

DropNAS: Grouped Operation Dropout for Differentiable Architecture Search*

Weijun Hong^{1†}, Guilin Li², Weinan Zhang^{1‡}, Ruiming Tang²,
Yunhe Wang², Zhenguo Li² and Yong Yu¹

¹Shanghai Jiao Tong University, China

²Huawei Noah’s Ark Lab, China

wiljohn@apex.sjtu.edu.cn, liguilin2@huawei.com, wnzhang@sjtu.edu.cn,
{tangruiming,yunhe.wang,li.zhenguo}@huawei.com, yyu@apex.sjtu.edu.cn

Abstract

Neural architecture search (NAS) has shown encouraging results in automating the architecture design. Recently, DARTS relaxes the search process with a differentiable formulation that leverages weight-sharing and SGD where all candidate operations are trained simultaneously. Our empirical results show that such procedure results in the *co-adaptation problem* and *Matthew Effect*: operations with fewer parameters would be trained maturely earlier. This causes two problems: firstly, the operations with more parameters may never have the chance to express the desired function since those with less have already done the job; secondly, the system will punish those underperforming operations by lowering their architecture parameter, and they will get smaller loss gradients, which causes the *Matthew Effect*. In this paper, we systematically study these problems and propose a novel grouped operation dropout algorithm named DropNAS to fix the problems with DARTS. Extensive experiments demonstrate that DropNAS solves the above issues and achieves promising performance. Specifically, DropNAS achieves 2.26% test error on CIFAR-10, 16.39% on CIFAR-100 and 23.4% on ImageNet (with the same training hyperparameters as DARTS for a fair comparison). It is also observed that DropNAS is robust across variants of the DARTS search space. Code is available at <https://github.com/huawei-noah>.

1 Introduction

With the rapid growth of deep learning in recent years [Krizhevsky *et al.*, 2012; Silver *et al.*, 2017; Chen *et al.*, 2019a], designing high-performance neural network architecture is attaining increasing attention. However, such architecture design processes involve a great amount of human expertise. More recently, automatic neural architecture

search (NAS) has been brought into focus and achieves state-of-the-art results on various tasks including image classification [Zoph and Le, 2017; Yang *et al.*, 2019], object detection [Zoph *et al.*, 2018], and recommender system [Liu *et al.*, 2020], outperforming human-designed architectures.

To reduce the evaluation cost of NAS, one promising search strategy is to leverage weight-sharing: a supernet containing all possible subnets is built, each of the subnets is optimized as the supernet trained [Bender *et al.*, 2018; Liu *et al.*, 2019; Cai *et al.*, 2019; Pham *et al.*, 2018]. The target subnet can be evaluated by inheriting the parameters from the supernet, which strikingly cuts down the search cost. DARTS built the supernet by introducing continuous architecture parameter. Using two-level optimization, the architecture parameter was trained alternatively with the network weights.

Many following works of DARTS have studied whether the architecture parameters can be properly learned with the current framework, including many questioning the convergence problem of the two-level optimization [Li *et al.*, 2019; He *et al.*, 2020; Zela *et al.*, 2020; Guo *et al.*, 2019] and the optimization gap induced by proxy network [Cai *et al.*, 2019; Chen *et al.*, 2019b; Li *et al.*, 2019]. However, very few have been explored about how well the candidate operations trained in the supernet with parameter sharing can reflect their stand-alone performance. [Xie *et al.*, 2019] raised the child network performance consistency problem, and used Gumble trick to improve the child network performance. However, they did not compare the performance rank of child networks estimated from the DARTS or SNAS framework to the true rank obtained by fully training the stand-alone child subnets. Moreover, SNAS did not manage to surpass the performance of DARTS.

In this work, we explore how well each candidate operation is trained in the supernet, and how we can balance between the supernet’s overall training stability and the individual training of each subnet. In DARTS, all candidate operations are trained simultaneously during the network weight training step. Our empirical results show that this training procedure leads to two problems:

- *The Co-adaptation Problem*: Operations with fewer parameters would be trained maturely with fewer epochs and express the desired function earlier than those with more parameters. In such cases, the operations with

*Sponsored by Huawei Innovation Research Program.

[†]This work is done when Weijun Hong worked as an intern at Huawei Noah’s Ark Lab.

[‡]The corresponding author is supported by NSFC 61702327.

more parameters may rarely have the chance to express the desired function. Therefore, those operations which take longer to converge may never have the chance to express what they could do if trained in a stand-alone model. This makes the system prefer operations that are easier to train.

- *The Matthew Effect*: The system will punish those underperforming operations by lowering their architecture parameters and backward smaller loss gradients, which causes the Matthew Effect. The Matthew Effect makes the case even worse for operations to take a longer time to train, where their architecture parameters assign them low scores at the very early stage of supernet training.

In this paper, we systematically study these problems and propose a novel grouped operation dropout algorithm named DropNAS to fix the problems with DARTS. The proposed DropNAS largely improves the DARTS framework, including the most state-of-the-art modified algorithms such as P-DARTS [Chen *et al.*, 2019b] and PC-DARTS [Xu *et al.*, 2020] on various datasets. It can also be verified that many previous differentiable NAS research works, including DARTS [Liu *et al.*, 2019], SNAS [Xie *et al.*, 2019] and ProxylessNAS [Cai *et al.*, 2019], are essentially special cases of DropNAS, inferior to DropNAS with optimized drop path rate.

In our experiments, we first show the architectures discovered by DropNAS achieve 97.74% test accuracy on CIFAR-10 and 83.61% on CIFAR-100, without additional training tricks (such as increasing channels/epochs or using auto-augmentation). For a fair comparison with other recent works, we also train the searched architectures with auto-augmentation: DropNAS can achieve 98.12% and 85.90% accuracy on CIFAR-10 and CIFAR-100 respectively. When transferred to ImageNet, our approach reaches only 23.4% test error. We also conduct experiments on variants of the DARTS search space and demonstrate that the proposed strategy can perform consistently well when a different set of operations are included in the search space.

In summary, our contributions can be listed as follows:

- We systematically studied *the co-adaption problem* of DARTS and present empirical evidence on how the performance of DARTS is degraded by this problem.
- We introduce grouped operation dropout, which is previously neglected in differentiable NAS community, to alleviate *the co-adaption problem*, meanwhile maintaining the training stability of the supernet.
- We unify various differentiable NAS approaches including DARTS, ProxylessNAS and SNAS, showing that all of them are special cases of DropNAS and inferior to it.
- We conduct sufficient experiments which show state-of-the-art performance on various benchmark datasets, and the found search scheme is robust across various search spaces and datasets.

2 Methodology

2.1 DARTS

In this work, we follow the DARTS framework [Liu *et al.*, 2019], whose objective is to search for the best cell that can be stacked to form the target neural networks. Each cell is defined as a directed acyclic graph (DAG) of N nodes $\{x_0, x_1, \dots, x_{N-1}\}$, each regarded as a layer of neural network. An edge $E_{(i,j)}$ connects the node x_i and node x_j , and consists of a set of candidate operations. We denote the operation space as \mathcal{O} , and follow the original DARTS setting where there are seven candidate operations $\{\text{Sep33}, \text{Sep55}, \text{Dil33}, \text{Dil55}, \text{Maxpool}, \text{Avepool}, \text{Identity}, \text{Zero}\}$ in it.

DARTS replaces the discrete operation selection with a weighted sum of all candidate operations, which can be formulated as:

$$\begin{aligned} \bar{o}^{(i,j)}(x_i) &= \sum_{o \in \mathcal{O}} p_o^{(i,j)} \cdot o^{(i,j)}(x_i) \\ &= \sum_{o \in \mathcal{O}} \frac{\exp(\alpha_o^{(i,j)})}{\sum_{o' \in \mathcal{O}} \exp(\alpha_{o'}^{(i,j)})} \cdot o^{(i,j)}(x_i). \end{aligned} \quad (1)$$

This formula explains how a feature mapping $\bar{o}^{(i,j)}(x_i)$ on edge $E_{(i,j)}$ is computed from the previous node x_i . Here $\alpha_o^{(i,j)}$ is the architecture parameter, and $p_o^{(i,j)}$ represents the relative contribution of each operation $o^{(i,j)} \in \mathcal{O}$. Within a cell, each immediate node x_j is represented as the sum of all the feature mappings on edges connecting to it: $x_j = \sum_{i < j} \bar{o}^{(i,j)}(x_i)$. In this work, we adopt the one-level optimization for stability and data efficiency similar as [Li *et al.*, 2019], where the update rule can be easily obtained by applying stochastic gradient descent on both w and α . After the architecture parameter α is obtained, we can derive the final searched architecture following these steps like DARTS: 1) replace the mixed operation on each edge by $o^{(i,j)} = \arg \max_{o \in \mathcal{O}, o \neq \text{zero}} p_o^{(i,j)}$, 2) retain two edges from different predecessor nodes with the largest $\max_{o \in \mathcal{O}, o \neq \text{zero}} p_o^{(i,j)}$.

2.2 The Co-adaption Problem and Matthew Effect

To explore the co-adaption phenomenon, we visualize the clustering of the feature maps generated from all the seven operations (except *zero*) on edge $E_{(0,2)}$ in Figure 1. We find that similar operations like convolutions with different kernel sizes generally produce similar feature maps, which means they serve as similar functions in the system. However, it is also common that convolutions with larger kernel size and more parameters stand far away from the other convolutions, which suggests that they are expressing something different from the others, while these standing out operations are always getting lower architecture value in the system. Then we check the stand-alone accuracy of these operations by training the corresponding architecture separately, and we find that large kernels usually perform better if trained properly, which contradicts the score they obtained from the DARTS system. This suggests that the system suffers from the co-adaption problem that those with fewer parameters would be trained maturely earlier and express the desired function

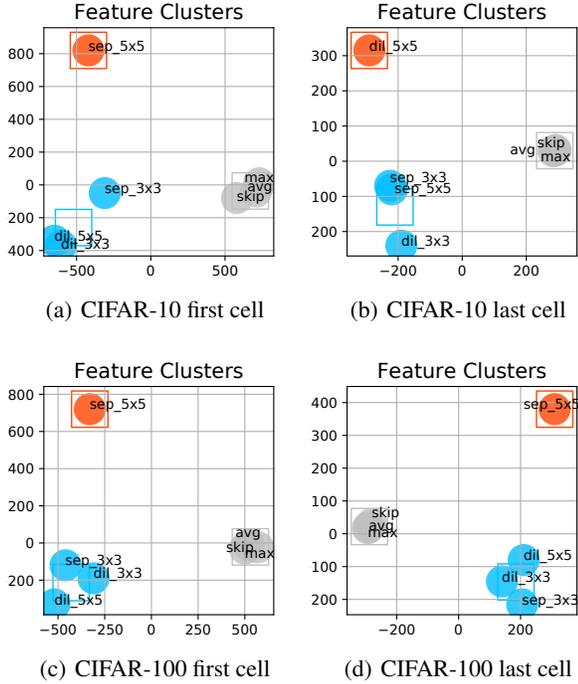


Figure 1: Feature clusters generated from different operations on edge $E_{(0,2)}$: following [Li *et al.*, 2019], we randomly select 1000 data, from CIFAR-10 and CIFAR-100 respectively, to generate the feature mappings, and apply K-means to cluster these mappings into 3 clusters to show the similarities between them, and finally use PCA to get a two-dimensional illustration. In (a) and (c) we select the edge in the first cell of the one-shot model, and in (b) and (d) we select the edge in the last cell.

more quickly, causing that 5×5 convs rarely have chance to express the desired function. This also causes their α to be smaller and less gradient to backward to them and consequently causes the Matthew Effect.

In this work, we will introduce operation dropout to DARTS based on one-level optimization to explore a more general search scheme, which actually unifies the existing differentiable methods and finds the best strategy for weight-sharing in the DARTS framework.

2.3 Grouped Operation Dropout

In this part, we propose a simple yet effective search scheme called *Grouped Operation Dropout* to break the correlation in the weight-sharing supernet.

Specifically, during the search stage, for each edge, we randomly and independently select a subset of the operations, and zero out their outputs, to make their α and w not updated during back-propagation. Such a strategy mitigates the co-adaptation among the parameterized operations since the under fitted operations have more chance to play an important role during the one-shot model training so that it can better fit the training target and benefit the α 's learning.

In practice, we partition the eight operations into two groups according to whether they are learnable, i.e. one parameterized group \mathcal{O}_p containing all the convolutional oper-

ations, and one non-parameterized group \mathcal{O}_{np} involving the remains. During the entire search procedure, for group \mathcal{O}_p as an example, we fix the probability of each operation to be dropped as $p_d = r^{1/|\mathcal{O}_p|}$, where $0 < r < 1$ is a hyperparameter called drop path rate. Note that $|\mathcal{O}_p| = |\mathcal{O}_{np}| = 4$, and the hyperparameter r denotes the probability of disabling all the operations in \mathcal{O}_p and \mathcal{O}_{np} in the DARTS search space. For example, if we set $r = 3 \times 10^{-5}$, then $p_d = r^{1/4} = 0.074$. Additionally, we also enforce at least one operation to remain in each group to further stabilize the training, which is realized by resampling if the operations on some edge happen to be all dropped. During the backpropagation period, the w and α of the dropped operations will receive no gradient. By enforcing to keep at least one operation in each group, the equivalent function of an edge is always a mixture of learnable and non-learnable operations, resulting in a relatively stable training environment for the architecture parameters α .

Note that operation dropout in one-level DARTS essentially unifies most existing differentiable NAS approaches: DARTS [Liu *et al.*, 2019] updates all the operations on an edge at once, which corresponds to the $p_d = 0$ case; SNAS [Xie *et al.*, 2019] and ProxylessNAS [Cai *et al.*, 2019] once samples only one and two operations to update, corresponding to $p_d = 0.875$ and $p_d = 0.75$ respectively in expectation. We will later show that all of them are actually inferior to the best p_d we find.

α -Adjust: Prevent Passive Update. Note that in DARTS, we measure the contribution $p_o^{(i,j)}$ of a certain operation $o^{(i,j)}$ on the edge $E_{(i,j)}$ via a softmax over all the learnable architecture parameters $\alpha_o^{(i,j)}$, as in Equation (1). As a result, the contribution $p_o^{(i,j)}$ of the dropped operations that do not receive any gradient during the backward pass, will get changed even though their corresponding $\alpha_o^{(i,j)}$ remain the same. In order to prevent the passive update of the dropped operations' $p_o^{(i,j)}$, we need to adjust the value of each $\alpha_o^{(i,j)}$ after applying the gradient. Our approach is to solve for an additional term x according to:

$$\frac{\sum_{o \in \mathcal{O}_d} \exp(\alpha_o^{old})}{\sum_{o \in \mathcal{O}_k} \exp(\alpha_o^{old})} = \frac{\sum_{o \in \mathcal{O}_d} \exp(\alpha_o^{new})}{\sum_{o \in \mathcal{O}_k} \exp(\alpha_o^{new} + x)} \quad (2)$$

where we omitted the subscript (i, j) , \mathcal{O}_d & \mathcal{O}_k refer to the operation sets that are dropped & kept on edge $E_{(i,j)}$, α_o^{old} & α_o^{new} means the value before & after backpropagation. With the additional term x to adjust the value of α_o^{new} for $o \in \mathcal{O}_k$, the contribution $p_o^{(i,j)}$ for $o \in \mathcal{O}_d$ remains unchanged. Note that $\alpha_o^{old} = \alpha_o^{new}$ for $o \in \mathcal{O}_d$, by solving Equation (2) we get:

$$x = \ln \left[\frac{\sum_{o \in \mathcal{O}_k} \exp(\alpha_o^{old})}{\sum_{o \in \mathcal{O}_k} \exp(\alpha_o^{new})} \right]. \quad (3)$$

Partial-Decay: Prevent Unnecessary Weight Decay. L2 regularization is employed during the search stage of original DARTS, and we also find it useful in one-level optimization. However, when applied with dropout, the parameters of \mathcal{O}_d will be regularized even when they are dropped. So in our implementation, we apply the L2 weight decay only to those w and α belonging to \mathcal{O}_k to prevent the over-regularization.

Architecture	Test Error (%)	Param (M)	Search Cost (GPU Days)
NASNet-A [Zoph and Le, 2017]	2.65	3.3	1800
AmoebaNet-B [Real <i>et al.</i> , 2019]	2.55 ± 0.05	2.8	3150
ENAS [Pham <i>et al.</i> , 2018]	2.89	4.6	0.5
DARTS [Liu <i>et al.</i> , 2019]	3.00	3.3	1.5
SNAS [Xie <i>et al.</i> , 2019]	2.85	2.8	1.5
ProxylessNAS [Cai <i>et al.</i> , 2019] ¹	2.08	5.7	4
P-DARTS [Chen <i>et al.</i> , 2019b]	2.50	3.4	0.3
DARTS+ [Liang <i>et al.</i> , 2019] ²	2.20(2.37 ± 0.13)	4.3	0.6
StacNAS [Li <i>et al.</i> , 2019]	2.33(2.48 ± 0.08)	3.9	0.8
ASAP [Noy <i>et al.</i> , 2019]	2.49 ± 0.04	2.5	0.2
PC-DARTS [Xu <i>et al.</i> , 2020]	2.57 ± 0.07	3.6	0.1
DropNAS³	2.26(2.58 ± 0.14)	4.1	0.6
DropNAS (Augmented)⁴	1.88	4.1	0.6

Table 1: Performance of different architectures on CIFAR-10. ProxylessNAS¹ uses a search space different from DARTS. DARTS+² trains the evaluation model for 2,000 epochs, while others just train 600 epochs. Our DropNAS³ reports both the mean and standard deviation with eight seeds, **by keeping training epochs or channels the same with the original DARTS for a fair comparison**. DropNAS (Augmented)⁴ denotes training with AutoAugment and 1,200 epochs.

Architecture	Test Error (%)	Param (M)	Search Cost (GPU Days)
DARTS [Liu <i>et al.</i> , 2019] ¹	17.76	3.3	1.5
P-DARTS [Chen <i>et al.</i> , 2019b]	15.92	3.6	0.3
DARTS+ [Liang <i>et al.</i> , 2019] ²	14.87(15.45 ± 0.30)	3.9	0.5
StacNAS [Li <i>et al.</i> , 2019]	15.90(16.11 ± 0.2)	4.3	0.8
ASAP [Noy <i>et al.</i> , 2019] ¹	15.6	2.5	0.2
DropNAS³	16.39(16.95 ± 0.41)	4.4	0.7
DropNAS (Augmented)⁴	14.10	4.4	0.7

Table 2: Results of different architectures on CIFAR-100. The results denoted with ¹ use the architectures found on CIFAR-10. The subscript ², ³ and ⁴ have the same meaning as in Table 1.

3 Related Work

In their scientific investigation work, [Bender *et al.*, 2018] explored path level dropout during the training of the supernet for NAS, concluding that a proper drop path rate is desired for reducing the co-adaptation problem and maintaining the stability of the training. However, their findings are largely neglected in the differentiable NAS community, where most of the current works including DARTS, P-DARTS and StacNAS train all operations simultaneously. [Cai *et al.*, 2019; Guo *et al.*, 2019] train the supernet by sampling one path with probability proportional to architecture or uniform distribution respectively, which is equivalent to drop path rate as $\frac{N-1}{N}$, where N is the total number of operations. However, our empirical results show that such a high drop path rate is not the best choice of training the supernet of DARTS search space: the system is instable where convolution is sometimes followed by pooling and sometimes a convolution.

In this work, we introduce dropout to the DARTS framework. By adopting a properly tuned drop path rate and leveraging the operation grouping and one-level optimization proposed by [Li *et al.*, 2019], we show that we could further improve the most SOTA results achieved before.

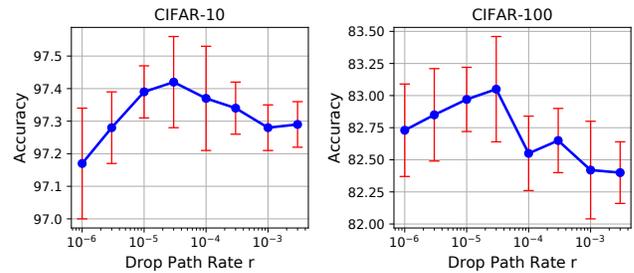


Figure 2: The impact of drop path rate reflected on the stand-alone model accuracy. The red error bars show the standard deviation of 8 repeated experiments.

4 Benchmark

4.1 Datasets

To benchmark our grouped operation dropout algorithm, extensive experiments are carried out on CIFAR-10, CIFAR-100 and ImageNet.

Both the CIFAR-10 and CIFAR-100 datasets contain 50K training images and 10K testing images, and the resolution of each image is 32×32 . All the images are equally partitioned into 10/100 categories in CIFAR-10/100.

ImageNet is a much larger dataset consisting of 1.3M images for training and 50K images for testing, equally distributed among 1,000 classes. In this paper, we use ImageNet to evaluate the transferability of our architectures found on CIFAR-10/100. We follow the conventions of [Liu *et al.*, 2019] that consider the mobile setting to fix the size of the input image to 224×224 and limit the multiply-add operations to be no more than 600M.

4.2 Implementation Details

Architecture Search As we have mentioned before, we leverage the DARTS search space with the same eight candidate operations. Since we use one-level optimization, the training images do not need to be split for another validation set, so the architecture search is conducted on CIFAR-10/100 with all the training images on a single Nvidia Tesla V100. We use 14 cells stacked with 16 channels to form the one-shot model, train the supernet for 76 epochs with batch size 96, and pick the architecture discovered in the final epoch. The model weights w are optimized by SGD with initial learning rate 0.0375, momentum 0.9, and weight decay 0.0003, and we clip the gradient norm of w to be less than 3 for each batch. The architecture parameters α are optimized by Adam, with initial learning rate 0.0003, momentum (0.5, 0.999) and weight decay 0.001. Drop path rate r is fixed to 3×10^{-5} .

Architecture Evaluation On CIFAR-10 and CIFAR-100, to fairly evaluate the discovered architectures, neither the initial channels nor the training epochs are increased for the evaluation network, compared with DARTS. 20 cells are stacked to form the evaluation network with 36 initial channels. The network is trained on a single Nvidia Tesla V100 for 600 epochs with batch size 192. The network parameters are optimized by SGD with learning rate 0.05, momentum 0.9 and weight decay 0.0003, and the gradient is clipped in the same way as in the search stage. The data augmentation

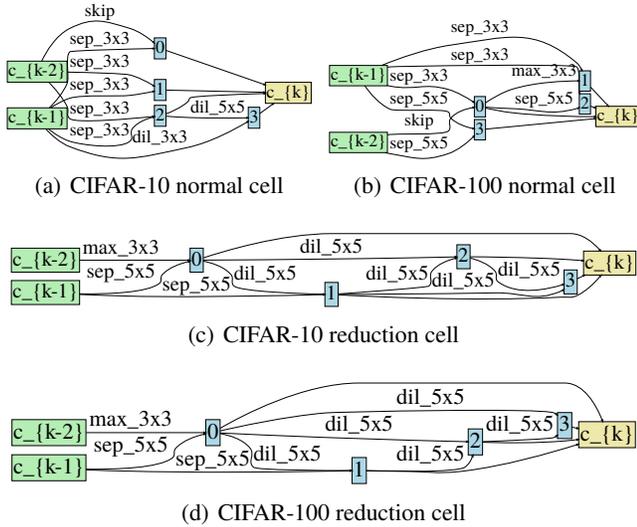


Figure 3: The found architectures on CIFAR-10 and CIFAR-100

method Cutout and an auxiliary tower with weight 0.4 are also employed as in DARTS. To exploit the potentials of the architectures, we additionally use AutoAugment to train the model for 1,200 epochs. The best architecture discovered are represented in Fig. 3 and their evaluation results are shown in Table 1 and 2. We can see that the best architecture discovered by DropNAS achieves the SOTA test error 2.26% on CIFAR-10, and on CIFAR-100 DropNAS still works well compared to DARTS, and largely surpasses the one-level version which prefers to ending up with many *skip-connect* in the final architecture if directly searched on CIFAR-100.

To test the transferability of our selected architectures, we adopt the best architectures found on CIFAR-10 and CIFAR-100 to form a 14-cell, 48-channel evaluation network to train on ImageNet. The network is trained for 600 epochs with batch size 2048 on 8 Nvidia Tesla V100 GPUs, optimized by SGD with initial learning rate 0.8, momentum 0.9, weight decay 3×10^{-5} , and gradient clipping 3.0. The additional enhancement approaches that we use include AutoAugment, mixup, SE module, auxiliary tower with loss weight 0.4, and label smoothing with $\epsilon = 0.1$. Table 3 shows that the architecture found by DropNAS is transferable and obtains encouraging result on ImageNet.

5 Diagnostic Experiments

5.1 Impact of Drop Path Rate

In DropNAS we introduce a new hyperparameter, i.e. the drop path rate r , whose value has a strong impact on the results since a higher drop path rate results in a lower correlation between the operations. To demonstrate its significance, we repeat the search and evaluation stages with varying drop path rates and report the stand-alone model accuracy in Fig. 2. The best results are achieved when $r = 3 \times 10^{-5}$ on both datasets, which indicates that the found best drop path rate is transferable to different datasets. Note that p_d is just 0.074 when $r = 3 \times 10^{-5}$, so the other cases like $p_d = 0.875, 0.75$

Architecture	Test Err. (%)		Params (M)	Search Days
	Top-1	Top-5		
NASNet-A [Zoph and Le, 2017]	26.0	8.4	5.3	1800
EfficientNet-B0 [Tan and Le, 2019]	23.7	6.8	5.3	-
DARTS [Liu <i>et al.</i> , 2019]	26.7	8.7	4.7	4.0
SNAS (mild) [Xie <i>et al.</i> , 2019]	27.3	9.2	4.3	1.5
ProxylessNAS [Cai <i>et al.</i> , 2019] ^{†*}	24.9	7.5	7.1	8.3
P-DARTS (C10) [Chen <i>et al.</i> , 2019b]	24.4	7.4	4.9	0.3
ASAP [Noy <i>et al.</i> , 2019]	26.7	-	-	0.2
XNAS [Nayman <i>et al.</i> , 2019]	24.0	-	5.2	0.3
PC-DARTS [Xu <i>et al.</i> , 2020] [†]	24.2	7.3	5.3	3.8
ScarletNAS[Chu <i>et al.</i> , 2019] ^{†*}	23.1	6.6	6.7	10
DARTS+[Liang <i>et al.</i> , 2019] [†]	23.9	7.4	5.1	6.8
StacNAS[Li <i>et al.</i> , 2019] [†]	24.3	6.4	5.7	20
Single-Path NAS [Stamoulis <i>et al.</i> , 2019] ^{†*}	25.0	7.8	-	0.16
DropNAS (CIFAR-10)	23.4	6.7	5.7	0.6
DropNAS (CIFAR-100)	23.5	6.8	6.1	0.7

 Table 3: Results of different architectures on ImageNet. The results denoted with [†] use the architectures directly searched on ImageNet, and those denoted with * use the backbone different from DARTS.

or 0 are all inferior to it, which correspond to the search scheme of SNAS, ProxylessNAS and DARTS, respectively.

5.2 Feature Clusters in DropNAS

For comparison we again draw the feature clusters in DropNAS with $r = 3 \times 10^{-5}$, following the same way in Fig. 1. The results are plotted in Fig. 4.

It is significant that the point of parameterized operation no longer shifts away from its similar partner, and there is no cluster containing only one single point anymore. So we claim that the severe co-adaption problem existing in the one-level DARTS has been greatly reduced by DropNAS.

5.3 Performance on Other Search Space

We are also interested in the adaptability of DropNAS in other search spaces. We purposely design two search spaces: in the first space we replace the original 3×3 *avg-pooling* and 3×3 *max-pooling* operations by *skip-connect*; And in the second space we remove the 3×3 *avg-pooling* and 3×3 *max-pooling* operations in \mathcal{O}_{np} . We again search on CIFAR-10 and evaluate the found architectures, report the mean accuracy and standard deviations of eight repeated runs.

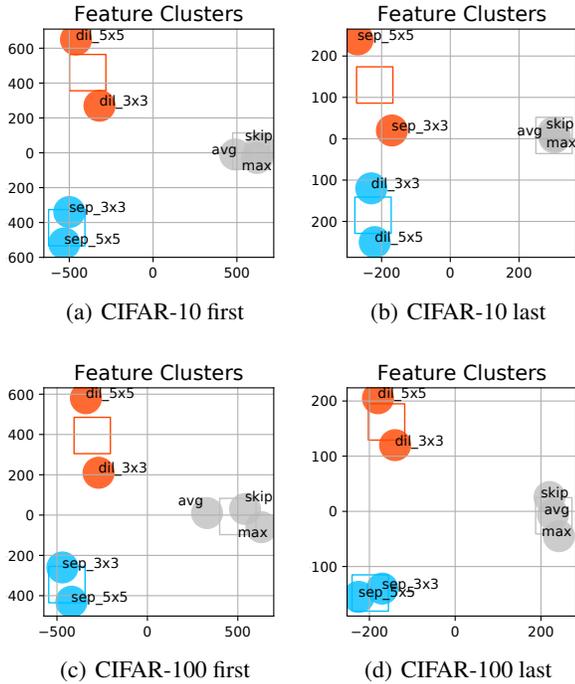
The results shown in Table 4 demonstrates that DropNAS is robust across variants of the DARTS search spaces in different datasets.

5.4 Impact of Drop Path Rates in Different Groups

As we mentioned in Section 2.3, one advantage of grouping in DropNAS is that we can apply different drop path rates to different operation groups. However, our architecture search is actually conducted with r fixed to 3×10^{-5} for both \mathcal{O}_p and \mathcal{O}_{np} . In fact, we have assigned \mathcal{O}_p and \mathcal{O}_{np} with different drop path rates around 3×10^{-5} , and the results are shown in Table 5, which means the best performance is achieved when the two groups share exactly the same rate.

5.5 Performance Correlation between Stand-Alone Model and Architecture Parameters

DropNAS is supposed to break the correlation between the operations, so that the architecture parameters α can represent

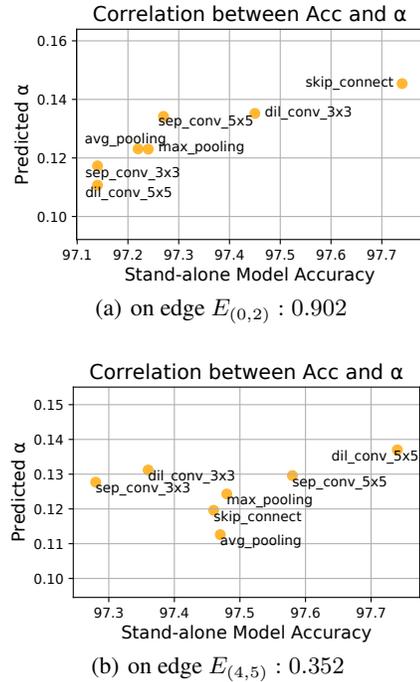

 Figure 4: Feature clusters of DropNAS on $E_{(0,2)}$

Dataset	Search Space	Test Error (%)	
		DropNAS	one-level DARTS
CIFAR-10	3-skip	2.68 ± 0.10	3.19 ± 0.18
	1-skip	2.67 ± 0.11	2.85 ± 0.12
	original	2.58 ± 0.14	2.90 ± 0.16
CIFAR-100	3-skip	16.97 ± 0.35	18.00 ± 0.34
	1-skip	16.47 ± 0.19	17.73 ± 0.25
	original	16.95 ± 0.41	17.27 ± 0.36

Table 4: The performance of DropNAS and one-level DARTS across different search spaces on CIFAR-10/100.

the real importance of each operation, and then we can easily select the best the architecture by ranking α . Fig. 5 shows the correlation between the architectures and their corresponding α on two representative edges in normal cell, $E_{(0,2)}$ and $E_{(4,5)}$, which are the first and the last edge within the cell. We claim that the α learned by DropNAS has a vigorous representative power of the accuracy of the stand-alone model, since the correlation coefficient between them is 0.902 on $E_{(0,2)}$, largely surpassing that of DARTS (0.2, reported in [Li *et al.*, 2019]), and 0.352 on $E_{(4,5)}$, where the choice of a specific operation is less significant.

\mathcal{O}_{np} \ \mathcal{O}_p	1×10^{-5}	3×10^{-5}	1×10^{-4}
1×10^{-5}	2.60 ± 0.16	2.72 ± 0.04	2.64 ± 0.12
3×10^{-5}	2.64 ± 0.11	2.58 ± 0.14	2.69 ± 0.05
1×10^{-4}	2.65 ± 0.07	2.69 ± 0.10	2.63 ± 0.16

 Table 5: The test error of DropNAS on CIFAR-10 when the operation groups \mathcal{O}_p and \mathcal{O}_{np} are applied with different drop path rates. The above results are obtained over 8 different seeds.

 Figure 5: Correlation coefficients between the accuracy of stand-alone model and their corresponding α . The results are obtained by first searching on CIFAR-10, figuring out the best architecture, then generating other 6 architectures by replacing the operation on edges $E_{(0,2)}$ and $E_{(4,5)}$ in the normal cell with other $o \in \mathcal{O}$, and finally the corresponding stand-alone models are trained from scratch.

	Test Err. (%)			
	DropNAS	No α -adjust	No partial-decay	No grouping
CIFAR-10	2.58 ± 0.14	2.75 ± 0.08	2.71 ± 0.06	2.74 ± 0.11
CIFAR-100	16.95 ± 0.41	17.40 ± 0.22	17.62 ± 0.37	17.98 ± 0.33

Table 6: Ablation study on CIFAR-10/100, averaged over 8 runs.

5.6 Ablation Study

To show the techniques we proposed in Section 2.3 really improve the DropNAS performance, we further conduct experiments for DropNAS with each of the techniques disabled. The results in Table 6 show that each component of DropNAS is indispensable for achieving good performance.

6 Conclusion

We propose DropNAS, a grouped operation dropout method for one-level DARTS, that greatly improves the DARTS performance over various benchmark datasets. We explore the co-adaptation problem of DARTS and present empirical evidence about how DARTS performance is degraded by this problem. It should be noticed that various differentiable NAS approaches are unified in our DropNAS framework, but fail to match the best drop path rate we find. Moreover, the found best drop path rate of DropNAS is transferable in different datasets and variants of DARTS search spaces, demonstrating its strong applicability in a wider range of tasks.

References

- [Bender *et al.*, 2018] Gabriel Bender, Pieter-Jan Kindermans, Barret Zoph, Vijay Vasudevan, and Quoc Le. Understanding and simplifying one-shot architecture search. In *ICML*, pages 549–558, 2018.
- [Cai *et al.*, 2019] Han Cai, Ligeng Zhu, and Song Han. ProxylessNAS: Direct neural architecture search on target task and hardware. In *ICLR*, 2019.
- [Chen *et al.*, 2019a] Hanting Chen, Yunhe Wang, Chang Xu, Zhaohui Yang, Chuanjian Liu, Boxin Shi, Chunjing Xu, Chao Xu, and Qi Tian. Data-free learning of student networks. In *CVPR*, pages 3514–3522, 2019.
- [Chen *et al.*, 2019b] Xin Chen, Lingxi Xie, Jun Wu, and Qi Tian. Progressive differentiable architecture search: Bridging the depth gap between search and evaluation. In *ICCV*, pages 1294–1303, 2019.
- [Chu *et al.*, 2019] Xiangxiang Chu, Bo Zhang, Jixiang Li, Qingyuan Li, and Ruijun Xu. Scarletnas: Bridging the gap between scalability and fairness in neural architecture search. *arXiv preprint arXiv:1908.06022*, 2019.
- [Guo *et al.*, 2019] Zichao Guo, Xiangyu Zhang, Haoyuan Mu, Wen Heng, Zechun Liu, Yichen Wei, and Jian Sun. Single path one-shot neural architecture search with uniform sampling. *arXiv preprint arXiv:1904.00420*, 2019.
- [He *et al.*, 2020] Chaoyang He, Haishan Ye, Li Shen, and Tong Zhang. Milenas: Efficient neural architecture search via mixed-level reformulation. In *CVPR*, 2020.
- [Krizhevsky *et al.*, 2012] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.
- [Li *et al.*, 2019] Guilin Li, Xing Zhang, Zitong Wang, Zhen-guo Li, and Tong Zhang. Stacnas: Towards stable and consistent optimization for differentiable neural architecture search. *arXiv preprint arXiv:1909.11926*, 2019.
- [Liang *et al.*, 2019] Hanwen Liang, Shifeng Zhang, Jiacheng Sun, Xingqiu He, Weiran Huang, Kechen Zhuang, and Zhenguo Li. Darts+: Improved differentiable architecture search with early stopping. *arXiv preprint arXiv:1909.06035*, 2019.
- [Liu *et al.*, 2019] Hanxiao Liu, Karen Simonyan, and Yiming Yang. DARTS: Differentiable architecture search. In *ICLR*, 2019.
- [Liu *et al.*, 2020] Bin Liu, Chenxu Zhu, Guilin Li, Weinan Zhang, Jincal Lai, Ruiming Tang, Xiuqiang He, Zhen-guo Li, and Yong Yu. Autofis: Automatic feature interaction selection in factorization models for click-through rate prediction. *arXiv preprint arXiv:2003.11235*, 2020.
- [Nayman *et al.*, 2019] Niv Nayman, Asaf Noy, Tal Ridnik, Itamar Friedman, Rong Jin, and Lihi Zelnik. Xnas: Neural architecture search with expert advice. In *Advances in Neural Information Processing Systems*, pages 1975–1985, 2019.
- [Noy *et al.*, 2019] Asaf Noy, Niv Nayman, Tal Ridnik, Nadav Zamir, Sivan Doveh, Itamar Friedman, Raja Giryes, and Lihi Zelnik-Manor. Asap: Architecture search, anneal and prune. *arXiv preprint arXiv:1904.04123*, 2019.
- [Pham *et al.*, 2018] Hieu Pham, Melody Guan, Barret Zoph, Quoc Le, and Jeff Dean. Efficient neural architecture search via parameters sharing. In *ICML*, pages 4095–4104, 2018.
- [Real *et al.*, 2019] Esteban Real, Alok Aggarwal, Yanping Huang, and Quoc V Le. Regularized evolution for image classifier architecture search. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 4780–4789, 2019.
- [Silver *et al.*, 2017] David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, et al. Mastering the game of go without human knowledge. *Nature*, 550(7676):354, 2017.
- [Stamoulis *et al.*, 2019] Dimitrios Stamoulis, Ruizhou Ding, Di Wang, Dimitrios Lymberopoulos, Bodhi Priyantha, Jie Liu, and Diana Marculescu. Single-path nas: Designing hardware-efficient convnets in less than 4 hours. *arXiv preprint arXiv:1904.02877*, 2019.
- [Tan and Le, 2019] Mingxing Tan and Quoc Le. Efficient-net: Rethinking model scaling for convolutional neural networks. In *ICML*, pages 6105–6114, 2019.
- [Xie *et al.*, 2019] Sirui Xie, Hehui Zheng, Chunxiao Liu, and Liang Lin. SNAS: stochastic neural architecture search. In *International Conference on Learning Representations*, 2019.
- [Xu *et al.*, 2020] Yuhui Xu, Lingxi Xie, Xiaopeng Zhang, Xin Chen, Guo-Jun Qi, Qi Tian, and Hongkai Xiong. Pc-darts: Partial channel connections for memory-efficient architecture search. In *International Conference on Learning Representations*, 2020.
- [Yang *et al.*, 2019] Zhaohui Yang, Yunhe Wang, Xinghao Chen, Boxin Shi, Chao Xu, Chunjing Xu, Qi Tian, and Chang Xu. Cars: Continuous evolution for efficient neural architecture search. *arXiv preprint arXiv:1909.04977*, 2019.
- [Zela *et al.*, 2020] Arber Zela, Thomas Elsken, Tonmoy Saikia, Yassine Marrakchi, Thomas Brox, and Frank Hutter. Understanding and robustifying differentiable architecture search. In *ICLR*, 2020.
- [Zoph and Le, 2017] Barret Zoph and Quoc V Le. Neural architecture search with reinforcement learning. In *ICLR*, 2017.
- [Zoph *et al.*, 2018] Barret Zoph, Vijay Vasudevan, Jonathon Shlens, and Quoc V Le. Learning transferable architectures for scalable image recognition. In *CVPR*, pages 8697–8710, 2018.