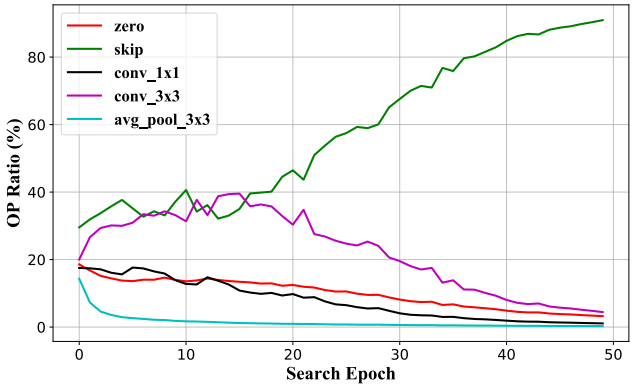
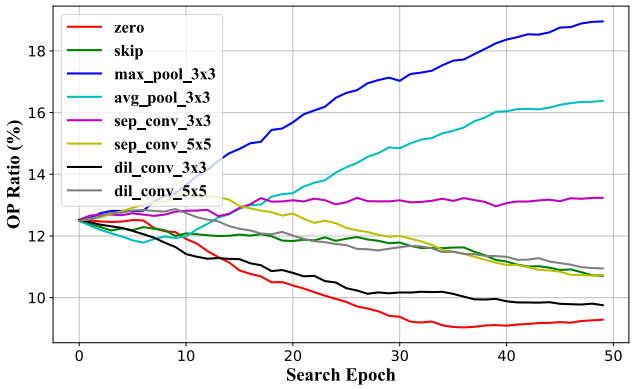
A. Irreversible Trend of Non-parametric Operations

The probability curves of architecture parameters ($\mathrm{softmax}(\alpha)$) during the whole search of DARTS on CIFAR-10 in (a) NAS-201 and (b) DARTS Search space. There is an irreversible trend that non-learnable operations ($e.g.$, skip, pool) surpass learnable operations.



(a) NAS-201 Search Space



(b) DARTS Search Space

Figure 4: Irreversible trend of non-parametric operations in different space

B. Search space details

NAS-Bench-201 [15] builds a cell-based search space, where one cell could be seen as a directed acyclic graph consisting of 4 nodes and 6 edges. Each network is stacked by 15 cells. Each edge represents an operation selected from (1) zero, (2) skip connection, (3) 1×1 convolution, (4) 3×3 convolution, and (5) 3×3 average pooling. The search space has 15,625 neural cell candidates in total. And all the candidates are given training accuracy/valid accuracy/test accuracy on three datasets: (1) CIFAR-10: In NAS-Bench-201, it separates the train set in original CIFAR-10 into two parts, one as the train set and the other as the validation set, each set contains 25K images with 10 classes, (2) CIFAR-100: For CIFAR-100, it separates the original test set to get the validation and test set, each set has 5K images, and (3) ImageNet-16-120. It downsamples ImageNet to 16×16 pixels and selects 120 classes. Totally, it involves 151.7K training images, 3K validation images, and 3K test images. We set Adam as architecture parameters' optimizer and SGD as network weights'.

DARTS [15] is also a cell-based search space. Each cell contains 6 nodes and each node has to select 2 edges to connect with the previous 2 nodes. Each edge has 8 operations: 3×3 and 5×5 separable convolution, 3×3 and 5×5 dilated separable convolution, 3×3 max-pooling, 3×3 average-pooling, skip-connect (identity), and zero (none). The stacked networks have normal cells and reduction cells. It contains 10^{18} candidates, which is quite large.

C. Proof of theorem 4.2

Theorem C.1 *For function* $\mathcal{L} = g(\sum_i \frac{exp(\alpha_i)}{\sum_j exp(\alpha_j)} x_i)$, *let* $p_i = \frac{exp(\alpha_i)}{\sum_j exp(\alpha_j)}$, $\bar{x} = \sum_i p_i x_i$, *if* $\frac{\partial \mathcal{L}}{\partial \bar{x}}^T x_j < \frac{\partial \mathcal{L}}{\partial \bar{x}}^T x_i$ *and* $\alpha_j \geq \alpha_i$, *then* $\frac{\partial \mathcal{L}}{\partial \alpha^j} \leq \frac{\partial \mathcal{L}}{\partial \alpha_i}$

Proof C.1 *It's easily to see that* $p_i \geq p_j$ *if* $\alpha_i \geq \alpha_j$ *and* $p_{i^*} \geq \frac{1}{n}$ *for* $\sum_i p_i = 1$. *Consider*

$$\begin{aligned}
\frac{\partial \mathcal{L}}{\partial \alpha_i} &= \frac{\partial \mathcal{L}}{\partial \bar{x}}^T (p_i(1-p_i)x_i - p_i \sum_{k \neq i} p_k x_k) \\
&= \frac{\partial \mathcal{L}}{\partial \bar{x}}^T \sum_{k \neq i} p_i p_k (x_i - x_k) \\
&= \frac{\partial \mathcal{L}}{\partial \bar{x}}^T \sum_{k} p_i p_k (x_i - x_k) \\
&= p_i \sum_k p_k \frac{\partial \mathcal{L}}{\partial \bar{x}}^T (x_i - x_k)
\end{aligned}$$

For $\frac{\partial \mathcal{L}}{\partial \bar{x}}^T x_j \leq \frac{\partial \mathcal{L}}{\partial \bar{x}}^T x_i$,

$$\begin{aligned}
&\frac{\partial \mathcal{L}}{\partial \alpha_i} - \frac{\partial \mathcal{L}}{\partial \alpha_j} \\
&= p_i \sum_k p_k \frac{\partial \mathcal{L}}{\partial \bar{x}}^T (x_i - x_k) - p_j \sum_k p_k \frac{\partial \mathcal{L}}{\partial \bar{x}}^T (x_j - x_k) \\
&= (p_i - p_j) \sum_k p_k \frac{\partial \mathcal{L}}{\partial \bar{x}} (x_i - x_k) + p_j (\sum_k p_k \frac{\partial \mathcal{L}}{\partial \bar{x}} (x_i - x_k) \\
&- \sum_k p_k \frac{\partial \mathcal{L}}{\partial \bar{x}} (x_j - x_k)) \\
&= (p_i - p_j) \sum_k p_k \frac{\partial \mathcal{L}}{\partial \bar{x}} (x_i - x_k) + p_j \sum_k p_k \frac{\partial \mathcal{L}}{\partial \bar{x}} (x_i - x_j)
\end{aligned}$$

As a result

$$
\begin{aligned}
&\frac{\partial l}{\partial \alpha_{i^*}} - \frac{\partial l}{\partial \alpha_i} \\
&= p_{i^*} \sum_j p_j \frac{\partial l}{\partial \boldsymbol{z}}(\boldsymbol{z}_{i^*} - \boldsymbol{z}_j) - p_i \sum_j p_j \frac{\partial l}{\partial \boldsymbol{z}}(\boldsymbol{z}_i - \boldsymbol{z}_j) \\
&= (p_{i^*} - p_i) \sum_j p_j \frac{\partial l}{\partial \boldsymbol{z}}(\boldsymbol{z}_{i^*} - \boldsymbol{z}_j) + p_i (\sum_j p_j \frac{\partial l}{\partial \boldsymbol{z}}(\boldsymbol{z}_{i^*} - \boldsymbol{z}_j) \\
&- \sum_j p_j \frac{\partial l}{\partial \boldsymbol{z}}(\boldsymbol{z}_i - \boldsymbol{z}_j)) \\
&= (p_{i^*} - p_i) \sum_j p_j \frac{\partial l}{\partial \boldsymbol{z}}(\boldsymbol{z}_{i^*} - \boldsymbol{z}_j) + p_i \sum_j p_j \frac{\partial l}{\partial \boldsymbol{z}}(\boldsymbol{z}_{i^*} - \boldsymbol{z}_i) \\
&\geq (p_{i^*} - p_i)\delta + p_i \delta \\
&= p_{i^*} \delta \\
&\geq \frac{\delta}{n}
\end{aligned}
\tag{17}
$$

Under gradient ascent, on update step t we have, for any $i \neq i^*$,

$$
\begin{aligned}
\alpha_{i^*}^t - \alpha_i^t &= \alpha_{i^*}^{t-1} - \alpha_i^{t-1} + \eta (\frac{dl_{t-1}}{d\alpha_{i^*}^{t-1}} - \frac{dl_{t-1}}{d\alpha_i^{t-1}}) \\
&\geq \alpha_{i^*}^{t-1} - \alpha_i^{t-1} + \eta p_{i^*}^t \delta_t \\
&\geq \alpha_{i^*}^0 - \alpha_i^0 + \eta \sum_t p_{i^*}^t \delta_t \\
&\geq \frac{\eta t \delta}{n}
\end{aligned}
\tag{18}
$$

When $t = \frac{n \ln((1-\epsilon)n)}{\eta \delta}$, we have

$$
\alpha_{i^*} - \alpha_i \geq \ln((1-\epsilon)n)
$$

Thus

$$
p_{i^*} = \frac{1}{\sum_i \exp(\alpha_i - \alpha_{i^*})} \geq \frac{1}{\sum_i \exp(-\ln((1-\epsilon)n))} = 1 - \epsilon \tag{19}
$$

Under gradient descent, for $i^* = \arg\min_i \frac{\partial l}{\partial \boldsymbol{z}} \boldsymbol{z}_i$, we have the same conclusion.

D. Searched results in DARTS

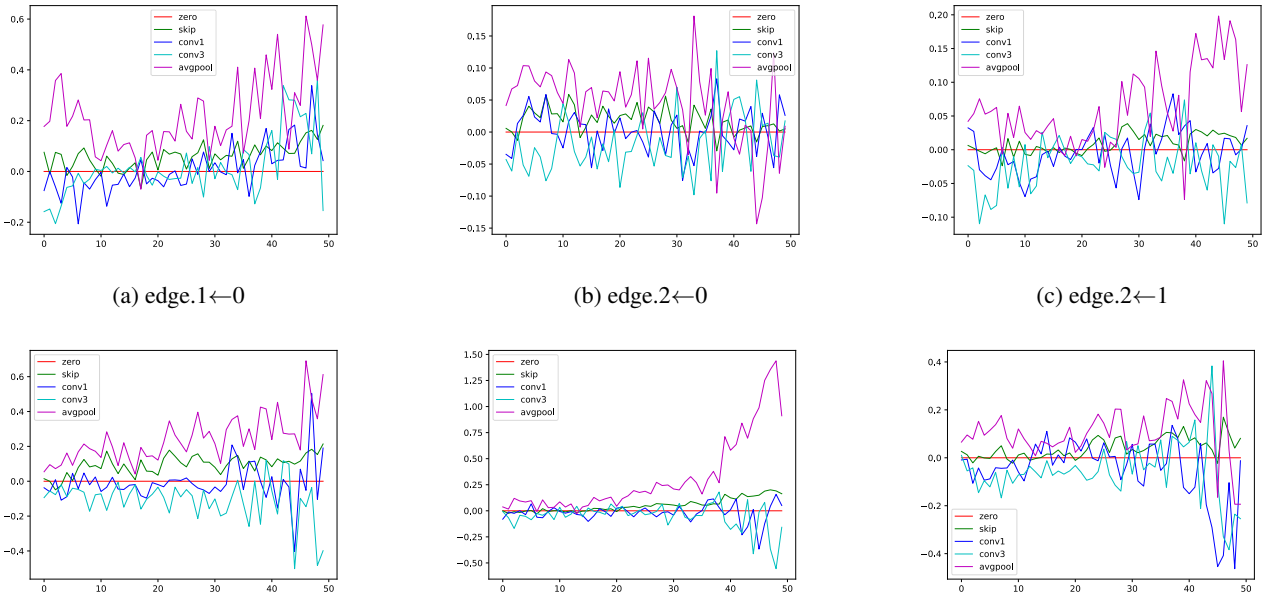
We use Single-DARTS to search directly on the ImageNet-1K dataset in DARTS space. Initializing α_i as $-\ln(7)$ ($\mathrm{softmax}(\alpha_i) = 0.125$) will improve the performance. The training setting follows PDARTS, without any additional tricks. In addition, for Single-DARTS, using half of the dataset also gains promising results.

E. More comparison of gradients of p_i in 5.2.2

F. Visualization of architectures

Table 6: $*$ denotes α is initialized as $-\ln(7)$. 'Data' means using full or half of data to search.

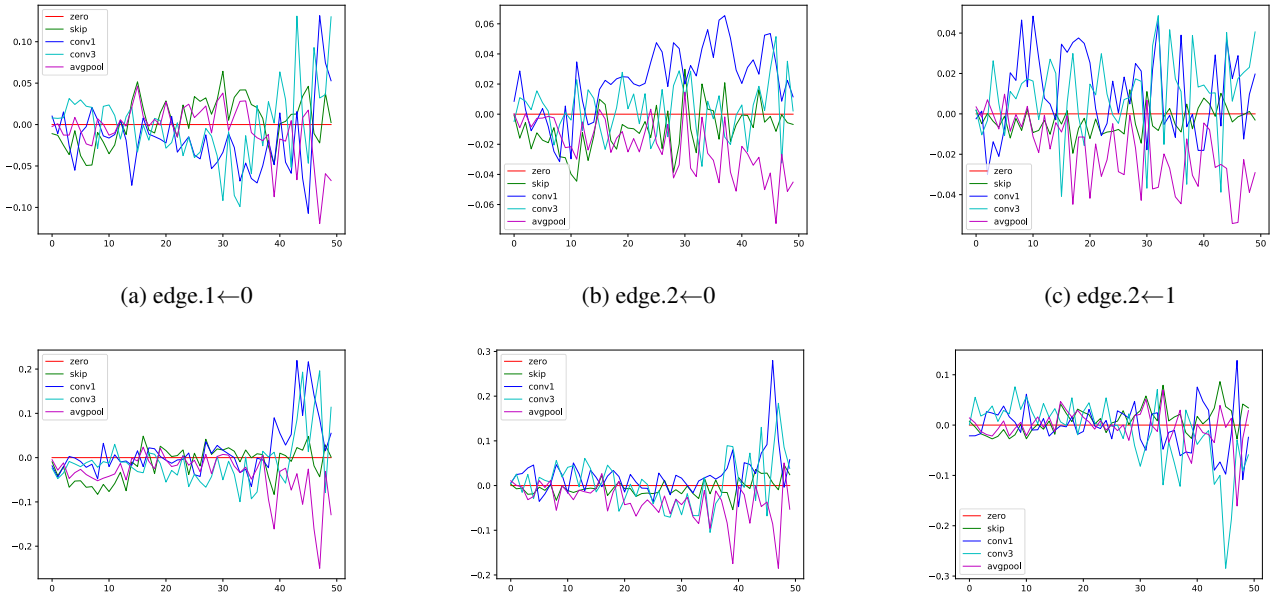
Activation	**Data** (M)	**Seed** (M)	**FLOPs** (%)	**Params**	**Top-1.**
softmax	full	0	714.72	6.58	76.28
softmax	full	1	712.92	6.56	76.71
softmax	full	2	722.25	6.61	76.27
sigmoid	full	0	738.21	6.69	76.12
sigmoid	full	1	738.21	6.69	76.58
sigmoid	full	2	721.35	6.60	76.29
sigmoid*	full	0	707.89	6.50	76.96
sigmoid*	full	1	721.35	6.60	77.0
sigmoid*	full	2	700.61	6.40	76.95
softmax	half	0	712.01	6.55	76.51
softmax	half	1	692.18	6.36	76.31
softmax	half	2	709.04	6.45	76.51
sigmoid*	half	0	720.44	6.59	76.67
sigmoid*	half	1	692.18	6.36	76.78
sigmoid*	half	2	707.89	6.50	76.54



(a) edge.1←0 (b) edge.2←0 (c) edge.2←1

(d) edge.3←0 (e) edge.3←1 (f) edge.3←2

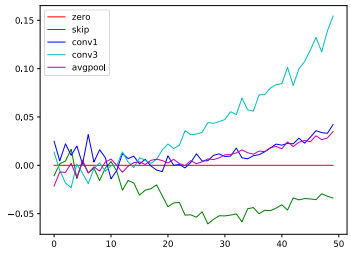
Figure 5: DARTS, the 0th cell.



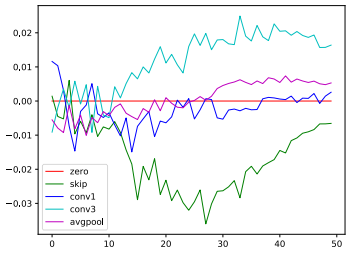
(a) edge.1←0 (b) edge.2←0 (c) edge.2←1

(d) edge.3←0 (e) edge.3←1 (f) edge.3←2

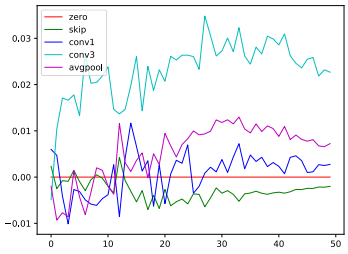
Figure 6: DARTS, the 8th cell.



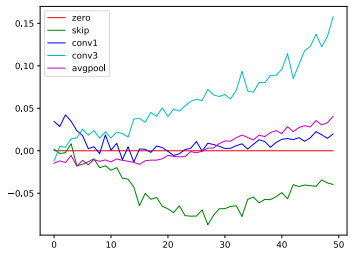
(a) edge.1←0



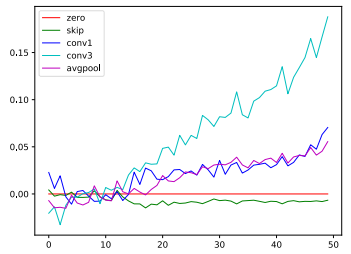
(b) edge.2←0



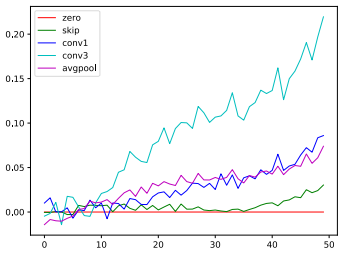
(c) edge.2←1



(d) edge.3←0

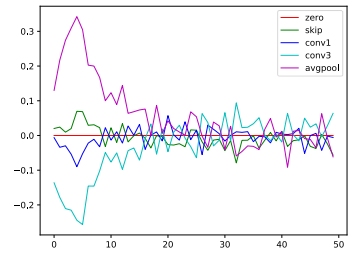


(e) edge.3←1

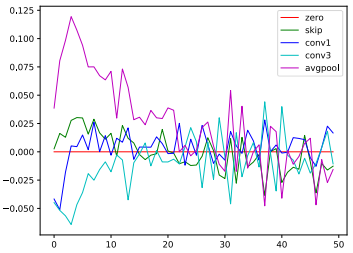


(f) edge.3←2

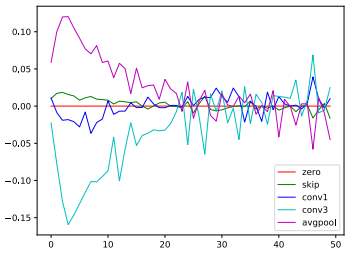
Figure 7: DARTS, the 16th cell.



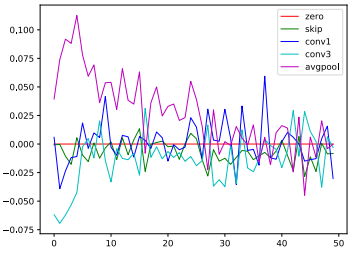
(a) edge.1←0



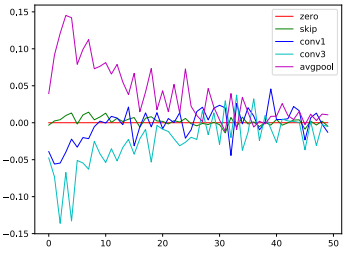
(b) edge.2←0



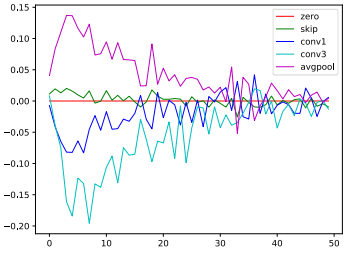
(c) edge.2←1



(d) edge.3←0

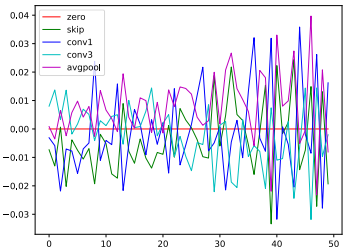


(e) edge.3←1

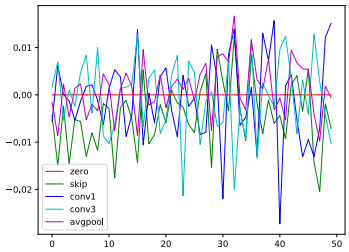


(f) edge.3←2

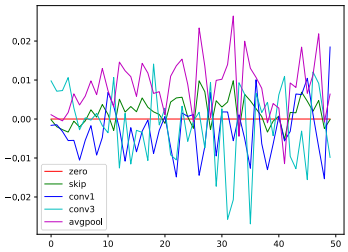
Figure 8: Single-DARTS, the 0th cell.



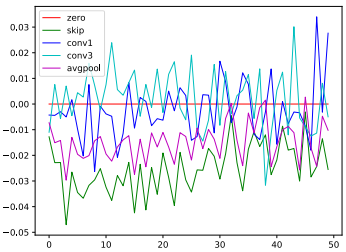
(a) edge.1←0



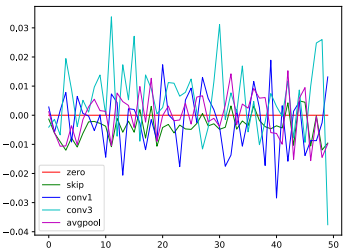
(b) edge.2←0



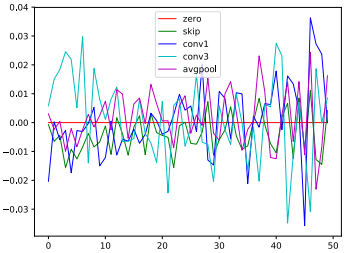
(c) edge.2←1



(d) edge.3←0

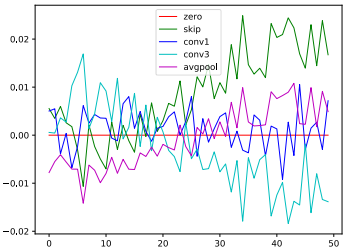


(e) edge.3←1

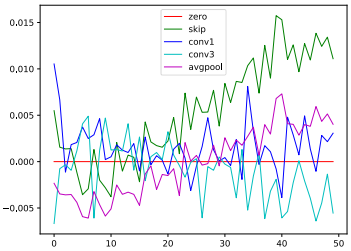


(f) edge.3←2

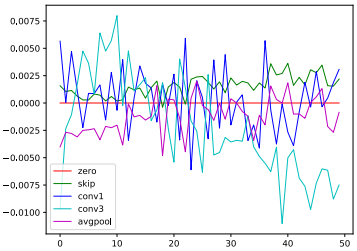
Figure 9: Single-DARTS, the 8th cell.



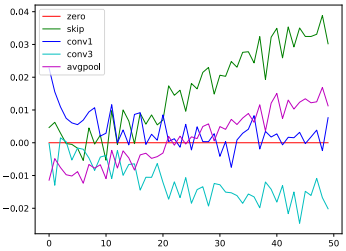
(a) edge.1←0



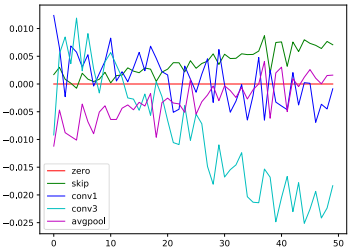
(b) edge.2←0



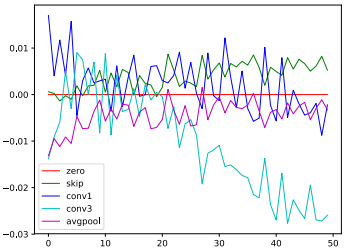
(c) edge.2←1



(d) edge.3←0

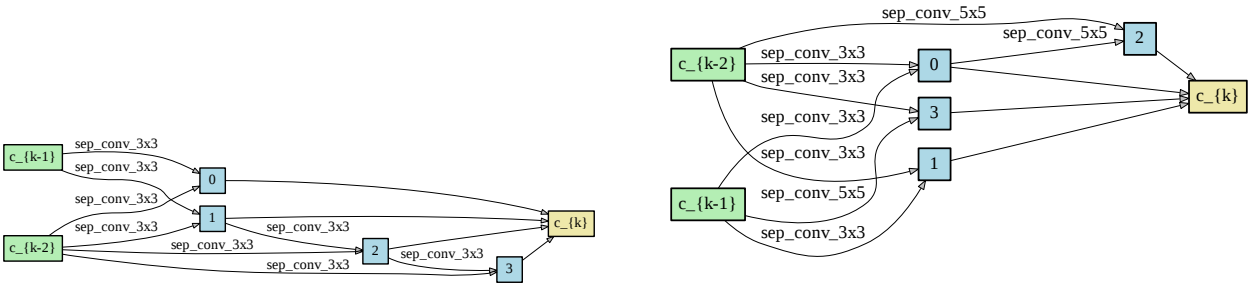


(e) edge.3←1



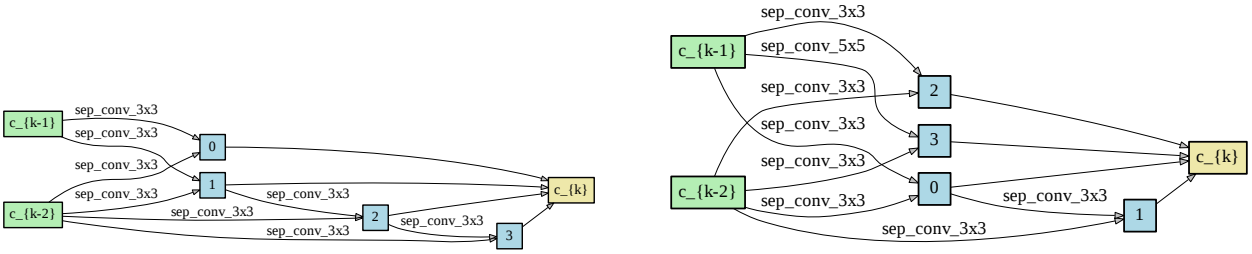
(f) edge.3←2

Figure 10: Single-DARTS, the 16th cell.



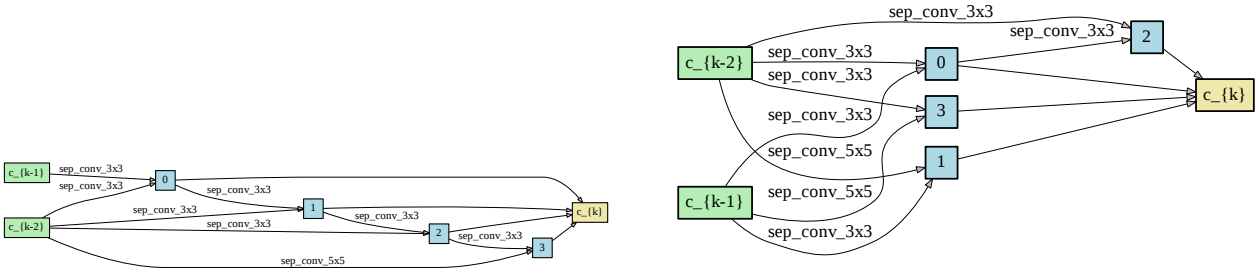
(a) normal (b) reduction

Figure 11: activation=softmax, data=full, seed=0



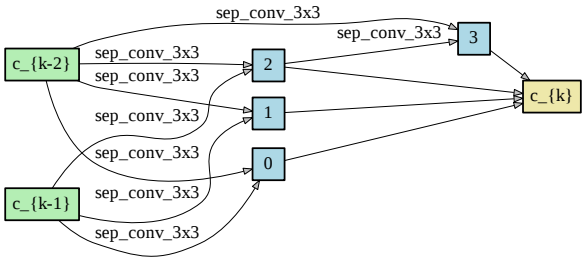
(a) normal (b) reduction

Figure 12: activation=softmax, data=full, seed=1

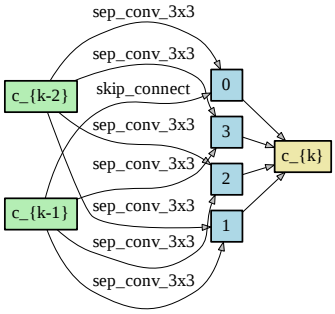


(a) normal (b) reduction

Figure 13: activation=softmax, data=full, seed=2

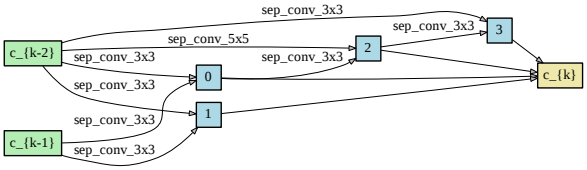


(a) normal cell

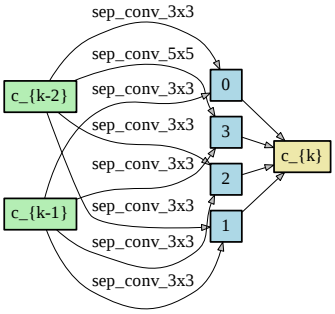


(b) reduction cell

Figure 14: activation=sigmoid*, data=full, seed=0

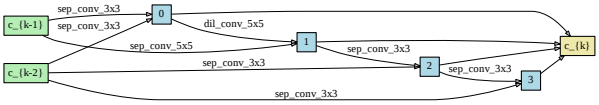


(a) normal

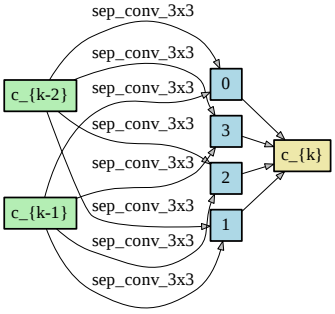


(b) reduction

Figure 15: activation=sigmoid*, data=full, seed=2



(a) normal



(b) reduction

Figure 16: activation=sigmoid*, data=full, seed=2