Problem

- Task: recognize objects in the image
- Challenges: visual separability between different object categories is highly uneven
- Solution: Hierarchical Deep Convolutional Neural Network (HD-CNN)

Visual Category Hierarchy

- Learn visual category hierarchy from data
- Coarse / Fine category classifiers
- Final prediction = a weighted combination of predictions from fine category classifiers

```
Root
  Coarse category 1:
    {giraffe, rhino, ...}
  Coarse category 2:
    {tiger, elephant, ...}
  Coarse category K:
    {deer, monkey, ...}
```

- Confusion matrix $F$ from validation set
- Transform $F$ into a distance matrix $D$
  
  \[
  D = 0.5 * \left( (I - F) + (I - F)^T \right)
  \]
- Perform spectral clustering on $D$ and obtain a disjoint many-to-one mapping $p_{\text{disjoint}}: \{1, C\} \rightarrow \{1, K\}$
- Add more fine categories to coarse categories, which become overlapping to each other

HD-CNN Testing

- HD-CNN scalability
- Conditional Executions
  - Only execute relevant fine category classifiers
- Parameter Compression
  - Parameters in Conv / FC layer are matrices
  - Production Quadrant: K-means clustering of matrix rows

Results on ImageNet

<table>
<thead>
<tr>
<th>Method</th>
<th>ImageNet-NIN baseline</th>
<th>HD-CNN Instantiations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Top-1, Top-5</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Results on CIFAR100

<table>
<thead>
<tr>
<th>Method</th>
<th>Model</th>
<th>Comparisons of Computational Cost</th>
</tr>
</thead>
<tbody>
<tr>
<td>Top-1, Top-5</td>
<td>Base: ImageNet-NIN</td>
<td>Base: CIFAR100-NIN</td>
</tr>
<tr>
<td></td>
<td>Model averaging (3 base nets)</td>
<td>HD-CNN with SL</td>
</tr>
<tr>
<td></td>
<td>HD-CNN, disjoint CC</td>
<td>HD-CNN</td>
</tr>
<tr>
<td></td>
<td>HD-CNN</td>
<td>HD-CNN+NCE</td>
</tr>
<tr>
<td></td>
<td>HD-CNN+NCE+PC</td>
<td>HD-CNN+NCE+PC</td>
</tr>
<tr>
<td></td>
<td>ImageNet-NIN</td>
<td>Base: ImageNet-NIN</td>
</tr>
<tr>
<td></td>
<td>ImageNet-VGG-16-layer</td>
<td>Base: ImageNet-VGG-16-layer</td>
</tr>
</tbody>
</table>

Case Study

ImageNet-VGG-16-layer baseline

<table>
<thead>
<tr>
<th>Method</th>
<th>ImageNet-NIN baseline</th>
<th>HD-CNN Instantiations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Top-1, Top-5</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Notations: C = Conditional Execution, P = Parameter Compression