Unsupervised Visual Representation Learning by Context Prediction

Berkan Demirel

Most slides in this representation are adopted from authors' original presentation at ICCV 2015
ImageNet + Deep Learning

- Image Retrieval
- Detection (RCNN)
- Segmentation (FCN)
- Depth Estimation
- ...
Do we need semantic labels?

Materials?
Parts?
Geometry?
Boundaries?
Pose?
Context as Supervision
[Collobert & Weston 2008; Mikolov et al. 2013]

house, where the professor lived without his wife and child; or so he said jokingly sometimes: “Here’s where I live. My house.” His daughter often added, without resentment, for the visitor’s information, “It started out to be for me, but it’s really his.” And she might reach in to bring forth an inch-high table lamp with fluted shade, or a blue dish the size of her little fingernail, marked “Kitty” and half full of eternal milk, but she was sure to replace these, after they had been admired, pretty near exactly where they had been. The little house was very orderly, and just big enough for all it contained, though to some tastes the bric-à-brac in the parlor might seem excessive. The daughter’s preference was for the store-bought gimmicks and appliances, the toasters and carpet sweepers of Lilliput, but she knew that most adult visitors would
Semantics from a non-semantic task
Relative Position Task

8 possible locations

Randomly Sample Patch
Sample Second Patch
Note: connects *across* instances!
Training requires Batch Normalization [Ioffe et al. 2015], but no other tricks.
Avoiding Trivial Shortcuts

Include a gap

Jitter the patch locations
A Not-So “Trivial” Shortcut

Position in Image
Chromatic Aberration
Solutions

Color Dropping
Randomly drop 2 of the 3 color channels from each patch. Then, replacing the dropped colors with Gaussian Noise (standard deviation ~1/100 the standard deviation of the remaining channel).

Projection
Shift green and magenta (red+blue) towards gray
Implementation Details

• Train on the ImageNet 2012 training set (1.3M images), using only the images and discarding the labels.
• Resize each image to between 150K and 450K total pixels, preserving the aspect-ratio.
• Sample patches at resolution 96-by-96.
• Sample the patches from a grid like pattern. Each sampled patch can participate in as many as 8 separate pairings.
• Allow a gap of 48 pixels between the sampled patches in the grid, but also jitter the location of each patch in the grid by −7 to 7 pixels in each direction.
• Preprocess patches by (1) mean subtraction, (2) projecting or dropping colors, (3) randomly downsampling some patches to as little as 100 total pixels, and then upsampling it, to build robustness to pixelation.
• Use batch normalization, without the scale and shift.
Experiments

- Chromatic Aberration
- Nearest-Neighbor Matching
- Object Detection
- Geometry Estimation
- Visual Data Mining
- Layout Prediction
Chromatic Aberration
Chromatic Aberration
Nearest-Neighbor Matching

- fc6 layer features and only one of the two stacks are used.
- fc7 and higher layers are removed.
- Normalized cross correlation is used to find similar patches
- Randomly selected 96x96 patches are used in the comparison.
What is learned?

Input | Ours | Random Initialization | ImageNet AlexNet

![Input Images](image1.png) ![Ours Images](image2.png) ![Random Initialization Images](image3.png) ![ImageNet AlexNet Images](image4.png)
Still don’t capture everything

You don’t always need to learn!
Object Detection

1. Input image
2. Extract region proposals (~2k)
3. Compute CNN features
4. Classify regions

Pre-train on relative-position task, w/o labels

[Girshick et al. 2014]
Object Detection

Figure 6. Our architecture for Pascal VOC detection. Layers from conv1 through pool5 are copied from our patch-based network (Figure 3). The new 'conv6' layer is created by converting the fc6 layer into a convolution layer. Kernel sizes, output units, and stride are given in parentheses, as in Figure 3.
Object Detection

<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>DPM-v5[17]</td>
<td>33.2</td>
<td>60.3</td>
<td>10.2</td>
<td>16.1</td>
<td>27.3</td>
<td>54.3</td>
<td>58.2</td>
<td>23.0</td>
<td>20.0</td>
<td>24.1</td>
<td>26.7</td>
<td>12.7</td>
<td>58.1</td>
<td>48.2</td>
<td>43.2</td>
<td>12.0</td>
<td>21.1</td>
<td>36.1</td>
<td>46.0</td>
<td>43.5</td>
<td>33.7</td>
</tr>
<tr>
<td>[8] w/o context</td>
<td>52.6</td>
<td>52.6</td>
<td>19.2</td>
<td>25.4</td>
<td>18.7</td>
<td>47.3</td>
<td>56.9</td>
<td>42.1</td>
<td>16.6</td>
<td>41.4</td>
<td>41.9</td>
<td>27.7</td>
<td>47.9</td>
<td>51.5</td>
<td>29.9</td>
<td>20.0</td>
<td>41.1</td>
<td>36.4</td>
<td>48.6</td>
<td>53.2</td>
<td>38.5</td>
</tr>
<tr>
<td>Regionlets[58]</td>
<td>54.2</td>
<td>52.0</td>
<td>20.3</td>
<td>24.0</td>
<td>20.1</td>
<td>55.5</td>
<td>68.7</td>
<td>42.6</td>
<td>19.2</td>
<td>44.2</td>
<td>49.1</td>
<td>26.6</td>
<td>57.0</td>
<td>54.5</td>
<td>43.4</td>
<td>16.4</td>
<td>36.6</td>
<td>37.7</td>
<td>59.4</td>
<td>52.3</td>
<td>41.7</td>
</tr>
<tr>
<td>Scratch-R-CNN[2]</td>
<td>49.9</td>
<td>60.6</td>
<td>24.7</td>
<td>23.7</td>
<td>20.3</td>
<td>52.5</td>
<td>64.8</td>
<td>32.9</td>
<td>20.4</td>
<td>43.5</td>
<td>34.2</td>
<td>29.9</td>
<td>49.0</td>
<td>60.4</td>
<td>47.5</td>
<td>28.0</td>
<td>42.3</td>
<td>28.6</td>
<td>51.2</td>
<td>50.0</td>
<td>40.7</td>
</tr>
<tr>
<td>Scratch-Ours</td>
<td>52.6</td>
<td>60.5</td>
<td>23.8</td>
<td>24.3</td>
<td>18.1</td>
<td>50.6</td>
<td>65.9</td>
<td>29.2</td>
<td>19.5</td>
<td>43.5</td>
<td>35.2</td>
<td>27.6</td>
<td>46.5</td>
<td>59.4</td>
<td>46.5</td>
<td>25.6</td>
<td>42.4</td>
<td>23.5</td>
<td>50.0</td>
<td>50.6</td>
<td>39.8</td>
</tr>
<tr>
<td>Ours-projection</td>
<td>58.4</td>
<td>62.8</td>
<td>33.5</td>
<td>27.7</td>
<td>24.4</td>
<td>58.5</td>
<td>68.5</td>
<td>41.2</td>
<td>26.3</td>
<td>49.5</td>
<td>42.6</td>
<td>37.3</td>
<td>55.7</td>
<td>62.5</td>
<td>49.4</td>
<td>29.0</td>
<td>47.5</td>
<td>28.4</td>
<td>54.7</td>
<td>56.8</td>
<td>45.7</td>
</tr>
<tr>
<td>Ours-color-dropping</td>
<td>60.5</td>
<td>66.5</td>
<td>29.6</td>
<td>28.5</td>
<td>26.3</td>
<td>56.1</td>
<td>70.4</td>
<td>44.8</td>
<td>24.6</td>
<td>45.5</td>
<td>45.4</td>
<td>35.1</td>
<td>52.2</td>
<td>60.2</td>
<td>50.0</td>
<td>28.1</td>
<td>46.7</td>
<td>42.6</td>
<td>54.8</td>
<td>58.6</td>
<td>46.3</td>
</tr>
<tr>
<td>Ours-Yahoo100m</td>
<td>56.2</td>
<td>63.9</td>
<td>29.8</td>
<td>27.8</td>
<td>23.9</td>
<td>57.4</td>
<td>69.8</td>
<td>35.6</td>
<td>23.7</td>
<td>47.4</td>
<td>43.0</td>
<td>29.5</td>
<td>52.9</td>
<td>62.0</td>
<td>48.7</td>
<td>28.4</td>
<td>45.1</td>
<td>33.6</td>
<td>49.0</td>
<td>55.5</td>
<td>44.2</td>
</tr>
<tr>
<td>ImageNet-R-CNN[21]</td>
<td>64.2</td>
<td>69.7</td>
<td>50</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>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>
<td>54.2</td>
</tr>
<tr>
<td>K-means-rescale [31]</td>
<td>55.7</td>
<td>60.9</td>
<td>27.9</td>
<td>30.9</td>
<td>12.0</td>
<td>59.1</td>
<td>63.7</td>
<td>47.0</td>
<td>21.4</td>
<td>45.2</td>
<td>55.8</td>
<td>40.3</td>
<td>67.5</td>
<td>61.2</td>
<td>48.3</td>
<td>21.9</td>
<td>32.8</td>
<td>46.9</td>
<td>61.6</td>
<td>51.7</td>
<td>45.6</td>
</tr>
<tr>
<td>Ours-rescale [31]</td>
<td>61.9</td>
<td>63.3</td>
<td>35.8</td>
<td>32.6</td>
<td>17.2</td>
<td>68.0</td>
<td>67.9</td>
<td>54.8</td>
<td>29.6</td>
<td>52.4</td>
<td>62.9</td>
<td>51.3</td>
<td>67.1</td>
<td>64.3</td>
<td>50.5</td>
<td>24.4</td>
<td>43.7</td>
<td>54.9</td>
<td>67.1</td>
<td>52.7</td>
<td>51.1</td>
</tr>
<tr>
<td>ImageNet-rescale[31]</td>
<td>64.0</td>
<td>69.6</td>
<td>53.2</td>
<td>44.4</td>
<td>24.9</td>
<td>65.7</td>
<td>69.6</td>
<td>69.2</td>
<td>28.9</td>
<td>63.6</td>
<td>62.8</td>
<td>63.9</td>
<td>73.3</td>
<td>64.6</td>
<td>55.8</td>
<td>25.7</td>
<td>50.5</td>
<td>55.4</td>
<td>69.3</td>
<td>56.4</td>
<td>56.5</td>
</tr>
<tr>
<td>VGG-K-means-rescale</td>
<td>56.1</td>
<td>58.6</td>
<td>23.3</td>
<td>25.7</td>
<td>12.8</td>
<td>57.8</td>
<td>61.2</td>
<td>45.2</td>
<td>21.4</td>
<td>47.1</td>
<td>39.5</td>
<td>35.6</td>
<td>60.1</td>
<td>61.4</td>
<td>44.9</td>
<td>17.3</td>
<td>33.2</td>
<td>57.9</td>
<td>51.2</td>
<td>42.4</td>
<td>42.4</td>
</tr>
<tr>
<td>VGG-Ours-rescale</td>
<td>71.1</td>
<td>72.4</td>
<td>54.1</td>
<td>48.2</td>
<td>29.9</td>
<td>75.2</td>
<td>78.0</td>
<td>71.9</td>
<td>38.3</td>
<td>60.5</td>
<td>62.3</td>
<td>68.1</td>
<td>74.3</td>
<td>74.2</td>
<td>64.8</td>
<td>32.6</td>
<td>56.5</td>
<td>66.4</td>
<td>74.0</td>
<td>60.3</td>
<td>61.7</td>
</tr>
<tr>
<td>VGG-ImageNet-rescale</td>
<td>76.6</td>
<td>79.6</td>
<td>68.5</td>
<td>57.4</td>
<td>40.8</td>
<td>79.9</td>
<td>78.4</td>
<td>85.4</td>
<td>41.7</td>
<td>77.0</td>
<td>69.3</td>
<td>80.1</td>
<td>78.6</td>
<td>74.6</td>
<td>70.1</td>
<td>37.5</td>
<td>66.0</td>
<td>67.5</td>
<td>77.4</td>
<td>64.9</td>
<td>68.6</td>
</tr>
</tbody>
</table>

Table 1. Mean Average Precision on VOC-2007.

[Girshick et al. 2014]
## Multi-Task Training?

<table>
<thead>
<tr>
<th></th>
<th>S</th>
<th>M</th>
<th>L</th>
</tr>
</thead>
<tbody>
<tr>
<td>multi-task training?</td>
<td>√</td>
<td>√</td>
<td>√</td>
</tr>
<tr>
<td>stage-wise training?</td>
<td>√</td>
<td></td>
<td>√</td>
</tr>
<tr>
<td>test-time bbox reg?</td>
<td>√</td>
<td>√</td>
<td>√</td>
</tr>
<tr>
<td>VOC07 mAP</td>
<td>52.2</td>
<td>53.3</td>
<td>54.6</td>
</tr>
</tbody>
</table>

Table 6. Multi-task training (forth column per group) improves mAP over piecewise training (third column per group).
# Surface-normal Estimation

<table>
<thead>
<tr>
<th>Method</th>
<th>Error (Lower Better)</th>
<th>% Good Pixels (Higher Better)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>Median</td>
</tr>
<tr>
<td>No Pretraining</td>
<td>38.6</td>
<td>26.5</td>
</tr>
<tr>
<td>Unsup. Track.</td>
<td>34.2</td>
<td>21.9</td>
</tr>
<tr>
<td>Ours</td>
<td><strong>33.2</strong></td>
<td><strong>21.3</strong></td>
</tr>
<tr>
<td>ImageNet Labels</td>
<td>33.3</td>
<td><strong>20.8</strong></td>
</tr>
</tbody>
</table>
Visual Data Mining

• Sample a constellation of four adjacent patches from an image (we use four to reduce the likelihood of a matching spatial arrangement happening by chance).
• Find top 100 images which have the strongest matches for all four patches, ignoring spatial layout.
• Use a type of a geometric verification to filter away the images where the four matches are not geometrically consistent.
• Apply the described mining algorithm to Pascal VOC 2011.
Visual Data Mining

Via Geometric Verification

Simplified from [Chum et al 2007]
Mined from Pascal VOC2011
Layout Prediction

Visual Data Mining Algorithm results for 15,000 Street View images from Paris
Purity Test

Figure 9. Purity vs coverage for objects discovered on a subset of Pascal VOC 2007. The numbers in the legend indicate area under the curve (AUC). In parentheses is the AUC up to a coverage of .5.
So, do we need semantic labels?
Source Code & Supplementary Materials

- Magic Init
- Unsupervised Visual Representation Learning by Context Prediction
- Visual Data Mining Results on unlabeled PASCAL VOC 2011 Images
- Nearest Neighbors on PASCAL VOC 2007
- More
THANK YOU!