Supplementary Materials for Neural Architecture Search for Joint Human Parsing and Pose Estimation

Dan Zeng\textsuperscript{1}  Yuhang Huang\textsuperscript{1}  Qian Bao\textsuperscript{2}  Junjie Zhang\textsuperscript{1}  Chi Su\textsuperscript{3}  Wu Liu\textsuperscript{2}
\textsuperscript{1}Shanghai University  \textsuperscript{2}JD AI Research  \textsuperscript{3}Kingsoft Cloud
dzeng@shu.edu.cn, huangyuhang@shu.edu.cn, baoqian@jd.com
junjie_zhang@shu.edu.cn, suchi@kingsoft.com, liuwul@jd.com

1. Loss Function

We use the Mean Square Error as the loss function for pose branch while the Cross-Entropy loss is employed for parsing branch. Therefore, the loss function of each branch can be calculated by:

\[
L_{\text{pos}} = L_{\text{MSE}}(P_{\text{pos}}, G_{\text{pos}}) + L_{\text{MSE}}(P_{\text{pos,aux}}, G_{\text{pos,aux}})
\]
\[
L_{\text{par}} = L_{\text{CE}}(P_{\text{par}}, G_{\text{par}}) + L_{\text{MSE}}(P_{\text{edg}}, G_{\text{edg}})
\]

where \( P \) denotes prediction and \( G \) denotes ground truth. Since the two loss values are quite different, referring to [1], we adopt the uncertainty loss to learn the weights for the two branches.

\[
L = e^{-\sigma_1} \cdot L_{\text{pos}} + \sigma_1 + e^{-\sigma_2} \cdot L_{\text{par}} + \sigma_2
\]

where \( \sigma_1 \) and \( \sigma_2 \) are learnable parameters that balance the two tasks.

Since we use supervision in both two stages of forward, in order to distinguish between the two losses, we denote the losses as \( L_1 \) and \( L_2 \) respectively, so the total loss \( L_{\text{total}} \) can be formulated as:

\[
L_{\text{total}} = L_1 + L_2
\]

2. Pruning Schemes

After searching, we combine two pruning schemes. First of all, since the search of encoder-decoder and feature fusion are both cell-based, we apply the pruning scheme in DARTS [3]. However, the search for the multi-scale feature interaction is not cell-based, and the connections are much denser. More valuable connections should be retained to ensure the efficiency of the network after pruning. Therefore, we design a new pruning scheme and its pseudo-code is in Algorithm 1.

Algorithm 1 Pruning algorithm used in multi-scale feature interaction search.

\textbf{Input:} The proportion of the operation \( o(\cdot) \) in connection from node \( j \) to node \( i \), \( \gamma_{i,j}^o \), \( j \leq i \);

\textbf{Output:} The remaining operations list, \( \text{remain}_o \);

1: Initialize \( s = 0, n = 0, \text{remain}_o = [ ] \);
2: while \( s \leq 0.7 \) and \( n \leq 4 \) do
3: \hspace{0.5cm} \( p = \max(\gamma_{i,j}) \);
4: \hspace{0.5cm} \( o', i', j' = p.\text{index}() \);
5: \hspace{0.5cm} \( \text{remain}_o.\text{append}(o \text{ from } j' \text{ to } i') \)
6: \hspace{0.5cm} \( s = s + p, n = n + 1 \)
7: end while
8: return \( \text{remain}_o \);

Reduction Cell 2, and Decoding Cell 2 are used in the parsing branch. The encoder-decoder of the pose branch has more dilated convolutions while the parsing branch prefers to use standard convolution. Moreover, the cells in the parsing branch are deeper. These all show there are differences between the pose branch and the parsing branch which suggests using different architecture to extract task-specific features for the two tasks. For feature fusion, the parsing feature is fused with both the pose feature and the pose auxiliary feature, which means the parsing branch assists the pose branch more. On the other hand, the pose feature is fused with the edge feature in the parsing cell indicates pose information may help enhance the boundary of each area.

Fig. 2 shows the connections of intermediate features between the two tasks. The pose branch absorbs more connections from the parsing branch while the parsing branch has fewer connections from another branch. This is consistent with our experiments that parsing information helps all types of key points while the promotion of human parsing is concentrated in a few special classes.

3. Architecture Detail

We design three search spaces, where encoder-decoder search and high-level feature search are cell-based, and multi-scale feature interaction search is not. We show the searched architecture in Fig. 1 and Fig. 2 respectively. In Fig. 1, the Normal Cell 1, Reduction Cell 1, and Decoding Cell 1 are used in the pose branch while the Normal Cell 2,
tational cost which shows the great potential of NPPNet to explore the interaction of the two branches. Note that the pose estimation networks are usually smaller than human parsing networks, and the MuLA [4] is mainly designed for pose estimation and its performance on human parsing is less satisfactory.

Table 1. Model size and computational cost on LIP dataset.

<table>
<thead>
<tr>
<th>Method</th>
<th>Param</th>
<th>FLOPs</th>
<th>mIOU</th>
<th>PCK</th>
<th>Task</th>
</tr>
</thead>
<tbody>
<tr>
<td>CNIF [6]</td>
<td>83.6M</td>
<td>300.0G</td>
<td>57.74</td>
<td>-</td>
<td>parsing only</td>
</tr>
<tr>
<td>HRNet [5]</td>
<td>60.6M</td>
<td>44.0G</td>
<td>-</td>
<td>88.0</td>
<td>pose only</td>
</tr>
<tr>
<td>JPPNet [2]</td>
<td>89.1M</td>
<td>263.5G</td>
<td>51.37</td>
<td>82.7</td>
<td>both tasks</td>
</tr>
<tr>
<td>MuLA [4]</td>
<td>44.4M</td>
<td>43.4G</td>
<td>49.30</td>
<td>-</td>
<td>both tasks</td>
</tr>
<tr>
<td>NPPNet(ours)</td>
<td>73.4M</td>
<td>113.7G</td>
<td>58.56</td>
<td>88.9</td>
<td>both tasks</td>
</tr>
</tbody>
</table>

References


