Unsupervised Visual Representation Learning by Context Prediction

Carl Doersch, Alexei A. Efros & Abhinav Gupta

Present: Yisha Sun
11/01/2016
ImageNet + Deep Learning

- Image Retrieval
- Detection (RCNN)
- Segmentation (FCN)
- Depth Estimation
- ...

Do we even need semantic labels?

Pose?

Boundaries?

Parts?

Geometry?

Materials?

ImageNet + Deep Learning

Do we need this task?

Do we even need semantic labels?

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 infor-
mation, “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 music, 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

Context Prediction for Images
Semantics from a non-semantic task

Relative Position Task

Randomly Sample Patch
Sample Second Patch

• The ultimate goal is to learn a feature embedding for *individual* patches, such that patches which are visually similar (across different images) would be close in the embedding space.
Architecture

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
Chromatic Aberration

Chromatic Aberration

What is learned?

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

Still don’t capture everything

<table>
<thead>
<tr>
<th>Input</th>
<th>Ours</th>
<th>Random Initialization</th>
<th>ImageNet AlexNet</th>
</tr>
</thead>
<tbody>
<tr>
<td>![Input Image]</td>
<td>![Ours Image]</td>
<td>![Random Initialization Image]</td>
<td>![ImageNet AlexNet Image]</td>
</tr>
<tr>
<td>![Input Image]</td>
<td>![Ours Image]</td>
<td>![Random Initialization Image]</td>
<td>![ImageNet AlexNet Image]</td>
</tr>
<tr>
<td>![Input Image]</td>
<td>![Ours Image]</td>
<td>![Random Initialization Image]</td>
<td>![ImageNet AlexNet Image]</td>
</tr>
<tr>
<td>![Input Image]</td>
<td>![Ours Image]</td>
<td>![Random Initialization Image]</td>
<td>![ImageNet AlexNet Image]</td>
</tr>
<tr>
<td>![Input Image]</td>
<td>![Ours Image]</td>
<td>![Random Initialization Image]</td>
<td>![ImageNet AlexNet Image]</td>
</tr>
<tr>
<td>![Input Image]</td>
<td>![Ours Image]</td>
<td>![Random Initialization Image]</td>
<td>![ImageNet AlexNet Image]</td>
</tr>
</tbody>
</table>

You don’t always need to learn!

<table>
<thead>
<tr>
<th>Input</th>
<th>Ours</th>
<th>Random Initialization</th>
<th>ImageNet AlexNet</th>
</tr>
</thead>
<tbody>
<tr>
<td>![Input Image]</td>
<td>![Ours Image]</td>
<td>![Random Initialization Image]</td>
<td>![ImageNet AlexNet Image]</td>
</tr>
<tr>
<td>![Input Image]</td>
<td>![Ours Image]</td>
<td>![Random Initialization Image]</td>
<td>![ImageNet AlexNet Image]</td>
</tr>
<tr>
<td>![Input Image]</td>
<td>![Ours Image]</td>
<td>![Random Initialization Image]</td>
<td>![ImageNet AlexNet Image]</td>
</tr>
<tr>
<td>![Input Image]</td>
<td>![Ours Image]</td>
<td>![Random Initialization Image]</td>
<td>![ImageNet AlexNet Image]</td>
</tr>
<tr>
<td>![Input Image]</td>
<td>![Ours Image]</td>
<td>![Random Initialization Image]</td>
<td>![ImageNet AlexNet Image]</td>
</tr>
<tr>
<td>![Input Image]</td>
<td>![Ours Image]</td>
<td>![Random Initialization Image]</td>
<td>![ImageNet AlexNet Image]</td>
</tr>
</tbody>
</table>

• The trained network is applied to two domains:
  • Unsupervised object detection
    • Find sets of images or image segments in a large image collection which depict the same object
  • Used a pre-training for standard vision task
    • R-CNN
Visual Data Mining

Via Geometric Verification
Simplified from [Chum et al 2007]

Mined from Pascal VOC2011

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

[Girshick et al. 2014]

VOC 2007 Performance
(pretraining for R-CNN)

% Average Precision

No Rescaling
Krähnenbühl et al. 2015
VGG + Krähnenbühl et al.

So, do we need semantic labels?
“Self-Supervision” and the Future

Ego-Motion

Video

Context

[Agrawal et al. 2015; Jayaraman et al. 2015] [Wang et al. 2015; Srivastava et al 2015; ...]

[Doersch et al. 2014; Pathak et al. 2015; Isola et al. 2015]

Strength and Weaknesses

• Unsupervised approach for feature learning
• High coverage rate of proposed objects, indicating learning invariances
• Used as pre-training, boost the R-CNN performance significantly
• Uncovers some connection of objects and respective surrounding

• Low performance on pretext task, while reason unclear
Possible Extensions

- Train on larger datasets
- Use DecovNets to uncover what is learned in each layer
- Trace the reasons for low performance by modifying the datasets, like the color aberration problem.
Thank you!