Snapshots Ensembles: Train 1, Get $M$ for Free

Motivation

Training neural network ensembles can be expensive:
- Training a single model can last for days or weeks even on GPUs
- Cost for ensembles increases linearly

Question: Can we approximate ensembles at no additional training cost?

Approach: Train a single neural network, let it converge and escape from several local minima along its optimization path and ensemble the last $m$ models.

Cyclic learning rate scheduling (Loshchilov and Hutter 2017)

$$\alpha(t) = \frac{\epsilon}{2} \left( \frac{\text{mod}(t - 1, \frac{T}{M}) + 1}{T/M} \right)$$

Method

- The ensemble prediction at test time is the average of the last $m$ models ($m \leq M$) model’s softmax outputs.
- Always ensemble the last $m$ models, as the models tend to have the lowest test error.

Results: ImageNet (ResNet-50, $m=M$)

<table>
<thead>
<tr>
<th>Method</th>
<th>C10</th>
<th>C100</th>
<th>SVHN</th>
<th>Tiny ImageNet</th>
</tr>
</thead>
<tbody>
<tr>
<td>Single model</td>
<td>5.42</td>
<td>28.02</td>
<td>1.96</td>
<td>46.50</td>
</tr>
<tr>
<td>NoCycle Snapshot Ensemble</td>
<td>5.49</td>
<td>26.97</td>
<td>1.78</td>
<td>43.69</td>
</tr>
<tr>
<td>SingleCycle Ensembles</td>
<td>6.66</td>
<td>24.54</td>
<td>1.74</td>
<td>42.60</td>
</tr>
<tr>
<td>Snapshot Ensemble ($\alpha_t = 0.1$)</td>
<td>5.73</td>
<td>25.55</td>
<td>1.63</td>
<td>40.54</td>
</tr>
<tr>
<td>Snapshot Ensemble ($\alpha_t = 0.2$)</td>
<td>5.32</td>
<td>24.19</td>
<td>1.66</td>
<td>39.40</td>
</tr>
</tbody>
</table>

Diversity of Model Ensembles: Parameters

- Sharp minima: mixing snapshot parameters worsens the performance.
- Flat minima: Snapshot parameters lie in the same minimum as the final model.

Diversity of Model Ensembles: Activations

- Large correlations between the last 3 snapshots of the non-cyclic training schedule (right) - less diversity
- Conversely, there is less correlation between all cyclic snapshots (left) - better balance between diversity and accuracy

Fix number of cycles $M = 6$ and vary training budget $B$ from 60 to 300.

Compute the pairwise correlation of softmax outputs for every pair of snapshots.