“Unsupervised” Deep Learning

James Hays
slides from Carl Doersch and Richard Zhang
Recap from Previous Lecture

• We saw two strategies to get **structured** output while using deep learning
  – With object detection, one strategy is brute force: detect everywhere at once
Recap from Previous Lecture

• We saw two strategies to get structured output while using deep learning
  – With pose estimation / keypoint detection, the network produces an image-based intermediate representation
Recap from Previous Lecture

• More generally, it can pay off to get creative. Even if Deep ConvNets aren’t a natural fit for an image-related task, they might be able to learn a subtask or create a useful intermediate representation.
Today’s Lecture

• Two methods for “unsupervised” deep learning
  – Context Prediction. Doersch et al. ICCV 2015

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

Carl Doersch
Joint work with Alexei A. Efros & Abhinav Gupta

ICCV 2015
ImageNet + Deep Learning

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

- Materials?
- Geometry?
- Parts?
- Boundaries?
- Pose?

ImageNet + Deep Learning

Beagle
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
Context Prediction for Images

A

B
Semantics from a non-semantic task
Relative Position Task

Randomly Sample Patch

Sample Second Patch

8 possible locations

CNN

Classifier
CNN Classifier

Patch Embedding

Input Nearest Neighbors

Note: connects *across* instances!
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
What is learned?

Input

Ours

Random Initialization

ImageNet AlexNet
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="Image" /></td>
<td><img src="image2" alt="Image" /></td>
<td><img src="image3" alt="Image" /></td>
<td><img src="image4" alt="Image" /></td>
</tr>
<tr>
<td><img src="image5" alt="Image" /></td>
<td><img src="image6" alt="Image" /></td>
<td><img src="image7" alt="Image" /></td>
<td><img src="image8" alt="Image" /></td>
</tr>
<tr>
<td><img src="image9" alt="Image" /></td>
<td><img src="image10" alt="Image" /></td>
<td><img src="image11" alt="Image" /></td>
<td><img src="image12" alt="Image" /></td>
</tr>
<tr>
<td><img src="image13" alt="Image" /></td>
<td><img src="image14" alt="Image" /></td>
<td><img src="image15" alt="Image" /></td>
<td><img src="image16" alt="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><img src="image17" alt="Image" /></td>
<td><img src="image18" alt="Image" /></td>
<td><img src="image19" alt="Image" /></td>
<td><img src="image20" alt="Image" /></td>
</tr>
<tr>
<td><img src="image21" alt="Image" /></td>
<td><img src="image22" alt="Image" /></td>
<td><img src="image23" alt="Image" /></td>
<td><img src="image24" alt="Image" /></td>
</tr>
<tr>
<td><img src="image25" alt="Image" /></td>
<td><img src="image26" alt="Image" /></td>
<td><img src="image27" alt="Image" /></td>
<td><img src="image28" alt="Image" /></td>
</tr>
<tr>
<td><img src="image29" alt="Image" /></td>
<td><img src="image30" alt="Image" /></td>
<td><img src="image31" alt="Image" /></td>
<td><img src="image32" alt="Image" /></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

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

ImageNet Labels

- Ours
- No Pretraining

- 68.6
- 56.8
- 54.2
- 51.1
- 46.3
- 45.6
- 42.4
- 40.7

[Krähenbühl, Doersch, Donahue & Darrell, “Data-dependent Initializations of CNNs”, 2015]
Capturing Geometry?
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
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

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

Concatenate $(L, ab)$

$(X, \hat{Y})$

“Free” supervisory signal

$\mathcal{F}$

Semantics? Higher-level abstraction?
Inherent Ambiguity

Grayscale
Inherent Ambiguity

Our Output

Ground Truth
Better Loss Function

- Regression with L2 loss inadequate

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

- 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
  \[ L(\hat{Z}, Z) = -\frac{1}{HW} \sum_{h,w} v(Z_{h,w}) \sum_q Z_{h,w,q} \log(\hat{Z}_{h,w,q}) \]


Hand-engineered Features


L2 Regression

Deep Networks


Classification


Larsson et al. In ECCV 2016. [Concurrent]
Network Architecture

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

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

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

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

Ground Truth  L2 Regression  Class w/ Rebalancing
Failure Cases
Biases
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>Low-level stimuli</td>
<td>PASCAL classification, detection, segmentation</td>
</tr>
<tr>
<td></td>
<td>Legacy grayscale photos</td>
<td>Hidden unit activations</td>
</tr>
<tr>
<td><strong>Qualitative</strong></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
## 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><strong>Perceptual realism</strong></td>
<td>ImageNet classification</td>
</tr>
<tr>
<td></td>
<td>Semantic interpretability</td>
<td>Task &amp; dataset generalization</td>
</tr>
<tr>
<td><strong>Qualitative</strong></td>
<td>Low-level stimuli</td>
<td>PASCAL classification, detection, segmentation</td>
</tr>
<tr>
<td></td>
<td>Legacy grayscale photos</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Hidden unit activations</td>
</tr>
</tbody>
</table>
Perceptual Realism / Amazon Mechanical Turk Test
clap if “fake”

clap if “fake”
clap if “fake”
clap if “fake”
Fake, 55% fooled
clap if “fake”
Fake, 58% fooled
from Reddit /u/SherySantucci
Photo taken by Reddit /u/Timteroo, Mural from street artist Eduardo Kobra
Perceptual Realism Test

AMT Labeled Real [%]

- **Ground Truth**: 50%
- **Ours (L2)**: 21.2%
- **Ours (class)**: 23.9%
- **Ours (full)**: 32.3%
- **Larsson et al.** (Concurrent): 27.2%
- **Random**: 13.0%

1600 images tested per algorithm
Predicting Labels from Data

Supervised training

Data $x$

ImageNet images

Learned feature hierarchy

Label $y$

ImageNet labels
Predicting Data from Data

Supervised training

Unsupervised/Self-supervised training

ImageNet images

Learned feature hierarchy

Label y

ImageNet labels
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
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**
Dataset & Task Generalization on PASCAL VOC

ImageNet Labels

% from Gaussian to ImageNet labels

Gaussian Initialization

Classification
Detection
Segmentation

Autoencoder
Wang & Gupta
Krähenbühl et al.
Doersch et al.
Agrawal et al.
Donahue et al.
Pathak et al.
Ours
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.
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

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

• Code
  – [https://github.com/richzhang/colorization](https://github.com/richzhang/colorization)

• Website – full paper, user examples, visualizations
For the full paper, additional examples and our model:
richzhang.github.io/colorization