Unsupervised Visual Representation Learning by Context Prediction

Carl Doersch, Alexei A. Efros, Abhinav Gupta

Presented by Maheen Rashid for ECS 289G
Motivation

• How can we scale to billions rather than millions of images?
  • Imagenet trained on ~1.2 million images
• Unsupervised learning
  • Problem - What should be represented?
Inspiration - Context

• Similar words appear in similar contexts
• Learn to relate a given word to its surrounding words
• Context prediction becomes a ‘pretext’ task
A simple way to learn feature vectors for words (Collobert and Weston, 2008)

- word at $t-2$
- word at $t-1$
- word at $t$ or random word
- word at $t+1$
- word at $t+2$
A simple way to learn feature vectors for words (Collobert and Weston, 2008)
A simple way to learn feature vectors for words (Collobert and Weston, 2008)

right or random?

units that learn to predict the output from features of the input words

word code

word at $t-2$

word code

word at $t-1$

word code

word at $t$ or random word

word code

word at $t+1$

word code

word at $t+2$

Slide from Geoff Hinton
Right or Random for Images?

Slide from Carl Doersch
Right or Random for Images?

Slide from Carl Doersch
Can you tell where B goes relative to A?
Answer:

Slide from Carl Doersch
Answer:

Doing this requires recognizing semantics!

Slide from Carl Doersch
Unlabeled training image

Randomly Sample Patch

Slide from Carl Doersch
Unlabeled training image

Randomly Sample Patch
Sample Second Patch

Slide from Carl Doersch
Train Deep Net to recover relative position

Slide from Carl Doersch
CNN

Slide from Carl Doersch
CNN

Patch Features

Slide from Carl Doersch
How to sample patches
How to sample patches

Slide from Carl Doersch
How to sample patches

Slide from Carl Doersch
How to sample patches

Include a gap

Slide from Carl Doersch
How to sample patches

- Include a gap
- Jitter the patch locations

Slide from Carl Doersch
Another trivial shortcut

- Chromatic Aberration
- Shift colors towards grey (Projection)
- Drop 2 out of three channels during training
Another trivial shortcut

• Chromatic Aberration
• Shift colors towards grey (Projection)
• Drop 2 out of three channels during training
Another trivial shortcut

- Chromatic Aberration
- Shift colors towards grey (Projection)
- Drop 2 out of three channels during training
Another trivial shortcut

- Chromatic Aberration
- Shift colors towards grey (Projection)
- Drop 2 out of three channels during training

Slide from Carl Doersch
Another trivial shortcut

- Chromatic Aberration
- Shift colors towards grey (Projection)
- Drop 2 out of three channels during training
What is learned?

Input  Ours

Slide from Carl Doersch
What is learned?

<table>
<thead>
<tr>
<th>Input</th>
<th>Ours</th>
<th>Random Initialization</th>
</tr>
</thead>
</table>

Slide from Carl Doersch
What is learned?

Input | Ours | Random Initialization | ImageNet AlexNet
--- | --- | --- | ---

Slide from Carl Doersch
<table>
<thead>
<tr>
<th>Input</th>
<th>Ours</th>
<th>Random Initialization</th>
<th>ImageNet AlexNet</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image1.jpg" alt="Images" /></td>
<td><img src="image2.jpg" alt="Images" /></td>
<td><img src="image3.jpg" alt="Images" /></td>
<td><img src="image4.jpg" alt="Images" /></td>
</tr>
</tbody>
</table>

Still don’t capture everything

Slide from Carl Doersch
Still don’t capture everything

You don’t always need to learn!

Slide from Carl Doersch
Pre-Training for R-CNN

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

Slide from Carl Doersch
Pre-Training for R-CNN

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

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

Slide from Carl Doersch
Details

- Use stack from patch context predictor before pool5
- Resize convolution layers to work on 227x227 instead of 96x96
- Use FC7 as the final representation
Architecture

fc9 (8)
fc8 (4096)

fc7 (4096)

fc6 (4096)
pool5 (3x3, 256, 2)
conv5 (3x3, 256, 1)
conv4 (3x3, 384, 1)
conv3 (3x3, 384, 1)
LRN2
pool2 (3x3, 384, 2)
conv2 (5x5, 384, 2)
LRN1
pool1 (3x3, 96, 2)
conv1 (11x11, 96, 4)

Patch 1

fc6 (4096)
pool5 (3x3, 256, 2)
conv5 (3x3, 256, 1)
conv4 (3x3, 384, 1)
conv3 (3x3, 384, 1)
LRN2
pool2 (3x3, 384, 2)
conv2 (5x5, 384, 2)
LRN1
pool1 (3x3, 96, 2)
conv1 (11x11, 96, 4)

Patch 2

Slide from Carl Doersch
VOC 2007 Performance
(pretraining for R-CNN)
VOC 2007 Performance
(pretraining for R-CNN)
VOC 2007 Performance
(pretraining for R-CNN)

Average Precision

<table>
<thead>
<tr>
<th>Pretraining</th>
<th>Precision</th>
</tr>
</thead>
<tbody>
<tr>
<td>No</td>
<td>40.7%</td>
</tr>
</tbody>
</table>
VOC 2007 Performance
(pretraining for R-CNN)

No Pretraining: 40.7%
Ours (No Labels): 46.3%

Slide from Carl Doersch
VOC 2007 Performance
(pretraining for R-CNN)

<table>
<thead>
<tr>
<th>Condition</th>
<th>Average Precision</th>
</tr>
</thead>
<tbody>
<tr>
<td>No Pretraining</td>
<td>40.7%</td>
</tr>
<tr>
<td>Ours (No Labels)</td>
<td>46.3%</td>
</tr>
<tr>
<td>ImageNet Labels</td>
<td>54.2%</td>
</tr>
</tbody>
</table>

Slide from Carl Doersch
Unsupervised Object Discovery

Slide from Carl Doersch
Unsupervised Object Discovery

Slide from Carl Doersch
Unsupervised Object Discovery

Slide from Carl Doersch
Unsupervised Object Discovery

Slide from Carl Doersch
Unsupervised Object Discovery

Slide from Carl Doersch
Algorithm

Slide from Carl Doersch
Algorithm

Slide from Carl Doersch
Algorithm

Slide from Carl Doersch
Purity vs Coverage

Purity-Coverage for Proposed Objects

Visual Words .63 (.37)
Russel et al. .66 (.38)
HOG Kmeans .70 (.40)
Singh et al. .83 (.47)
Doersch et al. .83 (.48)
Our Approach .87 (.48)
Pretext Task

- Performance on Pascal VOC is 38.4% (Chance is 12.5%)
- On ImageNet accuracy is 39.5% on training set, and 40.5% on validation
- On GT box patches - similar performance. 39.2% overall with 45.6% on cars
Questions?