Self-supervised Representation Learning

Computer Vision
Fall 2018
Columbia University
Project

• Presentation schedule posted on Piazza.

• Review it ASAP and let us know of any problems by Wednesday

• For those presenting on December 5: OK to have some experiments in progress

• Final reports due December 10 midnight — no extensions!
GPU Credits

- If you have not requested GPU credits, do so immediately.
- We are starting to give them away…
Homeworks

- HW3 is back: median is 100%!
- HW4 grades soon
- HW5 due today
Final Grades

• We will likely curve down, but we will guarantee:
  • 90% is at least A
  • 80% is at least B
  • 70% is at least C
Next Semester

• E6998 Advanced Computer Vision, offered Spring 2019

• Focuses on research frontier of computer vision and applied machine learning

• Make sure to fill out survey to get off wait list
Observed image

Drawn from memory

[Bartlett, 1932]

[Intraub & Richardson, 1989]
Observed image

Drawn from memory

[Bartlett, 1932]

[Intraub & Richardson, 1989]
“I stand at the window and see a house, trees, sky. Theoretically I might say there were 327 brightnesses and nuances of colour. Do I have "327"? No. I have sky, house, and trees.”

— Max Wertheimer, 1923
Representation learning

Image $X$ → Compact mental representation

“Coral” → “Fish”

Slide credit: Phillip Isola
Representation learning

Good representations are:

1. Compact
2. Explanatory
3. Disentangled
4. Interpretable

Slide credit: Phillip Isola
A CNN is a multiscale, hierarchical representation of data.
Training
Object recognition

Testing
Place recognition

Often, what we will be “tested” on is to learn to do a new thing.
Finetuning starts with the representation learned on a previous task, and adapts it to perform well on a new task.
If we keep on finetuning for every new datapoint or task that comes our way, we get **online learning**. Humans seem to do this, we never stop learning.
Supervised object recognition

image X \rightarrow \text{Learner} \rightarrow \text{“Fish”}

label Y

Slide credit: Phillip Isola
Supervised object recognition

image X → Learner → "Fish" → label Y

Slide credit: Phillip Isola
Supervised object recognition

image X \rightarrow \text{Learner} \rightarrow \text{“Fish”}

label Y

Slide credit: Phillip Isola
Supervised object recognition

image X \rightarrow \text{Learner} \rightarrow \text{“Fish”}

label Y

Slide credit: Phillip Isola
Kitten Carousel

Held and Hein, 1963

Fig. 1. Apparatus for equating motion and consequent visual feedback for an actively moving (A) and a passively moved (P) S.
Supervised computer vision

- Hand-curated training data
  - Informative
  - Expensive
  - Limited to teacher’s knowledge

Vision in nature

- Raw unlabeled training data
  - Cheap
  - Noisy
  - Harder to interpret
Learning from examples
(aka supervised learning)

Training data

\[
\{x_1, y_1\} \quad \rightarrow \quad \text{Learner} \quad \rightarrow \quad f : X \rightarrow Y
\]

\[
f^* = \arg \min_{f \in \mathcal{F}} \sum_{i=1}^{N} \mathcal{L}(f(x_i), y_i)
\]

Slide credit: Phillip Isola
Learning without examples
(includes unsupervised learning and reinforcement learning)

Data

\{x_1\} \rightarrow \text{Learner} \rightarrow ?
Unsupervised Representation Learning

Data

\{x_1\}
\{x_2\}
\{x_3\}
\ldots

Learner

\rightarrow \text{Representations}

Slide credit: Phillip Isola
Unsupervised Representation Learning

X

Image

“Fish”

“Coral”

Compact mental representation

Slide credit: Phillip Isola
Unsupervised Representation Learning

Image

\[ X \] compressed image
code (vector \( z \))

Slide credit: Phillip Isola
Unsupervised Representation Learning

Image $X$ → compressed image code (vector $z$) → Reconstructed image $\hat{X}$

“Autoencoder”

[e.g., Hinton & Salakhutdinov, Science 2006]
Autoencoder

\[ \text{arg min}_{\mathcal{F}} \mathbb{E}_X [\| \mathcal{F}(X) - X \|] \]

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

[e.g., Hinton & Salakhutdinov, Science 2006]
\[
\hat{X} = \mathcal{F}(X)
\]

[e.g., Hinton & Salakhutdinov, Science 2006]
Data compression

\[ \text{Data} \xrightarrow{\text{method}} \hat{\text{Data}} \]

[e.g., Hinton & Salakhutdinov, Science 2006]
Data prediction

\[ X_1 \rightarrow \text{Some data} \rightarrow \text{Other data} \rightarrow \hat{X}_2 \]

Slide credit: Phillip Isola
Grayscale image: L channel

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

Color information: ab channels

\[ \hat{Y} \in \mathbb{R}^{H \times W \times 2} \]

[Zhang, Isola, Efros, ECCV 2016]
Ansel Adams, Yosemite Valley Bridge
Result of [Zhang et al., ECCV 2016]
Image colorization

Input $\mathbf{x}$ \hspace{1cm} Output $\mathbf{y}$

$\{x, y\}$

$\{\text{Training data}\}$

$\arg\min_{f \in \mathcal{F}} \sum_{i=1}^{N} \mathcal{L}(f(x_i), y_i)$

[Zhang et al., ECCV 2016]
Choosing loss and representation

\[
\mathcal{L}(f(x), y) = \| f(x) - y \|^2_2
\]

Slide credit: Phillip Isola
\[ \mathcal{L}(f(x), y) = \| f(x) - y \|_2^2 \]
\( \mathbf{y} \in \mathbb{R}^{H \times W \times 2} \)

\[
\mathcal{L}(f(x), y) = \| f(x) - y \|_2^2
\]

Prediction for a single pixel \(i,j\)

Slide credit: Phillip Isola
\[ y \in \mathbb{R}^{H \times W \times 2} \]
\( y \in \mathbb{R}^{H \times W \times K} \)

\[ \mathcal{L}(y, f_{\theta}(x)) = H(y, \text{softmax}(f_{\theta}(x))) \]
\[ L(x, y) = H(y, \text{softmax}(f_\theta(x))) \]
- Continuous-valued prediction
- (Usually) models unimodal distribution

- Discrete-valued prediction
- Models multimodal distribution
Instructive failure
Instructive failure
Deep Net “Electrophysiology”

[Zeiler & Fergus, ECCV 2014]
[Zhou et al., ICLR 2015]
Stimuli that drive selected neurons (conv5 layer)

- faces
- dog
- faces
- flowers
Is the code informative about object class $\mathbf{y}$?

Logistic regression:

$$y = \sigma(Wz + b)$$
Convolutional Neural Network (CNN) architecture with layers `conv1`, `conv2`, `conv3`, `conv4`, `conv5`, and `pool1`, `pool2`, `pool5`. The figure shows the process from raw data to reconstructed data and predicted color channels.

**Raw Data**:
- Input: $x$
- Process:
  - $x_1$
  - $x_2$
  - $\hat{x}$
- Output: $\hat{x}$

**Classification Performance**:
- Accuracy graph comparing autoencoder and colorization tasks.
- Layers: `conv1`, `pool1`, `conv2`, `pool2`, `conv3`, `conv4`, `conv5`, `pool5`.
- The graph shows an increase in accuracy with each layer, indicating better performance.

Task from [Russakovsky et al. 2015]
Self-supervised learning

Common trick:

- Convert “unsupervised” problem into “supervised” empirical risk minimization
- Do so by cooking up “labels” (prediction targets) from the raw data itself
Multisensory self-supervision

Virginia de Sa. Learning Classification with Unlabeled Data. NIPS 1994. [see also “Six lessons from babies”, Smith and Gasser 2005]
Ambient Sound Provides Supervision for Visual Learning

Andrew Owens   Jiajun Wu   Josh McDermott
William Freeman   Antonio Torralba

MIT, Google Research
Audio is invariant to many visual transformations.
Audio is invariant to many visual transformations.
Audio is invariant to many visual transformations
Audio is invariant to many visual transformations.
Audio is invariant to many visual transformations.

Image space

Audio space

Slide credit: Andrew Owens
Predicting sound

- Flickr video dataset.
- 180K videos, 10 random frames from each.
- Trained from scratch

Slide credit: Andrew Owens

Sound texture
[McDermott & Simoncelli 2011]
Predicting sound

Video frame

ConvNet

Sound feature

Slide credit: Andrew Owens
Predicting sound

Video frame

ConvNet

Audio

K-means or

Slide credit: Andrew Owens
Top audio clips for one cluster (of 30).
Top audio clips for one cluster (of 30).
What did the network learn?
What did the network learn?

PASCAL VOC 2007

Slide credit: Andrew Owens
PASCAL VOC Classification

% mAP

Sound
[Doersch15]

Context
[Wang15]

Tracking
[Agrawal15]

Egomotion Spectrum
only

Visual clusters
[ImageNet][Krizhevsky12]

46.1
42.2
31.2
44.0
37.5
65.5

Slide credit: Andrew Owens
SUN397 Scene Recognition

<table>
<thead>
<tr>
<th>Method</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sound</td>
<td>22.5%</td>
</tr>
<tr>
<td>Context</td>
<td>22.2%</td>
</tr>
<tr>
<td>Tracking</td>
<td>18.7%</td>
</tr>
<tr>
<td>Egomotion</td>
<td>11.3%</td>
</tr>
<tr>
<td>Places</td>
<td>42.1%</td>
</tr>
</tbody>
</table>

Zhou14

Slide credit: Andrew Owens
What did the network learn?
Unit visualizations

Top responses (unit #90)

Audio label
Unit visualizations

Audio label

256

conv

13

Slide credit: Andrew Owens
Unit visualizations

Slide credit: Andrew Owens
Unsupervised visual representation learning by context prediction
[Carl Doersch, Abhinav Gupta, Alexei A. Efros, ICCV 2015]
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 ashes, or 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
Randomly Sample Patch
Sample Second Patch

Relative Position Task

<8 possible locations

Classifier

CNN

CNN

[Slide credit: Carl Doersch]
Patch Embedding (representation)

Input  Nearest Neighbors

Note: connects *across* instances!

[Slide credit: Carl Doersch]
Learning by Rotating

Unsupervised Representation Learning by Predicting Image Rotations
Spyros Gidaris, Praveer Singh, Nikos Komodakis
## How are we doing?

<table>
<thead>
<tr>
<th></th>
<th>Classification</th>
<th>Detection</th>
<th>Segmentation</th>
</tr>
</thead>
<tbody>
<tr>
<td>ImageNet</td>
<td>78.2%</td>
<td>56.8%</td>
<td>48.0%</td>
</tr>
<tr>
<td>Context</td>
<td>55.3%</td>
<td>46.6%</td>
<td>-</td>
</tr>
<tr>
<td>Jigsaw Puzzle</td>
<td>67.6%</td>
<td>53.2%</td>
<td>37.6%</td>
</tr>
<tr>
<td>Inpainting</td>
<td>56.5%</td>
<td>44.5%</td>
<td>30.0%</td>
</tr>
<tr>
<td>Colorization</td>
<td>61.5%</td>
<td>46.9%</td>
<td>35.6%</td>
</tr>
<tr>
<td>Tracking</td>
<td>58.7%</td>
<td>47.4%</td>
<td>-</td>
</tr>
<tr>
<td>Counting</td>
<td>67.7%</td>
<td>51.4%</td>
<td>36.6%</td>
</tr>
<tr>
<td>Rotation</td>
<td>72.9%</td>
<td>54.4%</td>
<td>39.1%</td>
</tr>
</tbody>
</table>

PASCAL VOC 2007
1. To survive, biological agents are constantly trying to anticipate, to predict sensations.

2. This trains up representations useful for prediction — surfaces, objects, events!

Prediction hypothesis
Yann LeCun’s cake:

1. Cake is unsupervised representation learning
2. Frosting is supervised transfer learning
3. Cherry on top is reinforcement learning (model-based RL)