Dynamic Slimmable Network
Supplementary Material

Changlin Li\textsuperscript{1}  Guangrun Wang\textsuperscript{2}  Bing Wang\textsuperscript{3}  Xiaodan Liang\textsuperscript{4}  Zhihui Li\textsuperscript{5}  Xiaojun Chang\textsuperscript{1}*

\textsuperscript{1}GORSE Lab, Dept. of DSAI, Monash University  \textsuperscript{2}University of Oxford  \textsuperscript{3}Alibaba Group  \\textsuperscript{4}Sun Yat-Sen University  \textsuperscript{5}Shandong Artificial Intelligence, Qilu University of Technology

\texttt{changlin.li@monash.edu, wanggrun@gmail.com, fengquan.wb@alibaba-inc.com, xdliang328@gmail.com, zhihuilics@gmail.com, xiaojun.chang@monash.edu}

Appendix

A. Implementation Details

Losses in Stage II. Complexity penalty loss $L_{cplx}$ is used to increase the model efficiency in training stage II. To provide a stable and fair constant, we use the number of multiply-adds on the fly, $\text{MAdds}(X, \theta)$, as the metrics of model complexity. Specifically, the complexity penalty is given by:

$$L_{cplx}(X, \theta) = \left(\frac{\text{MAdds}(X, \theta)}{T}\right)^2,$$

where $T$ is a normalize factor set to the total MAdds of the supernet in our implementation. Note that this loss term always pushes the gate to route towards a faster architecture, towards an architecture with target MAdds, which can effectively prevent routing easy and hard instances to the same architecture.

Overall, the slimming gate can be optimized with a joint loss function:

$$L(X, \theta) = \lambda_1 L_{cls} + \lambda_2 L_{cplx} + \lambda_3 L_{SGS}.$$  \hspace{1cm} (2)

The three balancing factors are set to $\lambda_1 = 1$, $\lambda_2 = 0.5$, $\lambda_3 = 1$ in our experiments. Different target MAdds is reached by adjusting the routing space during gate training. For instance, when training the gate of DS-MBNet-S, we set $\rho \in [0.35 : 0.05 : 0.5]$ to prevent routing to heavier sub-networks.

Equispaced channel group. Following previous works [14, 13], we set the the smallest division of channel number to 8. When using 0.05 as the interval of $\rho$, rounding channels by 8 may result in different intervals, which could lead to training failure when using Group Normalization [11]. To prevent such problem, we always adopt a consistent interval ($e.g.$ 8, 16, 32) in a single layer, instead of multiplying $\rho$ and rounding the channel. This results in a difference of the slimming ratio between our implemented architecture and our design.

Additional details. Weight decay is set to $1^{-4}$ in all of our experiments on ImageNet. To stabilize the optimization, weight decay of all the layers in the dynamic gate is removed. The weight $\gamma$ of the last normalization layer of each residual block is initialized to zeros following [15]. The weight of the fully-connected layer in channel attention head, $W_3$ of the main text, is also zero-initialized to ease the optimization following [12]. Additional training techniques include [2, 1]. We do not use label smoothing [7], DropPath [6] and RMSProp [10], which are popularly used in previous works [9, 4, 13, 14].

B. Experiments on EfficientNet

We also applied our method on EfficientNet [9], a state-of-the-art network family with high efficiency. Similar to our DS-MBNet, Dynamic Slimmable EfficientNet-B0 (DS-EffNet) has only one slimming gate after its 8-th inverted residual block, controlling the rest 8 blocks. The fixed slimming ratio for the first 8 blocks is 0.5, while a uniform dynamic slimming ratio $\rho \in [0.75 : 0.05 : 1.75]$ is used for the last 8 blocks. This supernet with 20 paths in total is trained with a similar config with the supernet of DS-ResNet and DS-MBNet.

We train the supernet with 512 total batch size using 0.2 learning rate that decays with a cosine scheduler in 150 epochs. To enable direct comparision, we opt to reproduce the EfficientNet results using our training setup, with a 150 epoch schedule and no extra enhancement of DropPath [6], RMSProp [10], etc.

The result is shown in Tab. 1. DS-EffNet outperforms the original EfficientNet-B0 by 0.7% and 0.8%, proving its efficacy on recent methods with inverted bottleneck blocks [8] and Squeeze-and-Excitation module [5].

C. Additional Ablations

Slimming gate. We analysis the improvement brought by slimming gate by comparing the performance of DS-Net and its supernet. As shown in Tab. 2, slimming gate boosts the
Table 1. Comparison of EfficientNet-B0 and DS-EffNet on ImageNet.

<table>
<thead>
<tr>
<th>Method</th>
<th>MAdds</th>
<th>Top-1 Acc.</th>
</tr>
</thead>
<tbody>
<tr>
<td>EffNet-B0 [9] (repro.)</td>
<td>399M</td>
<td>76.0</td>
</tr>
<tr>
<td>DS-EffNet-L (Ours)</td>
<td>400M</td>
<td>76.7</td>
</tr>
<tr>
<td>EffNet-B0 0.75×[9]</td>
<td>267M</td>
<td>74.6</td>
</tr>
<tr>
<td>DS-EffNet-S (Ours)</td>
<td>270M</td>
<td>75.4</td>
</tr>
</tbody>
</table>

Table 2. Ablation analysis of slimming gate.

<table>
<thead>
<tr>
<th>model</th>
<th>MAdds</th>
<th>Top-1 Acc.</th>
</tr>
</thead>
<tbody>
<tr>
<td>supernet (DS-MBNet)</td>
<td>140M</td>
<td>69.3</td>
</tr>
<tr>
<td>DS-MBNet-S</td>
<td>153M</td>
<td>70.1</td>
</tr>
<tr>
<td>supernet (DS-ResNet)</td>
<td>1.1B</td>
<td>73.4</td>
</tr>
<tr>
<td>DS-ResNet-S</td>
<td>1.2B</td>
<td>74.6</td>
</tr>
</tbody>
</table>

Table 3. Ablation analysis of distillation temperature $\tau$ (40 epochs).

<table>
<thead>
<tr>
<th>$\tau$</th>
<th>slimmest</th>
<th>widest</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>59.2</td>
<td>65.6</td>
</tr>
<tr>
<td>4</td>
<td>49.0</td>
<td>67.6</td>
</tr>
</tbody>
</table>

Performance of DS-MBNet-S and DS-ResNet-S by 0.8% and 1.2% respectively, comparing to sub-networks with similar sizes in their supernet.

**Distillation temperature.** Temperature $\tau$ in distillation loss was first introduced in [3] to control the smoothness of the target. Using a properly larger $\tau$ usually yields better performance of the student. Surprisingly, we find a huge performance degradation in the slimmest sub-network when using larger $\tau$ in in-place distillation. We test $\tau = 4$ with DS-MBNet for 40 epochs and compare the it with the performance of default setting ($\tau = 1$). As shown in Tab. 3, the performance of the slimmest sub-network decrease by 10.2% after applying the temperature $\tau = 4$.

References