Supplementary Materials for

*DyRep: Bootstrapping Training with Dynamic Re-parameterization*

A. More Ablation Studies

A.1. Performance on different operation spaces

Compared to the DBB [2] using 4 branches to build the augmented network, DyRep further adds 3 operations for more flexible structures. We now conduct experiments to compare our DyRep with DBB using the same operation spaces for fair comparisons. As summarized in Table 8, we train ResNet-50 using DyRep and DBB with 4 operations and 7 operations on ImageNet and report their validation accuracies. When using the same operation space as DBB, our DyRep can also enjoy significant efficiency and performance improvements. Besides, if we adopt DBB on our larger operation space, its training cost will be higher, and our superiority will be more significant. Comparing 4 branches and 7 branches, the larger one achieves better accuracy because of more diverse representations.

<table>
<thead>
<tr>
<th>Rep method</th>
<th>#operations</th>
<th>FLOPs</th>
<th>Params</th>
<th>Training cost</th>
<th>ACC (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Origin</td>
<td>1</td>
<td>4.09</td>
<td>25.6</td>
<td>7.5</td>
<td>76.14</td>
</tr>
<tr>
<td>DBB</td>
<td>4</td>
<td>6.79</td>
<td>40.7</td>
<td>13.7</td>
<td>76.71</td>
</tr>
<tr>
<td>DyRep</td>
<td>4</td>
<td>4.93</td>
<td>-27.4%</td>
<td>30.2 (-25.8%)</td>
<td>8.1 (-40.9%)</td>
</tr>
<tr>
<td>DBB</td>
<td>7</td>
<td>8.02</td>
<td>48.3</td>
<td>17.3</td>
<td>76.87</td>
</tr>
<tr>
<td>DyRep</td>
<td>7</td>
<td>5.05</td>
<td>-37.0%</td>
<td>31.5 (-34.8%)</td>
<td>8.5 (-50.9%)</td>
</tr>
</tbody>
</table>

A.2. Effects of different update intervals \(t\) of DyRep

We update the structures of the network in every \(t\) epoch. If the structures are expanded more frequently, the final network will be larger. As summarized in Table 9, we train the models with different update intervals \(t\), and the results show that for a small \(t\), the accuracy will be further improved, but the training cost also increases accordingly. For efficiency consideration, we choose a moderate frequency \(t = 15\) on CIFAR-10.

<table>
<thead>
<tr>
<th>Rep method</th>
<th>Update interval (t)</th>
<th>Cost (GPU hours)</th>
<th>FLOPs (M)</th>
<th>Params (M)</th>
<th>ACC (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>origin</td>
<td>-</td>
<td>2.4</td>
<td>313</td>
<td>15.0</td>
<td>94.68</td>
</tr>
<tr>
<td>DBB</td>
<td>-</td>
<td>9.4</td>
<td>728</td>
<td>34.7</td>
<td>94.97</td>
</tr>
<tr>
<td>DyRep</td>
<td>5</td>
<td>13.3</td>
<td>1575</td>
<td>83.4</td>
<td>95.39</td>
</tr>
<tr>
<td>DyRep</td>
<td>10</td>
<td>8.8</td>
<td>992</td>
<td>33.6</td>
<td>95.33</td>
</tr>
<tr>
<td>DyRep</td>
<td>15</td>
<td>6.9</td>
<td>597</td>
<td>26.4</td>
<td>95.22</td>
</tr>
<tr>
<td>DyRep</td>
<td>30</td>
<td>5.8</td>
<td>522</td>
<td>23.7</td>
<td>94.91</td>
</tr>
<tr>
<td>DyRep</td>
<td>50</td>
<td>4.1</td>
<td>430</td>
<td>20.3</td>
<td>94.82</td>
</tr>
</tbody>
</table>

A.3. Effects of different scoring metrics in our Rep

Many metrics [14, 24, 25] are proposed to measure the saliency score of weights in network pruning. In our paper, to choose the most suitable metric, we conduct experiments to evaluate these scoring metrics in DyRep. The experimented scoring metrics are summarized as follows.

For one operation with weights \(\theta\), its saliency score can be represented as following metrics:
• **random**: the score of each operation is generated randomly.

\[ S_o(\theta) \overset{iid}{\sim} \mathbb{R}^1. \] (11)

• **grad\_norm**: A simple baseline of summing the Euclidean norm of the gradients.

\[ S_o(\theta) = \left\| \frac{\partial L}{\partial \theta} \right\|_2. \] (12)

• **snip [14]**:

\[ S_o(\theta) = \frac{1}{n} \sum_{i} \left| \frac{\partial L}{\partial \theta_i} \circ \theta_i \right|. \] (13)

• **grasp**: [25] aims to improve **snip** by approximating the change in gradient norm (instead of loss), therefore its **grasp** metric is computed as

\[ S_o(\theta) = \frac{1}{n} \sum_{i} \left( \frac{\partial L}{\partial \theta_i} \circ \theta_i \right). \] (14)

where \( H \) denotes the Hessian matrix.

• **synflow**: SynFlow [24] proposes a modified version (**synflow**) to avoid layer collapse when performing parameter pruning.

\[ S_o(\theta) = \frac{1}{n} \sum_{i} \left| \frac{\partial L}{\partial \theta_i} \circ \theta_i \right|. \] (15)

• **vote**: Inspired by [?], which leverages the above metrics to vote the decisions, we also provide a result of picking operations with the most votes.

We measure all the metrics above on CIFAR-10 dataset, as summarized in Table 10. The random baseline even worsens the performance as it could introduce some unnecessary disturbances to the original weights, showing that it is important in choosing operations. Besides, **synflow** achieves the best performance compared to other metrics, and we thus adopt it as the scoring metric in our DyRep.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>origin</th>
<th>random</th>
<th>grad_norm</th>
<th>snip</th>
<th>grasp</th>
<th>synflow</th>
<th>vote</th>
</tr>
</thead>
<tbody>
<tr>
<td>CIFAR-10</td>
<td>94.68</td>
<td>94.25</td>
<td>94.71</td>
<td>94.97</td>
<td>94.82</td>
<td>95.22</td>
<td>95.03</td>
</tr>
<tr>
<td>CIFAR-100</td>
<td>74.10</td>
<td>74.03</td>
<td>74.19</td>
<td>74.56</td>
<td>74.73</td>
<td>74.91</td>
<td>74.79</td>
</tr>
</tbody>
</table>

**A.4. Effects of training with DyRep for different epochs**

To validate the effectiveness of DyRep in boosting training, we adopt DyRep for the first 20,40,60,80, and 100 epochs, then train the remained epochs with fixed structures. As illustrated in Figure 7, our DyRep can dynamically adapt the structures thus is more steady compared to training with fixed structures. Besides, adopting Rep can always obtain higher accuracy than fixed structures, showing that DyRep can improve performance in the whole training process.
Figure 7. Training curves of adopting DyRep for different epochs on CIFAR-100. DyRep-$N$ denotes training with DyRep for the first $N$ epochs then fixing the structure for the latter $100 - N$ epochs. DyRep with red line means adopting DyRep in the whole training.

References


