NAS-Bench-201: Extending the Scope of Reproducible Neural Architecture Search
Xuanyi Dong, Yi Yang

INTRODUCTION
We propose NAS-Bench-201, a new NAS benchmark.

Motivation:
Different NAS methods have different setups, which raises a comparability problem when comparing their performance.

Highlighted features:
- Loss and accuracy on training / validation / test sets are provided for every epoch.
- Results of 15625 architectures on three datasets are provided.
- Results trained with two different kinds of hyper-parameters are provided.
- The architecture space is agnostic to all NAS algorithms.
- The weights of all trained architectures are provided.
- 10 NAS algorithms are open sourced in one code base.

ARCHITECTURE SPACE
Top: the macro skeleton of each architecture candidate. Bottom-left: examples of neural cell with 4 nodes. Each cell is a directed acyclic graph, where each edge is associated with an operation selected from a predefined operation set.

BENCHMARK 10 NAS METHODS

<table>
<thead>
<tr>
<th>Method</th>
<th>Search (seconds)</th>
<th>CIFAR-10 validation</th>
<th>CIFAR-10 test</th>
<th>ImageNet-16-120 validation</th>
<th>ImageNet-16-120 test</th>
</tr>
</thead>
<tbody>
<tr>
<td>RSPS</td>
<td>8007.13</td>
<td>80.42±1.35±5.88</td>
<td>84.07±3.61</td>
<td>52.12±5.55</td>
<td>52.31±5.77</td>
</tr>
<tr>
<td>DARTS-V1</td>
<td>11625.77</td>
<td>39.77±1.00</td>
<td>54.30±1.00</td>
<td>15.03±0.00</td>
<td>15.61±0.00</td>
</tr>
<tr>
<td>DARTS-V2</td>
<td>35781.80</td>
<td>39.77±1.00</td>
<td>54.30±1.00</td>
<td>15.03±0.00</td>
<td>15.61±0.00</td>
</tr>
<tr>
<td>GDAS</td>
<td>31609.80</td>
<td>89.89±0.08</td>
<td>93.61±0.09</td>
<td>71.34±0.04</td>
<td>70.70±0.30</td>
</tr>
<tr>
<td>SETN</td>
<td>34139.53</td>
<td>84.04±1.28</td>
<td>87.64±1.00</td>
<td>58.86±1.06</td>
<td>59.05±0.24</td>
</tr>
<tr>
<td>ENAS</td>
<td>14058.80</td>
<td>37.51±1.19</td>
<td>53.89±0.58</td>
<td>13.37±2.35</td>
<td>13.96±2.33</td>
</tr>
</tbody>
</table>

TAKE AWAY
(1) NAS-Bench-201 helps you fairly and quickly compare your NAS method with others.
(2) The correlation of the model’s performance between different datasets is not high.
(3) Using the batch mean and var of BN for searching algorithm instead of accumulated mean and var.

Searching results of 10 NAS methods on 3 datasets