**Motivation**

**Problem Description:**
- Automatically design CNN architectures that surpass human expert designs: Neural Architecture Search (NAS)
- Part of the AutoML initiative
- Modern CNNs usually follow a two-level hierarchy:
  - Inner cell level governs specific layer-wise computations
  - Outer network level controls spatial resolution changes
- Most previous approaches:
  - Focused on image classification
  - Search inner cell level; hand-specify outer network level

Our Goal:
- NAS for dense image prediction: semantic segmentation
- Challenge 1: Search outer network level
- Challenge 2: Computationally friendly

**Architecture Search Space**

**Inner Cell Level**
- $10^{14}$ different architectures
  - Same as the one used in NASNet, PNASNet, DARTS...
  - Each cell consists of $B = 5$ blocks

**Outer Network Level (NEW)**
- $10^5$ different architectures
  - Next layer is either twice as large, or twice as small, or same
  - The smallest spatial resolution is downsampled by 32
  - NAS = find a good path in this $L$-layer trellis

**Method**

**Continuous Relaxation of Architectures**

- **Cell Architecture**
  
  \[ O_{j \to i}(H') = \sum_{O' \in O} a_{j \to i} O'(H') \]  

  where $a_{j \to i}$ are normalized scalars, implemented as softmax.

  The cell level update may be summarized as:

  \[ H^l = \text{Cell}(H^{l-1}, \beta; \alpha) \]  

- **Network Architecture**
  Associated a scalar $\beta$ with each gray arrow:

  \[ + \beta_{2 \to s} \text{Cell}(s H^{l-1}, s H^{l-2}; \alpha) \]

  where $s = 4, 8, 16, 32$ and $l = 1, 2, \ldots, L$. The scalars $\beta$ are normalized such that $\beta_{2 \to s} + \beta_{s \to 2s} = 1$ for $s, l$.

**Optimization**

1. Update network weights $w$ by

   \[ \nabla_w L_{\text{trainA}}(w; \alpha, \beta) \]

2. Update architecture $\alpha, \beta$ by

   \[ \nabla_{\alpha, \beta} L_{\text{trainB}}(w; \alpha, \beta) \]

**Decoding Discrete Architectures**

- **Cell Architecture:** Greedy argmax
- **Network Architecture:** Viterbi algorithm

**Found Architecture: Auto-DeepLab**

**Experiments & Results**

**About the Auto-DeepLab Architecture**

- Downsample in the first 3/4 layers; upsample in the last 1/4
- Atrous conv often used; learned the importance of context

**Results on Cityscapes**

<table>
<thead>
<tr>
<th>Method</th>
<th>ImageNet</th>
<th>Multi-Adds</th>
<th>Params</th>
<th>mIOU (val)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Auto-DeepLab-L</td>
<td></td>
<td>333.25B</td>
<td>10.15M</td>
<td>79.74</td>
</tr>
<tr>
<td>Auto-DeepLab-M</td>
<td></td>
<td>461.93B</td>
<td>21.62M</td>
<td>80.04</td>
</tr>
<tr>
<td>Auto-DeepLab-S</td>
<td></td>
<td>695.03B</td>
<td>44.42M</td>
<td>80.33</td>
</tr>
<tr>
<td>FRRN-A</td>
<td></td>
<td></td>
<td>17.76M</td>
<td></td>
</tr>
<tr>
<td>FRRN-B</td>
<td></td>
<td></td>
<td>24.78M</td>
<td></td>
</tr>
<tr>
<td>DeepLabv3+</td>
<td>✔</td>
<td>1551.05B</td>
<td>43.48M</td>
<td>79.55</td>
</tr>
</tbody>
</table>

**Results on PASCAL VOC 2012 (test set)**

<table>
<thead>
<tr>
<th>Method</th>
<th>ImageNet</th>
<th>COCO</th>
<th>mIOU (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Auto-DeepLab-L</td>
<td>✔</td>
<td></td>
<td>82.5</td>
</tr>
<tr>
<td>Auto-DeepLab-M</td>
<td>✔</td>
<td></td>
<td>84.1</td>
</tr>
<tr>
<td>Auto-DeepLab-S</td>
<td>✔</td>
<td></td>
<td>85.6</td>
</tr>
<tr>
<td>PSPNet</td>
<td>✔</td>
<td>✔</td>
<td>85.4</td>
</tr>
<tr>
<td>DeepLabv3+</td>
<td>✔</td>
<td>✔</td>
<td>87.8</td>
</tr>
</tbody>
</table>

**Results on ADE20K (val set)**

<table>
<thead>
<tr>
<th>Method</th>
<th>ImageNet</th>
<th>mIOU (%)</th>
<th>Pixel-Acc (%)</th>
<th>Avg (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Auto-DeepLab-L</td>
<td></td>
<td>40.69</td>
<td>80.60</td>
<td>60.65</td>
</tr>
<tr>
<td>Auto-DeepLab-M</td>
<td></td>
<td>42.19</td>
<td>81.09</td>
<td>61.64</td>
</tr>
<tr>
<td>Auto-DeepLab-S</td>
<td></td>
<td>43.98</td>
<td>81.72</td>
<td>62.85</td>
</tr>
<tr>
<td>PSPNet</td>
<td>✔</td>
<td>43.51</td>
<td>81.38</td>
<td>62.45</td>
</tr>
<tr>
<td>DeepLabv3+</td>
<td>✔</td>
<td>45.65</td>
<td>82.52</td>
<td>64.09</td>
</tr>
</tbody>
</table>

**Conclusion**

- **NOVEL:** One of the first attempts to extend NAS beyond image classification to dense image prediction
- **CHALLENGING:** A network level search space that augments and complements the cell level one; joint, hierarchical search
- **EFFICIENT:** 3 GPU days on $321 \times 321$ Cityscapes image crops
- **COMPETITIVE:** Auto-DeepLab (always trained from scratch) outperforms other models trained from scratch significantly, and is comparable with other ImageNet-pretrained models