Appendices for DSA: More Efficient Budgeted Pruning via Differentiable Sparsity Allocation

Xuefei Ning[1⋆], Tianchen Zhao[2*], Wenshuo Li[1], Peng Lei[2], Yu Wang[1], and Huazhong Yang[1]

[1] Department of Electronic Engineering, Tsinghua University
[2] Department of Electronic Engineering, Beihang University
foxdoraame@gmail.com, ztc16@buaa.edu.cn, yu-wang@tsinghua.edu.cn

1 Topological Grouping and Budget Model $\mathcal{F}$

Alg. 1 describes the procedure of topological grouping to handle the topological constraints introduced by shortcuts, depthwise convolutions and so on. An example of grouping convolutions in two consecutive residual blocks is shown in Fig. 1. All the incoming connections of the normal convolutions are removed, and then the convolutions in each connected component belong to the same topological group (i.e., share the same keep ratio $\alpha^{(k)}$ and masks $m^{(k)}_{i=1,\cdots,C}$).

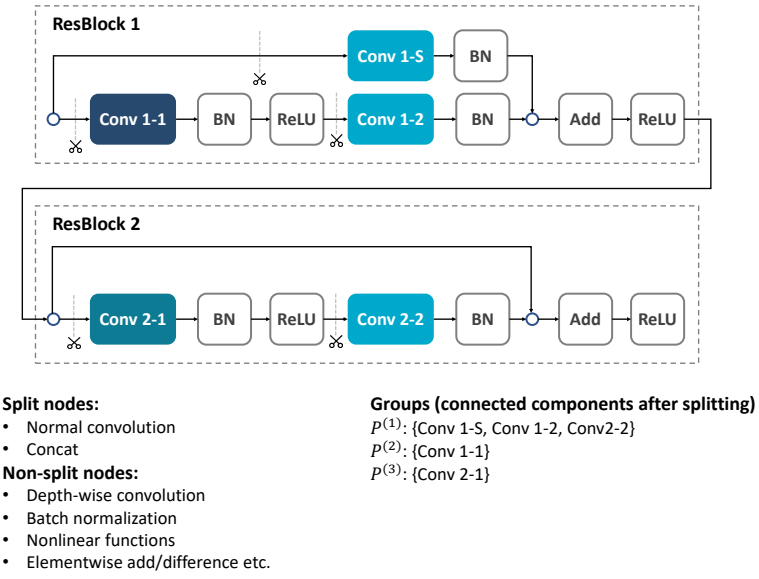


Fig. 1. An example of the topological grouping procedure

⋆ Both authors contributed equally to this work.

Algorithm 1 Topological grouping procedure

1: Construct the computational directed acyclic graph G
2: Removing all the incoming connections at split nodes (operation nodes that is not a channel-wise operation): normal convolution, concat operation.
NOTE: Non-split nodes include all channel-wise operations: depthwise convolution, element-wise add, ReLU, batch normalization, etc.
3: Find the connected components $\{P^{(k)}\}_{k=1,\cdots,K}$. All the convolution layers in each $P^{(k)}$ share the same keep ratio $\alpha^{(k)}$ and masks $m^{(k)}_{i=1,\cdots,C}$

A budget model $\mathcal{F}$ is needed for measuring the resource consumption $\mathcal{F}(\mathcal{A})$ corresponding to the sparsity allocation $\mathcal{A}$. Taking FLOPs as an example, $\mathcal{F}(\mathcal{A})$ could be represented as $\mathcal{F}(\mathcal{A}) = \mathcal{A}^T\mathcal{F}_A\mathcal{A} + \mathcal{F}_B^T\mathcal{A}$. We summarize the calculation procedure of $\mathcal{F}_A$ and $\mathcal{F}_B$ in Alg. 2.

Algorithm 2 Calculation of $\mathcal{F}_A, \mathcal{F}_B$ ($\mathcal{F}$ for FLOPs resource)

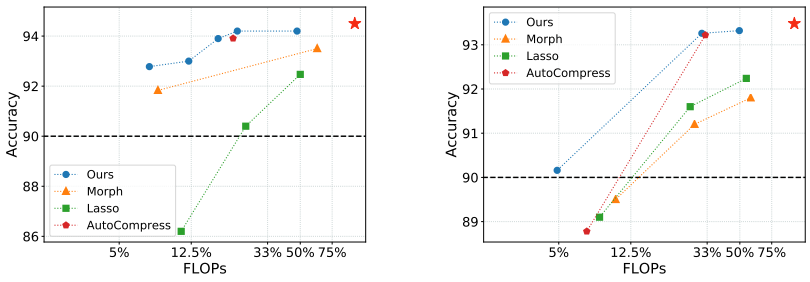
G: the directed acyclic graph of operations
K: number of connected components
$\mathcal{F}_A = \mathbf{0}_{K\times K}, \mathcal{F}_B = \mathbf{0}_K$
Convolution node attributes: 1) C: output channel number; 2) k: the index of the connected components that the convolution node belongs to; 3) kss: kernel spatial size (e.g. $3 \times 3 = 9$); 4) oss: output spatial size (e.g. $16 \times 16 = 256$)
for all convolution node M in G **do**
  if n.type == DEPTHWISE_CONV **then**
    $\mathcal{F}_B[M.k] += 2 \times M.c \times M.\text{kss} \times M.\text{oss}$
  else
    stack = [predecessor(M)]
    while stack **do**
      n = stack.pop()
      if n.type == CONCAT **then**
        for pn in predecessor(n) **do**
          stack.push(pn)
        end for
      else if n.type in ELEMENTWISE_OPs (e.g. ADD, ReLU) **then**
        stack.push(predecessor(n)[0])
      else if n.type == NORMAL_CONV **then**
        $\mathcal{F}_A[M.k, n.k] += 2 \times n.C \times M.C \times M.\text{kss} \times M.\text{oss}$
      else if n.type == INPUT **then**
        $\mathcal{F}_B[M.k] += 2 \times n.C \times M.C \times M.\text{kss} \times M.\text{oss}$
      else
        stack.push(n)
      end if
    end while
  end if
end for
return $\mathcal{F}_A, \mathcal{F}_B$

2 Additional Results

Table. 1 and Fig. 2 show the results of pruning ResNet-18 and VGG-16 on CIFAR-10.

Table 1. Pruning results of ResNet-18 and VGG-16 on CIFAR-10

Method	ResNet-18 (94.0%)		VGG-16 (93.48%)	
	Acc.	FLOPs ratio	Acc.	FLOPs ratio
AutoCompress	93.91%	4.7× (21.28%)	93.22%	3.1× (32.26%)
	-	-	88.78%	14.0× (7.14%)
DSA (Ours)	**94.19%**	4.5× (22.46%)	**93.26%**	3.2× (30.85%)
	93.90%	**5.7× (17.54%)**	90.16%	**20.4× (4.90%)**



(a) ResNet-18 (b) VGG-16

Fig. 2. Pruning results of ResNet-18, VGG-16 on CIFAR-10