Task: Fine-grained recognition

- **Example**: distinguish the two bird species
  - California gull
  - Ringed beak gull

- **Challenge**: intra-category variation vs. inter-category variation
  - Location, pose, viewpoint, background, lighting, gender, etc

Approach 1: Part-based models

- **Localize parts and compare** corresponding locations

Approach 2: Texture-based models

- **Image as a collection of patches**

Goal: combine the best of both approaches

Proposed approach: Bilinear CNN model

- **Bilinear model** is a four tuple:
  \[ f : \mathcal{L} \times I \rightarrow R^{D} \]
  \[ B = (f_A, f_B, f_C, f_D) \]

  - **Feature extractor**
  - **Pooling**
  - **Classification**

- **Classification pipeline**:
  1. For each location \( i \), extract features \( f_A(l, I) \) and \( f_B(l, I) \)
  2. Take the outer product: \( \text{bilinear}(l, I) = f_A(l, I)^T \cdot f_B(l, I) \)
  3. Pool across locations: \( \Phi(I) = \sum \text{bilinear}(l, I) \)
  4. Predict class probability: \( \text{softmax}(\Phi(I)) \)

- **Motivation**:
  - Model pairwise feature interactions in a translationally invariant manner
  - Compositional features — \( O(n^2) \) representation with \( O(n) \) features
  - End-to-end learning of parameters

Experiments

- **Classification accuracy**:
  - Using image labels only (no part or bounding-box annotations)
  - CNNs used: VGG-M [M] (Chatfield et al.) and VGG-VERYDEEP-16 [D] (Simonyan et al.)

<table>
<thead>
<tr>
<th>Method</th>
<th>CUB-200-2011</th>
<th>FGVC-Aircraft</th>
<th>Stanford Cars</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fisher vector SIFT</td>
<td>18.8 w/ ft</td>
<td>61.0 w/ ft</td>
<td>59.2 w/ ft</td>
</tr>
<tr>
<td>Fully connected CNN</td>
<td>52.7 w/ ft</td>
<td>44.4 w/ ft</td>
<td>37.3 w/ ft</td>
</tr>
<tr>
<td>FC-CNN [M]</td>
<td>61.0 w/ ft</td>
<td>45.0 w/ ft</td>
<td>36.5 w/ ft</td>
</tr>
<tr>
<td>FC-CNN [D]</td>
<td>61.0 w/ ft</td>
<td>45.0 w/ ft</td>
<td>36.5 w/ ft</td>
</tr>
<tr>
<td>Fisher vector CNN</td>
<td>61.1 w/ ft</td>
<td>64.3 w/ ft</td>
<td>70.8 w/ ft</td>
</tr>
<tr>
<td>FC-CNN [M]</td>
<td>71.3 w/ ft</td>
<td>70.4 w/ ft</td>
<td>75.2 w/ ft</td>
</tr>
<tr>
<td>FC-CNN [D]</td>
<td>72.0 w/ ft</td>
<td>72.7 w/ ft</td>
<td>77.8 w/ ft</td>
</tr>
<tr>
<td>Bilinear CNN</td>
<td>80.1 w/ ft</td>
<td>78.4 w/ ft</td>
<td>83.9 w/ ft</td>
</tr>
<tr>
<td>Bilinear CNN</td>
<td>80.1 w/ ft</td>
<td>84.0 w/ ft</td>
<td>84.1 w/ ft</td>
</tr>
</tbody>
</table>

- **Previous Work**: 84.1 [w/o b-box], 73.9 [w/o b-box], 75.7 [w/o b-box]

- **Visualizations on the CUB dataset**
  - Top activations of various conv5 filters in the fine-tuned B-CNN [D,M] model

- **Effect of fine-tuning**:
  - FC-CNN: big improvements on aircrafts (29%) and cars (43%)
  - FV-CNN: indirect fine-tuning, i.e. using fine-tuned FC-CNN for FV-CNN, leads to 3-10% improvement
  - B-CNN: Fine-tuning improves results (4-7%)
  - Fairly efficient during testing: 10 fps for the B-CNN [D,D] model
  - Translational invariance: 84.1% (w/o b-box) vs. 85.1% (w/o b-box)

References