CS 2770: Computer Vision

Self-Supervised Learning & Visual Discovery

Prof. Adriana Kovashka

University of Pittsburgh

April 10, 2017
Motivation

• So far we’ve assumed access to plentiful labeled data
• How is this data obtained?
Crowdsourcing

Task: Dog?

Answer: Yes

Pay: $0.01

Is this a dog?

- Yes
- No

Broker

www.mturk.com

Workers

Alex Sorokin
Crowdsourcing via games

• The ESP Game
  – Two-player online game
  – Partners don’t know each other and can’t communicate
  – Objective of the game: type the same word
  – The only thing in common is an image

THE ESP GAME

PLAYER 1

GUESSING: CAR
GUESSING: HAT
GUESSING: KID
SUCCESS!
YOU AGREE ON CAR

PLAYER 2

GUESSING: BOY
GUESSING: CAR
SUCCESS!
YOU AGREE ON CAR

Motivation

• So far we’ve assumed access to plentiful labeled data
• **What if we have limited or no labeled data?**
• One approach: learn from unlabeled data (unsupervised learning)
  – Mine for interesting patterns (discovery)
  – Use supervision (labels) inherent in the data (self-supervised learning)
• Another approach (not discussed): carefully choose which data to label
  – Active learning, human-in-the-loop
Plan for this lecture

• Self-supervised learning
  – From surrounding regions (ICCV 2015)
  – From temporal order (ECCV 2016)
  – From ego-motion (ICCV 2015)

• Visual discovery
  – From surrounding objects (CVPR 2010)
  – Elements of architectural styles (SIGGRAPH 2012)
  – Gradual style changes in time/space (ICCV 2013)
Unsupervised Visual Representation Learning by Context Prediction

Carl Doersch, Alexei Efros and Abhinav Gupta

ICCV 2015
ImageNet + Deep Learning

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

ImageNet + Deep Learning

Do we even need semantic labels?

Materials?
Parts?
Geometry?
Boundaries?
Pose?

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 mink, 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

8 possible locations

Architecture

Softmax loss
Fully connected

Fully connected

Fully connected
Max Pooling
Convolution
Convolution
Convolution
LRN
Max Pooling
Convolution
Convolution
LRN
Max Pooling
Convolution

Tied Weights

Patch 1
Patch 2

What is learned?

<table>
<thead>
<tr>
<th>Input</th>
<th>Ours</th>
<th>ImageNet AlexNet</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image1" alt="Input Image" /></td>
<td><img src="image2" alt="Ours Image" /></td>
<td><img src="image3" alt="ImageNet AlexNet Image" /></td>
</tr>
<tr>
<td><img src="image4" alt="Input Image" /></td>
<td><img src="image5" alt="Ours Image" /></td>
<td><img src="image6" alt="ImageNet AlexNet Image" /></td>
</tr>
<tr>
<td><img src="image7" alt="Input Image" /></td>
<td><img src="image8" alt="Ours Image" /></td>
<td><img src="image9" alt="ImageNet AlexNet Image" /></td>
</tr>
<tr>
<td><img src="image10" alt="Input Image" /></td>
<td><img src="image11" alt="Ours Image" /></td>
<td><img src="image12" alt="ImageNet AlexNet Image" /></td>
</tr>
<tr>
<td><img src="image13" alt="Input Image" /></td>
<td><img src="image14" alt="Ours Image" /></td>
<td><img src="image15" alt="ImageNet AlexNet Image" /></td>
</tr>
<tr>
<td><img src="image16" alt="Input Image" /></td>
<td><img src="image17" alt="Ours Image" /></td>
<td><img src="image18" alt="ImageNet AlexNet Image" /></td>
</tr>
<tr>
<td><img src="image19" alt="Input Image" /></td>
<td><img src="image20" alt="Ours Image" /></td>
<td><img src="image21" alt="ImageNet AlexNet Image" /></td>
</tr>
</tbody>
</table>

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: 54.2
Ours: 46.3
No Pretraining: 40.7

Shuffle and Learn: Unsupervised Learning using Temporal Order Verification

Ishan Misra, C. Lawrence Zitnick, and Martial Hebert

ECCV 2016
Fig. 1: (a) A video imposes a natural temporal structure for visual data. In many cases, one can easily verify whether frames are in the correct temporal order (shuffled or not). Such a simple sequential verification task captures important spatiotemporal signals in videos. We use this task for unsupervised pre-training of a Convolutional Neural Network (CNN). (b) Some examples of the automatically extracted positive and negative tuples used to formulate a classification task for a CNN.
Fig. 2: (a) We sample tuples of frames from high motion windows in a video. We form positive and negative tuples based on whether the three input frames are in the correct temporal order. (b) Our triplet Siamese network architecture has three parallel network stacks with shared weights up to the $fc7$ layer. Each stack takes a frame as input, and produces a representation at the $fc7$ layer. The concatenated $fc7$ representations are used to predict whether the input tuple is in the correct temporal order.

Table 2: Mean classification accuracies over the 3 splits of UCF101 and HMDB51 datasets. We compare different initializations and finetune them for action recognition.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Initialization</th>
<th>Mean Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>UCF101</td>
<td>Random</td>
<td>38.6</td>
</tr>
<tr>
<td></td>
<td>(Ours) Tuple verification</td>
<td><strong>50.2</strong></td>
</tr>
<tr>
<td>HMDB51</td>
<td>Random</td>
<td>13.3</td>
</tr>
<tr>
<td></td>
<td>UCF Supervised</td>
<td>15.2</td>
</tr>
<tr>
<td></td>
<td>(Ours) Tuple verification</td>
<td><strong>18.1</strong></td>
</tr>
</tbody>
</table>
Learning image representations tied to ego-motion

Dinesh Jayaraman and Kristen Grauman
ICCV 2015
The kitten carousel experiment
[Held & Hein, 1963]

active kitten

passive kitten

Key to perceptual development:
**self-generated motion + visual feedback**
Problem with today’s visual learning

**Status quo:** Learn from “disembodied” bag of labeled snapshots.

**Our goal:** Learn in the context of acting and moving in the world.

Our idea: Ego-motion $\leftrightarrow$ vision

Goal: Teach computer vision system the connection: “how I move” $\leftrightarrow$ “how my visual surroundings change”

Ego-motion motor signals + Unlabeled video

Ego-motion $\leftrightarrow$ vision: view prediction

After moving:

Ego-motion ↔ vision for recognition

Learning this connection requires:

- Depth, 3D geometry
- Semantics
- Context

Also key to recognition!

Can be learned without manual labels!

Our approach: unsupervised feature learning using egocentric video + motor signals

Approach idea: Ego-motion equivariance

**Invariant features**: unresponsive to some classes of transformations

\[ z(gx) \approx z(x) \]

**Equivariant features**: predictably responsive to some classes of transformations, through simple mappings (e.g., linear)

\[ z(gx) \approx M_g z(x) \]

Invariance **discards** information; equivariance **organizes** it.

Approach idea: Ego-motion equivariance

**Training data**
Unlabeled video +
motor signals

**Equivariant embedding**
organized by ego-motions

Pairs of frames related by
similar ego-motion should
be related by same
feature transformation

---

Approach overview

Our approach: unsupervised feature learning using egocentric video + motor signals

1. Extract training frame pairs from video
2. Learn ego-motion-equivariant image features
3. Train on target recognition task in parallel

Training frame pair mining

Discovery of ego-motion clusters

forward distance

yaw change

Ego-motion equivariant feature learning

Given:

Desired: for all motions $g$ and all images $x$,

$$z_\theta(gx) \approx M_g z_\theta(x)$$

Unsupervised training

Supervised training

Jayaraman and Grauman, "Learning image representations tied to ego-motion", ICCV 2015
Ego-motion training pairs

Neural network training

Equivariant embedding

APPROACH

Scene and object recognition

Next-best view selection

Football field?
Pagoda?
Airport?
Cathedral?
Army base?

cup
frying pan

Results: Recognition

Learn from **unlabeled car video** (KITTI)

Exploit features for **static scene classification** (SUN, 397 classes)

Geiger et al, IJRR ’13

**Results: Recognition**

Do ego-motion equivariant features improve recognition?

Up to 30% accuracy increase over state of the art!

Discussion

• Many types of supervision cues have been tried
• What other types of supervision “for free” can we use?
• How do we know if a certain supervision type would work?
• Can we make this type of learning perform on par with supervised learning?
Plan for next two lectures

• Self-supervised learning
  – From surrounding regions (ICCV 2015)
  – From temporal order (ECCV 2016)
  – From ego-motion (ICCV 2015)

• Visual discovery
  – From surrounding objects (CVPR 2010)
  – Elements of architectural styles (SIGGRAPH 2012)
  – Gradual style changes in time/space (ICCV 2013)
Object-Graphs for Context-Aware Category Discovery

Yong Jae Lee and Kristen Grauman

CVPR 2010
Goal

- Discover \textit{new} object categories, based on their relation to categories for which we have trained models

Lee and Grauman, “Object-Graphs for Context-Aware Category Discovery”, CVPR 2010
Existing approaches

Previous work treats unsupervised visual discovery as an appearance-grouping problem.

Can you identify the recurring pattern?

Lee and Grauman, “Object-Graphs for Context-Aware Category Discovery”, CVPR 2010
Our idea

How can seeing previously learned objects in novel images help to discover *new* categories?

Can you identify the recurring pattern?

Lee and Grauman, “Object-Graphs for Context-Aware Category Discovery”, CVPR 2010
Our idea

Discover visual categories within unlabeled images by modeling interactions between the unfamiliar regions and familiar objects.

Can you identify the recurring pattern?

Lee and Grauman, “Object-Graphs for Context-Aware Category Discovery”, CVPR 2010
Context-aware visual discovery

Lee and Grauman, “Object-Graphs for Context-Aware Category Discovery”, CVPR 2010
Learn “Known” Categories

- Offline: Train region-based classifiers for $N$ “known” categories using labeled training data.

Lee and Grauman, “Object-Graphs for Context-Aware Category Discovery”, CVPR 2010
Identifying Unknown Objects

Input: unlabeled pool of novel images

Compute multiple segmentations for each unlabeled image

Lee and Grauman, “Object-Graphs for Context-Aware Category Discovery”, CVPR 2010
Identifying Unknown Objects

- For all segments, use classifiers to compute posteriors for the $N$ “known” categories.
- Deem each segment as “known” or “unknown” based on resulting entropy.

Lee and Grauman, “Object-Graphs for Context-Aware Category Discovery”, CVPR 2010
• Model the topology of category predictions relative to the unknown (unfamiliar) region.
• Incorporate uncertainty from classifiers.

Lee and Grauman, “Object-Graphs for Context-Aware Category Discovery”, CVPR 2010
Object-Graphs

An unknown region within an image

Closest nodes in its object-graph

- Consider spatially near regions *above* and *below*, record distributions for each known class.

\[
g(s) = \begin{bmatrix}
0_{self} & 0_{self} & 1_{above} & 1_{below} & R_{above} & R_{below} \\
\end{bmatrix}
\]

Lee and Grauman, “Object-Graphs for Context-Aware Category Discovery”, CVPR 2010
Object-Graph matching

- Object-graphs are very similar $\rightarrow$ produces a strong match

Lee and Grauman, “Object-Graphs for Context-Aware Category Discovery”, CVPR 2010
Object-Graph matching

- Object-graphs are partially similar $\rightarrow$ produces a fair match

Lee and Grauman, “Object-Graphs for Context-Aware Category Discovery”, CVPR 2010
Clusters from region-region affinities

\[ K(s_i, s_j) = K_{app}(s_i, s_j) + K_{obj-graph}(s_i, s_j) \]

Lee and Grauman, “Object-Graphs for Context-Aware Category Discovery”, CVPR 2010
Object Discovery Accuracy

Lee and Grauman, “Object-Graphs for Context-Aware Category Discovery”, CVPR 2010
Examples of Discovered Categories

Lee and Grauman, “Object-Graphs for Context-Aware Category Discovery”, CVPR 2010
What Makes Paris Look like Paris?

Carl Doersch, Saurabh Singh, Abhinav Gupta, Josef Sivic, Alexei Efros
SIGGRAPH 2012
One of these is from Paris
Raise your hand if...

...this is Paris

Raise your hand if...

We showed 20 subjects:
- 100 Random Street View Images
- 50 from Paris
- They classified Paris vs non-Paris
- Accuracy: 79%

How do they know?

We showed 20 subjects:
- 100 Random Street View Images
- 50 from Paris
- They classified Paris non-Paris
- Accuracy: 79%

How do they know?

Our Goal:

*Given a large geo-tagged image dataset, we automatically discover visual elements that characterize a geographic location*

*Why might this be a useful task?*

Our Hypothesis

• The visual elements that capture Paris:
  – Frequent: Occur often in Paris
  – Discriminative: Are not found outside Paris
Step 1: Nearest Neighbors for Every Patch
Using normalized correlation of HOG features as a distance metric

Step 2: Find the Parisian Clusters by Sorting

Approach Summary

1. A Cluster for Every Patch

1. A cluster for every patch
2. Find clusters that are mostly Parisian

Approach Summary

1. A cluster for every patch
2. Find clusters that are mostly Parisian
3. Refine clusters by making them more Parisian

<table>
<thead>
<tr>
<th>Elements from Prague</th>
<th>Elements from London</th>
<th>Elements from Barcelona</th>
</tr>
</thead>
</table>

In the U.S.

Elements from San Francisco

Elements from Boston

Style-aware Mid-level Representation for Discovering Visual Connections in Space and Time

Yong Jae Lee, Alexei A. Efros, and Martial Hebert

ICCV 2013
Our Goal

- Mine mid-level visual elements in temporally- and spatially-varying data and model their “visual style”

*when?*  
Historical dating of cars

*where?*  
Geolocalization of StreetView images

Lee et al., “Style-aware Mid-level Representation for Discovering Visual Connections in Space and Time”, ICCV 2013
Key Idea

1) Establish connections

2) Model style-specific differences

“closed-world”

Lee et al., “Style-aware Mid-level Representation for Discovering Visual Connections in Space and Time”, ICCV 2013
Mining style-sensitive elements

Patch

Nearest neighbors

Lee et al., “Style-aware Mid-level Representation for Discovering Visual Connections in Space and Time”, ICCV 2013
Mining style-sensitive elements

Lee et al., “Style-aware Mid-level Representation for Discovering Visual Connections in Space and Time”, ICCV 2013
(a) Peaky (low-entropy) clusters

Lee et al., “Style-aware Mid-level Representation for Discovering Visual Connections in Space and Time”, ICCV 2013
Uninformative elements

(b) Uniform (high-entropy) clusters

Lee et al., “Style-aware Mid-level Representation for Discovering Visual Connections in Space and Time”, ICCV 2013