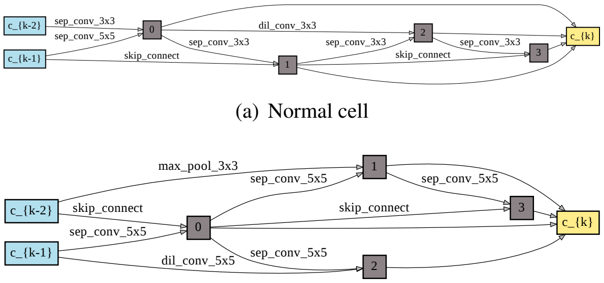
Supplementary Material

A. Implementation Details

A.1 Architecture Search on CIFAR-10

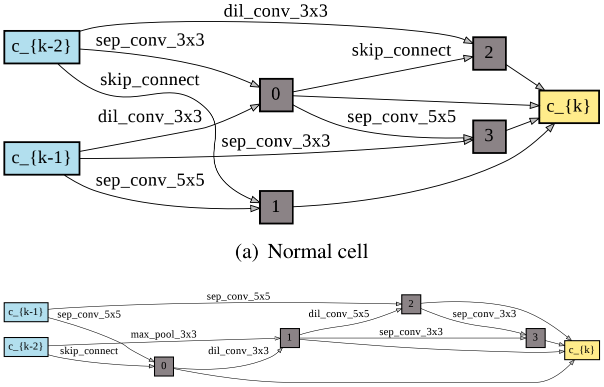
Following DARTS [1], the super network is constructed by stacking 6 normal cells and 2 reduction cells. Each cell contains seven nodes, including two input nodes, four intermediate nodes and one output node. The outputs of four intermediate nodes are concatenated as the input to the output node. Each cell has 14 candidate edges, with 8 candidate operations for each edge. The candidate operations are in accordance with DARTS, consisting of 3×3 and 5×5 separable convolution, 3×3 and 5×5 dilated separable convolution, 3×3 max and average pooling, skip-connect and none operation.



(a) Normal cell

(b) Reduction cell

Figure 1. Searched cells of VIM-NAS-Large on CIFAR-10



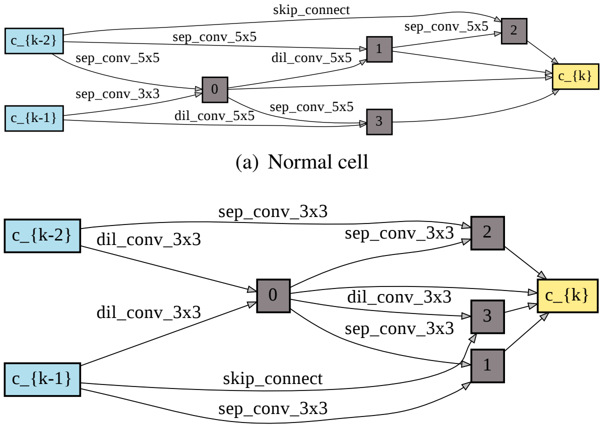
(a) Normal cell

(b) Reduction cell

Figure 2. Searched cells of VIM-NAS-Small on CIFAR-10

A.2 Architecture Search on ImageNet

Following PC-DARTS [2], we use a momentum SGD with an initial learning rate of 0.5 (annealed down to zero following a cosine schedule without restart), a momentum of 0.9, and a weight decay of 3×10^{-5}. For hyperparameters, we use the Adam optimizer with a fixed learning rate of 6×10^{-3}, a momentum (0.5, 0.999) and a weight decay of 10^{-3}. We use one 1080 Ti GPU for search, and the total batch size is 64. The entire search process takes on epoch to converge.



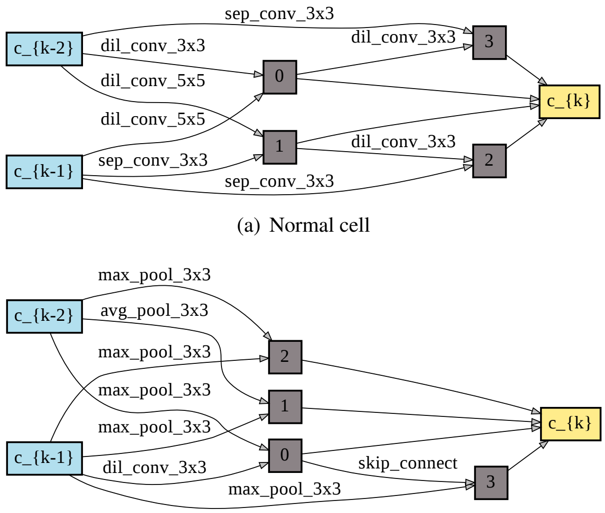
(a) Normal cell

(b) Reduction cell

Figure 3. Searched cells of VIM-NAS on ImageNet

A.3 Architecture Evaluation on CIFAR-10/100

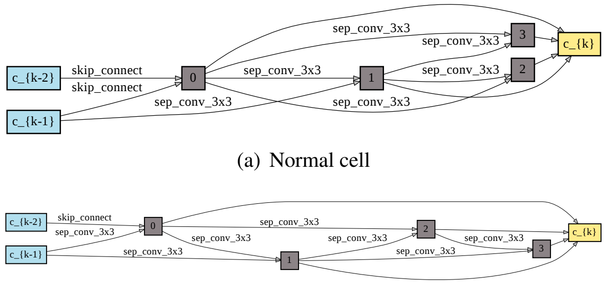
A large network with 20 cells (*i.e.*, 18 normal cells and 2 reduction cells) is constructed by stacking the searched normal cell and reduction cell. Following the evaluation of PC-DARTS, we use the same hyperparameters to train the network using all the 50K training images from scratch for 600 epochs with a batch size of 96. The initial number of channels is set to 36. We use the SGD optimizer with an initial learning rate of 0.025 (annealed down to zero following a cosine schedule without restart), a momentum of 0.9, a weight decay of $3 \times 10^{-4}/5 \times 10^{-4}$ and a norm gradient clipping at 5. Moreover, we set cutout with size 16, path dropout with the probability of 0.3 and auxiliary towers with weight 0.4.



(a) Normal cell

(b) Reduction cell

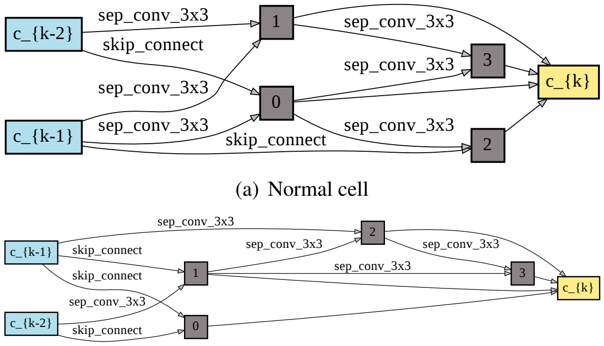
Figure 4. Searched cells of VIM-NAS on S1 search space



(a) Normal cell

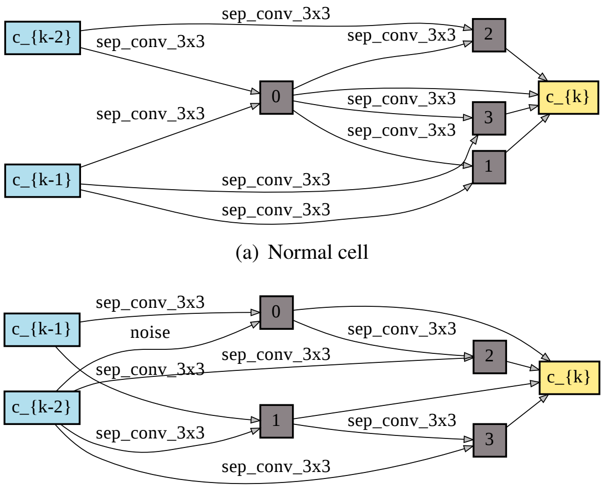
(b) Reduction cell

Figure 5. Searched cells of VIM-NAS on S2 search space



(a) Normal cell

Figure 6. Searched cells of VIM-NAS on S3 search space



(a) Normal cell

(b) Reduction cell

Figure 7. Searched cells of VIM-NAS on S4 search space

A.4 Architecture Evaluation on ImageNet

The network of 12 normal cells and 2 reduction cells is trained from scratch for 250 epochs with a batch size of 1024. The initial number of channels is set to 48. We use the SGD optimizer with a momentum of 0.9, an initial learning rate of 0.5 (decayed down to zero linearly) and a weight decay of 3×10^{-5}. Label smoothing and auxiliary loss tower are also used during training.

B. Architectural Neural Network

The detailed structures of architectural neural networks for VIM-NAS-Small, VIM-NAS and VIM-NAS-Large are presented in Table 1, respectively.

layer	output	VIM-NAS-S	VIM-NAS	VIM-NAS-L
1	14×8	CRB(3,1,3)	CRB(3,14,3)	CRB(3,14,3)
2	14×8	-	CRB(14,1,3)	CRB(14,14,3)
3	14×8	-	-	CRB(14,14,3)
4	14×8	-	-	CRB(14,14,3)
5	14×8	-	-	CRB(14,1,3)

Table 1. Detailed structures of architectural neural networks of VIM-NAS-Small, VIM-NAS and VIM-NAS-Large. CRB denotes the ConvReLUBN module stacked with convolution, relu and batch normalization, and the following three numbers denote the input channel number, output channel number and kernel size, respectively.

C. Ablation Study on the Learning Rate of the Parameters of Architectural Neural Network

We experiment with a small learning rate (0.001) of the parameters of the architectural neural network as shown in Figure 8. The search process with learning rate 0.001 exhibits a lower convergent speed (20 epochs). Then, we test our searched architecture and get a $2.52\pm0.05\%$ top 1 error rate with 4.0M parameters. This result exhibits that smaller learning rate will lead to slower convergent speed without performance improvement (comparable performance). Comparatively, our method with a learning rate of 0.025 can reach a satisfactory local minimum at the extremely fast speed.

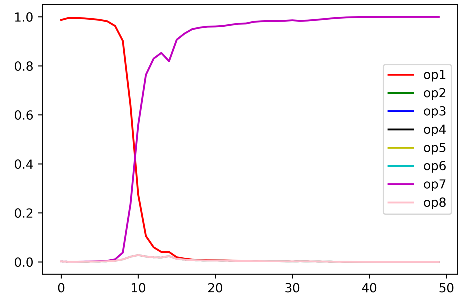


Figure 8. Anytime architectural weights on DARTS search space of VIM-NAS with 0.001 learning rate.

D. Searched Architectures

D.1 VIM-NAS on ImageNet

Searched cells of VIM-NAS on ImageNet are shown in Figures 3.

D.2 VIM-NAS-Large

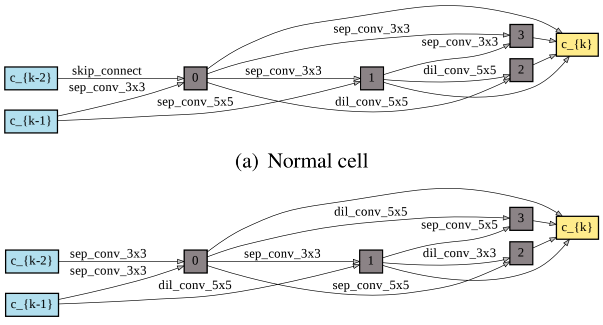
Searched cells of VIM-NAS-Large are shown in Figure 1.

D.3 VIM-NAS-Small

Searched cells of VIM-NAS-Small are shown in Figure 2.

D.4 VIM-NAS with a Small Learning Rate of 0.001

Searched cells of VIM-NAS with a small learning rate of 0.001 on CIFAR-10 are shown in Figure 9.



(a) Normal cell

Figure 9. Searched cells of VIM-NAS with a small learning rate of 0.001 on CIFAR-10.

D.5 VIM-NAS on simplified search spaces S1-S4

Searched cells of VIM-NAS on four simplified search spaces S1-S4 are shown in Figures 4, 5, 6, and 7, respectively.

References

[1] Hanxiao Liu, Karen Simonyan, and Yiming Yang. DARTS: Differentiable architecture search. In *7th International Conference on Learning Representations*, New Orleans, LA, USA, May 2019.

[2] 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*, 2019.