“Unsupervised” Self Supervised Deep Learning

James Hays
slides from Carl Doersch and Richard Zhang
Recap

Big Data
  – The Unreasonable Effectiveness of Data
  – Scene Completion
  – Im2gps
  – Recognition via Tiny Images

Crowdsourcing
  – "Wisdom of the Crowds" / consensus
  – Find good annotators through grading
  – Pricing affects throughput but not quality
  – User interface and instructions matter a lot
Today’s Lecture

• Three methods for “unsupervised” deep learning
  – Context Prediction. Doersch et al. ICCV 2015
  – SimCLR. Chen et al. ICML 2020

• Big picture: do we need big, labeled datasets like ImageNet to make deep learning worthwhile? Can we learn from something else?
Unsupervised Visual Representation Learning by Context Prediction

Carl Doersch, Alexei A. Efros, and Abhinav Gupta

ICCV 2015
ImageNet + Deep Learning

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

Beagle
ImageNet + Deep Learning

Do we need semantic labels?

Do we need this task?"
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 india, 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

A

B
Semantics from a non-semantic task
Relative Position Task

8 possible locations

Randomly Sample Patch
Sample Second Patch
Patch Embedding

Input

CNN

Nearest Neighbors

Note: connects *across* instances!
Architecture

Patch 1

Patch 2

Training requires Batch Normalization [Ioffe et al. 2015], but no other tricks

Architecture
Avoiding Trivial Shortcuts

Include a gap

Jitter the patch locations
A Not-So “Trivial” Shortcut
Chromatic Aberration
Chromatic Aberration
<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>
</tbody>
</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><img src="image1" alt="Input" /></td>
<td><img src="image2" alt="Ours" /></td>
<td><img src="image3" alt="Random Initialization" /></td>
<td><img src="image4" alt="ImageNet AlexNet" /></td>
</tr>
<tr>
<td><img src="image5" alt="Input" /></td>
<td><img src="image6" alt="Ours" /></td>
<td><img src="image7" alt="Random Initialization" /></td>
<td><img src="image8" alt="ImageNet AlexNet" /></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><img src="image9" alt="Input" /></td>
<td><img src="image10" alt="Ours" /></td>
<td><img src="image11" alt="Random Initialization" /></td>
<td><img src="image12" alt="ImageNet AlexNet" /></td>
</tr>
<tr>
<td><img src="image13" alt="Input" /></td>
<td><img src="image14" alt="Ours" /></td>
<td><img src="image15" alt="Random Initialization" /></td>
<td><img src="image16" alt="ImageNet AlexNet" /></td>
</tr>
<tr>
<td><img src="image17" alt="Input" /></td>
<td><img src="image18" alt="Ours" /></td>
<td><img src="image19" alt="Random Initialization" /></td>
<td><img src="image20" alt="ImageNet AlexNet" /></td>
</tr>
</tbody>
</table>
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

ImageNet Labels  Ours  No Pretraining

54.2  56.8  51.1
46.3  45.6  42.4
40.7  45.6  42.4

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

[Krähenbühl, Doersch, Donahue & Darrell, “Data-dependent Initializations of CNNs”, 2015]
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]
Colorful Image Colorization
Richard Zhang, Phillip Isola, Alexei (Alyosha) Efros
richzhang.github.io/colorization
Ansel Adams, Yosemite Valley
Bridge
Ansel Adams, Yosemite Valley Bridge – Our Result
Grayscale image: $L$ channel

$X \in \mathbb{R}^{H \times W \times 1}$

Color information: $ab$ channels

$\hat{Y} \in \mathbb{R}^{H \times W \times 2}$
Grayscale image: $L$ channel $X \in \mathbb{R}^{H \times W}$

$\mathcal{F}$

Concatenate $(L, ab)$

$(X, \hat{Y})$

Semantics? Higher-level abstraction?

“Free” supervisory signal
Inherent Ambiguity

Grayscale
Inherent Ambiguity

Our Output

Ground Truth
Better Loss Function

- Regression with L2 loss inadequate

\[ L_2(\hat{Y}, Y) = \frac{1}{2} \sum_{h,w} ||Y_{h,w} - \hat{Y}_{h,w}||_2^2 \]

Colors in \( ab \) space

(continuous)
Better Loss Function

- Regression with L2 loss inadequate
  \[ L_2(\hat{Y}, Y) = \frac{1}{2} \sum_{h,w} \| Y_{h,w} - \hat{Y}_{h,w} \|^2 \]

- Use **multinomial classification**
  \[ L(\hat{Z}, Z) = -\frac{1}{HW} \sum_{h,w} \sum_{a} Z_{h,w,q} \log(\hat{Z}_{h,w,q}) \]
Better Loss Function

• Regression with L2 loss inadequate
  \[ L_2(\hat{Y}, Y) = \frac{1}{2} \sum_{h,w} \| Y_{h,w} - \hat{Y}_{h,w} \|^2 \]

• Use multinomial classification
  \[ L(\hat{Z}, Z) = -\frac{1}{HW} \sum_{h,w} \sum_q Z_{h,w,q} \log(\hat{Z}_{h,w,q}) \]
Better Loss Function

- Regression with L2 loss inadequate
  \[ L_2(\hat{Y}, Y) = \frac{1}{2} \sum_{h,w} \| Y_{h,w} - \hat{Y}_{h,w} \|_2^2 \]

- Use **multinomial classification**
  \[ L(\hat{Z}, Z) = -\frac{1}{HW} \sum_{h,w} \sum_{q} Z_{h,w,q} \log(\hat{Z}_{h,w,q}) \]

- **Class rebalancing** to encourage learning of *rare* colors
  \[ L(\hat{Z}, Z) = -\frac{1}{HW} \sum_{h,w} v(Z_{h,w}) \sum_{q} Z_{h,w,q} \log(\hat{Z}_{h,w,q}) \]
Network Architecture

\[ \hat{Y} = F(X) \]

\[ \hat{Z} \in [0, 1]^{H \times W \times Q} \]
Network Architecture

$\hat{Z} = G(X)$

$\hat{Y} = H(\hat{Z})$

Failure Cases
Evaluation

Visual Quality

Quantitative

Per-pixel accuracy
## Evaluation

<table>
<thead>
<tr>
<th></th>
<th>Visual Quality</th>
<th>Representation Learning</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Quantitative</strong></td>
<td>Per-pixel accuracy</td>
<td>Task generalization</td>
</tr>
<tr>
<td></td>
<td>Perceptual realism</td>
<td>ImageNet classification</td>
</tr>
<tr>
<td></td>
<td>Semantic interpretability</td>
<td>Task &amp; dataset generalization</td>
</tr>
<tr>
<td></td>
<td></td>
<td>PASCAL classification, detection, segmentation</td>
</tr>
<tr>
<td><strong>Qualitative</strong></td>
<td>Low-level stimuli</td>
<td>Hidden unit activations</td>
</tr>
<tr>
<td></td>
<td>Legacy grayscale photos</td>
<td></td>
</tr>
</tbody>
</table>

- Hidden unit activations
# Evaluation

<table>
<thead>
<tr>
<th></th>
<th>Visual Quality</th>
<th>Representation Learning</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Quantitative</strong></td>
<td>Per-pixel accuracy</td>
<td>Task generalization</td>
</tr>
<tr>
<td></td>
<td>Perceptual realism</td>
<td>ImageNet classification</td>
</tr>
<tr>
<td></td>
<td>Semantic interpretability</td>
<td>Task &amp; dataset generalization</td>
</tr>
<tr>
<td></td>
<td>Low-level stimuli</td>
<td>PASCAL classification, detection, segmentation</td>
</tr>
<tr>
<td><strong>Qualitative</strong></td>
<td>Legacy grayscale photos</td>
<td>Hidden unit activations</td>
</tr>
</tbody>
</table>
Perceptual Realism / Amazon Mechanical Turk Test
clap if “fake”

clap if “fake”
Fake, 0% fooled
clap if “fake”

clap if “fake”
Fake, 55% fooled
clap if “fake”

clap if “fake”
Fake, 58% fooled
from Reddit /u/SherySantucci
Recolorized by Reddit ColorizeBot
Recolorized by Reddit ColorizeBot
Perceptual Realism Test

AMT Labeled Real [%]

<table>
<thead>
<tr>
<th>Method</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random</td>
<td>13.0</td>
</tr>
<tr>
<td>Ours (L2)</td>
<td>21.2</td>
</tr>
<tr>
<td>Ours (class)</td>
<td>23.9</td>
</tr>
<tr>
<td>Ours (full)</td>
<td>32.3</td>
</tr>
<tr>
<td>Larsson et al.</td>
<td>27.2</td>
</tr>
<tr>
<td>Ground Truth</td>
<td>50</td>
</tr>
</tbody>
</table>

1600 images tested per algorithm

[Bar chart showing the comparison of different methods in terms of perceived realism.]
Predicting Labels from Data

Supervised training

Data $x$

Learned feature hierarchy

Label $y$

ImageNet images

ImageNet labels
Predicting Data from Data

Supervised training

ImageNet images

Learned feature hierarchy

Label y

ImageNet labels

Unsupervised/ Self-supervised training

Learned feature hierarchy

$x_0$ $x_1$
Cross-Channel Encoder

Hidden Unit Activations

\[ \hat{Z} = \mathcal{G}(X) \]

\[ \hat{Y} = \mathcal{H}(\hat{Z}) \]

Task Generalization: ILSVRC linear classification

Are semantic classes *linearly separable* in the learned feature space?
Task Generalization: ILSVRC linear classification
Task Generalization: ILSVRC linear classification
Task: Generalization: ILSVRC linear classification
Hidden Unit (conv5) Activations

sky

trees

water
Hidden Unit (conv5) Activations

faces

dog faces

flowers
Dataset & Task Generalization on PASCAL VOC

Does the feature representation transfer to other datasets and tasks?

**Classification**

**Detection**

**Segmentation**
Does the method work on *legacy* black and white photos?
Thylacine, Dr. David Fleay, extinct in 1936.
Thylacine, Dr. David Fleay, extinct in 1936.
Amateur Family Photo, 1956.
Amateur Family Photo,
1956.
Henri Cartier-Bresson, Sunday on the Banks of the River Seine, 1938.
Henri Cartier-Bresson, Sunday on the Banks of the River Seine, 1938.
Dorothea Lange, Migrant Mother, 1936.
Dorothea Lange, Migrant Mother, 1936.
Additional Information

• Demo
  – http://demos.algorithmia.com/colorize-photos/

• Reddit ColorizeBot
  – Type “colorizebot” under any image post

• Code
  – https://github.com/richzhang/colorization

• Website – full paper, user examples, visualizations
  – http://richzhang.github.io/colorization
For the full paper, additional examples and our model:
richzhang.github.io/colorization
A Simple Framework for Contrastive Learning of Visual Representations

Ting Chen ¹  Simon Kornblith ¹  Mohammad Norouzi ¹  Geoffrey Hinton ¹

SimCLR, IMCL 2020
Figure 1. ImageNet Top-1 accuracy of linear classifiers trained on representations learned with different self-supervised methods (pretrained on ImageNet). Gray cross indicates supervised ResNet-50. Our method, SimCLR, is shown in bold.
(a) Original  (b) Crop and resize  (c) Crop, resize (and flip)  (d) Color distort. (drop)  (e) Color distort. (jitter)

(f) Rotate $\{90^\circ, 180^\circ, 270^\circ\}$  (g) Cutout  (h) Gaussian noise  (i) Gaussian blur  (j) Sobel filtering
How Useful is Self-Supervised Pretraining for Visual Tasks?

Alejandro Newell    Jia Deng
Princeton University
{anewell, jiadeng}@cs.princeton.edu

Figure 1. We highlight three possible outcomes when using self-supervised pretraining, the pretrained model either: a) always provides an improvement over the the model trained from scratch even as the amount of labeled data increases, b) reaches higher accuracy with fewer labels but plateaus to the same accuracy as the baseline, c) converges to baseline performance before accuracy plateaus. In our experiments we find option (c) to be the most common outcome.