# A proper way of regularizing DARTS

Department of Computer Engineering

Dongseo University

Jie Yong Shin,

2023.01.26

# CONTENTS

1. Highlights (1)
2. Problem of DARTS (6)
3. Beta-Decay DARTS (6)
4. Benefits of Beta-Decay DARTS (6)
5. Experimental Results (5)
6. Conclusion (1)

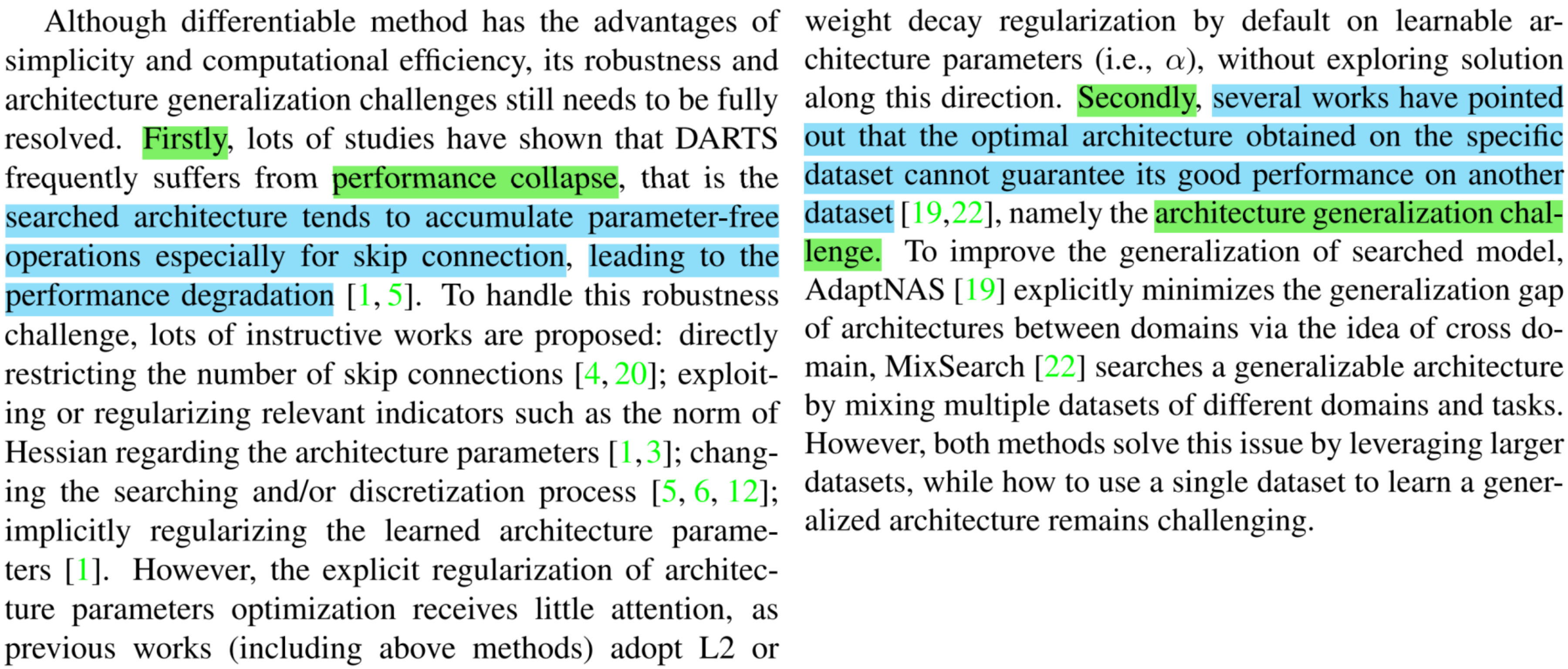
# 1. Highlights

1. Find out the reason why TD-DARTS couldn't improve much compared to DARTS on accuracy
   - Prevent skip connection with Regularization

2. Learn an effective way to regularize Differentiable Architecture Search
   - Based on Beta-Decay DARTS

3. Observe the difference between Beta-Decay DARTS and other regularization methods

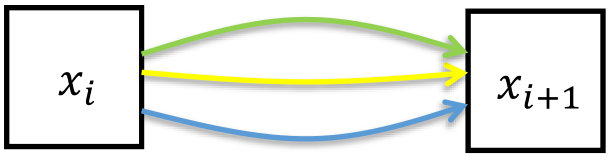
# 2. Problems of DARTS



- **Problems**

1. Discrepancy: Between Continuously encoded architecture and Discrete architecture.
2. Performance Collapse: Searched architecture tends to select skip connection.
3. Architecture Generalization Challenge: Architecture searched in the specific dataset cannot guarantee it is good also on another dataset.

# 2. Problem of DARTS




$$\bar{o}^{(i,j)}(x_i) = \sum_{o \in \mathcal{O}} \frac{\exp\left(\alpha_o^{(i,j)}\right)}{\sum_{o' \in \mathcal{O}} \exp\left(\alpha_{o'}^{(i,j)}\right)} o(x_i)$$

- Let, Operation Pool = {Convolution(), Pooling(), Skip()}

- One-hot encoding:

Operation 1 : Convolution() = [1, 0, 0]
Operation 2 : Pooling() = [0, 1, 0]
Operation 3 : Skip() = [0, 0, 1]

- DARTS Assumes that,

$\alpha_{op1}$ ≈ Convolution() ≈ [1, 0, 0]
$\alpha_{op2}$ ≈ Pooling() ≈ [0, 1, 0]
$\alpha_{op3}$ ≈ Skip() ≈ [0, 0, 1]

# 2. Problem of DARTS

$$\bar{o}^{(i,j)}(x_i) = \sum_{o \in \mathcal{O}} \frac{\exp\left(\alpha_o^{(i,j)}\right)}{\sum_{o' \in \mathcal{O}} \exp\left(\alpha_{o'}^{(i,j)}\right)} o(x_i)$$

$$\bar{O}^{(i,j)}(x) = \sum_{k=1}^{|\mathcal{O}|} \beta_k^{(i,j)} O_k(x)$$

$$\beta_k^{(i,j)} = \frac{\exp\left(\alpha_k^{(i,j)}\right)}{\sum_{k'=1}^{|\mathcal{O}|} \exp\left(\alpha_{k'}^{(i,j)}\right)}$$

# 2. Problem of DARTS



$$\alpha = [\alpha_1, \alpha_2, \alpha_3, \cdots, \alpha_k]$$

$$\alpha_1 \approx OP_1$$

$$\alpha_2 \approx OP_2$$

$$OP_1 = [1, 0, 0, \cdots, 0]$$

$$OP_2 = [0, 1, 0, \cdots, 0]$$

$\max(\alpha)$ = optimal operation

# 2. Problem of DARTS



$$\beta = \text{softmax}(\alpha)$$

$$\beta_k = \frac{e^{\alpha_k}}{\sum_{k'=1}^{101} e^{\alpha_{k'}}}$$

$$\beta = [\beta_1, \beta_2, \beta_3, \cdots, \beta_k]$$

# 2. Problem of DARTS

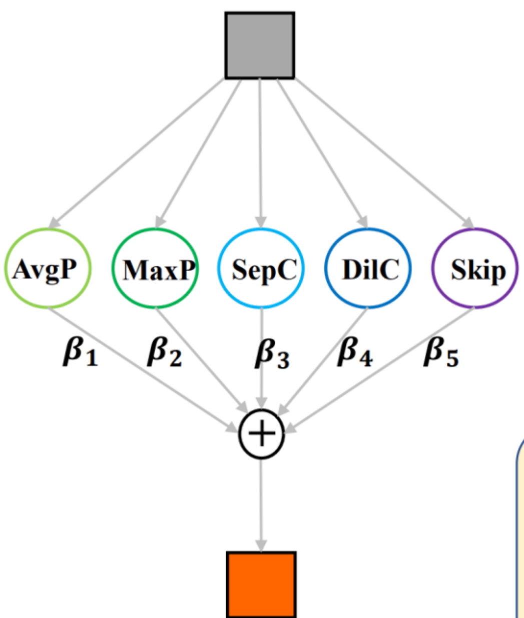
ex) Let. $\alpha = [\alpha_1, \alpha_2, \alpha_3]$

$$\beta = \left[ \beta_1 = \frac{e^{\alpha_1}}{e^{\alpha_1} + e^{\alpha_2} + e^{\alpha_3}},\ \beta_2 = \frac{e^{\alpha_2}}{e^{\alpha_1} + e^{\alpha_2} + e^{\alpha_3}},\ \beta_3 = \frac{e^{\alpha_3}}{e^{\alpha_1} + e^{\alpha_2} + e^{\alpha_3}} \right]$$

$$\therefore \sum \beta = 1$$

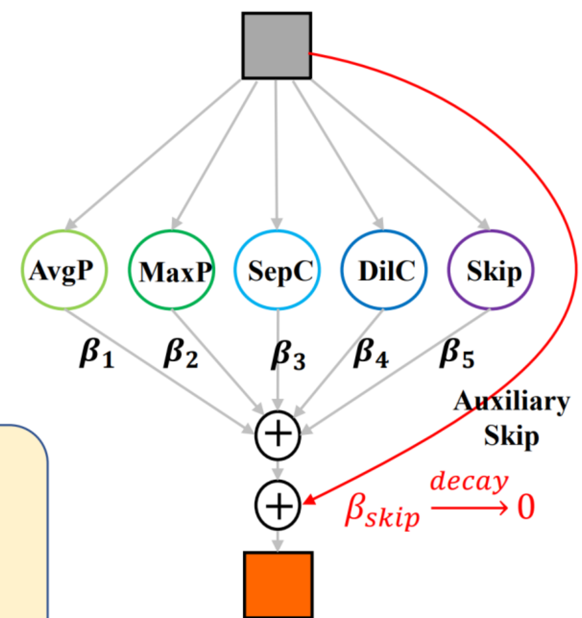
# $\beta$-DARTS: Beta-Decay Regularization for Differentiable Architecture Search

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
yepeng20@fudan.edu.cn




$$\beta_k = \frac{exp(\alpha_k)}{\sum_{k'=1}^{|O|} exp(\alpha_{k'})}$$

**DARTS and DARTS-:**
**Regularize $\alpha$**
**$\beta$-DARTS:**
**Regularize $\beta$**



# 3. Beta-Decay DARTS

**Algorithm 1** PyTorch Implementation in DARTS

1: $\mathcal{L}_{Beta}$ = torch.mean(torch.logsumexp(
self.model._arch_parameters, dim=-1))
2: loss = self._val_loss(self.model, input_valid,
target_valid)+$\lambda\mathcal{L}_{Beta}$

**Algorithm 2** $\beta$-DARTS

**Require:**
Architecture parameters $\alpha$; Network weights $w$; Number of search epochs $E$; Regularization coefficient adjustment scheme $\lambda_e, e \in \{1, 2, \ldots, E\}$.
1: Construct a supernet and initialize architecture parameters $\alpha$ and supernet weights $w$
2: For each $e \in [1, E]$ do
3: Update architecture parameters $\alpha$ by descending $\nabla_\alpha \mathcal{L}_{val} + \lambda_e \mathcal{L}_{Beta}$
4: Update network weights w by descending $\nabla_w \mathcal{L}_{train}$
5: Derive the final architecture based on the learned $\alpha$.

# 3. Beta-Decay DARTS



- How DARTS update

$$\alpha_k^{t+1} \leftarrow \alpha_k^t - \eta_\alpha \cdot \nabla_{\alpha_k} \mathcal{L}_{val}$$

- Temperature Decay DARTS update

$$\alpha_k^{t+1} \leftarrow \frac{\alpha_k^t - \nabla_{\alpha_k} \mathcal{L}_{val}}{T^{t+1}}$$

- Commonly used regularization method on DARTS

$$\bar{\alpha}_k^{t+1} \leftarrow \alpha_k^t - \eta_\alpha \cdot \nabla_{\alpha_k} \mathcal{L}_{val} - \eta_\alpha \lambda \mathcal{N}\left(\alpha_k^t\right)$$

- What Beta-Decay DARTS suggests

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

# 3. Beta-Decay DARTS

- Constrain the value of Beta from changing too much

$$\bar{\beta}_k^{t+1} = \theta_k^{t+1}\left(\alpha_k^t\right)\beta_k^{t+1}$$

Note that, $\beta_k^{(i,j)} = \frac{\exp\left(\alpha_k^{(i,j)}\right)}{\sum_{k'=1}^{|\mathcal{O}|}\exp\left(\alpha_{k'}^{(i,j)}\right)}$

- Influence of Beta-Decay DARTS

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

Where, $F\left(\alpha_k\right) = \frac{\exp\left(\alpha_k\right)}{\sum_{k'=1}^{|\mathcal{O}|}\exp\left(\alpha_{k'}\right)}$

# 3. Beta-Decay DARTS

- The mapping function F has to meet following two points:
  1) F is not affected by the amplitude of $\alpha$
     - To avoid invalid regularization and optimization difficulties
  1) F can refplect the relative amplitude of $\alpha$
     - To impose more penalty on larger amplitude

- So, they have chosen F as

$$F\left(\alpha_k\right)=\frac{\exp\left(\alpha_k\right)}{\sum_{k'=1}^{|\mathcal{O}|}\exp\left(\alpha_{k'}\right)}$$

- Therefore,

$$\theta_k^{t+1}\left(\alpha_k^t\right)=\frac{\sum_{k'=1}^{|\mathcal{O}|}\exp\left(\alpha_{k'}^{t+1}\right)}{\sum_{k'=1}^{|\mathcal{O}|}\left[\exp\left(\frac{\exp(\alpha_k^t)-\exp(\alpha_{k'}^t)}{\sum_{k''=1}^{|\mathcal{O}|}\exp\left(\alpha_{k''}^t\right)}\right)\right]^{\lambda\eta_\alpha}\exp\left(\alpha_{k'}^{t+1}\right)}$$

# 3. Beta-Decay DARTS



- And to make the loss function's gradient respective to $\alpha$ equals F($\alpha$),

$$\mathcal{L}_{Beta} = \log\left(\sum_{k=1}^{|\mathcal{O}|} e^{\alpha_k}\right) = \operatorname{smoothmax}\left(\{\alpha_k\}\right)$$

$$\theta_k^{t+1}\left(\alpha_k^t\right) = \frac{\sum_{k'=1}^{|\mathcal{O}|} \exp\left(\alpha_{k'}^{t+1}\right)}{\sum_{k'=1}^{|\mathcal{O}|}\left[\exp\left(\frac{\exp(\alpha_k^t)-\exp(\alpha_{k'}^t)}{\sum_{k''=1}^{|\mathcal{O}|}\exp\left(\alpha_{k''}^t\right)}\right)\right]^{\lambda\eta_\alpha} \exp\left(\alpha_{k'}^{t+1}\right)}$$

- "As a result, the variance of $\beta$ is constrained to be smaller, and the value of $\beta$ is constrained to be closer to its mean, achieving the effect similar to weight decay, thus called Beta-Decay regularization"

# 4. Benefits of Beta-Decay DARTS

## Benefit 1) Probability to choose skip connection decreases

- According to theorem revealed by recent work, the convergence of network weights $\omega$ can heavily rely on $\beta_{skip}$ in the supernet.
- Loss can be reduced by ratio $(1 - \eta_\omega \varphi / 4)$ with probability of at leat 1-$\sigma$, where $\eta_\omega$ is the corresponding learning rate and will be bounded by $\sigma$, and $\varphi$ obeys.

1) Probability to choose skip connection in DARTS

$$\varphi \propto \sum_{i=0}^{h-2} \left[ \left( \beta_{conv}^{(i,h-1)} \right)^2 \prod_{t=0}^{i-1} \left( \beta_{skip}^{(t,i)} \right)^2 \right]$$

where,
$h$ is the number of supernet layers

2) Probability to choose skip connection in Beta-Decay DARTS

$$\varphi \propto \sum_{i=0}^{h-2} \left[ \left( \theta_{conv}^{(i,h-1)} \beta_{conv}^{(i,h-1)} \right)^2 \prod_{t=0}^{i-1} \left( \theta_{skip}^{(i,h-1)} \beta_{skip}^{(t,i)} \right)^2 \right]$$

- Note that, $\theta$ becomes smaller when $\beta$ is larger

## Benefit 2) Stronger generalization

1. Lipschitz constraint

Let, model = $f_\omega(x)$
When $\|x_1 - x_2\|$ is very small, a well-trained model should meet the following constraints.

$$\|f_w(x_1) - f_w(x_2)\| \leq C(w) \cdot \|x_1 - x_2\|$$

Where, $C(\omega)$ is the Lipschitz constant

- The smaller the constant is, the trained model will be less sensitive to input disturbances and have better generalization ability.

## Benefit 2) Stronger generalization

1. Cauchy's inequality

Let,
Operation set: $F_{(x)} = (f_1(x), f_2(x), f_3(x))$
Architecture parameters $\beta = (\beta_1, \beta_2, \beta_3)$
Then, according to Cauchy's inequality,

$$\left\| \beta F^{\mathrm{T}}(x_1) - \beta F^{\mathrm{T}}(x_2) \right\| \leq \|\beta\| \left\| F^{\mathrm{T}}(x_1) - F^{\mathrm{T}}(x_2) \right\|$$

Where, $\|\beta\| = \sqrt{\Sigma \beta_i^2}$ as Lipschitz constant

and, $\Sigma \beta_i = 1$

- As a result, the smaller the measure $\|\beta\|$ is, the supernet will be less sensitive to the impact of input on the operation set, and the searched architecture will have better generalization ability.

## Benefit 3) Comparison with commonly-used regularization

1. L2 regularization and weight decay regularization

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

- When values in $\alpha$ are all around one, regularization has little effect on Beta.
- When the median of $\alpha$ is equal to 0, has the same effect with Beta regularization.
- Large variance of $\alpha$ makes optimization process more sensitive to the hyperparameter $\lambda$ and $\eta_\alpha$ .

## Benefit 3) Comparison with commonly-used regularization

2. Beta-Decay regularization

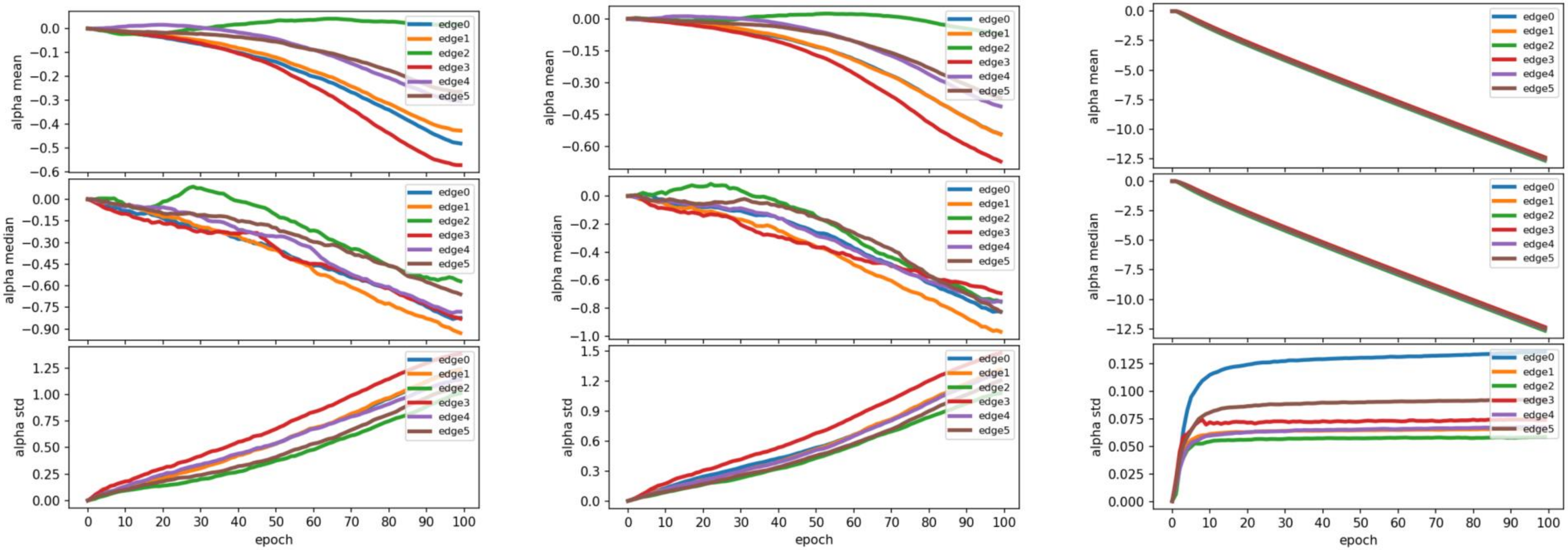
$$\frac{\bar{\beta}_k^{t+1}}{\beta_k^{t+1}} = \frac{\sum_{k'=1}^{|\mathcal{O}|} \exp\left(\alpha_{k'}^{t+1}\right)}{\sum_{k'=1}^{|\mathcal{O}|} \left[\exp\left(\alpha_k^t - \alpha_{k'}^t\right)\right]^{\lambda\eta_\alpha} \exp\left(\alpha_{k'}^{t+1}\right)}$$

- The mean and median of $\alpha$ are basically equal.
- When the standard deviation of $\alpha$ increases to a certain extent, it will remain unchanged.

# 4. Benefits of Beta-Decay DARTS



## Benefit 3) Comparison with commonly-used regularization




**(a) L2 regularization** **(b) weight decay regularization** **(c) Beta-Decay regularization**

- Search on NAS-Bench-201 CIFAR-10 dataset, Adapt on NAS-Bench-201 CIFAR-10, NAS-Bench-201 CIFAR-100, NAS-Bench-201 ImageNet16-120

Table 1. Performance comparison on NAS-Bench-201 benchmark [9]. Note that $\beta$-DARTS only searches on CIFAR-10 dataset, but can robustly achieve new SOTA on CIFAR-10, CIFAR-100 and ImageNet16-120. Averaged on 4 independent runs of searching.

| Methods | Cost (hours) | CIFAR-10 | | CIFAR-100 | | ImageNet16-120 | |
|---|---|---|---|---|---|---|---|
| | | valid | test | valid | test | valid | test |
| DARTS(1st) [21] | 3.2 | 39.77±0.00 | 54.30±0.00 | 15.03±0.00 | 15.61±0.00 | 16.43±0.00 | 16.32±0.00 |
| DARTS(2nd) [21] | 10.2 | 39.77±0.00 | 54.30±0.00 | 15.03±0.00 | 15.61±0.00 | 16.43±0.00 | 16.32±0.00 |
| GDAS [8] | 8.7 | 89.89±0.08 | 93.61±0.09 | 71.34±0.04 | 70.70±0.30 | 41.59±1.33 | 41.71±0.98 |
| SNAS [28] | - | 90.10±1.04 | 92.77±0.83 | 69.69±2.39 | 69.34±1.98 | 42.84±1.79 | 43.16±2.64 |
| DSNAS [16] | - | 89.66±0.29 | 93.08±0.13 | 30.87±16.40 | 31.01±16.38 | 40.61±0.09 | 41.07±0.09 |
| PC-DARTS [29] | - | 89.96±0.15 | 93.41±0.30 | 67.12±0.39 | 67.48±0.89 | 40.83±0.08 | 41.31±0.22 |
| iDARTS [31] | - | 89.86±0.60 | 93.58±0.32 | 70.57±0.24 | 70.83±0.48 | 40.38±0.59 | 40.89±0.68 |
| DARTS- [5] | 3.2 | 91.03±0.44 | 93.80±0.40 | 71.36±1.51 | 71.53±1.51 | 44.87±1.46 | 45.12±0.82 |
| $\beta$-DARTS | 3.2 | **91.55±0.00** | **94.36±0.00** | **73.49±0.00** | **73.51±0.00** | **46.37±0.00** | **46.34±0.00** |
| optimal | - | 91.61 | 94.37 | 73.49 | 73.51 | 46.77 | 47.31 |

- Search on NAS-Bench-201 CIFAR-10 dataset, Adapt on NAS-Bench-201 CIFAR-10, NAS-Bench-201 CIFAR-100, NAS-Bench-201 ImageNet16-120

Table 1. Performance comparison on NAS-Bench-201 benchmark [9]. Note that $\beta$-DARTS only searches on CIFAR-10 dataset, but can robustly achieve new SOTA on CIFAR-10, CIFAR-100 and ImageNet16-120. Averaged on 4 independent runs of searching.

| Methods | Cost (hours) | CIFAR-10 valid | CIFAR-10 test | CIFAR-100 valid | CIFAR-100 test | ImageNet16-120 valid | ImageNet16-120 test |
|---|---|---|---|---|---|---|---|
| DARTS(1st) [21] | 3.2 | 39.77±0.00 | 54.30±0.00 | 15.03±0.00 | 15.61±0.00 | 16.43±0.00 | 16.32±0.00 |
| DARTS(2nd) [21] | 10.2 | 39.77±0.00 | 54.30±0.00 | 15.03±0.00 | 15.61±0.00 | 16.43±0.00 | 16.32±0.00 |
| GDAS [8] | 8.7 | 89.89±0.08 | 93.61±0.09 | 71.34±0.04 | 70.70±0.30 | 41.59±1.33 | 41.71±0.98 |
| SNAS [28] | - | 90.10±1.04 | 92.77±0.83 | 69.69±2.39 | 69.34±1.98 | 42.84±1.79 | 43.16±2.64 |
| DSNAS [16] | - | 89.66±0.29 | 93.08±0.13 | 30.87±16.40 | 31.01±16.38 | 40.61±0.09 | 41.07±0.09 |
| PC-DARTS [29] | - | 89.96±0.15 | 93.41±0.30 | 67.12±0.39 | 67.48±0.89 | 40.83±0.08 | 41.31±0.22 |
| iDARTS [31] | - | 89.86±0.60 | 93.58±0.32 | 70.57±0.24 | 70.83±0.48 | 40.38±0.59 | 40.89±0.68 |
| DARTS- [5] | 3.2 | 91.03±0.44 | 93.80±0.40 | 71.36±1.51 | 71.53±1.51 | 44.87±1.46 | 45.12±0.82 |
| $\beta$-DARTS | 3.2 | **91.55±0.00** | **94.36±0.00** | **73.49±0.00** | **73.51±0.00** | **46.37±0.00** | **46.34±0.00** |
| optimal | - | 91.61 | 94.37 | 73.49 | 73.51 | 46.77 | 47.31 |

# 5. Experimental Results



- ++ : Search on CIFAR-100, test on CIFAR-10 and CIFAR-100
- + : Search on CIFAR-10, test on CIFAR-10 and CIFAR-100
- Search on ImageNet / Cross Domain(CIFAR-10 and part of ImageNet) / CIFAR-10 or CIFAR-100

| Method | GPU (Days) | CIFAR-10 | | CIFAR-100 | |
|---|---|---|---|---|---|
| | | Params(M) | Acc(%) | Params(M) | Acc(%) |
| NASNet-A [33] | 2000 | 3.3 | 97.35 | 3.3 | 83.18 |
| DARTS(1st) [21] | 0.4 | 3.4 | 97.00±0.14 | 3.4 | 82.46 |
| DARTS(2nd) [21] | 1 | 3.3 | 97.24±0.09 | - | - |
| SNAS [28] | 1.5 | 2.8 | 97.15±0.02 | 2.8 | 82.45 |
| GDAS [8] | 0.2 | 3.4 | 97.07 | 3.4 | 81.62 |
| P-DARTS [4] | 0.3 | 3.4 | 97.50 | 3.6 | 82.51 |
| PC-DARTS [29] | 0.1 | 3.6 | 97.43±0.07 | 3.6 | 83.10 |
| P-DARTS [4] | 0.3 | 3.3±0.21 | 97.19±0.14 | - | - |
| R-DARTS(L2) [1] | 1.6 | - | 97.05±0.21 | - | 81.99±0.26 |
| SDARTS-ADV [3] | 1.3 | 3.3 | 97.39±0.02 | - | - |
| DOTS [12] | 0.3 | 3.5 | 97.51±0.06 | 4.1 | 83.52±0.13 |
| DARTS+PT [27] | 0.8 | 3.0 | 97.39±0.08 | - | - |
| DARTS- [5] | 0.4 | 3.5±0.13 | 97.41±0.08 | 3.4 | 82.49±0.25 |
| $\beta$-DARTS‡ | 0.4 | 3.78±0.08 | 97.49±0.07 | 3.83±0.08 | 83.48±0.03 |
| $\beta$-DARTS† | 0.4 | 3.75±0.15 | 97.47±0.08 | 3.80±0.15 | 83.76±0.22 |

| Method | GPU (Days) | Params (M) | FLOPs (M) | Top1 (%) | Top5 (%) |
|---|---|---|---|---|---|
| MnasNet-92*(Img.) [26] | 1667 | 4.4 | 388 | 74.8 | 92.0 |
| FairDARTS*(Img.) [6] | 3 | 4.3 | 440 | 75.6 | 92.6 |
| PC-DARTS(Img.) [29] | 3.8 | 5.3 | 597 | 75.8 | 92.7 |
| DOTS(Img.) [12] | 1.3 | 5.3 | 596 | 76.0 | 92.8 |
| DARTS-*(Img.) [5] | 4.5 | 4.9 | 467 | 76.2 | 93.0 |
| AdaptNAS-S(CD.) [19] | 1.8 | 5.0 | 552 | 74.7 | 92.2 |
| AdaptNAS-C(CD.) [19] | 2.0 | 5.3 | 583 | 75.8 | 92.6 |
| AmoebaNet-C(C10) [25] | 3150 | 6.4 | 570 | 75.7 | 92.4 |
| SNAS(C10) [28] | 1.5 | 4.3 | 522 | 72.7 | 90.8 |
| P-DARTS(C100) [4] | 0.3 | 5.1 | 577 | 75.3 | 92.5 |
| SDARTS-ADV(C10) [3] | 1.3 | 5.4 | 594 | 74.8 | 92.2 |
| DOTS(C10) [12] | 0.3 | 5.2 | 581 | 75.7 | 92.6 |
| DARTS+PT(C10) [27] | 0.8 | 4.6 | - | 74.5 | 92.0 |
| $\beta$-DARTS(C100) | 0.4 | 5.4 | 597 | 75.8 | 92.9 |
| $\beta$-DARTS(C10) | 0.4 | 5.5 | 609 | 76.1 | 93.0 |

# 5. Experimental Results

- Search on Different Weighting Scheme (Increase)

Table 3. Influence of different weighting schemes on $\beta$-DARTS.

| Weighting Scheme | CIFAR-10 valid | CIFAR-10 test |
|---|---|---|
| 0-15/25/50/100 | 91.21/91.55/91.55/91.55 | 93.83/94.36/94.36/94.36 |
| 5/10/15/25 | 84.96/90.59/91.55/90.59 | 88.02/93.31/94.36/93.31 |
| 25-15/10/5/0 | 90.59/87.30/73.58/39.77 | 93.31/90.65/76.88/54.30 |

# 5. Experimental Results

- Comparison of different weighting schemes

Table 4. The results of different Beta regularization loss with different weighting schemes on NAS-Bench-201 benchmark. Note that we only search on CIFAR-10 dataset, and perform 2 runs of searching under different random seeds.

| Methods | Weighting Scheme | CIFAR-10 | | CIFAR-100 | | ImageNet16-120 | |
|---|---|---|---|---|---|---|---|
| | | valid | test | valid | test | valid | test |
| DARTS(1st) [21] | 3.2 | 39.77±0.00 | 54.30±0.00 | 15.03±0.00 | 15.61±0.00 | 16.43±0.00 | 16.32±0.00 |
| Beta-Global | 0-25 | 91.55/91.55 | 94.36/94.36 | 73.49/73.49 | 73.51/73.51 | 46.37/46.37 | 46.34/46.34 |
| Beta-Global | 0-50 | 91.55/91.55 | 94.36/94.36 | 73.49/73.49 | 73.51/73.51 | 46.37/46.37 | 46.34/46.34 |
| Beta-Global | 0-75 | 91.55/91.55 | 94.36/94.36 | 73.49/73.49 | 73.51/73.51 | 46.37/46.37 | 46.34/46.34 |
| Beta-Global | 0-100 | 91.21/91.55 | 93.83/94.36 | 71.60/73.49 | 71.88/73.51 | 45.75/46.37 | 44.65/46.34 |
| Beta-Zero | 0-25 | 91.21/90.97 | 93.83/93.91 | 71.60/70.41 | 71.88/70.78 | 45.75/43.77 | 44.65/44.78 |
| Beta-Zero | 0-50 | 91.55/91.21 | 94.36/93.83 | 73.49/71.60 | 73.51/71.88 | 46.37/45.74 | 46.34/44.65 |
| Beta-Zero | 0-75 | 91.61/91.05 | 94.37/93.66 | 72.75/71.02 | 73.22/71.38 | 45.56/45.23 | 46.71/44.70 |
| Beta-Zero | 0-100 | 91.21/91.21 | 93.83/93.83 | 71.60/71.60 | 71.88/71.88 | 45.75/45.75 | 44.65/44.65 |

# 4. Conclusion

1. Change the Loss function
   - Examples from Beta-Decay DARTS paper

$$\mathcal{L}_{Beta-Global} = \text{smoothmax}\left( \alpha_1^1, \cdots, \alpha_{|\mathcal{O}|}^L \right) = \log\left(\sum_{l=1}^{L}\sum_{k=1}^{|\mathcal{O}|} e^{\alpha_k^l}\right)$$

$$\mathcal{L}_{Beta-Zero} = \text{smoothmax}\left( 0, \alpha_k^l \right) = -\log\left(1 + e^{-\alpha_k^l}\right)$$

2. Use different Weighting Scheme
3. Make $\theta$ as learnable parameter

# References


- **DARTS: Differentiable Architecture Search** https://arxiv.org/abs/1806.09055
- **β-DARTS: Beta-Decay Regularization for Differentiable Architecture Search** https://arxiv.org/abs/2203.01665
- **Theory-Inspired Path-Regularized Differential Network Architecture Search** https://arxiv.org/abs/2006.16537

# Thank you

## Q & A