Internal Representations in Deep Networks

Alan Yuille
Dept. Cognitive Science and Computer Science
Johns Hopkins University
Visual Concepts: Internal Representations

• We study internal representations within Deep Nets.
• We restrict ourselves to study vehicles at fixed scale from the Pascal3D+ dataset.
• We showed that visual concepts, encoded by feature populations, represented subparts of the vehicles.
• We quantified the visual concepts for a series of tasks including semantic part detection under occlusion.
Background

• It has been shown (e.g., B. Zhou et al. ICLR 2015) that deep nets contain internal representations represented by neural features. The findings included:

• (I) If Deep Nets are trained to perform scene recognition, then the internal representations correspond to objects.

• (II) If Deep Nets are trained to perform object recognition, then the internal representations correspond to object parts.
Visual Concepts

• We conjectured that subparts of objects are encoded by populations of feature vectors – instead of by features themselves.

• These *visual concepts* were found by clustering the feature vectors. We restricted ourselves to vehicles from Pascal3D+ and fixed the scale of the objects.
Visual Concepts: Clustering

• The clustering was done using k-means with k=200 (alternative clustering methods, and alternative values of k gave similar results).

• The clustering was done at different levels of the Deep Net. E.g., Pool3, Pool4, Pool5. Results were similar for AlexNet and VGG.

• Visual Concepts correspond to parts of objects. VCs at higher layers correspond to larger parts (e.g., Pool4 wheel, Pool3 wheel-part).
Visual Concepts: Perceptually Tight

- **Findings 1:** The visual concepts were perceptually tight. Image patches corresponding to the same visual concept are very similar.

- We show the closest 6 image patches (left), a random sample of 6 patches from the top 500 image patches (center), and the mean of the edge map and of the patches of the top 500 patches (right).
Visual Concepts: Coverage of the Object

- *Visual Concepts respond to (cover) almost all parts of the object.*
- Here are 44 (out of 170) VCs for cars.
- This can be quantified, by showing that the objects could be represented in terms of VCs by binary encoding (see later).
To Explore: We Annotate Semantic Parts.

- We annotated the vehicles in PASCAL 3D+.
  To create the *Vehicle Semantic Part dataset*. 
VCs as Key-Point, Semantic Part Detectors.

• VCs were fairly good for detecting key-points and semantic parts of the Vehicles. But much worse than supervised models.

• Key-Points.
  13 K-Ps for Bike.

<table>
<thead>
<tr>
<th>Bike</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
</tr>
</thead>
<tbody>
<tr>
<td>SF</td>
<td>.77</td>
<td>.84</td>
<td>.89</td>
<td>.91</td>
<td>.94</td>
<td>.92</td>
<td>.94</td>
<td>.91</td>
<td>.91</td>
<td>.56</td>
<td>.53</td>
<td>.15</td>
<td>.40</td>
</tr>
<tr>
<td>VC</td>
<td>.91</td>
<td>.95</td>
<td>.98</td>
<td>.96</td>
<td>.96</td>
<td>.96</td>
<td>.97</td>
<td>.96</td>
<td>.97</td>
<td>.73</td>
<td>.69</td>
<td>.19</td>
<td>.50</td>
</tr>
</tbody>
</table>

• Semantic Parts.
  Yellow bars show the best APs for each VC.
VCs detect subparts of Semantic Parts

• VCs can act as unsupervised detectors for key-points and semantic-parts. Their Average Precisions (APs) are weaker than supervised methods.

• We observe that most VCs respond to several different semantic parts (typically 1-4). The VCs correspond to subparts of semantic parts (which are shared).
Combining VCs to detect Semantic Parts

- We design a compositional model for detecting semantic parts. Each model consists of a set of VCs which fire in different spatial positions. (Illustrated for object – car – instead of semantic part).
- Compositional Voting: each VC votes for the semantic part (depending on spatial position).
Semantic Part Detection with Occlusion

• We introduce occlusion to make semantic part detection more challenging. *Vehicle Occlusion Dataset.*

• Our intuition is that Deep Nets have difficulty with occlusion. *But compositional voting is likely to be most robust.* The occluded VC will not respond, but the un-occluded VCs will still vote.

• *Compositional voting* also includes context, image information outside the semantic part, because this is also robust.
Detecting Semantic Parts with Occlusion

- In the occlusion dataset semantic parts can be: (i) fully occluded (red) (ii) partially occluded (blue) (iii) un-occluded (yellow).

- Compositional voting uses VCs on and off the semantic parts. If a VC is detected (green) then it votes for the semantic part. If a VC is occluded (red) then it gives no vote.

- Note: a semantic part can be detected even if it is fully occluded.
Compositional Voting: Detect Semantic Parts

- The compositional voting method (VT) outperforms alternatives like Deep Nets if there is significant occlusion.

- **Main idea:** explicit representation of subparts (by VC) enables the algorithm to switch them on and off automatically. This makes them robust to occlusion.

<table>
<thead>
<tr>
<th>Object</th>
<th>2 Occ's, 0.2 ≤ r &lt; 0.4</th>
<th>3 Occ's, 0.4 ≤ r &lt; 0.6</th>
<th>4 Occ's, 0.6 ≤ r &lt; 0.8</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>SV</td>
<td>FR</td>
<td>VT</td>
</tr>
<tr>
<td>airplane</td>
<td>12.0</td>
<td>26.8</td>
<td>23.2</td>
</tr>
<tr>
<td>bicycle</td>
<td>44.6</td>
<td>65.7</td>
<td>71.7</td>
</tr>
<tr>
<td>bus</td>
<td>12.3</td>
<td>41.3</td>
<td>31.3</td>
</tr>
<tr>
<td>car</td>
<td>13.4</td>
<td>35.9</td>
<td>35.9</td>
</tr>
<tr>
<td>motorbike</td>
<td>11.4</td>
<td>35.9</td>
<td>44.1</td>
</tr>
<tr>
<td>train</td>
<td>4.6</td>
<td>20.0</td>
<td>21.7</td>
</tr>
<tr>
<td>mean</td>
<td>16.4</td>
<td>37.6</td>
<td>38.0</td>
</tr>
</tbody>
</table>

Visual Concepts: Summary

• The Deep Nets encode representations of the parts. These are stored by the activity patterns of the feature vectors (individual features were less successful – quantitatively). Note: vehicles only (rigid classes) and fixed scale.

• Making this representation explicit – e.g., by compositional voting – enables us to detect semantic parts despite heavy occlusion. The algorithm can automatically switch off subparts (VCs) if they are not detected in the correct locations.

• It is harder for Deep Nets to deal with occluders, because their representations are not explicit, so it is difficult to switch parts off.

• Can we extend this too classify objects?