Improving AutoTVM Efficiency by Schedule Sharing

AWS AI

Cody Yu
AWS Deep Engine Science
Motivation

- Tuning MobileNet V2 1.0 on Nvidia V100 for 500 trails

31 tasks, including:
- conv2d
- depthwise conv2d

One task took about 40 mins $\rightarrow$ ~19 hours in total
Motivation

• Observations
  • AutoTVM tunes every task in a model from scratch
  • The high quality schedules of some tasks in a model are similar

• Proposed Solution
  • Only tune representative tasks and share their best schedules to others
Schedule Sharing among AutoTVM Tasks

- **Idea**
  - Two tasks can share schedules if their tuning spaces are similar

- **Approach**
  - Cluster tasks with a distance metric
  - Identify a centric (representative) task in each cluster and only tune them
  - Share their best schedules to other tasks in the same cluster

[Diagram showing nodes and edges with clustering and determination of representative tasks]
Evaluation Results and RFC

- Selected 7 models from Gluon CV model zoo
- Tuned selected tasks (~5-7) for 3,000 trials on Nvidia V100
- Achieve on average 84% performance with only 28% tuning time

<table>
<thead>
<tr>
<th>Model</th>
<th>Tuning Time w/o Sharing (mins)</th>
<th>Perf. w/o Sharing (ms)</th>
<th>Tuning Time w. Sharing (mins)</th>
<th>Perf. w. Sharing (ms)</th>
<th>Used Time</th>
<th>Achieve Perf.</th>
</tr>
</thead>
<tbody>
<tr>
<td>MobileNet V2 1.0</td>
<td>1185</td>
<td>0.74</td>
<td>404</td>
<td>0.78</td>
<td>34%</td>
<td>95%</td>
</tr>
<tr>
<td>ResNet 50 V1</td>
<td>1666</td>
<td>2.27</td>
<td>358</td>
<td>3.7</td>
<td>21%</td>
<td>61%</td>
</tr>
<tr>
<td>VGG 19 BN</td>
<td>479</td>
<td>5.08</td>
<td>169</td>
<td>6.36</td>
<td>35%</td>
<td>80%</td>
</tr>
<tr>
<td>SqueezeNet 1.1</td>
<td>574</td>
<td>0.54</td>
<td>167</td>
<td>0.5</td>
<td>29%</td>
<td>108%</td>
</tr>
<tr>
<td>DenseNet 121</td>
<td>2670</td>
<td>2.99</td>
<td>377</td>
<td>3.02</td>
<td>14%</td>
<td>99%</td>
</tr>
<tr>
<td>Yolo3 MobileNet1.0 voc</td>
<td>2784</td>
<td>5.4</td>
<td>774</td>
<td>7.16</td>
<td>28%</td>
<td>75%</td>
</tr>
<tr>
<td>SSD512 ResNet50 V1 voc</td>
<td>3426</td>
<td>8.47</td>
<td>1150</td>
<td>5.65</td>
<td>34%</td>
<td>67%</td>
</tr>
</tbody>
</table>

Average 28% 84%

⇧ RFC ⇧
Welcome to Participate!