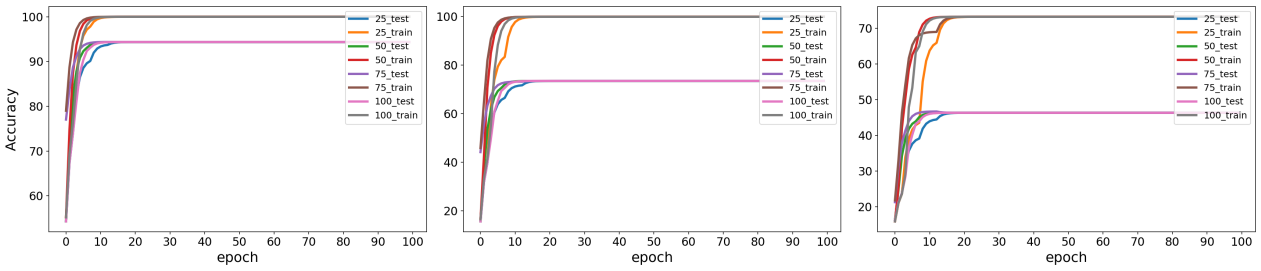
# Supplementary File for Beta-Decay Regularization for Differentiable Architecture Search

Peng Ye[1]*, Baopu Li[2], Yikang Li[3], Tao Chen[1]†, Jiayuan Fan[1], Wanli Ouyang[4]
[1]Fudan University, [2]BAIDU USA LLC, [3]Shanghai AI Laboratory,
[4]The University of Sydney, SenseTime Computer Vision Group, Australia




(a) CIFAR10 (b) CIFAR100 (c) ImageNet16

Figure 1. Accuracy of different datasets of $\beta$-DARTS with different max weights (i.e., 25/50/75/100) on NAS-Bench-201 benchmark. The curve is smoothed with a coefficient of 0.5. Note that we only search once on CIFAR-10 dataset and report the results of different datasets.

## A. Related Details

### A.1. More Results about The Search Trajectories.

In Fig. 1, we present the search trajectories of different datasets of $\beta$-DARTS with different max weights (i.e., 25/50/75/100) on NAS-Bench-201 benchmark. Similarly, we can see that: (1) the performance collapse issue is well solved and $\beta$-DARTS has a stable search process; (2) the architecture found on CIFAR-10 performs well on CIFAR-10, CIFAR-100 and ImageNet; (3) the search process of different datasets reach the optimal point at an early stage but in different epochs; (4) different runs of searching under different max weight always find the same optimal solution.

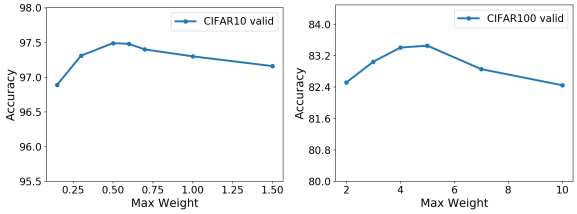
### A.2. Wide Range of The Optimal Weight.

In Fig. 2, we further show the influence of different max weights on the searching results of $\beta$-DARTS on common DARTS search space on CIFAR-10 and CIFAR-100. Also, the performance of original DARTS (i.e., 97.00 on CIFAR-10 and 82.46 on CIFAR-100) is improved on a wide range of max weights and the optimal searching result is obtained on multiple values of max weights.

### A.3. Derivation of Eq. (8) in The Main Text.

For the optimization process of architecture parameters, we utilize the following unified formulas to represent the single-step update with/without regularization.

$$\begin{aligned} \alpha_k^{t+1} &\leftarrow \alpha_k^t - \eta_\alpha \nabla_{\alpha_k} \mathcal{L}_{val} \\ \bar{\alpha}_k^{t+1} &\leftarrow \alpha_k^t - \eta_\alpha \nabla_{\alpha_k} \mathcal{L}_{val} - \eta_\alpha \lambda F\left(\alpha_k^t\right) \end{aligned} \tag{1}$$




(a) CIFAR-10 (b) CIFAR-100

Figure 2. The effects of different max weight of linear increase weighting on the searching results of CIFAR-10 and CIFAR-100.

For the single-step update without regularization, via $\alpha$ in Eq. (1), we can compute the softmax-activated architecture parameters $\beta$ as

$$\beta_k^{t+1} = \frac{\exp\left(\alpha_k^{t+1}\right)}{\sum_{k'=1}^{|\mathcal{O}|} \exp\left(\alpha_{k'}^{t+1}\right)} \tag{2}$$

For the single-step update with regularization, based on $\bar{\alpha}$ in Eq. (1), we can further compute the softmax-activated architecture parameters $\bar{\beta}$ as

$$\begin{aligned} \bar{\beta}_k^{t+1} &= \frac{\exp\left(\bar{\alpha}_k^{t+1}\right)}{\sum_{k'=1}^{|\mathcal{O}|} \exp\left(\bar{\alpha}_{k'}^{t+1}\right)} \\ &= \frac{\exp\left(-\lambda\eta_\alpha F(\alpha_k^t)\right)\exp\left(\alpha_k^{t+1}\right)}{\sum_{k'=1}^{|\mathcal{O}|} \exp\left(-\lambda\eta_\alpha F(\alpha_{k'}^t)\right)\exp\left(\alpha_{k'}^{t+1}\right)} \\ &= \frac{\exp\left(\alpha_k^{t+1}\right)}{\sum_{k'=1}^{|\mathcal{O}|} \frac{\exp\left(-\lambda\eta_\alpha F(\alpha_{k'}^t)\right)}{\exp\left(-\lambda\eta_\alpha F(\alpha_k^t)\right)} \exp\left(\alpha_{k'}^{t+1}\right)} \\ &= \frac{\exp\left(\alpha_k^{t+1}\right)}{\sum_{k'=1}^{|\mathcal{O}|} \left[\exp\left(F(\alpha_k^t) - F(\alpha_{k'}^t)\right)\right]^{\lambda\eta_\alpha} \exp\left(\alpha_{k'}^{t+1}\right)} \end{aligned} \tag{3}$$

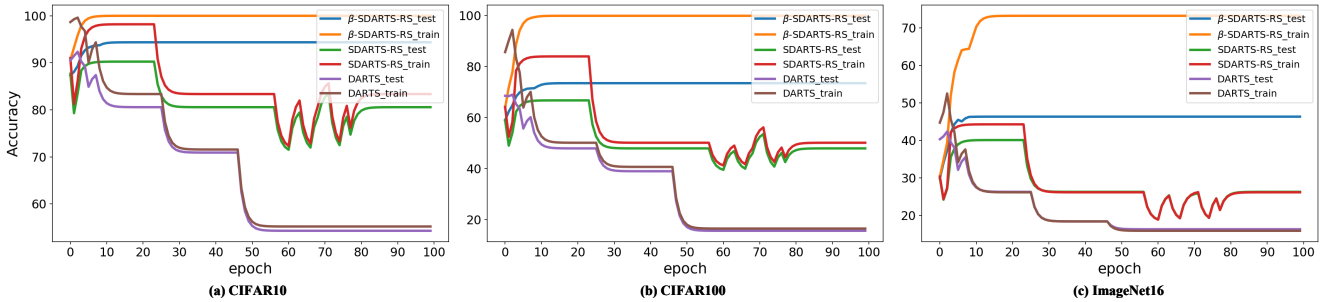



Figure 3. Accuracy of different datasets of DARTS, SDARTS-RS [2] and $\beta$-SDARTS-RS on NAS-Bench-201 benchmark. The curve is smoothed with a coefficient of 0.5. Note that we only search once on CIFAR-10 dataset and report the results of different datasets.

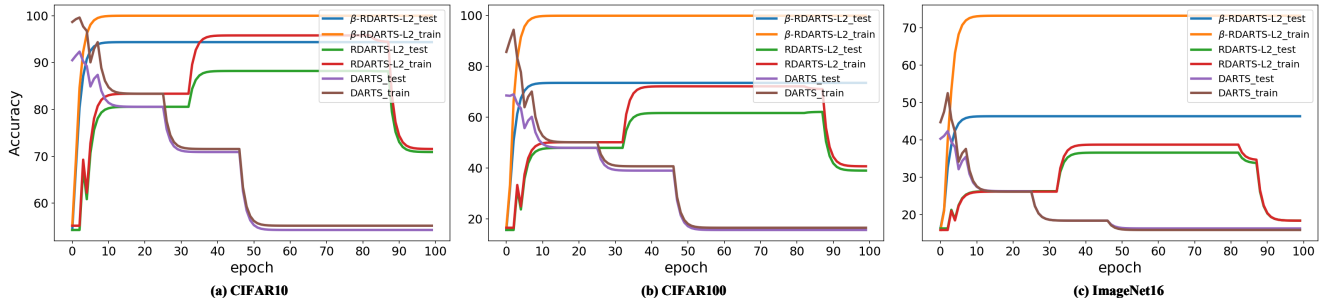



Figure 4. Accuracy of different datasets of DARTS, RDARTS-L2 [1] and $\beta$-RDARTS-L2 on NAS-Bench-201 benchmark. The curve is smoothed with a coefficient of 0.5. Note that we only search once on CIFAR-10 dataset and report the results of different datasets.

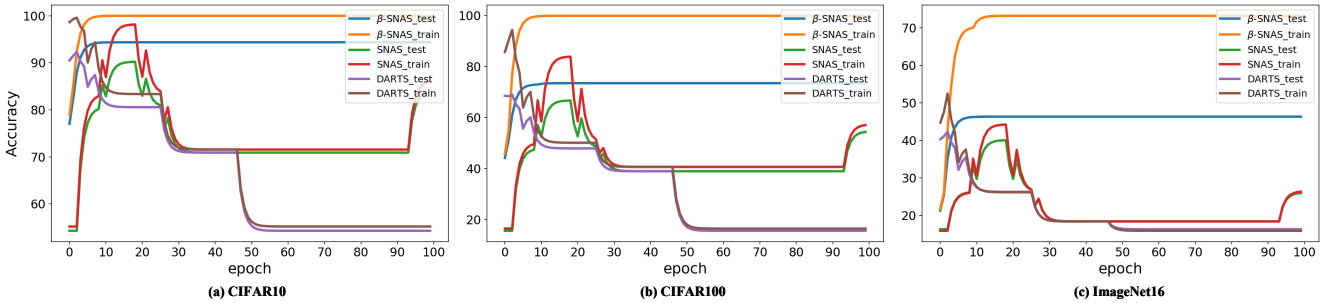



Figure 5. Accuracy of different datasets of DARTS, SNAS [3] and $\beta$-SNAS on NAS-Bench-201 benchmark. The curve is smoothed with a coefficient of 0.5. Note that we only search once on CIFAR-10 dataset and report the results of different datasets.

Table 1. The results of different DARTS variants and their Beta Decay regularization improved versions on NAS-Bench-201 benchmark. Note that we only search on CIFAR-10 dataset, and perform 3 runs of searching under different random seeds.

| Methods | CIFAR-10 | | CIFAR-100 | | ImageNet16-120 | |
|---|---|---|---|---|---|---|
| | valid | test | valid | test | valid | test |
| SDARTS-RS | 75.21/68.29/75.21 | 80.57/70.92/80.57 | 47.51/38.57/47.51 | 47.93/38.97/47.93 | 27.79/18.87/27.79 | 26.29/18.41/26.29 |
| **$\beta$-SDARTS-RS** | **91.55/91.61/91.61** | **94.36/94.37/94.37** | **73.49/72.75/72.75** | **73.51/73.22/73.22** | **46.37/45.56/45.56** | **46.34/46.71/46.71** |
| RDARTS-L2 | 68.29/68.29/39.77 | 70.92/70.92/54.30 | 38.57/38.57/15.03 | 38.97/38.97/15.61 | 18.87/18.87/16.43 | 18.41/18.41/16.32 |
| **$\beta$-RDARTS-L2** | **91.55/91.28/91.61** | **94.36/93.79/94.37** | **73.49/71.88/72.75** | **73.51/71.60/73.22** | **46.37/46.40/45.56** | **46.34/46.67/46.71** |
| SNAS | 82.39/89.07/91.14 | 84.16/91.89/93.60 | 54.57/67.11/71.38 | 54.64/66.99/70.74 | 27.17/39.98/44.10 | 26.10/39.13/45.03 |
| **$\beta$-SNAS** | **91.55/91.55/91.55** | **94.36/94.36/94.36** | **73.49/73.49/73.49** | **73.51/73.51/73.51** | **46.37/46.37/46.37** | **46.34/46.34/46.34** |
| DARTS(1st) | 39.77 | 54.30 | 15.03 | 15.61 | 16.43 | 16.32 |
| **optimal** | **91.61** | **94.37** | **74.49** | **73.51** | **46.77** | **47.31** |

Then, dividing Eq. (3) by Eq. (2), we can obtain the influence of $\alpha$ with regularization on $\beta$.

$$\frac{\bar{\beta}_k^{t+1}}{\beta_k^{t+1}} = \frac{\sum_{k'=1}^{|\mathcal{O}|} \exp\left(\alpha_{k'}^{t+1}\right)}{\sum_{k'=1}^{|\mathcal{O}|} \left[\exp\left(F(\alpha_k^t) - F(\alpha_{k'}^t)\right)\right]^{\lambda\eta_\alpha} \exp\left(\alpha_{k'}^{t+1}\right)} \quad (4)$$

## B. Combination with Other Variants

The proposed Beta Decay regularization can easily combine with other DARTS variants for improving both the robustness of the search process and the generalization abil-

Table 2. The results of $\beta$-DARTS (partial data) that uses partial data (e.g. 75%, 50% and 25% of CIFAR-10) for searching on NAS-Bench-201 benchmark. Note that we only search on CIFAR-10 dataset, and perform 3 runs of searching under different random seeds.

| $\beta$-DARTS | Cost | Weighting | CIFAR-10 | | CIFAR-100 | | ImageNet16-120 | |
|---|---|---|---|---|---|---|---|---|
| (partial data) | (hours) | scheme | valid | test | valid | test | valid | test |
| 100% | 3.2 | 0-50 | 91.55/91.55/91.55 | 94.36/94.36/94.36 | 73.49/73.49/73.49 | 73.51/73.51/73.51 | 46.37/46.37/46.37 | 46.34/46.34/46.34 |
| 75% | 2.4 | 0-50 | 91.55/91.55/91.50 | 94.36/94.36/94.37 | 73.49/73.49/73.31 | 73.51/73.51/73.09 | 46.37/46.37/45.59 | 46.34/46.34/46.33 |
| 50% | 1.6 | 0-100 | 91.55/91.44/91.55 | 94.36/94.34/94.36 | 73.49/72.74/73.49 | 73.51/72.75/73.51 | 46.37/46.56/46.37 | 46.34/46.59/46.34 |
| 25% | 0.8 | 0-100 | 91.35/91.40/91.42 | 94.30/93.88/93.81 | 72.77/72.42/72.40 | 72.30/73.16/73.26 | 45.53/45.77/46.50 | 46.44/45.67/46.50 |
| **optimal** | - | - | **91.61** | **94.37** | **74.49** | **73.51** | **46.77** | **47.31** |

ity of the searched architecture. To show this, we also test the proposed method on SDARTS-RS [2], RDARTS-L2 [1] and SNAS [3] on NAS-Bench-201 benchmark (shown as $\beta$-SDARTS-RS, $\beta$-RDARTS-L2 and $\beta$-SNAS). The implementation is consistent with their papers or open source code. When adapting Beta Decay regularization, we set the weighting schemes of SDARTS-RS, RDARTS-L2 and SNAS as 0-50, 0-50 and 0-5 respectivley.

In Fig. 3, Fig. 4 and Fig. 5, we firstly compare the search trajectories between different baselines and their Beta Decay regularization improved versions. The search trajectory of original DARTS is also shown in each figure. As we can see, although all these variants can improve the final result of original DARTS, they still suffer from the performance degradation issue. As a comparison, with the help of Beta Decay regularization, all the improved versions have continuously rising performance and reach much higher accuracy than the original variants. In addition, we can see that the architectures searched by all these variants have a poor generalization ability, while Beta Decay regularization has the ability to relieve this problem.

In Table. 1, we further compare the final searching performance between different baselines and their Beta Decay regularization improved versions. The searching results of original DARTS and the optimal results of NAS-Bench-201 are also shown in the table. As we can see, Beta Decay regularization can boost the performance of all these variants by a large margin, close to the optimal results, across different datasets. More importantly, with Beta Decay regularization, multiple runs of searching have relatively low variance and always find the satisfactory solution, meanwhile the architectures searched on CIFAR-10 can perform well on CIFAR-10, CIFAR-100 and ImageNet.

## C. Searching with Partial Data

In Table. 2, we show the results of $\beta$-DARTS (partial data), which uses partial data (e.g. 75%, 50% and 25% samples of CIFAR-10 dataset) for searching on NAS-Bench-201 benchmark. Here, we also only search once on CIFAR-10 dataset. As we can see, even using partial data for searching, $\beta$-DARTS can still maintain a SOTA performance, such results further verify the outstanding property of our search scheme, namely being less dependent on training data. As a byproduct, when we use 25% samples of CIFAR-10 dataset for searching, the time cost will be reduced by 4 times.

## D. Visualization of Searched Genotypes

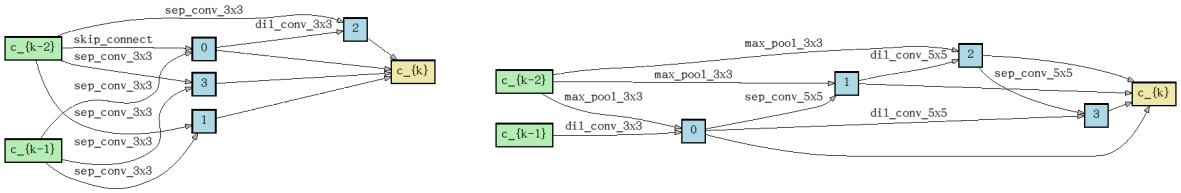
In Fig. 6 and Fig. 7, we visualize the genotypes of normal and reduction cells searched on common DARTS search space on CIFAR-10 and CIFAR-100 datasets respectively.

## References

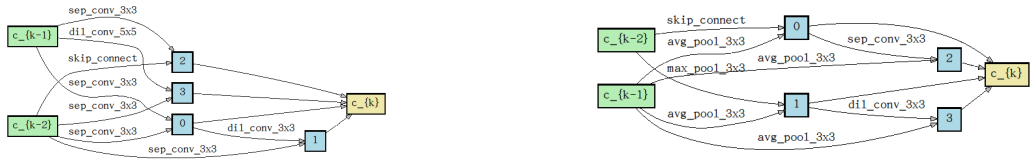

[1] Thomas Elsken Arber Zela, Tonmoy Saikia, Yassine Marrakchi, Thomas Brox, and Frank Hutter. Understanding and robustifying differentiable architecture search. In *International Conference on Learning Representations*, volume 3, page 7, 2020. 2, 3

[2] Xiangning Chen and Cho-Jui Hsieh. Stabilizing differentiable architecture search via perturbation-based regularization. In *International Conference on Machine Learning*, pages 1554–1565. PMLR, 2020. 2, 3

[3] Sirui Xie, Hehui Zheng, Chunxiao Liu, and Liang Lin. Snas: stochastic neural architecture search. *arXiv preprint arXiv:1812.09926*, 2018. 2, 3



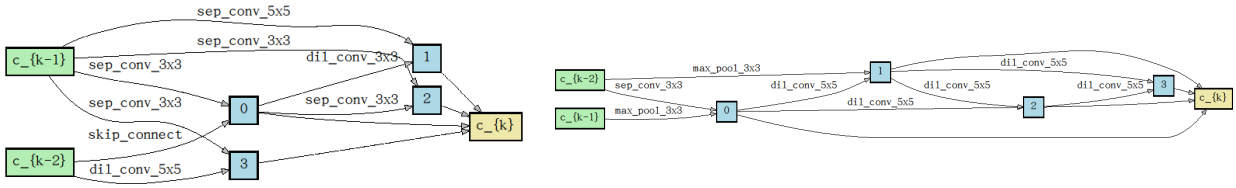

**(a) Normal Cell and Reduction Cell**



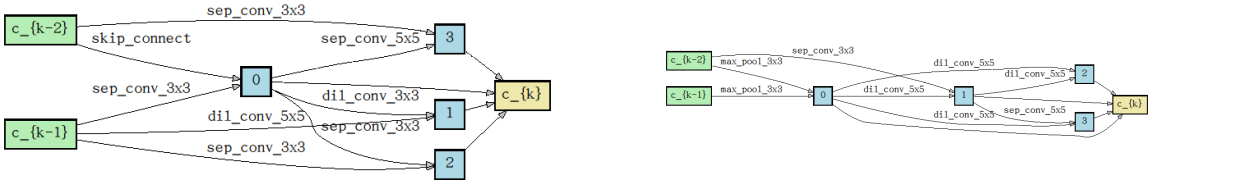

**(b) Normal Cell and Reduction Cell**


**(c) Normal Cell and Reduction Cell**

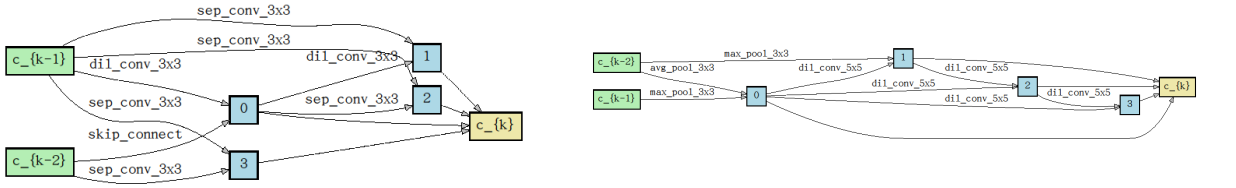
Figure 6. Normal cell (left) and Reduction cell (right) discovered by $\beta$-DARTS on common DARTS search space on CIFAR-10 dataset. (a), (b) and (c) denote the found genotypes of the 3 runs of independent searching.




**(a) Normal Cell and Reduction Cell**




**(b) Normal Cell and Reduction Cell**




**(c) Normal Cell and Reduction Cell**

Figure 7. Normal cell (left) and Reduction cell (right) discovered by $\beta$-DARTS on common DARTS search space on CIFAR-100. (a), (b) and (c) denote the found genotypes of 3 runs of independent searching.