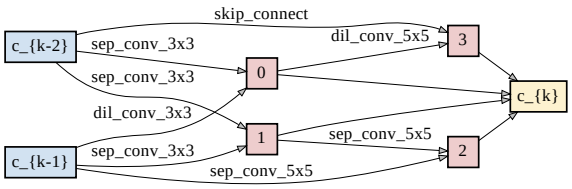
[7] Chenxi Liu, Barret Zoph, Maxim Neumann, Jonathon Shlens, Wei Hua, Li-Jia Li, Li Fei-Fei, Alan Yuille, Jonathan Huang, and Kevin Murphy. Progressive neural architecture search. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 19–34, 2018. 1

[8] Charles Ruizhongtai Qi, Hao Su, Kaichun Mo, and Leonidas J. Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 77–85, 2016. 1

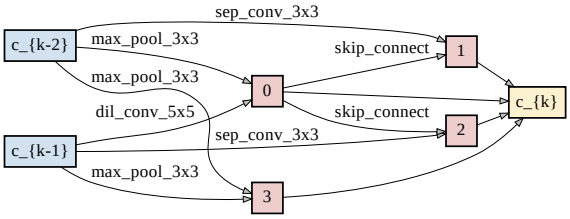
[9] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. Graph attention networks. *arXiv preprint arXiv:1710.10903*, 2017. 1, 2

[10] Yue Wang, Yongbin Sun, Ziwei Liu, Sanjay E. Sarma, Michael M. Bronstein, and Justin M. Solomon. Dynamic graph cnn for learning on point clouds. *ACM Transactions on Graphics (TOG)*, 2019. 1, 2

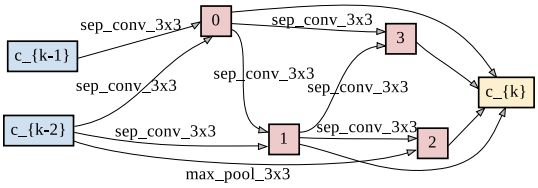
[11] Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. How powerful are graph neural networks? *ArXiv*, abs/1810.00826, 2018. 1, 2



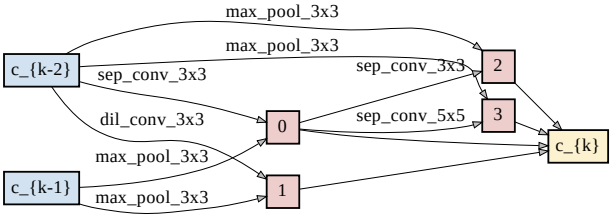
(a) Normal cell of the best model with SGAS *Cri. 1* on CIFAR-10 and ImageNet



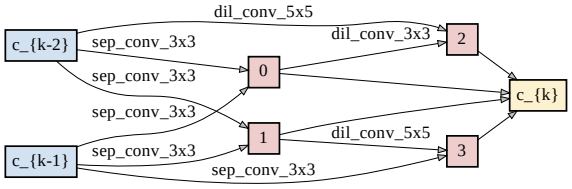
(b) Reduction cell of the best model with SGAS *Cri. 1* on CIFAR-10 and ImageNet



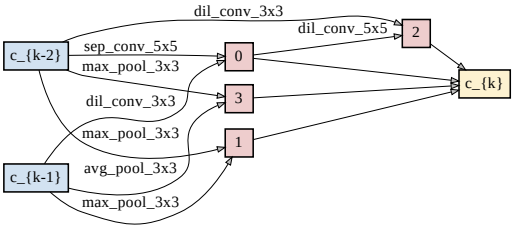
(c) Normal cell of the best model with SGAS *Cri. 2* on CIFAR-10



(d) Reduction cell of the best model with SGAS *Cri. 2* on CIFAR-10

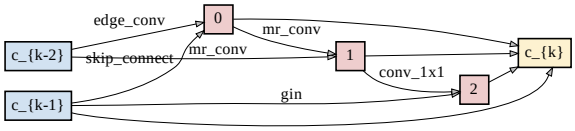


(e) Normal cell of the best model with SGAS *Cri. 2* on ImageNet

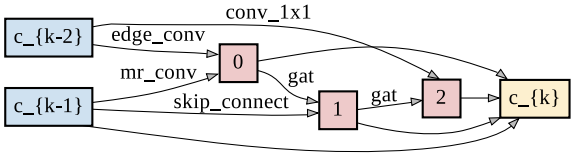


(f) Reduction cell of the best model with SGAS *Cri. 2* on ImageNet

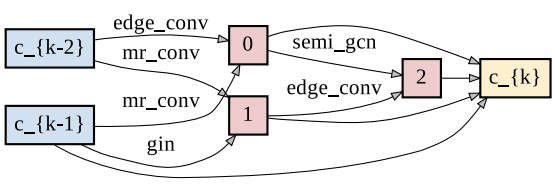
Figure 1. Best cell architecture for image classification tasks



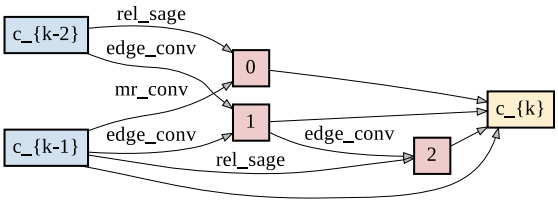
(a) Normal cell of the best model with SGAS *Cri. 1* on ModelNet40



(b) Normal cell of the best model with SGAS *Cri. 2* on ModelNet40



(c) Normal cell of the best model with SGAS *Cri. 1* on PPI



(d) Normal cell of the best model with SGAS *Cri. 2* on PPI

Figure 2. Best cell architectures on ModelNet40 and PPI with each *Criterion*

Experiment	Validation error (%)	Params (M)	Test error (%)	Evaluation ranking
Cri.1_CIFAR_1	16.94	3.75	2.44	2
Cri.1_CIFAR_2	17.33	3.73	2.50	3
Cri.1_CIFAR_3	17.90	3.80	2.39	1
Cri.1_CIFAR_4	17.90	3.32	2.63	6
Cri.1_CIFAR_5	17.99	3.45	2.78	8
Cri.1_CIFAR_6	18.43	3.47	2.68	7
Cri.1_CIFAR_7	18.72	3.83	2.51	4
Cri.1_CIFAR_8	19.82	3.66	2.61	5
Cri.1_CIFAR_9	19.93	3.98	3.18	10
Cri.1_CIFAR_10	21.53	3.61	2.87	9
Average	18.65±1.4	3.66±0.2	2.66±0.24	**Kendall** τ
Best Model	17.90	3.80	2.39	0.56

Table 3. Results of SGAS *Criterion 1* on CIFAR-10

Experiment	Validation error (%)	Params (M)	Test error (%)	Evaluation ranking
Cri.2_CIFAR_1	16.48	4.14	2.57	4
Cri.2_CIFAR_2	17.26	3.88	2.60	6
Cri.2_CIFAR_3	17.31	4.09	2.44	1
Cri.2_CIFAR_4	17.47	3.91	2.49	2
Cri.2_CIFAR_5	17.53	3.69	2.52	3
Cri.2_CIFAR_6	17.98	3.95	3.12	10
Cri.2_CIFAR_7	18.28	3.69	2.58	5
Cri.2_CIFAR_8	18.28	4.33	2.85	8
Cri.2_CIFAR_9	19.48	3.73	2.85	9
Cri.2_CIFAR_10	19.98	3.68	2.66	7
Average	18.00±1.06	3.91±0.22	2.67±0.21	**Kendall** τ
Best Model	17.31	4.09	2.44	0.42

Table 4. Results of SGAS *Criterion 2* on CIFAR-10

Experiment	Validation error (%)	Params (M)	Test error (%)	Evaluation ranking
DARTS_1st_CIFAR_1	11.37	3.27	2.83	4
DARTS_1st_CIFAR_2	11.45	3.65	2.57	2
DARTS_1st_CIFAR_3	11.47	2.29	2.94	7
DARTS_1st_CIFAR_4	11.48	2.65	2.96	8
DARTS_1st_CIFAR_5	11.65	3.09	2.50	1
DARTS_1st_CIFAR_6	11.75	2.86	2.84	5
DARTS_1st_CIFAR_7	11.77	2.09	3.06	10
DARTS_1st_CIFAR_8	11.81	2.52	3.01	9
DARTS_1st_CIFAR_9	11.82	2.65	2.94	6
DARTS_1st_CIFAR_10	11.94	3.27	2.82	3
Average	11.65±0.19	2.84±0.49	2.85±0.18	**Kendall** τ
Best Model	11.65	3.09	2.50	0.16

Table 5. Results of DARTS 1st order on CIFAR-10

Experiment	Validation error (%)	Params (M)	Test error (%)	Evaluation ranking
DARTS_2nd_CIFAR_1	11.35	2.91	2.96	8
DARTS_2nd_CIFAR_2	11.51	2.93	2.73	5
DARTS_2nd_CIFAR_3	11.68	2.20	3.01	9
DARTS_2nd_CIFAR_4	11.76	2.66	2.75	6
DARTS_2nd_CIFAR_5	11.80	3.09	2.72	4
DARTS_2nd_CIFAR_6	11.82	3.40	2.62	3
DARTS_2nd_CIFAR_7	11.83	2.91	2.82	7
DARTS_2nd_CIFAR_8	11.93	3.20	2.51	1
DARTS_2nd_CIFAR_9	11.95	2.14	3.48	10
DARTS_2nd_CIFAR_10	12.03	2.55	2.62	2
Average	11.77±0.21	2.8±0.41	2.82±0.28	**Kendall τ**
Best Model	11.93	3.20	2.51	-0.29

Table 6. Results of DARTS 2nd order on CIFAR-10

Experiment	Test error top-1 (%)	Test error top-5 (%)	Params (M)	$\times+$
Cri.1_ImageNet_1	24.47	7.23	5.25	578
Cri.1_ImageNet_2	24.53	7.40	5.23	574
Cri.1_ImageNet_3	24.22	7.25	5.29	585
Average	24.41±0.16	7.29±0.09	5.25±0.03	579
Best Model	24.22	7.25	5.29	585

Table 7. Results of SGAS *Criterion 1* on ImageNet. Note that the chosen architectures are the three best ones from the results obtained on CIFAR-10.

Experiment	Test error top-1 (%)	Test error top-5 (%)	Params (M)	$\times+$
Cri.2_ImageNet_1	24.44	7.41	5.70	621
Cri.2_ImageNet_2	24.13	7.31	5.44	598
Cri.2_ImageNet_3	24.55	7.44	5.20	571
Average	24.38±0.22	7.39±0.07	5.44±0.25	597
Best Model	24.13	7.31	5.44	598

Table 8. Results of SGAS with *Criterion 2* on ImageNet. Note that chosen the architectures are the three best ones from the results obtained on CIFAR-10.

Experiment	Params (M)	Test OA (%)
Cri.1_ModelNet_1	8.79	92.71
Cri.1_ModelNet_2	9.23	92.83
Cri.1_ModelNet_3	8.79	92.79
Cri.1_ModelNet_4	8.78	92.34
Cri.1_ModelNet_5	8.93	92.79
Cri.1_ModelNet_6	8.19	92.30
Cri.1_ModelNet_7	8.63	92.83
Cri.1_ModelNet_8	8.63	92.71
Cri.1_ModelNet_9	8.63	92.87
Cri.1_ModelNet_10	9.23	92.79
Average	8.78±0.30	92.69±0.20
Best Model	8.63	92.87

Table 9. Results of SGAS with *Criterion 1* on ModelNet40. Architectures are formed by stacking 9 cells with 128 channel size.

Experiment	Params (M)	Test OA (%)
Cri.2_ModelNet_1	8.78	92.91
Cri.2_ModelNet_2	8.78	92.67
Cri.2_ModelNet_3	9.08	92.79
Cri.2_ModelNet_4	8.49	93.23
Cri.2_ModelNet_5	9.08	93.03
Cri.2_ModelNet_6	9.08	93.07
Cri.2_ModelNet_7	8.78	93.11
Cri.2_ModelNet_8	8.63	92.67
Cri.2_ModelNet_9	8.78	92.83
Cri.2_ModelNet_10	9.23	92.95
Average	8.87±0.23	92.93±0.19
Best Model	8.49	93.23

Table 10. Results of SGAS with *Criterion 2* on ModelNet40. Architectures are formed by stacking 9 cells with 128 channel size.

Experiment	Params (M)	Test OA (%)
Random_ModelNet_1	9.22	92.79
Random_ModelNet_2	8.93	92.67
Random_ModelNet_3	9.08	92.71
Random_ModelNet_4	8.78	92.46
Random_ModelNet_5	8.19	92.79
Random_ModelNet_6	8.63	92.54
Random_ModelNet_7	8.93	91.94
Random_ModelNet_8	8.63	92.99
Random_ModelNet_9	8.79	93.15
Random_ModelNet_10	8.49	92.46
Average	8.77±0.30	92.65±0.33
Best Model	8.79	93.15

Table 11. Results of random search on ModelNet40. Architectures are formed by stacking 9 cells with 128 channel size.

Experiment	Params (M)	Test micro-F1 (%)
Cri.1_PPI_1	27.11	99.45
Cri.1_PPI_2	23.18	99.42
Cri.1_PPI_3	25.80	98.91
Cri.1_PPI_4	25.80	99.38
Cri.1_PPI_5	24.49	99.44
Cri.1_PPI_6	29.73	99.44
Cri.1_PPI_7	24.50	99.44
Cri.1_PPI_8	21.87	99.43
Cri.1_PPI_9	24.49	99.44
Cri.1_PPI_10	23.18	99.46
Average	25.01±2.24	99.38±0.17
Best Model	23.18	99.46

Table 12. Results of SGAS with *Criterion 1* on PPI.

Experiment	Params (M)	Test micro-F1 (%)
Cri.2_PPI_1	25.80	99.17
Cri.2_PPI_2	28.42	99.46
Cri.2_PPI_3	20.55	99.40
Cri.2_PPI_4	21.87	99.43
Cri.2_PPI_5	24.49	99.44
Cri.2_PPI_6	28.42	99.43
Cri.2_PPI_7	29.73	99.42
Cri.2_PPI_8	25.80	99.45
Cri.2_PPI_9	28.42	99.44
Cri.2_PPI_10	25.79	99.41
Average	25.93±2.99	99.40±0.09
Best Model	28.42	99.46

Table 13. Results of SGAS with *Criterion 2* on PPI.

Experiment	Params (M)	Test micro-F1 (%)
Random_PPI_1	20.57	99.27
Random_PPI_2	24.48	99.37
Random_PPI_3	24.49	99.40
Random_PPI_4	19.24	99.40
Random_PPI_5	24.48	99.37
Random_PPI_6	21.85	99.36
Random_PPI_7	27.11	99.32
Random_PPI_8	23.17	99.40
Random_PPI_9	24.48	99.39
Random_PPI_10	27.11	99.38
Average	23.7±2.56	99.36±0.04
Best Model	23.17	99.40

Table 14. Results of random search on PPI.