Parameter efficient training of deep convolutional neural networks by dynamic sparse reparameterization

Hesham Mostafa (Intel AI)
Xin Wang (Intel AI, Cerebras Systems)
**Easy**: post-training (sparse) compression

**Hard**: direct training of sparse networks
“Winning lottery tickets” (Frankle & Carbin 2018): post hoc identification of trainable sparse nets
Dynamic sparse reparameterization (ours): training-time structural exploration
Direct training sparse nets to generalize as well as post-training compression: 

*is this possible?*  - **YES**

Directly trained sparse nets:

*are they “winning lottery tickets”?*  - **NO**
Dynamic sparse reparameterization

1. for each sparse parameter tensor $W_i$ do
2.   $(W_i, k_i) \leftarrow \text{prune}\_\text{by}\_\text{threshold}(W_i, H)$  \hfill $\triangleright k_i$ is the number of pruned weights
3.   $l_i \leftarrow \text{number}\_\text{of}\_\text{nonzero}\_\text{entries}(W_i)$  \hfill $\triangleright$ Number of surviving weights after pruning
4. end for
5. $(K, L) \leftarrow (\sum_i k_i , \sum_i l_i)$  \hfill $\triangleright$ Total number of pruned and surviving weights
6. $H \leftarrow \text{adjust}\_\text{pruning}\_\text{threshold}(H, K, \delta)$  \hfill $\triangleright$ Adjust pruning threshold
7. for each sparse parameter tensor $W_i$ do
8.   $W_i \leftarrow \text{grow}\_\text{back}(W_i, \frac{L}{L} K)$  \hfill $\triangleright$ Grow $\frac{L}{L} K$ zero-initialized weights at random in $W_i$
9. end for
Closed gap between post-training compression and direct training of sparse nets

WRN-28-2 on CIFAR10

Global sparsity

Test accuracy%

Number of parameters (K)

Resnet-50 on Imagenet

<table>
<thead>
<tr>
<th>Sparsity (# Param)</th>
<th>0.8 (7.3M)</th>
<th>0.9 (5.1M)</th>
<th>0.0 (25.6M)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Thin dense</td>
<td>72.4</td>
<td>90.9</td>
<td></td>
</tr>
<tr>
<td></td>
<td>[-2.5]</td>
<td>[-1.5]</td>
<td></td>
</tr>
<tr>
<td>Static sparse</td>
<td>71.6</td>
<td>90.4</td>
<td>67.8</td>
</tr>
<tr>
<td></td>
<td>[-3.3]</td>
<td>[-2.0]</td>
<td>[-7.1]</td>
</tr>
<tr>
<td>DeepR</td>
<td>71.7</td>
<td>90.6</td>
<td>70.2</td>
</tr>
<tr>
<td>(Bellec et al., 2017)</td>
<td>[-3.2]</td>
<td>[-1.8]</td>
<td>[-4.7]</td>
</tr>
<tr>
<td>SET</td>
<td>72.6</td>
<td>91.2</td>
<td>70.4</td>
</tr>
<tr>
<td>(Mocanu et al., 2018)</td>
<td>[-2.3]</td>
<td>[-1.2]</td>
<td>[-4.5]</td>
</tr>
<tr>
<td>Dynamic sparse</td>
<td>73.3</td>
<td>92.4</td>
<td>71.6</td>
</tr>
<tr>
<td>(Ours)</td>
<td>[-1.6]</td>
<td>[0.0]</td>
<td>[-3.3]</td>
</tr>
<tr>
<td>Compressed sparse</td>
<td>73.2</td>
<td>91.5</td>
<td>70.3</td>
</tr>
<tr>
<td>(Zhu &amp; Gupta, 2017)</td>
<td>[-1.7]</td>
<td>[-0.9]</td>
<td>[-4.6]</td>
</tr>
</tbody>
</table>

Full dense
Compressed sparse
Dynamic sparse
Static sparse
DeepR
SET
Thin dense
Global sparsity
Directly trained sparse nets are not “winning tickets”: exploration of structural degrees of freedom is crucial.
Visit our poster:
Wednesday, Pacific Ballroom #248