Webly Supervised Learning of Convolutional Network

Author: Xinlei Chen, Abhinav Gupta
Presented by: Vicky Xuening Wang

ECS 289G, Nov 2015, UC Davis
Motivation

• Deeper Network + More Training Data => Better CNN

• Inspired by curriculum learning

• Two-step approach for learning CNNs:
  1. Train an initial visual representation with easy images
  2. Fine-tuning the CNN with harder images
Web Data

• Easy Images:
  • e.g. results of Google search engine
  • Biased

• Harder Images
  • e.g. Flickr style images
  • Noisy
Main Idea

BLVC in Caffe
Main idea

• First, bootstrap CNN training with easy examples

• Then, conduct a more extensive and comprehensive learning procedure with similarity constraints to learn visual representation

• Constraints: Relationship graph learned from easy examples
Network Set-up

• Initialization:

• Category - a combination of categorical lists from ImageNet Challenge, SUN database and NEIL knowledge base

• Query from Google Image search engine
  • ~600 per category -> exclude unreadable

• Query from Flickr with tag search
  • ~1500 categories close to uniform distribution
Similarity constraint

- A relationship graph represented by a matrix
- Learned from the first stage, and applied to the second stage
- Remove noisy example mistake
• confusion matrix

\[ R_{ij} = P(i|j) = \frac{\sum_{k \in C_i} CNN(j|I_k)}{|C_i|}, \quad (1) \]

• Rij is used to characterize the label-flip noise

• Softmax loss function:

\[ L = \sum_{k} \sum_{i} R_{il_k} \log(CNN(i|I_k)), \quad (2) \]

• context of a category is obtained
Noise

Figure 1: A toy classification example with 3 classes, illustrating the two types of label noise encountered on real datasets. In the label flip case, the images all belong to the 3 classes, but sometimes the labels are confused between them. In the outlier case, some images are unrelated to the classification task but possess one of the 3 labels.

Localizing Objects

- **Seed**: full images returned from Google as seed bounding box

- **Nearest Neighbor propagation**: 
  - train Exemplar-LDA detector using fc7 features
  - fire on remaining images to find top k nearest neighbors
  - use EdgeBox instead of sliding window for efficiency

- **Clustering into subdirectories**
  - based on E-LDA similarity scores and density estimation
  - iteratively merge neighbors from bottom up
Localizing Objects con’d
Moreover

- train R-CNN detector for each categories
- negative examples: random patches from YFCC
- positive examples: hundreds of instances per category from original training dataset
- increase positive bounding boxes
  - **EdgeBox Augmentation** (EA): add IoU > 0.5
  - **Category Expansion** (CE): verify with semantics
**Experiment**

2,240 objects, 89 attributes, 874 scenes

Two networks are trained:

- The object-attribute network (GoogleO), output dimension: 2,329
  - ~1.5 million images downloaded from Google Image Search
  - ~1.2 million Flickr images for fine-tuning

- All included network (GoogleA), output dimension: 3,203
  - plus scene images = ~2.1 million images
Experiment - 2 stages

Learning:
- batch size: 256
- learning rate: $0.01 / 10^i$, per 150K iterations
- termination: 450K iterations

Fine-tuning:
- step size: 30K
- termination: 100K iterations
Experiment

#1. PASCAL VOC Object Detection
#2. Clean up web images
#3. Failure Modes for Webly Trained Detectors
#4. Scene Classification
Result - confusion matrix
#1. PASCAL VOC Object Detection

- VOC 2007
  - fine-tuning with trainval images
  - step size: 20K
  - total: 100K iterations
- VOC 2012
  - step size: 50K
  - total: 200K iterations
- Did not tune R-CNN parameters, same SVM training settings with ImageNet
- GoogleO-FT, GoogleA-FT, Flickr-FT
### Result #1

VOC 2007 and 2012 dataset

<table>
<thead>
<tr>
<th>VOC 2007 test</th>
<th>aero</th>
<th>bike</th>
<th>bird</th>
<th>boat</th>
<th>bottle</th>
<th>bus</th>
<th>car</th>
<th>cat</th>
<th>chair</th>
<th>cow</th>
<th>table</th>
<th>dog</th>
<th>horse</th>
<th>mbike</th>
<th>pers</th>
<th>plant</th>
<th>sheep</th>
<th>sofa</th>
<th>train</th>
<th>tv</th>
<th>mAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>ImageNet [20]</td>
<td>57.6</td>
<td>57.9</td>
<td>38.5</td>
<td>31.8</td>
<td>23.7</td>
<td>51.2</td>
<td>58.9</td>
<td>51.4</td>
<td>20.0</td>
<td>50.5</td>
<td>46.0</td>
<td>51.6</td>
<td>55.9</td>
<td>43.3</td>
<td>23.3</td>
<td>48.1</td>
<td>35.3</td>
<td>51.0</td>
<td>57.4</td>
<td>44.7</td>
<td></td>
</tr>
<tr>
<td>GoogleO [Obj.]</td>
<td>57.1</td>
<td>59.9</td>
<td>35.4</td>
<td>30.5</td>
<td>21.9</td>
<td>53.9</td>
<td>59.5</td>
<td>40.7</td>
<td>18.6</td>
<td>43.3</td>
<td>37.5</td>
<td>41.9</td>
<td>49.6</td>
<td>57.7</td>
<td>38.4</td>
<td>22.8</td>
<td>45.2</td>
<td>37.1</td>
<td>48.0</td>
<td>54.5</td>
<td></td>
</tr>
<tr>
<td>GoogleA [Obj. + Sce.]</td>
<td>54.9</td>
<td>58.2</td>
<td>35.7</td>
<td>30.7</td>
<td>22.0</td>
<td>54.5</td>
<td>59.9</td>
<td>44.7</td>
<td>19.9</td>
<td>41.0</td>
<td>34.5</td>
<td>40.1</td>
<td>46.8</td>
<td>56.2</td>
<td>40.0</td>
<td>22.2</td>
<td>45.8</td>
<td>36.3</td>
<td>47.5</td>
<td>54.2</td>
<td></td>
</tr>
<tr>
<td>FlickrS [Flickr Obj.]</td>
<td>50.0</td>
<td>55.9</td>
<td>29.6</td>
<td>26.8</td>
<td>18.7</td>
<td>47.6</td>
<td>56.3</td>
<td>34.4</td>
<td>14.5</td>
<td>35.9</td>
<td>33.3</td>
<td>34.2</td>
<td>43.2</td>
<td>52.2</td>
<td>36.7</td>
<td>21.5</td>
<td>43.3</td>
<td>31.6</td>
<td>48.5</td>
<td>48.4</td>
<td></td>
</tr>
<tr>
<td>GFAll [All Obj., 1-stage]</td>
<td>52.1</td>
<td>57.8</td>
<td>38.1</td>
<td>25.6</td>
<td>21.2</td>
<td>47.6</td>
<td>56.4</td>
<td>43.8</td>
<td>19.6</td>
<td>42.6</td>
<td>30.3</td>
<td>37.6</td>
<td>45.1</td>
<td>50.8</td>
<td>39.3</td>
<td>22.9</td>
<td>43.5</td>
<td>34.2</td>
<td>48.3</td>
<td>52.2</td>
<td></td>
</tr>
<tr>
<td>FlickrF [2-stage]</td>
<td>53.9</td>
<td>60.7</td>
<td>37.0</td>
<td>31.6</td>
<td>23.8</td>
<td>55.0</td>
<td>60.8</td>
<td>44.1</td>
<td>20.3</td>
<td>46.5</td>
<td>31.5</td>
<td>39.8</td>
<td>49.7</td>
<td>59.0</td>
<td>41.6</td>
<td>23.0</td>
<td>44.4</td>
<td>36.2</td>
<td>49.9</td>
<td>56.2</td>
<td></td>
</tr>
<tr>
<td>FlickrG [2-stage, Graph]</td>
<td>55.3</td>
<td><strong>61.9</strong></td>
<td><strong>39.1</strong></td>
<td>29.5</td>
<td><strong>24.8</strong></td>
<td>55.1</td>
<td><strong>62.7</strong></td>
<td>43.5</td>
<td><strong>22.7</strong></td>
<td>49.3</td>
<td>36.6</td>
<td>42.7</td>
<td>48.9</td>
<td><strong>59.7</strong></td>
<td>41.2</td>
<td><strong>25.4</strong></td>
<td>47.7</td>
<td><strong>41.9</strong></td>
<td>48.8</td>
<td>56.8</td>
<td></td>
</tr>
</tbody>
</table>

| VOC-Scratch [2]| 49.9 | 60.6 | 24.7 | 23.7 | 20.3   | 52.5| 64.8| 32.9| 20.4  | 43.5| 34.2  | 29.9| 49.0  | 60.4  | 47.5 | 28.0  | 42.3 | 28.6  | 51.2| 50.0 |
| ImageNet [20]| 64.2 | **69.7** | **50.0** | **41.9** | **32.0** | 62.6| 71.0| **60.7** | **32.7** | 58.5| 46.5  | **56.1** | 60.6 | 66.8  | 54.2  | 31.5  | 52.8 | 48.9 | 57.9 | 64.7 | 54.2 |

<table>
<thead>
<tr>
<th>VOC 2012 test</th>
<th>aero</th>
<th>bike</th>
<th>bird</th>
<th>boat</th>
<th>bottle</th>
<th>bus</th>
<th>car</th>
<th>cat</th>
<th>chair</th>
<th>cow</th>
<th>table</th>
<th>dog</th>
<th>horse</th>
<th>mbike</th>
<th>person</th>
<th>plant</th>
<th>sheep</th>
<th>sofa</th>
<th>train</th>
<th>tv</th>
<th>mAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>ImageNet [20]</td>
<td>68.1</td>
<td>63.8</td>
<td>46.1</td>
<td>29.4</td>
<td>27.9</td>
<td>56.6</td>
<td>57.0</td>
<td>65.9</td>
<td>26.5</td>
<td>48.7</td>
<td>39.5</td>
<td>66.2</td>
<td>57.3</td>
<td>65.4</td>
<td>53.2</td>
<td>26.2</td>
<td>54.5</td>
<td>38.1</td>
<td>50.6</td>
<td>51.6</td>
<td></td>
</tr>
<tr>
<td>ImageNet-TV</td>
<td>73.3</td>
<td>67.1</td>
<td>46.3</td>
<td>31.7</td>
<td>30.6</td>
<td>59.4</td>
<td>61.0</td>
<td><strong>67.9</strong></td>
<td>27.3</td>
<td>53.1</td>
<td>39.1</td>
<td>64.1</td>
<td>60.5</td>
<td>70.9</td>
<td>57.2</td>
<td>26.1</td>
<td>59.0</td>
<td>40.1</td>
<td>56.2</td>
<td><strong>54.9</strong></td>
<td>52.3</td>
</tr>
<tr>
<td>GoogleO</td>
<td>72.2</td>
<td>67.3</td>
<td>46.0</td>
<td><strong>32.3</strong></td>
<td><strong>31.6</strong></td>
<td><strong>62.6</strong></td>
<td>62.5</td>
<td>66.5</td>
<td>27.3</td>
<td>52.1</td>
<td>38.9</td>
<td>64.0</td>
<td>59.1</td>
<td>71.6</td>
<td>58.0</td>
<td>27.2</td>
<td>57.6</td>
<td>41.3</td>
<td>56.3</td>
<td>53.7</td>
<td>52.4</td>
</tr>
<tr>
<td>FlickrG</td>
<td><strong>72.7</strong></td>
<td><strong>68.2</strong></td>
<td><strong>47.3</strong></td>
<td>32.2</td>
<td>30.6</td>
<td><strong>62.6</strong></td>
<td><strong>65.9</strong></td>
<td><strong>28.1</strong></td>
<td><strong>52.2</strong></td>
<td><strong>39.5</strong></td>
<td><strong>65.1</strong></td>
<td>60.0</td>
<td><strong>71.7</strong></td>
<td><strong>58.2</strong></td>
<td><strong>27.3</strong></td>
<td><strong>58.0</strong></td>
<td><strong>41.5</strong></td>
<td><strong>57.2</strong></td>
<td><strong>53.8</strong></td>
<td><strong>52.7</strong></td>
<td></td>
</tr>
</tbody>
</table>

Table 1. Results on VOC 2007 (PASCAL data used). Please see Section 4.1 for more details.

Table 2. Results on VOC 2012. Since [20] only fine-tuned on the train set, we also report results on trainval (ImageNet-TV) for fairness.
Result #1
VOC 2007 and 2012 dataset

<table>
<thead>
<tr>
<th>VOC 2007 test</th>
<th>aeros</th>
<th>bike</th>
<th>bird</th>
<th>boat</th>
<th>bottle</th>
<th>bus</th>
<th>car</th>
<th>cat</th>
<th>chair</th>
<th>cow</th>
<th>table</th>
<th>dog</th>
<th>horse</th>
<th>mbike</th>
<th>pers</th>
<th>plant</th>
<th>sheep</th>
<th>sofa</th>
<th>train</th>
<th>tv</th>
<th>mAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>ImageNet [20]</td>
<td>57.6</td>
<td>57.9</td>
<td>38.5</td>
<td>31.8</td>
<td>23.7</td>
<td>51.2</td>
<td>58.9</td>
<td>51.4</td>
<td>20.0</td>
<td>50.5</td>
<td>40.9</td>
<td>46.0</td>
<td>51.6</td>
<td>55.9</td>
<td>43.3</td>
<td>23.3</td>
<td>48.1</td>
<td>35.3</td>
<td>51.0</td>
<td>57.4</td>
<td>44.7</td>
</tr>
<tr>
<td>GoogleO [Obj.]</td>
<td>57.1</td>
<td>59.9</td>
<td>35.4</td>
<td>30.5</td>
<td>21.9</td>
<td>53.9</td>
<td>59.5</td>
<td>40.7</td>
<td>18.6</td>
<td>43.3</td>
<td>37.5</td>
<td>41.9</td>
<td>49.6</td>
<td>57.7</td>
<td>38.4</td>
<td>22.8</td>
<td>45.2</td>
<td>37.1</td>
<td>48.0</td>
<td>54.5</td>
<td>42.7</td>
</tr>
<tr>
<td>GoogleA [Obj. + Sce.]</td>
<td>54.9</td>
<td>58.2</td>
<td>35.7</td>
<td>30.7</td>
<td>22.0</td>
<td>54.5</td>
<td>59.9</td>
<td>44.7</td>
<td>19.9</td>
<td>41.0</td>
<td>34.5</td>
<td>40.1</td>
<td>46.8</td>
<td>56.2</td>
<td>40.0</td>
<td>22.2</td>
<td>45.8</td>
<td>36.3</td>
<td>47.5</td>
<td>54.2</td>
<td>42.8</td>
</tr>
<tr>
<td>FlickrS [Flickr Obj.]</td>
<td>50.0</td>
<td>55.9</td>
<td>29.6</td>
<td>26.8</td>
<td>18.7</td>
<td>47.6</td>
<td>56.3</td>
<td>34.4</td>
<td>14.5</td>
<td>35.9</td>
<td>33.3</td>
<td>34.2</td>
<td>43.2</td>
<td>52.2</td>
<td>36.7</td>
<td>21.5</td>
<td>43.3</td>
<td>31.6</td>
<td>48.5</td>
<td>48.4</td>
<td>38.1</td>
</tr>
<tr>
<td>GFAll [All Obj., 1-stage]</td>
<td>52.1</td>
<td>57.8</td>
<td>38.1</td>
<td>25.6</td>
<td>21.2</td>
<td>47.6</td>
<td>56.4</td>
<td>43.8</td>
<td>19.6</td>
<td>42.6</td>
<td>30.3</td>
<td>37.6</td>
<td>45.1</td>
<td>50.8</td>
<td>39.3</td>
<td>22.9</td>
<td>43.5</td>
<td>34.2</td>
<td>48.3</td>
<td>52.2</td>
<td>40.5</td>
</tr>
<tr>
<td>FlickrF [2-stage]</td>
<td>53.9</td>
<td>60.7</td>
<td>37.0</td>
<td>31.6</td>
<td>23.8</td>
<td>57.7</td>
<td>60.8</td>
<td>44.1</td>
<td>20.3</td>
<td>46.5</td>
<td>31.5</td>
<td>39.8</td>
<td>49.7</td>
<td>59.0</td>
<td>41.6</td>
<td>23.0</td>
<td>44.4</td>
<td>36.2</td>
<td>49.9</td>
<td>56.2</td>
<td>43.4</td>
</tr>
<tr>
<td>FlickrG [2-stage, Graph]</td>
<td>55.3</td>
<td>61.9</td>
<td>39.1</td>
<td>29.5</td>
<td><strong>24.8</strong></td>
<td>55.1</td>
<td><strong>62.7</strong></td>
<td>43.5</td>
<td>22.7</td>
<td>49.3</td>
<td>36.7</td>
<td>42.9</td>
<td>48.9</td>
<td><strong>59.7</strong></td>
<td>41.2</td>
<td>25.4</td>
<td>47.7</td>
<td><strong>41.9</strong></td>
<td>48.8</td>
<td>55.6</td>
<td>44.7</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>VOC-Scratch [2]</th>
<th>49.9</th>
<th>60.6</th>
<th>24.7</th>
<th>23.7</th>
<th>20.3</th>
<th>52.5</th>
<th>64.8</th>
<th>32.9</th>
<th>20.4</th>
<th>43.5</th>
<th>34.2</th>
<th>29.9</th>
<th>49.0</th>
<th>60.4</th>
<th>47.5</th>
<th>28.0</th>
<th>42.3</th>
<th>28.6</th>
<th>51.2</th>
<th>50.0</th>
<th>40.7</th>
</tr>
</thead>
<tbody>
<tr>
<td>ImageNet [20]</td>
<td>64.2</td>
<td>69.7</td>
<td>50.0</td>
<td>41.9</td>
<td>32.0</td>
<td>62.6</td>
<td>71.0</td>
<td>60.7</td>
<td>32.7</td>
<td>58.5</td>
<td>46.5</td>
<td>56.1</td>
<td>60.6</td>
<td>66.8</td>
<td>54.2</td>
<td>31.5</td>
<td>52.8</td>
<td>48.9</td>
<td>57.9</td>
<td>64.7</td>
<td>54.2</td>
</tr>
<tr>
<td>GoogleO</td>
<td>65.0</td>
<td>68.1</td>
<td>45.2</td>
<td>37.0</td>
<td>29.6</td>
<td>65.4</td>
<td>73.8</td>
<td>54.0</td>
<td>30.4</td>
<td>57.8</td>
<td>48.7</td>
<td>51.9</td>
<td><strong>64.1</strong></td>
<td>64.7</td>
<td>54.0</td>
<td><strong>32.0</strong></td>
<td>54.9</td>
<td>44.5</td>
<td>57.0</td>
<td>64.0</td>
<td>53.1</td>
</tr>
<tr>
<td>GoogleA</td>
<td>64.2</td>
<td>68.3</td>
<td>42.7</td>
<td>38.7</td>
<td>26.5</td>
<td>65.1</td>
<td>72.4</td>
<td>50.7</td>
<td>28.5</td>
<td><strong>60.9</strong></td>
<td>48.8</td>
<td>51.2</td>
<td>60.2</td>
<td>65.5</td>
<td><strong>54.5</strong></td>
<td>31.1</td>
<td>50.5</td>
<td><strong>48.5</strong></td>
<td>56.3</td>
<td>60.3</td>
<td>52.3</td>
</tr>
<tr>
<td>FlickrG</td>
<td>63.7</td>
<td>68.5</td>
<td>46.2</td>
<td>36.4</td>
<td>30.2</td>
<td><strong>68.4</strong></td>
<td>73.9</td>
<td>56.9</td>
<td>31.4</td>
<td>59.1</td>
<td>46.7</td>
<td>52.4</td>
<td>61.5</td>
<td><strong>69.2</strong></td>
<td>53.6</td>
<td>31.6</td>
<td>53.8</td>
<td>44.5</td>
<td><strong>58.1</strong></td>
<td>59.6</td>
<td>53.3</td>
</tr>
</tbody>
</table>

Table 1. Results on VOC 2007 (PASCAL data used). Please see Section 4.1 for more details.

<table>
<thead>
<tr>
<th>VOC 2012 test</th>
<th>aeros</th>
<th>bike</th>
<th>bird</th>
<th>boat</th>
<th>bottle</th>
<th>bus</th>
<th>car</th>
<th>cat</th>
<th>chair</th>
<th>cow</th>
<th>table</th>
<th>dog</th>
<th>horse</th>
<th>mbike</th>
<th>person</th>
<th>plant</th>
<th>sheep</th>
<th>sofa</th>
<th>train</th>
<th>tv</th>
<th>mAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>ImageNet [20]</td>
<td>68.1</td>
<td>63.8</td>
<td>46.1</td>
<td>29.4</td>
<td>27.9</td>
<td>56.6</td>
<td>57.0</td>
<td>65.9</td>
<td>26.5</td>
<td>48.7</td>
<td>39.5</td>
<td><strong>66.2</strong></td>
<td>57.3</td>
<td>65.4</td>
<td>53.2</td>
<td>26.2</td>
<td>54.5</td>
<td>38.1</td>
<td>50.6</td>
<td>51.6</td>
<td>49.6</td>
</tr>
<tr>
<td>ImageNet-TV</td>
<td><strong>73.3</strong></td>
<td>67.1</td>
<td>46.3</td>
<td>31.7</td>
<td>30.6</td>
<td>59.4</td>
<td>61.0</td>
<td><strong>67.9</strong></td>
<td>27.3</td>
<td><strong>53.1</strong></td>
<td>39.1</td>
<td>64.1</td>
<td><strong>60.5</strong></td>
<td>70.9</td>
<td>57.2</td>
<td>26.1</td>
<td><strong>59.0</strong></td>
<td>40.1</td>
<td>56.2</td>
<td><strong>54.9</strong></td>
<td>52.3</td>
</tr>
<tr>
<td>GoogleO</td>
<td>72.2</td>
<td>67.3</td>
<td>46.0</td>
<td><strong>32.3</strong></td>
<td><strong>31.6</strong></td>
<td><strong>62.6</strong></td>
<td>62.5</td>
<td>66.5</td>
<td>27.3</td>
<td>52.1</td>
<td>38.9</td>
<td>64.0</td>
<td>59.1</td>
<td>71.6</td>
<td>58.0</td>
<td>27.2</td>
<td>57.6</td>
<td>41.3</td>
<td>56.3</td>
<td>53.7</td>
<td>52.4</td>
</tr>
<tr>
<td>FlickrG</td>
<td>72.7</td>
<td><strong>68.2</strong></td>
<td><strong>47.3</strong></td>
<td>32.2</td>
<td>30.6</td>
<td>62.3</td>
<td><strong>62.6</strong></td>
<td>65.9</td>
<td><strong>28.1</strong></td>
<td>52.2</td>
<td><strong>39.5</strong></td>
<td>65.1</td>
<td>60.0</td>
<td><strong>71.7</strong></td>
<td>58.2</td>
<td><strong>27.3</strong></td>
<td>58.0</td>
<td><strong>41.5</strong></td>
<td>57.2</td>
<td>53.8</td>
<td>52.7</td>
</tr>
</tbody>
</table>

Table 2. Results on VOC 2012. Since [20] only fine-tuned on the train set, we also report results on trainval (ImageNet-TV) for fairness.
#2. Discover subdirectories and localize the object in the web images

- Train R-CNN based detector with VOC 2007
- compared with Google-Ngrams
#2. Discover subdirectories and localize the object in the web images

<table>
<thead>
<tr>
<th></th>
<th>feature</th>
<th>bounding box</th>
</tr>
</thead>
<tbody>
<tr>
<td>GoogleO</td>
<td>GoogleO CNN</td>
<td>easy images</td>
</tr>
<tr>
<td>GoogleA</td>
<td>GoogleA CNN</td>
<td>easy images</td>
</tr>
<tr>
<td>Flickr</td>
<td>2-stage CNN</td>
<td>easy images</td>
</tr>
<tr>
<td>Flickr-M</td>
<td>2-stage CNN</td>
<td>easy+hard images</td>
</tr>
<tr>
<td>Flickr-C</td>
<td>2-stage CNN</td>
<td>original and related categories</td>
</tr>
</tbody>
</table>
### Result #2

<table>
<thead>
<tr>
<th>VOC 2007 test</th>
<th>aero</th>
<th>bike</th>
<th>bird</th>
<th>boat</th>
<th>bottle</th>
<th>bus</th>
<th>car</th>
<th>cat</th>
<th>chair</th>
<th>cow</th>
<th>table</th>
<th>dog</th>
<th>horse</th>
<th>mbike</th>
<th>person</th>
<th>plant</th>
<th>sheep</th>
<th>sofa</th>
<th>train</th>
<th>tv</th>
<th>mAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>LEVAN [14]</td>
<td>14.0</td>
<td>36.2</td>
<td>12.5</td>
<td>10.3</td>
<td>9.2</td>
<td>35.0</td>
<td>8.4</td>
<td>10.0</td>
<td>17.5</td>
<td>6.5</td>
<td>12.9</td>
<td>30.6</td>
<td>27.5</td>
<td>6.0</td>
<td>1.5</td>
<td>18.8</td>
<td>10.3</td>
<td>23.5</td>
<td>16.4</td>
<td>17.1</td>
<td></td>
</tr>
<tr>
<td>GoogleO</td>
<td>30.2</td>
<td>34.3</td>
<td>16.7</td>
<td>13.3</td>
<td>6.1</td>
<td>43.6</td>
<td>27.4</td>
<td>22.6</td>
<td>16.4</td>
<td>10.0</td>
<td>21.3</td>
<td>25.0</td>
<td>35.9</td>
<td>7.6</td>
<td>9.3</td>
<td>21.8</td>
<td>17.3</td>
<td>31.0</td>
<td>18.1</td>
<td>20.7</td>
<td></td>
</tr>
<tr>
<td>GoogleA</td>
<td>29.5</td>
<td>38.3</td>
<td>15.1</td>
<td>14.0</td>
<td>9.1</td>
<td>44.3</td>
<td>29.3</td>
<td>24.9</td>
<td>6.9</td>
<td>15.8</td>
<td>9.7</td>
<td>22.6</td>
<td>34.3</td>
<td>9.7</td>
<td>12.7</td>
<td>21.4</td>
<td>15.8</td>
<td>33.4</td>
<td>19.4</td>
<td>21.5</td>
<td></td>
</tr>
<tr>
<td>FlickrG</td>
<td>32.6</td>
<td>42.8</td>
<td>19.3</td>
<td>13.9</td>
<td>9.2</td>
<td>46.6</td>
<td>29.6</td>
<td>20.6</td>
<td>6.8</td>
<td>17.8</td>
<td>10.2</td>
<td>22.4</td>
<td>26.7</td>
<td>11.7</td>
<td>14.0</td>
<td>19.0</td>
<td>19.0</td>
<td>34.0</td>
<td>21.9</td>
<td>22.9</td>
<td></td>
</tr>
<tr>
<td>FlickrG-EA</td>
<td>32.7</td>
<td>44.3</td>
<td>17.9</td>
<td>14.0</td>
<td>9.3</td>
<td>47.1</td>
<td>26.6</td>
<td>19.2</td>
<td>8.2</td>
<td>18.3</td>
<td>10.0</td>
<td>22.7</td>
<td>25.0</td>
<td>12.0</td>
<td>12.7</td>
<td>22.2</td>
<td>20.9</td>
<td>35.6</td>
<td>18.2</td>
<td>23.0</td>
<td></td>
</tr>
<tr>
<td>FlickrG-CE</td>
<td>30.2</td>
<td>41.3</td>
<td>21.7</td>
<td>18.3</td>
<td>9.2</td>
<td>44.3</td>
<td>32.2</td>
<td>25.5</td>
<td>9.8</td>
<td>21.5</td>
<td>10.4</td>
<td>26.7</td>
<td>27.3</td>
<td>12.6</td>
<td>13.3</td>
<td>20.4</td>
<td>20.9</td>
<td>36.2</td>
<td>22.8</td>
<td>24.4</td>
<td></td>
</tr>
</tbody>
</table>
#2 Object localization results

- **LEVAN**: 17.1
- **GoogleO**: 20.7
- **GoogleA**: 21.5
- **FlickrG**: 22.9
- **FlickrG-EA**: 23.0
- **FlickrG-CE**: 24.4

mAP values shown in the chart.
#3. Failure Modes for Webly Trained Detectors

- Localization error - false positive
  - background
  - multiple instances
  - spatial invariant
- Semantic drift between Google categories and PASCAL categories
  - sense disambiguation
Result #3
#4. Scene Classification

- MIT Indoor-67 dataset
- compute the fc7 feature vector: 4096 dimensions
- normalize to unit l2 length
- fixed SVM parameter (C=1)
Result #4

Scene classification results on MIT Indoor-67

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>ImageNet</td>
<td>56.8</td>
</tr>
<tr>
<td>OverFeat</td>
<td>58.4</td>
</tr>
<tr>
<td>GoogleO</td>
<td>58.1</td>
</tr>
<tr>
<td>FlickrG</td>
<td>59.2</td>
</tr>
<tr>
<td>GoogleA[Obj.+Sce.]</td>
<td>66.5</td>
</tr>
</tbody>
</table>
Conclusion
Conclusion

- The two-stage CNN comes close to ImageNet pertained architecture on VOC 2007, outperforms in VOC 2012

- only cost is providing a category to query from
Discussion

• What are good sources of data?
• What types of data can be exploited?
• How do we exploit the data?
• What should we learn from Web data?
Discussion

- What are good sources of data?
  - search engine + everyday images

- What types of data can be exploited?
  - images-only is enough

- How do we exploit the data?
  - learn from the queried domain from search engine
  - fine-tuning with everyday images

- What should we learn from Web data?
  - intra-category similarity relationship based on easy images (remove noise)
  - low-level features supplemented by Flickr style images (remove bias)