



Small Temperature is All You Need for Differentiable Architecture Search

Jiuling Zhang^{1,2} and Zhiming Ding^{1,2(✉)}

¹ University of Chinese Academy of Sciences, Beijing 100049, China
zhangjiuling19@mails.ucas.ac.cn, zhiming@iscas.ac.cn

² Institute of Software, Chinese Academy of Sciences, Beijing 100190, China

Abstract. Differentiable architecture search (DARTS) yields highly efficient gradient-based neural architecture search (NAS) by relaxing the discrete operation selection to optimize continuous architecture parameters that maps NAS from the discrete optimization to a continuous problem. DARTS then remaps the relaxed supernet back to the discrete space by one-off post-search pruning to obtain the final architecture (finalnet). Some emerging works argue that this remap is inherently prone to mismatch the network between training and evaluation which leads to performance discrepancy and even model collapse in extreme cases. We propose to close the gap between the relaxed supernet in training and the pruned finalnet in evaluation through utilizing small temperature to sparsify the continuous distribution in the training phase. To this end, we first formulate sparse-noisy softmax to get around gradient saturation. We then propose an exponential temperature schedule to better control the outbound distribution and elaborate an entropy-based adaptive scheme to finally achieve the enhancement. We conduct extensive experiments to verify the efficiency and efficacy of our method.

Keywords: Deep learning architecture • Neural architecture search

1 Introduction

DARTS abstracts the search space as a cell-based directed acyclic graph composed by V nodes $H = \{h_1, h_2, \dots, h_V\}$ and compound edges $C = \{c_{1,2}, \dots, c_{V-1,V}\}$. Every node represents feature maps and each edge subsumes all operation candidates to express the transformations between nodes. Compound edge $c_{u,v}$ connects node u to v and associates three attributes: candidate operation set $O_c = \{o_c^1, o_c^2, \dots, o_c^M\}$, corresponding operation parameter set $A_c = \{a_c^1, a_c^2, \dots, a_c^M\}$, probability distribution of the parameters $\beta_c = \text{softmax}(A_c)$. Every intermediate node is densely-connected to all its predecessor through an edge $h_v = c_{u,v}(h_u)$ and weighted product sum $c_{u,v}(h_u) = \langle \beta_c, O_c(h_u) \rangle$ for $u < v$. Generally, a unified set of operation candidates $O = \{o^1, o^2, \dots, o^M\}$ is defined for all edges in the space. The network that encodes all architectural candidates is termed supernet. In the training phase, DARTS first divides data into training

and validation sets and then formulates a bilevel objective depicted in Eq. (1) to alternately optimize architecture parameter a and operation weight ω on the validation set and training set respectively. We refer to [13] for more details of DARTS.

$$\min_a L_{val}(\omega^*(a), a) \quad \text{s.t. } \omega^* = \arg \min_{\omega} L_{train}(\omega, a) \quad (1)$$

Henceforth, we abbreviate operation weight ω as weights and architecture parameter $a \in A$ as parameters. After the training phase, DARTS selects the operation associated with the largest entry (probability component after normalized by softmax) through a post-search pruning depicted in Eq. (2) to obtain the final architecture (finalnet) for evaluation. In sum, the supernet is trained in a multi-path manner while the finalnet is evaluated in a single-path manner [14].

$$o_c^i \in O \text{ for } c \in C \text{ where } i = \arg \max_{c \in \{1, \dots, M\}} \beta_c, \quad \beta_c = \text{softmax}(A_c) \quad (2)$$

where the operation o_c^i is selected on edge c due to the largest entry β_c^i in β_c . However, the parameter-value-based post-search pruning depicted in Eq. (2) inevitably risks mismatching the architecture between supernet and finalnet which ultimately leads to performance discrepancy between training and evaluation in DARTS. Gradient-based NAS methods are deemed to favor architectures that are easier to be trained [15]. More skip connections can obviously help the network to converge faster which is considered as an unfair advantage in an exclusive competition [7, 8]. This unfairness leads to an abnormal preference for the skip connection in some cases during optimization. The architecture mismatch exacerbates this issue since the one-off post-pruning generally discards all operations except the dominant one ($O \setminus o^{\arg \max_i(\beta)}$). In extreme cases, the pruning removes all operations except the skip connection which causes collapse or the catastrophic failure of DARTS [1, 5, 10, 12].

Both Fair-DARTS [8] and GAEA [11] pointed out that the sparse parameters are crucial to alleviating the architecture mismatch. In particular, GAEA emphasized that obtaining sparse final architecture parameters is critical for good performance, both for the mixture relaxation, where it alleviates the effect of overfitting, and for the stochastic relaxation, where it reduces noise when sampling architectures. In DARTS, *the sparse parameters refer to the sparse distribution β (low entropy) after the parameters normalized by softmax*, i.e. $\beta = \text{softmax}(A)$. Sparse parameters intrinsically reduce the gap between multi-path supernet and single-path finalnet as shown in Fig. 1. [17] proposed to combine single-path and multi-path space by a Sparse Group Lasso constraint but **impose non-trivial additional time consumption** due to it's harder to converge. Temperature coefficient is widely used to control the **sparseness** and **smoothness** of the (gumbel) softmax output. In this paper, we propose to achieve sparse training straightforwardly by employing a small temperature on softmax. Our contributions can be summarized:

- We propose sparse-noisy softmax (sn-softmax) to alleviate the gradient saturation which causes **premature** convergence of the training of parameters while utilizing small temperature in the training phase;

- We propose exponential temperature schedule (ETS) and entropy-based adaptive scheme to maintain and better control the sparsity of β that inherently narrows the gap between the relaxed supernet in training and the pruned finalnet in evaluation;
- We carry out extensive evaluations on multiple spaces and datasets and conduct further **ablations** to show the effect of different hyperparameter choices.

2 Methodology

Utilizing small temperature to sparsify β in training is non-trivial for DARTS because the gradient saturation will **impede** the propagation when softmax converges. Likewise, the operation weights will not be updated either when the operation output are weighted by a zero entry in β . If that happens in the middle training of DARTS, supernet converges prematurely. In this section, we start with formulating sparse-noisy softmax to alleviate the gradient saturation when the outbound β converges. After that, we propose an exponential temperature schedule (ETS) to better control the temperature t to smooth the swing of the outbound β in training. We provide an entropy-based **dynamic decay** (EDD) to finally realize a flexible and robust enhancement for DARTS.

2.1 Sparse-Noisy Softmax

Softmax normalizes the input vector $A = \{a^1, \dots, a^M\}$ to a probability distribution $\beta = \{\beta^1, \dots, \beta^M\}$ depicted in Eq. (3).

$$\text{softmax}\left(\frac{A}{t}\right) = \beta_t \quad \text{where } \beta_t^i = \frac{\exp(a^i/t)}{\sum_{j=1}^M \exp(a^j/t)} \quad (3)$$

where t is the temperature coefficient and M is the total entries of the input of softmax in DARTS. The smaller the temperature t , the sharper the outbound β is and the closer the distribution compared to post-search one-hot argmax pruning in Eq. (2). The derivative of softmax can be gotten through:

$$\frac{\partial \beta_t^i}{\partial a^i} = \begin{cases} \frac{\beta_t^i(1-\beta_t^i)}{t} & \text{for } i = j \\ \frac{-\beta_t^i \beta_t^j}{t} & \text{for } i \neq j \end{cases} \quad (4)$$

by which we can get the Jacobian matrix as:

$$\frac{\partial \beta_t}{\partial A} = \frac{1}{t} \begin{bmatrix} \beta_t^1 - [\beta_t^1]^2 & -\beta_t^1 \beta_t^2 & -\beta_t^1 \beta_t^3 & \cdots & -\beta_t^1 \beta_t^M \\ -\beta_t^2 \beta_t^1 & \beta_t^2 - [\beta_t^2]^2 & -\beta_t^2 \beta_t^3 & \cdots & -\beta_t^2 \beta_t^M \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ -\beta_t^M \beta_t^1 & -\beta_t^M \beta_t^2 & -\beta_t^M \beta_t^3 & \cdots & \beta_t^M - [\beta_t^M]^2 \end{bmatrix} \quad (5)$$

Let \mathcal{A}_t denotes the rightmost matrix in Eq. (5) and \mathcal{A}_t/t is the Jacobian with the temperature t . Let l indicates the loss actually used, then the backpropagation to a can be written as:

$$\frac{\partial \beta_t}{\partial A} \frac{\partial l}{\partial \beta_t} = \frac{\mathcal{A}_t}{t} \frac{\partial l}{\partial \beta_t} \quad (6)$$

By leveraging a smaller temperature t to get a sparser β_t , only one single entry in β_t approaches one ($\beta_t^i \rightarrow 1$) and the others are thereby close to zero ($\beta_t^{j \neq i} \rightarrow 0$), Jacobian matrix is then overall trapped in zero $\lim_{\substack{\beta_t^i \rightarrow 1 \\ \beta_t^{j \neq i} \rightarrow 0}} \partial \beta_t / \partial A = 0$ and the gra-

dient saturates. When softmax is saturated, back propagation depicted in Eq. (6) stops propagating gradients through softmax to parameters thus the parameters stop updating. To deal with this problem, ideally, we hope that our softmax outputs a sparse β in feedforward as the middle pane of Fig. 1, but the back-propagation is smoother (not saturated). To this end, we propose **sn-softmax** (sn-softmax), summarized in Algorithm 1, to approximate the ideal case by *combining different temperature for feedforward and backpropagation* respectively. Our goal is to keep the Jacobian matrix not zero even softmax converges

$$\lim_{\substack{\beta_t^i \rightarrow 1 \\ \beta_t^{j \neq i} \rightarrow 0}} \partial \beta_t / \partial A \neq 0, \text{ which is different to the previous research that injects noise}$$

to postpone convergence [2].

Set the forward-pass temperature as t , we formulate the Jacobian matrix of sn-softmax as:

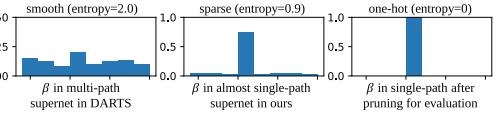


Fig. 1. Multi-path, sparse and single-path.

$$\frac{\partial \beta_t}{\partial A} = \frac{\mathcal{A}_t}{t} + \frac{\mathcal{A}_{st}}{st} \quad \text{for } s > 1 \quad (7)$$

where \mathcal{A}_{st}/st is the Jacobian of β_{st} . The backpropagation from the loss l w.r.t A can be written as:

$$\frac{\partial \beta_t}{\partial A} \frac{\partial l}{\partial \beta_t} = \left(\frac{\mathcal{A}_t}{t} + \frac{\mathcal{A}_{st}}{st} \right) \frac{\partial l}{\partial \beta_t} \quad (8)$$

Since the equal sign can not be guaranteed:

$$\text{sign}\left(\frac{\mathcal{A}_t}{t} \frac{\partial l}{\partial \beta_t}\right) \neq \text{sign}\left(\frac{\mathcal{A}_{st}}{st} \frac{\partial l}{\partial \beta_t}\right) \quad \text{for } s > 1 \quad (9)$$

Sn-softmax is equivalent to adding noise into the backpropagation by which the Jacobian matrix is not absolute zero after the forward term \mathcal{A}_t/t converges. s acts as a scaling factor in Eq. (8) to tune the noise intensity. A smaller s brings stronger gradient noise, but is easier to saturate either as the training progresses. We can also keep st as a constant and determine s accordingly and dynamically and thereby get out of the saturation in the whole training phase.

In practice, sn-softmax leverages a small temperature value t for feedforward to obtain a sparse output β_t while setting $s \gg 1$ (generally $s > 50$) thus get $1/st \ll 1/t$. When the feedforward does not converge ($\mathcal{A}_t \neq 0$), \mathcal{A}_t/t plays a leading role in Eq. (7) because of $1/st \ll 1/t$ and $\mathcal{A}_{st}/st \ll \mathcal{A}_t/t$. After the feedforward convergence, the gradient saturates ($\mathcal{A}_t = 0$, $\mathcal{A}_t/t = 0$). Since the scaling factor $s \gg 1$, β_{st} is much smoother than β_t which makes \mathcal{A}_{st} less convergent ($\mathcal{A}_{st} \neq 0$). As a result, \mathcal{A}_{st}/st supersedes the first term \mathcal{A}_t/t in the Jacobian matrix and comes into play in the backpropagation of sn-softmax depicted in

Eq.(8). This way, when sn-softmax does not converge, the normal term repels the noise term and dominate the update direction due to $\mathcal{A}_{st}/st \ll \mathcal{A}_t/t$ but allows sn-softmax to have a chance to escape the premature convergence after it's saturated. We schematically visualize the backpropagation dynamic of sn-softmax in terms of the gradient norm on different values of s in Fig. 2A.

Sn-softmax only needs *feed-forward and backpropagation once like softmax* to get $\partial l/\partial \beta_t$. Then it calculates the smoother Jacobian based on temperature st through Eq. (5). The calculation of another Jacobian and the final addition depicted in Eq. (7) only takes negligible additional budget in the whole training.

Algorithm 1. Sparse-noisy (sn) softmax

Input: logits $A = \{a^1, \dots, a^M\}$, feedforward temperature t , scaling factor s .

Output: normalized distribution β_t

Feedforward: $\text{softmax}(\frac{A}{t}) = \beta_t$ for $\beta_t^i = \frac{\exp(a^i/t)}{\sum_j^M \exp(a^j/t)}$

Backpropagation: $\frac{\partial \beta_t}{\partial A} = \frac{1}{t} \mathcal{A}_t + \frac{1}{st} \mathcal{A}_{st}$ for $s > 1$

2.2 Exponential Temperature Schedule

In general, the temperature t is gradually decayed to drive β converge to a sparse solution. As shown in Eq. (3), t is in the exponential term and non-linearly affects the transformation of $a \rightarrow \beta$ in softmax. Linearly scheduling t swings β nonlinearly and results in that the early temperature decay has little effect on the β while decay in the later stage of training has too much impact on the sparsity and causes the training converges precipitately (see LTS in Fig. 2B and LPCD-DARTS in Fig. 4A). Therefore, the naive linear temperature schedule is inappropriate for the training of DARTS. In this section, we focus on temperature t so let $t_a^{\text{exp}} = \exp(a/t)$ refers to the exponential function with e as the base and t, a as the exponent where we add the superscript “exp” in t_a^{exp} to emphasize the value after the exponential transformation $\exp(a/t)$. We then apply variable substitution in softmax as:

$$\beta^i = \frac{t_{a^i}^{\text{exp}}}{\sum_{j=1}^M t_{a^j}^{\text{exp}}} \quad \text{where } t_{a^*}^{\text{exp}} = \exp\left(\frac{a^*}{t}\right) \quad (10)$$

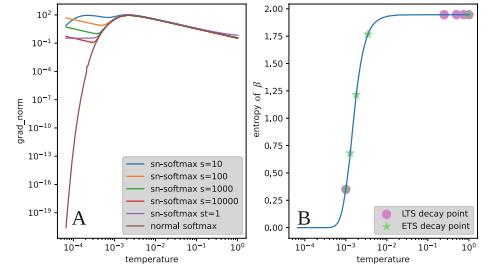


Fig. 2. Visualize the empirical analysis of sn-softmax and ETS. Tests are based on the order of magnitude of a initialized in DARTS. (A). By setting $s \gg 1$, gradients of sn-softmax is consistent with normal softmax before saturated. After saturated, we can get different backpropagation dynamics by setting different s . (B). For equidistant decay, ETS yields much smoother swings than the linear counterpart in terms of the entropy of outbound distribution β .

where the transformation from t_a^{\exp} to β^i is linear. We can then linearly schedule t_a^{\exp} instead of t to better control the sparsity of β . We call this design of scheduling t_a^{\exp} after the exponent substitution as exponential temperature schedule (ETS).

To calculate t_a^{\exp} , we first approximate a by the order of its expectation $E(a)$, $a \in A$ and get the temperature $t_{E(a)}^{\exp}$. From now on, we omit $E(a)$ in the subscript of $t_{E(a)}^{\exp}$ since it's fixed in this analysis. We then specify the initial temperature t_0 and the decay target t_N and get $t_0^{\exp} = e^{E(a)/t_0}$ and $t_N^{\exp} = e^{E(a)/t_N}$ for the start and target in the exponential space (after $\exp(a/t)$ transformation) respectively. Since t act as the denominator in the exponential term $\exp(a/t)$ of softmax, the temperature decay from t_0 to t_N corresponds to an ascent in the exponential space from t_0^{\exp} to t_N^{\exp} . For the equidistant temperature decay, the decay strength d^{\exp} indicates the variation amplitude within $[t_0^{\exp}, t_N^{\exp}]$ which can be calculated statically via $d^{\exp} = \frac{t_N^{\exp} - t_0^{\exp}}{N}$ where N is the number of decay points. All decay points together with t_0^{\exp} form a list L^{\exp} as

$$L^{\exp} = [t_n^{\exp} | t_n^{\exp} = t_0^{\exp} + nd^{\exp}, d^{\exp} = \frac{t_N^{\exp} - t_0^{\exp}}{N}, n = 0, 1, \dots, N] \quad (11)$$

where $t_0^{\exp} = t_0^{\exp}$, $t_N^{\exp} = t_N^{\exp}$. The temperature value corresponding to each decay point in L^{\exp} can be inversely solved by Eq. (12) where $t_n > 0$ when $t_n^{\exp} > 1$.

$$t_n = \frac{E(a)}{\ln(t_n^{\exp})} \text{ for } t_n^{\exp} \in L^{\exp}, n = 0, 1, \dots, N \quad (12)$$

We provide a quantitative example in Fig. 3 to show the clear usage of ETS and more details on the calculation are given below. Firstly, as in A are initialized by sampling from $N(0, 1)$ and scaling the samples by $1e-3$ in DARTS. When $a \leq 0$, the variation of t has little effect on the value of t_a^{\exp} , so we only consider $E(a)$ for $a > 0$. The scaled expectation of a can be get through $E(a) = (1e-3)E(N(0, 1)) = (1e-3)\sqrt{2}/\pi/2 \approx 4e-4$ for $a > 0$. By specifying the initial temperature $t_0 = 1$ and the decay target $t_N = 1e-3$, we can then get $t_0^{\exp} = \exp(E(a)/t_0) = e^{4e-4} \approx 1$ and $t_N^{\exp} = \exp(E(a)/t_N) = e^{4e-1} \approx 1.492$ in the exponential space respectively. The temperature decay from t_0 to t_N ($1 \rightarrow 1e-3$) thus corresponds to an ascent from t_0^{\exp} to t_N^{\exp} ($1 \rightarrow 1.492$) in the exponential space. For the equidistant temperature decay, we preset $N = 4$ and $t_4^{\exp} = t_N^{\exp} = 1.492$, then $[1, 1.492]$ is equidistantly divided into 4 segments by the other 3 decay points. We get the decay strength accordingly as $d^{\exp} = \frac{1.492-1}{4} = 0.123$ and determine the remaining three decay points as $1.123, 1.246, 1.369$ and form $L^{\exp} = [1, 1.123, 1.246, 1.369, 1.492]$ by Eq. (11). After that, we can get the temperatures w.r.t $t_{1.123}^{\exp}$ as $t_{1.123} = 4e-4/\ln(1.123) \approx 0.00345$ by Eq. (12). Similarly,

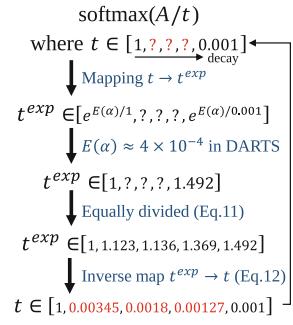


Fig. 3. Diagram of the quantitative example.

the temperature of the other 3 decay points are $t_{1.246} = 0.0018$, $t_{1.369} = 0.00127$, $t_{1.492} = 0.001$ from which we can clearly find that the temperature decay scheduled equidistantly in the exponential space $(1, 0.00345, 0.0018, 0.00127, 0.001)$ leads to a radical difference from the typical linear temperature scheduling LTS $(1, 0.75, 0.5, 0.25, 0.001)$. We visualize the effect of the differences between ETS and LTS in terms of the entropy of β in a more sensible way in Fig. 2B.

For linear temperature schedule, we always need to preset t_0 , t_N , N and calculate the decay strength statically to *prevent temperature from decreasing to less than 0*. In contrast, Eq. (12) is always greater than 0 for $t^{\exp} > 1$ which yields added flexibility for the design of ETS-based training scheme. We can calculate d^{\exp} and L^{\exp} after we specified $[t_0^{\exp}, t_N^{\exp}]$ and N . Alternatively, we can first determine t_0^{\exp} , d^{\exp} and get t_N^{\exp} or N by $t_N^{\exp} = t_0^{\exp} + Nd^{\exp}$, so that we can still build L^{\exp} through Eq. (11).

2.3 Entropy-Based Adaptive Scheme

If operation output $o_c^i(h)$ on edge c is zero-weighted by as $\beta_c^i o_c^i(h)$ for $\beta_c^i \in \beta_c$ and $\beta_c^i = 0$, both forward and backward paths of the operation o_c^i are blocked so that the operation weights ω within o_c^i cannot obtain effective gradients. Employing a small temperature at the beginning of training will lead to exaggerated swings of β and finally biases the search result. For an appropriate scheme, presetting fixed t_0 , t_N , d^{\exp} and updating $E(x)$ and L^{\exp} every epoch is ***cumbersome*** and ***inflexible***. Empirically, we also observe that the same architecture under different initialization exhibits various optimization dynamics. Some lead to strong convergence that the parameters of supernet converge quickly under mild temperature decay. In other cases, the search is ***indecisive*** among two or three operations. In sum, a ***fixed decay strength d^{\exp} is not robust in practice***.

Being equipped with the additional design freedom supported by ETS, we further propose **entropy-based dynamic decay (EDD)** to adaptively determine d^{\exp} in terms of both the sparsity of β and the current epoch k by which *we need only to tune one single hyperparameter λ to control the training process*. We first define the expectation of entropies of β s as $E(H(\beta_c))$, $c \in C$ in Eq. (13) to represent the sparsity of β s over all compound edges in the cell space.

$$E(H(\beta_c)) = \frac{-\sum_c^{|C|} \sum_i^M \beta_c^i \log \beta_c^i}{|C|} \quad \text{where } \beta_c = \text{softmax}(A_c), c \in C \quad (13)$$

Algorithm 2. EDD sparse training scheme

Input: s for sn-softmax, λ for EDD, $t_0 = 1$ (mostly)
Get the expectation of parameter $E(a)$
Get $t_0^{\exp} = e^{E(a)/t_0}$
while training epoch k **do**
 Training a and ω by Algorithm 1 in [13] with $t^{(k)}$, s .
 Update $[d^{\exp}]^{(k)}$ by Eq.(14).
 Update $t^{(k)}$ via $[d^{\exp}]^{(k)}$ by Eq.(15).
end while

Table 1. Evaluation results on NB201&C10. “clip” refers to the gradient clip on a to alleviate the effect of the noisy none (zero) operation which has been identified as unsearchable in DARTS [13].

Search dataset	Method	Search (seconds)	CIFAR-10		CIFAR-100		ImageNet-16-120	
			validation	test	validation	test	validation	test
			39.77±0.00	54.30±0.00	15.03±0.00	15.61±0.00	16.43±0.00	16.32±0.00
C10	DARTS-V2 [13]	22323	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-V1 [13]	7253	39.77±0.00	54.30±0.00	15.03±0.00	15.61±0.00	16.43±0.00	16.32±0.00
	GDAAS [9]	19720	90.00±0.21	93.51±0.13	71.14±0.27	70.61±0.26	41.70±1.26	41.84±0.90
	ETS-GDAS	19755	90.57±0.35	93.75±0.22	71.45±0.64	71.39±0.66	42.96±0.93	42.92±0.87
	DrNAS [4]	7544	90.15±0.10	93.74±0.03	70.82±0.27	71.07±0.08	40.76±0.05	41.37±0.17
	GAEA-Bilevel [11]	8280	82.80±1.01	84.64±1.00	55.24±1.47	55.35±1.72	27.72±1.35	26.40±0.85
	GAEA-ERM [11]	14464	84.59±0.00 (91.50±0.06)	86.59±0.00 (94.34±0.06)	58.12±0.00 (73.12±0.26)	58.43±0.00 (73.11±0.06)	29.54±0.00 (45.71±0.28)	28.19±0.00 (46.38±0.18)
	GibbsNAS [19]	-	90.02±0.60	92.72±0.60	68.88±1.43	69.20±1.44	42.31±1.69	42.08±1.95
	DARTS- [7]	-	91.03±0.44	93.80±0.40	71.36±1.51	71.53±1.51	44.87±1.46	45.12±0.82
	SurgeNAS [14]	-	90.2	93.7	71.2	71.6	44.5	45.2
EDD-DARTS	7392	90.95±0.44	93.80±0.33	71.44±1.21	71.42±1.25	45.14±0.78	45.12±0.56	
	EDD-DARTS (clip)	7400	91.12±0.19	94.05±0.11	72.59±0.62	72.43±0.64	45.89±0.41	45.80±0.29

Table 2. Experimental results on S1~S4&C100 and S1~S4&SVHN.

Dataset	Space	DARTS	PC-DARTS	DARTS-ES	R-DARTS (DP/L2)	SDARTS (RS/ADV)	DARTS+PT (unfixed/fixed)	EDD-DARTS
C100	S1	29.46	24.69	28.37	25.93/24.25	23.51/22.33	24.48/24.40	22.27
	S2	26.05	22.48	23.25	22.30/22.44	22.28/20.56	23.16/23.30	21.73
	S3	28.90	21.69	23.73	22.36/23.99	21.09/21.08	22.03/21.94	21.08
	S4	22.85	21.50	21.26	22.18/21.94	21.46/21.25	20.80/ 20.66	20.66
SVHN	S1	4.58	2.47	2.72	2.55/4.79	2.35/2.29	2.62/2.39	2.23
	S2	3.53	2.42	2.60	2.52/2.51	2.39/2.35	2.53/2.32	2.30
	S3	3.41	2.41	2.50	2.49/2.48	2.36/2.40	2.42/ 2.32	2.32
	S4	3.05	2.43	2.51	2.61/2.50	2.46/2.42	2.42/ 2.39	2.45

where M operation candidates and $|C|$ compound edges in search space. β_c^i denotes the i th entry in the β_c that used to weight the feature maps from operation o^i on edge c . The design principles can be summarized as follows:

- d^{exp} is stronger when $E(H(\beta))$ is higher (smoother β in softmax);
- Gradually increase d^{exp} w.r.t the training epoch k .

Based on the above design principles, We update d^{exp} according to Eq. (14) for epoch k .

$$[d^{\text{exp}}]^{(k)} = \lambda(1 - \rho)E(H(\beta)) + \rho[d^{\text{exp}}]^{(k-1)} \quad (14)$$

where we set $[d^{\text{exp}}]^{(0)} = 0$ in practice and keep an exponentially moving average with a momentum $\rho \equiv 0.5$ to avoid oscillations. λ is the hyperparameter to determine the influence of $E(H(\beta))$ on d^{exp} . Since d^{exp} is adaptively decided through Eq. (14), L^{exp} cannot be calculated in advance. After k th update of d^{exp} , we can get temperature $t^{(k)}$ accordingly by Eq. (15)

$$t^{(k)} = \frac{E(a)}{\ln(t_0^{\text{exp}} + k[d^{\text{exp}}]^{(k)})} \quad (15)$$

results in that d^{exp} is proportional to the training epoch k . EDD is summarised in Algorithm 2 where λ is the only hyperparameter for the scheme.

3 Evaluations

We evaluate EDD, namely EDD-DARTS, on CIFAR-10 (C10), CIFAR-100 (C100), ImageNet-1k (IN-1k), SVHN and multiple spaces: NAS-BENCH-201 (NB201), DARTS space (DS), S1~S4 [1].

Evaluations on NB201: NB201 supports three datasets (C10, C100, ImageNet-16 [6]) and has a unified cell-based search space with 15,625 architectures. We refer their paper [10] for more details of the search space. All our baselines come from recent top venues. Experiments of DrNAS and GAEA are both based on the released codes. We provide extra results for GAEA-ERM in parentheses by excluding the none operation since it's particularly fragile for "none" in NB201. The experimental results shown in Table 1 demonstrates that the our enhancement effectively eliminates the performance collapse of DARTS. Remarkably, EDD-DARTS claims superior searching on both C10 and C100 on the standard space (include none) of NB201.

Evaluations on DS: We employ the same search recipe as on NB201. Our evaluation is based on the source code released by DrNAS. We keep the hyper-parameter settings unchanged except for replacing the cell genotypes. As [4, 11], we repeat the search 3 times under different seeds, evaluate each result independently and report the mean accuracies and standard deviations in Table 3.

Transfer to ImageNet-1k: As a common practice, we transfer the most prominent architecture on C10 to ImageNet-1k (IN1K) for additional performance evaluation. As shown in Table 3 and Table 4, the EDD itself is already very competitive, and the combination of the EDD with partial channel trick [18] can deliver a clear new art scores with the lowest budget on both C10 and ImageNet-1k.

Comparison with Regularization: We evaluate EDD on four specially designed search spaces S1~S4 by [1] which are particularly challenging for DARTS-based methods. Unregularized DARTS is always prone to make a wrong choice of non-parametric operations on these spaces. We refer to [1] for more details of S1~S4. We

Table 3. Search and evaluate on DS&C10.

Method	Error (%)	Params (M)	GPU days
PC-DARTS [18]	2.57±0.07	3.6	0.1
GAEA+PC-DARTS [11]	2.50±0.06	3.7	0.1
DARTS+PT [16]	2.61±0.08	3.0	0.8
SDARTS-RS+PT [16]	2.54±0.10	3.3	0.8
SGAS+PT [16]	2.56±0.10	3.9	0.29
DrNAS [4]	2.54±0.03	4.0	0.4
DARTS- [7]	2.59±0.08	3.5	0.4
GibbsNAS [19]	2.53±0.02	4.1	0.5
SparseNAS [17]	2.69±0.03	4.2	1
β -DARTS [20]	2.53±0.08	3.7	0.4
EDD-DARTS	2.52±0.10	3.6	0.4
EDD-PC-DARTS	2.47±0.06	4.2	0.1

Table 4. Transfer to evaluate on IN1K.

Method	Top-1	Top-5	Params (M)	GPU days
PC-DARTS [18]	25.1	7.8	5.3	0.1
GAEA+PC-DARTS [11]	24.3	7.3	5.6	0.1
GibbsNAS [19]	24.6	-	5.1	0.5
DrNAS [4]	24.2	7.3	5.2	3.9*
DARTS- [7]	24.8	7	4.9	4.5*
SparseNAS [17]	24.6	7.6	5.7	-
β -DARTS [20]	24.2	7.1	5.4	0.4
EDD-DARTS	24.6	7.4	5.0	0.4
EDD-PC-DARTS	24.0	7.2	5.6	0.1

* Search on Imagenet.

evaluate our method against two strong baselines Smooth DARTS [3] and DARTS+PT [16] on this benchmark on C10, C100 and SVHN respectively. As illustrated in Table 2, EDD-DARTS performs well under all datasets and search spaces and outperforms baselines in aggregate.

4 Further Experiments, Analyses and Conclusion

To understand what makes EDD effective, we conduct ablation studies along two axes, s and λ , on both NB201 and DS. We come up with and elaborate another ETS-based baseline to validate our claim of superior robustness of EDD.

Table 5. Ablate s and λ on NB201&C10.

Search dataset	Softmax	s	λ	CIFAR-10		CIFAR-100		ImageNet-16-120	
				validation	test	validation	test	validation	test
C10	Normal softmax	-	0.06	90.56±0.75	93.29±0.70	70.03±1.48	70.21±1.54	43.59±2.05	43.33±1.80
			0.12	90.74±0.31	93.60±0.29	70.46±0.84	70.45±0.96	44.17±0.92	44.34±1.23
			0.24	89.93±1.44	92.82±1.16	69.38±2.05	69.60±1.49	43.01±2.16	43.07±2.14
	sn-softmax	$st \equiv 1$	0.06	90.95±0.44	93.80±0.33	71.44±1.21	71.42±1.25	45.14±0.78	45.12±0.56
			0.12	90.57±0.58	93.38±0.52	70.06±1.46	69.86±1.42	43.87±1.92	43.89±2.33
			0.24	90.06±0.81	92.83±0.52	69.07±1.35	69.26±1.29	42.28±2.07	42.43±1.83
	$s = 100$	$st \equiv 1$	0.06	90.68±0.71	93.44±0.54	70.45±1.69	70.43±1.61	43.98±1.85	44.08±2.17
			0.12	91.02±0.30	93.75±0.31	71.36±0.96	71.33±1.21	44.73±0.64	44.95±0.62
			0.24	90.54±0.74	93.22±0.58	69.90±1.60	70.19±1.88	43.43±1.82	43.49±1.68

Ablations on NB201: We ablate the impact of three configurations of softmax and λ on NB201. The experimental results on C10 are shown in Table 5. For the searching on C10, sn-softmax is helpful where both the highest and second-highest accuracies in the experiments of EDD-DARTS come from the results equipped with sn-softmax ($st \equiv 1$ or $s = 100$).

Ablations on DS: We ablate EDD-DARTS further on DS and the results are shown in Table 6. We find that a larger λ brings stronger temperature decay, EDD-DARTS tends to

find higher capacity architectures some of which are tricky to be trained properly but sn-softmax alleviates this trend and ensures that EDD keeps delivering efficient results.

Table 6. Ablate s and λ on DS&C10.

Search dataset	λ	normal softmax	sn-softmax	$st \equiv 1$	$s = 100$
		Test Error (%)	Test Error (%)	Test Error (%)	Test Error (%)
C10	0.06	2.60 ± 0.10	2.55 ± 0.10	2.52 ± 0.10	
	0.12	2.63 ± 0.13	2.56 ± 0.10	2.54 ± 0.09	
	0.24	2.65 ± 0.13	2.59 ± 0.12	2.55 ± 0.10	

Robustness Validations: To validate the robustness of our adaptive scheme EDD, we elaborate another scheme i.e. periodic cyclic decay (PCD) which excluding EDD and determine L^{exp} statically before training as the additional tailored baseline. We first finetune the hyperparameters of both EDD and PCD on NB201&C10 and then transfer the settings and recipes exactly to all other search spaces and datasets. we finetune PCD even marginally surpass EDD in Table 7. On the contrary, by transferring the configuration of NB201&C10, EDD starkly outperforms PCD on most other cases especially on NB201&C100, DS&C10 and DS&IN-1k. This results underpin the robustness virtue of EDD over less flexible decay scheme PCD. Further investigation identify that PCD is more brittle than EDD to the warmup epoch and preseted decay strength on NB201 and DS respectively. We also note that simply increase these two values can recover the performance.

Table 7. Finetune (FT) on NB201&C10 and transfer to other datasets and search spaces to evaluate the robustness.

Search space	Search dataset	Evaluation dataset	PCD-DARTS		EDD-DARTS	
			validation	test	validation	test
NB201	C10 (FT)	C10	91.06±0.49	93.78±0.43	90.95±0.44	93.80±0.33
		C100	71.61±1.30	72.00±1.48	71.44±1.21	71.42±1.25
		IN-16	45.17±0.84	45.63±0.86	45.14±0.78	45.12±0.56
	C100	C10	90.04±1.35	92.97±1.78	90.27±0.81	93.44±0.49
		C100	69.97±1.82	70.02±2.00	70.46±1.52	70.57±1.40
		IN-16	42.34±2.16	42.79±2.01	42.43±1.80	42.88±1.81
			[Error (%)]		[Param (M)]	
DS	C10	C10	2.57±0.09	3.0	2.55±0.10	3.6
		IN-1k	26.5	4.2	24.6	5.1
				Error (%)		Error (%)
S1		C10		2.75		2.77
		C100		22.35		22.27
		SVHN		2.27		2.23
S2		C10		2.56		2.54
		C100		21.44		21.73
		SVHN		2.33		2.30
S3		C10		2.50		2.49
		C100		21.05		21.08
		SVHN		2.32		2.32
S4		C10		2.95		2.61
		C100		21.48		20.66
		SVHN		2.44		2.45

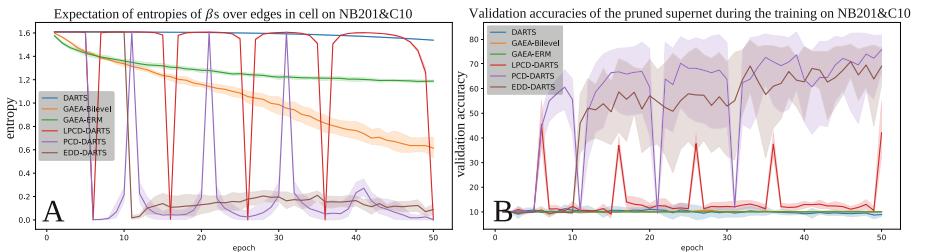


Fig. 4. (A). Trajectories of the expectation of entropies of β s during training on NB201&C10. (B). Discretized accuracies on validation set in training on NB201&C10.

Analyses: According to [11], the entropy of β and the discretized accuracy of the pruned finalnet on validation set are two main measurements for evaluating the impact of the sparsity on method. We illustrate the dynamics of the expectation of entropies of β s over edges in space during the training on NB201&C10 in Fig. 4A. in which the distribution entropy of EDD are much lower than other baselines. This validates our proposal of employing small temperature to sparsify β in training. To better characterize the effect of the sparse β on alleviating the mismatch, we also illustrate the corresponding discretized validation

accuracies during the training on NB201&C10 in Fig. 4B. Both DARTS and GEAE in Fig. 4B are basically stuck at the random discretized accuracies which is obviously due to the insufficient sparsity of parameters shown in Fig. 4A. In contrast, EDD-DARTS, shown by the brown line in Fig. 4B, maintains an appropriate sparsity of β and steadily improve the discretized accuracies throughout the training. We observe the similar phenomenon on DARTS space. The drift of $E(a)$ can be seen in Fig. 4A in which the three decay cycles of PCD exhibit slightly different dynamics of β for the same temperature sequence L^{\exp} . This is the downside of PCD since it determines the whole L^{\exp} beforehand and fixes it during training. In contrast, EDD adaptively gets t^{\exp} directly based on the sparsity depicted in Eq. (15), thereby finds the appropriate t timely to compensate that drift of the expectation.

5 Conclusion

In this paper, we focus on sparsifying the β via utilizing and scheduling small temperature in DARTS. We first propose sn-softmax to alleviate the gradient saturation of the premature convergence. Next, we propose ETS to better control the sparsity of β and we elaborate an entropy-based adaptive scheme EDD to finally deliver the effective enhancement in DARTS.

Acknowledgment. National Key R&D Program of China (2022YFF0503900).

References

- Arber Zela, T.E., Saikia, T., Marrakchi, Y., Brox, T., Hutter, F.: Understanding and robustifying differentiable architecture search. In: International Conference on Learning Representations, vol. 3, p. 7 (2020)
- Chen, B., Deng, W., Du, J.: Noisy softmax: Improving the generalization ability of DCNN via postponing the early softmax saturation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5372–5381 (2017)
- Chen, X., Hsieh, C.J.: Stabilizing differentiable architecture search via perturbation-based regularization. In: International Conference on Machine Learning, pp. 1554–1565. PMLR (2020)
- Chen, X., Wang, R., Cheng, M., Tang, X., Hsieh, C.J.: DrNAS: Dirichlet neural architecture search. In: International Conference on Learning Representations (2021). <https://openreview.net/forum?id=9FWas6YbmB3>
- Cheng, X., et al.: Hierarchical neural architecture search for deep stereo matching. Adv. Neural. Inf. Process. Syst. **33**, 22158–22169 (2020)
- Chrzaszcz, P., Loshchilov, I., Hutter, F.: A downsampled variant of imagenet as an alternative to the CIFAR datasets. arXiv preprint [arXiv:1707.08819](https://arxiv.org/abs/1707.08819) (2017)
- Chu, X., Wang, X., Zhang, B., Lu, S., Wei, X., Yan, J.: Darts-: Robustly stepping out of performance collapse without indicators. In: International Conference on Learning Representations (2021)
- Chu, X., Zhou, T., Zhang, B., Li, J.: Fair DARTS: eliminating unfair advantages in differentiable architecture search. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) ECCV 2020. LNCS, vol. 12360, pp. 465–480. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58555-6_28

9. Dong, X., Yang, Y.: Searching for a robust neural architecture in four gpu hours. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1761–1770 (2019)
10. Dong, X., Yang, Y.: NAS-bench-201: extending the scope of reproducible neural architecture search. In: International Conference on Learning Representations (2020)
11. Li, L., Khodak, M., Balcan, M.F., Talwalkar, A.: Geometry-aware gradient algorithms for neural architecture search. In: International Conference on Learning Representations (2021). <https://openreview.net/forum?id=MuSYkd1hxRP>
12. Liang, H., et al.: Darts+: improved differentiable architecture search with early stopping. arXiv preprint [arXiv:1909.06035](https://arxiv.org/abs/1909.06035) (2019)
13. Liu, H., Simonyan, K., Yang, Y.: Darts: differentiable architecture search. In: International Conference on Learning Representations (2019)
14. Luo, X., Liu, D., Kong, H., Huai, S., Chen, H., Liu, W.: SurgeNAS: a comprehensive surgery on hardware-aware differentiable neural architecture search. IEEE Transactions on Computers (2022)
15. Shu, Y., Wang, W., Cai, S.: Understanding architectures learnt by cell-based neural architecture search. In: International Conference on Learning Representations (2019)
16. Wang, R., Cheng, M., Chen, X., Tang, X., Hsieh, C.J.: Rethinking architecture selection in differentiable NAS. In: International Conference on Learning Representations (2021)
17. Wu, Y., Liu, A., Huang, Z., Zhang, S., Van Gool, L.: Neural architecture search as sparse supernet. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 35, pp. 10379–10387 (2021)
18. Xu, Y., et al.: PC-DARTS: Partial channel connections for memory-efficient architecture search. In: International Conference on Learning Representations (2020)
19. Xue, C., Wang, X., Yan, J., Hu, Y., Yang, X., Sun, K.: Rethinking bi-level optimization in neural architecture search: A gibbs sampling perspective. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 35, pp. 10551–10559 (2021)
20. Ye, P., Li, B., Li, Y., Chen, T., Fan, J., Ouyang, W.: β -darts: Beta-decay regularization for differentiable architecture search. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10874–10883 (2022)