Object Recognition

Computer Vision
Fall 2018
Columbia University
Project Proposals and Homework 3

• How’s it going?
• Questions?
Quick Experiment

- Get pen and paper
- Draw a coffee cup
Two Extremes of Vision

Extrapolation problem
Generalization
Diagnostic features

Interpolation problem
Correspondence
Finding the differences

Number of training samples

1 10 $10^2$ $10^3$ $10^4$ $10^5$ $10^6$
Tiny Images

80 million tiny images: a large dataset for non-parametric object and scene recognition

c) Segmentation of 32x32 images
Given a benchmark, resolution and human scene recognition accuracy increase to a limit
Humans vs. Computers: Car Classification

Humans for 32 pixel tall images

Various computer vision algorithms for *full* resolution images

Example test images

Torralba et al.
Powers of 10

Number of images on my hard drive: $10^6$

Number of images seen during my first 10 years: $10^8$
(3 images/second * 60 * 60 * 16 * 365 * 10 = 630,720,000)

Number of images seen by all humanity: $10^{20}$
106,456,367,669 humans * 60 years * 3 images/second * 60 * 60 * 16 * 365 =

Number of photons in the universe: $10^{88}$

Number of all 32x32 images: $10^{7373}$
256 $32^3$ ~ $10^{7373}$
But not all scenes are so original
Lots
Of
Images

A. Torralba, R. Fergus, W.T.Freeman. PAMI 2008
Lots Of Images
Lots Of Images
Application: Automatic Colorization

Input

Color Transfer

Color Transfer

Matches (gray)

Matches (w/ color)

Avg Color of Match
Application: Automatic Colorization
Person Recognition

(a) Example images with different difficulty levels:
- 20-100% difficulty
- 5-20% difficulty
- 1-5% difficulty
- <1% difficulty

(b) ROC curves showing detection rate vs. false alarm rate for different difficulty levels:
- Red: 20-100%
- Green: 5-20%
- Blue: 1-5%
- Yellow: <1%
Exploring the Limits of Weakly Supervised Pretraining

Laurens van der Maaten

ECCV 2018

Research question

Can we use large amounts of weakly supervised images for pretraining?

Highlights

- We pretrain models by predicting relevant hashtags for images
- We pretrain models to predict 17.5K hashtags for 3.5B images
- After finetuning, we beat the state-of-the-art on, e.g., ImageNet
Hashtag Supervision

- It is easy to get billions of public images and hashtags
- Hashtags are more structured than captions
- Hashtags were often assigned to make images “searchable”

#cheesecake #birthday
Hashtag Supervision

- But hashtags are not perfect supervision
- Some hashtags are not visually relevant
- Other hashtags are not in the photo
- And there are many false negatives
Experiments

- Select a set of hashtags
- Download all public Instagram images that has at least one of these hashtags
- Use WordNet synsets to merge hashtags into canonical form (merge #brownbear and #ursusarctos)
- The final list has 17,517 hashtags

3,500,000,000 images!
Experiments

- Select a set of hashtags

- Download all public Instagram images that has at least one of these hashtags

- Use WordNet synsets to merge hashtags into canonical form (merge #brownbear and #ursusarctos)

- Final dataset has ~3.5 billion images
Experiments

- Select a set of hashtags
- Download all public Instagram images that has at least one of these hashtags
- Use WordNet synsets to merge hashtags into canonical form (merge #brownbear and #ursusarctos)
- Final dataset has ~3.5 billion images
Experiments

- Train ResNeXt-32xCd convolutional networks
- Use c-of-K vector to represent multiple labels
- Train to minimize multi-class logistic loss
- Distribute training batches across 336 GPUs
- Scale learning rate by batch size \(N=8,064\) after learning rate "warm-up" (Goyal et al., 2017)
Fix Model; Vary Data

- Pretrain model on ImageNet or Instagram
- Finetune on ImageNet
Fix Model;
Vary Data

- Pretrain model on ImageNet or Instagram
- Finetune on ImageNet

"standard" ImageNet training
Fix Model; Vary Data

- Pretrain model on ImageNet or Instagram
- Finetune on ImageNet

pre-training on 1B Instagram images, selected to match ImageNet classes
Fix Model; Vary Data

- Pretrain model on ImageNet or Instagram
- Finetune on ImageNet

pretraining on 1-3.5B Instagram images, without selection
Fix Model; Vary Data

- Pretrain model on ImageNet or Instagram
- Finetune on ImageNet
- Similar results on larger versions of ImageNet
Two Extremes of Vision

Extrapolation problem
Generalization
Diagnostic features

Interpolation problem
Correspondence
Finding the differences

Slide credit: Aude Oliva
Exemplar-SVMs

- Learn a separate linear SVM for each instance (exemplar) in the dataset (PASCAL VOC)
- Each Exemplar-SVM is trained with a **single** positive instance
- Each Exemplar-SVM is more defined by “what it is not” vs. “what it is similar to”
Large-scale training

- Each exemplar performs its own hard negative mining
- Solve many convex learning problems
- Parallel training on cluster
Exemplar

Detector w

Appearance

Meta-data

Person
What’s this?

A bird

Typical member of a basic-level category are categorized at the expected level.

Atypical members tend to be classified at a subordinate level.

(Jolicoeur, Gluck, Kosslyn 1984)
What’s this?

A bird

An ostrich
Entry-level categories
(Jolicoeur, Gluck, Kosslyn 1984)

- Typical member of a basic-level category are categorized at the expected level.
- Atypical members tend to be classified at a subordinate level.

A bird

An ostrich
Classical Categorization

- Group objects by common properties
- What are birds?
  - animals, has wings, has feathers, can fly, chirps
Prototype Theory

Rosch and Lakoff

- According to the prototype view, an object will be classified as an instance of a category if it is sufficiently similar to the prototype.
- **Evidence for Prototype:**
  - **Typicality ratings:** how good are robins as an example of birds
  - **Production order of exemplars:** Name all the kinds of bird you can think of
  - **Time to verify categorical statements:** True or false: a robin is a bird

*Figure 7.3.* Schematic of the prototype model. Although many exemplars are seen, only the prototype is stored. The prototype is updated continually to incorporate more experience with new exemplars.
Canonical Perspective

The “best,” most easily identified view of an object.
(Palmer, Rosch & Chase, 1981)
Dyirbal Indigenous People
The perception of function

• Direct perception (affordances): Gibson
  
  - Flat surface
  - Horizontal
  - Knee-high
  - ... → Sittable upon

• Mediated perception (Categorization)
  
  - Flat surface
  - Horizontal
  - Knee-high
  - ... → Chair → Sittable upon

Diagram:
- Chair
- Chair
- Chair?
Direct perception

Some aspects of an object function can be perceived directly

- Functional form: Some forms clearly indicate to a function ("sittable-upon", container, cutting device, ...)

It does not seem easy to sit-upon this...
Direct perception
Some aspects of an object function can be perceived directly
• Observer relativity: Function is observer dependent
Limitations of Direct Perception

Objects of similar structure might have very different functions.

Not all functions seem to be available from direct visual information only.

The functions are the same at some level of description: we can put things inside in both and somebody will come later to empty them. However, we are not expected to put inside the same kinds of things…
Segmentation:
Where *really* are the people?
Semantic segmentation using convolutional networks
Semantic segmentation using convolutional networks
Semantic segmentation using convolutional networks

Slide credit: Bharath Hariharan
Semantic segmentation using convolutional networks

Slide credit: Bharath Hariharan
Semantic segmentation using convolutional networks

Convolve with \#classes 1x1 filters

Slide credit: Bharath Hariharan
Semantic segmentation using convolutional networks
Solution 1: Image pyramids


Slide credit: Bharath Hariharan
Solution 2: Skip connections
Skip connections

Skip connections

• Problem: early layers not semantic

Solution 3: Dilation

• Need subsampling to allow convolutional layers to capture large regions with small filters
  • Can we do this without subsampling?
Solution 3: Dilation

• Need subsampling to allow convolutional layers to capture large regions with small filters
  • Can we do this without subsampling?
Solution 3: Dilation

- Need subsampling to allow convolutional layers to capture large regions with small filters
  - Can we do this without subsampling?
Solution 4: Conditional Random Fields

• Idea: take convolutional network prediction and sharpen using classic techniques

• Conditional Random Field

\[ y^* = \arg \min_y \sum_{(i,j)} E_{data}(y(i, j)) + \sum_{(i,j),(k,l) \in N} E_{smooth}(y(i, j), y(k, l)) \]

\[ E_{smooth}(y(i, j), y(k, l)) = \mathbb{I}(y(i, j) \neq y(k, l))w(i, j, k, l) \]
Fully Connected CRFs

• Typically, only adjacent pixels connected
  • Fewer connections => Easier to optimize
• Dense connectivity: every pixel connected to everything else
• Intractable to optimize except if pairwise potential takes specific form

\[ E_{\text{smooth}}(y(i, j), y(k, l)) = \mathbb{I}(y(i, j) \neq y(k, l))w(i, j, k, l) \]
\[ w(i, j, k, l) = \sum_m w_m e^{-\| \mathbf{f}_m(i, j) - \mathbf{f}_m(k, l) \|^2} \]

What is the mustache made of?

http://www.visualqa.org/challenge.html
Fig. 1: Examples of free-form, open-ended questions collected for images via Amazon Mechanical Turk. Note that commonsense knowledge is needed along with a visual understanding of the scene to answer many questions.
# Questions and answers collected with Amazon Mechanical Turk

<table>
<thead>
<tr>
<th>Question</th>
<th>Yes</th>
<th>No</th>
<th>Sometimes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Is something under the sink broken?</td>
<td>yes</td>
<td>no</td>
<td>no</td>
</tr>
<tr>
<td>What number do you see?</td>
<td>33</td>
<td>5</td>
<td>6</td>
</tr>
<tr>
<td>Does this man have children?</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>Is this man crying?</td>
<td>no</td>
<td>no</td>
<td>yes</td>
</tr>
<tr>
<td>Can you park here?</td>
<td>no</td>
<td>no</td>
<td>no</td>
</tr>
<tr>
<td>What color is the hydrant?</td>
<td>white and orange</td>
<td>red</td>
<td>yellow</td>
</tr>
<tr>
<td>Has the pizza been baked?</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>What kind of cheese is topped on this pizza?</td>
<td>feta</td>
<td>feta</td>
<td>mozzarella</td>
</tr>
<tr>
<td>What kind of store is this?</td>
<td>bakery</td>
<td>bakery</td>
<td>art supplies</td>
</tr>
<tr>
<td>Is the display case as full as it could be?</td>
<td>no</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>How many pickles are on the plate?</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>What is the shape of the plate?</td>
<td>circle</td>
<td>circle</td>
<td>circle</td>
</tr>
</tbody>
</table>

Fig. 2: Examples of questions (black), (a subset of the) answers given when looking at the image (green), and answers given when not looking at the image (blue) for numerous representative examples of the dataset. See the appendix for more examples.
Words

- Need ways to compare words

Next to the 'sofa' is a desk, and a 'person' is sitting behind it.

'armchair'
'bench'
'chair'
'deck chair'
'ottoman'
'seat'
'stool'
'swivel chair'
'loveseat'
...

'man'
'woman'
'child'
'teenager'
'girl'
'boy'
'baby'
'daughter'
'son'
...

word2vec

I parked the **car** in a nearby street. It is a red **car** with two doors, ...

I parked the **vehicle** in a nearby street...
I parked the car in a nearby street. It is a red car with two doors, ...

word2vec

Hidden layer

Soft-max classifier

Encoder

Decoder

Word = ‘car’

Output prob. That each word is in the context of the input word

Algebraic operations with the vector representation of words

\[ X = \text{Vector(“Paris”)} - \text{vector(“France”)} + \text{vector(“Italy”)} \]

Closest nearest neighbor to X is vector(“Rome”)}
Architecture

Image → [Rectangle]

Question → [Rectangle]

Answer → [Rectangle]
There are 1000 possible answers in this system. Questions are unlimited.
Fig. 27: Random examples of questions (black), (a subset of the) answers given when looking at the image (green), and answers given when not looking at the image (blue) for numerous representative examples of the real image dataset.
what is on the ground?
what is on the ground?

Predicted top-5 answers with confidence:

- sand: 90.748%
- snow: 2.858%
- beach: 1.418%
- surfboards: 0.677%
- water: 0.528%
what color is the umbrella?
what color is the umbrella?

Predicted top-5 answers with confidence:

- **yellow**: 95.090%
- **white**: 81.11%
- **black**: 66.3%
- **blue**: 54.1%
- **gray**: 36.2%
are we alone in the universe?
are we alone in the universe?

Predicted top-5 answers with confidence:

- yes: 21.763%
- people: 0.001%
- birds: 0.000%
- out: 0.000%
- no: 78.234%
what is the meaning of life?
what is the meaning of life?

Predicted top-5 answers with confidence:

- beach: 15.262%
- sand: 8.537%
- seagull: 4.708%
- tower: 2.393%
- rocks: 1.746%
What is the yellow thing?

Predicted top-5 answers with confidence:

- frisbee: 79.844%
- surfboard: 7.319%
- banana: 2.844%
- lemon: 2.438%
- surfboards: 1.252%
how many trains are in the picture?

Predicted top-5 answers with confidence:

3
30.233%

5
18.270%

4
17.000%

2
11.343%

6
7.806%
are these people family?
### Table 2: Performance comparison on test-standard.

<table>
<thead>
<tr>
<th></th>
<th>Open-Ended</th>
<th></th>
<th></th>
<th></th>
<th>Multiple-Choice</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Overall</td>
<td>yes/no</td>
<td>number</td>
<td>others</td>
<td>Overall</td>
<td>yes/no</td>
<td>number</td>
<td>others</td>
<td></td>
</tr>
<tr>
<td>LSTMIMG [2]</td>
<td>54.06</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>NMN+LSTM [1]</td>
<td>55.10</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>ACK [16]</td>
<td>55.98</td>
<td>79.05</td>
<td>36.10</td>
<td>40.61</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>DPPnet [11]</td>
<td><strong>57.36</strong></td>
<td><strong>80.28</strong></td>
<td><strong>36.92</strong></td>
<td><strong>42.24</strong></td>
<td><strong>62.69</strong></td>
<td>80.35</td>
<td>38.79</td>
<td>52.79</td>
<td></td>
</tr>
<tr>
<td>iBOWIMG</td>
<td>55.89</td>
<td>76.76</td>
<td>34.98</td>
<td>42.62</td>
<td>61.97</td>
<td>76.86</td>
<td>37.30</td>
<td>54.60</td>
<td></td>
</tr>
</tbody>
</table>

The **bold** values indicate the best performance in each category.
Two Extremes of Vision

**Extrapolation problem**
- Generalization
- Diagnostic features

**Interpolation problem**
- Correspondence
- Finding the differences

Number of training samples