Visual recognition at the Large-Scale

Fei-Fei Li
(publish under L. Fei-Fei)

Computer Science Dept.
Psychology Dept.
Stanford University
ImageNet is an image database organized according to the WordNet hierarchy (currently only the nouns), in which each node of the hierarchy is depicted by hundreds and thousands of images. Currently we have an average of over five hundred images per node. We hope ImageNet will become a useful resource for researchers, educators, students and all of you who share our passion for pictures. Click Here to learn more about ImageNet, Click Here to join ImageNet mailing list.

What do these images have in common? Find out!

Update Notice: ImageNet 2010 Spring Version will be released in April, 2010
IMAGENET is a knowledge ontology

- Taxonomy

- S: (n) Eskimo dog, husky (breed of heavy-coated Arctic sled dog)
  - direct hypernym / inherited hypernym / sister-term
  - S: (n) working dog (any of several breeds of usually large powerful dogs bred to work as draft animals and guard and guide dogs)
    - S: (n) dog, domestic dog, Canis familiaris (a member of the genus Canis (probably descended from the common wolf) that has been domesticated by man since prehistoric times; occurs in many breeds) "the dog barked all night"
    - S: (n) canine, canid (any of various fissiped mammals with nonretractile claws and typically long muzzles)
      - S: (n) carnivore (a terrestrial or aquatic flesh-eating mammal) "terrestrial carnivores have four or five clawed digits on each limb"
    - S: (n) placental, placental mammal, eutherian, eutherian mammal (mammals having a placenta; all mammals except monotremes and marsupials)
      - S: (n) mammal, mammalian (any warm-blooded vertebrate having the skin more or less covered with hair; young are born alive except for the small subclass of monotremes and nourished with milk)
    - S: (n) vertebrate, cranate (animals having a bony or cartilaginous skeleton with a segmented spinal column and a large brain enclosed in a skull or cranium)
      - S: (n) chordate (any animal of the phylum Chordata having a notochord or spinal column)
      - S: (n) animal, animate being, beast, brute, creature, fauna (a living organism characterized by voluntary movement)
      - S: (n) organism, being (a living thing that has (or can develop) the ability to act or function independently)
      - S: (n) living thing, animate thing (a living or once-living entity)
        - S: (n) whole, unit (an assemblage of parts that is regarded as a single entity) "how big is that part compared to the whole?"; "the team is a unit"
        - S: (n) object, physical object (a tangible and visible entity; an entity that can cast a shadow) "it was full of rackets, balls and other objects"
      - S: (n) physical entity (an entity that has physical existence)
        - S: (n) entry (that which is perceived or known or inferred to have its own distinct existence (living or nonliving))
IMAGENET is a knowledge ontology

- Taxonomy
- Partonomy
IMAGENET is a knowledge ontology

- Taxonomy
- Parthood
- The “social network” of visual concepts
  - Prior knowledge
  - Context
  - Hidden knowledge and structure among visual concepts
outline

• Construction of ImageNet
  – 2-step process
  – Crowdsourcing: Amazon Mechanical Turk (AMT)
  – Properties of ImageNet

• Benchmarking: what does classifying 10k+ image categories tell us?
  – Computation matters
  – Size matters
  – Density matters
  – Hierarchy matters

• A “semanticvisual” hierarchy for personal albums
  – Building it from Flickr images and user tags
  – Using the hierarchy for visual recognition tasks
outline

• Construction of ImageNet
  – 2-step process
  – Crowdsourcing: Amazon Mechanical Turk (AMT)
  – Properties of ImageNet

• Benchmarking: what does classifying 10k+ image categories tell us?
  – Computation matters
  – Size matters
  – Density matters
  – Hierarchy matters

• A “semanticvisual” hierarchy for personal albums
  – Building it from Flickr images and user tags
  – Using the hierarchy for visual recognition tasks
Constructing IMAGENET

Step 1: Collect candidate images via the Internet

Step 2: Clean up the candidate Images by humans
Step 1: Collect Candidate Images from the Internet

- **Query expansion**
  - Synonyms: *German shepherd, German police dog, German shepherd dog, Alsatian*
  - Appending words from ancestors: *sheepdog, dog*
- **Multiple languages**
  - Italian, Dutch, Spanish, Chinese
    - *e.g. ovejero alemán, pastore tedesco,德国牧羊犬*
- **More engines**
- **Parallel downloading**
Step 1: Collect Candidate Images from the Internet

• “Mammal” subtree (1180 synsets)
  – Average # of images per synset: 10.5K

<table>
<thead>
<tr>
<th>Most populated</th>
<th>Least populated</th>
</tr>
</thead>
<tbody>
<tr>
<td>Humankind (118.5k)</td>
<td>Algeripithecus minutus (90)</td>
</tr>
<tr>
<td>Kitty, kitty-cat (69k)</td>
<td>Striped muishond (107)</td>
</tr>
<tr>
<td>Cattle, cows (65k)</td>
<td>Mylodonitid (127)</td>
</tr>
<tr>
<td>Pooch, doggie (62k)</td>
<td>Greater pichiciego (128)</td>
</tr>
<tr>
<td>Cougar, puma (57k)</td>
<td>Damaraland mole rat (188)</td>
</tr>
<tr>
<td>Frog, toad (53k)</td>
<td>Western pipistrel (196)</td>
</tr>
<tr>
<td>Hack, jade, nag (50k)</td>
<td>Muishond (215)</td>
</tr>
</tbody>
</table>
Step 1: Collect Candidate Images from the Internet

- “Mammal” subtree (1180 synsets)
  - Average accuracy per synset: 26%

![Histogram of synset precision](image)

<table>
<thead>
<tr>
<th>Most accurate</th>
<th>Least accurate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bottlenose dolpin (80%)</td>
<td>Fanaloka (1%)</td>
</tr>
<tr>
<td>Meerkat (74%)</td>
<td>Pallid bat (3%)</td>
</tr>
<tr>
<td>Burmese cat (74%)</td>
<td>Vaquita (3%)</td>
</tr>
<tr>
<td>Humpback whale (69%)</td>
<td>Fisher cat (3%)</td>
</tr>
<tr>
<td>African elephant (63%)</td>
<td>Walrus (4%)</td>
</tr>
<tr>
<td>Squirrel (60%)</td>
<td>Grison (4%)</td>
</tr>
<tr>
<td>Domestic cat (59%)</td>
<td>Pika, Mouse hare (4%)</td>
</tr>
</tbody>
</table>
Step 2: verifying the images by humans

- # of synsets: 40,000 (subject to: imageability analysis)
- # of candidate images to label per synset: 10,000
- # of people needed to verify: 2-5
- Speed of human labeling: 2 images/sec (one fixation: ~200msec)

\[
40,000 \times 10,000 \times 3 / 2 = 600,000,000 \text{ sec} \approx 19 \text{ years}
\]

Moral of the story:
no graduate students would want to do this project!
In summer 2008, we discovered crowdsourcing
Mechanical Turk is a marketplace for work.
We give businesses and developers access to an on-demand, scalable workforce.
Workers select from thousands of tasks and work whenever it's convenient.

149,499 HITs available. View them now.

Make Money by working on HITs

HITs - Human Intelligence Tasks - are individual tasks that you work on. Find HITs now.

As a Mechanical Turk Worker you:

- Can work from home
- Choose your own work hours
- Get paid for doing good work

Find an interesting task  Work  Earn money

or learn more about being a Worker

Get Results from Mechanical Turk Workers

Ask workers to complete HITs - Human Intelligence Tasks - and get results using Mechanical Turk. Register Now

As a Mechanical Turk Requester you:

- Have access to a global, on-demand, 24 x 7 workforce
- Get thousands of HITs completed in minutes
- Pay only when you're satisfied with the results

Fund your account  Load your tasks  Get results

Get Started
Step 2: verifying the images by humans

- # of synsets: 40,000 (subject to: imageability analysis)
- # of candidate images to label per synset: 10,000
- # of people needed to verify: 2-5
- Speed of human labeling: 2 images/sec (one fixation: ~200msec)
- Massive parallelism (N ~ 10^2-3)

\[
40,000 \times 10,000 \times 3 / 2 = 600,000,000 \text{sec} \approx 19 \text{years}
\]
Click on the good images.
First time workers please click here for instructions.

Click on the photos that contain the concept of: **delta**: a low triangular area of alluvial deposits where a river divides before entering a larger body of water; "the Mississippi River delta"; "the Nile delta". (PLEASE READ DEFINITION CAREFULLY)

Pick as many as possible. **PHOTOS ONLY, NO PAINTINGS, DRAWINGS, etc.** It’s OK to have other objects, multiple instances, occlusion or text in the image.

Do not use back or forward button of your browser. OCCASIONALLY THERE MIGHT BE ADULT OR DISTURBING CONTENT.
So are we exploiting chained prisoners?
Demography of AMT workers

United States: 46.80%
India: 34.00%
Miscellaneous: 19.20%

Panos Ipeirotis, NYU, Feb, 2010
Demography of AMT workers

Panos Ipeirotis, NYU, Feb, 2010
Demography of AMT workers

Panos Ipeirotis, NYU, Feb, 2010
U.S. economy 2008 - 2009

Personal Dimension, Gallup Index of Investor Optimism, November 2008-September 2009

IMAGENET hired more than 25,000 AMT workers in this period of time!!
outline

• Construction of ImageNet
  – 2-step process
  – Crowdsourcing: Amazon Mechanical Turk (AMT)
  – Properties of ImageNet

• Benchmarking: what does classifying 10k+ image categories tell us?
  – Computation matters
  – Size matters
  – Density matters
  – Hierarchy matters

• A “semanticvisual” hierarchy for personal albums
  – Building it from Flickr images and user tags
  – Using the hierarchy for visual recognition tasks
Datasets and computer vision

UIUC Cars (2004)
S. Agarwal, A. Awan, D. Roth

CMU/VASC Faces (1998)
H. Rowley, S. Baluja, T. Kanade

FERET Faces (1998)
P. Phillips, H. Wechsler, J. Huang, P. Raus

COIL Objects (1996)
S. Nene, S. Nayar, H. Murase

MNIST digits (1998-10)
Y LeCun & C. Cortes

KTH human action (2004)
I. Leptev & B. Caputo

Sign Language (2008)
P. Buehler, M. Everingham, A. Zisserman

Segmentation (2001)

3D Textures (2005)
S. Lazebnik, C. Schmid, J. Ponce

CuRRET Textures (1999)
K. Dana B. Van Ginneken S. Nayar J. Koenderink

CAVIAR Tracking (2005)
R. Fisher, J. Santos-Victor J. Crowley

Middlebury Stereo (2002)
D. Scharstein R. Szeliski
Object Recognition

Fergus, Perona, Zisserman, CVPR 2003
Object Recognition

Fergus, Perona, Zisserman, CVPR 2003
Holub, et al. ICCV 2005; Sivic et al. ICCV 2005
Object Recognition

PASCAL
[Everingham et al., 2009]

MSRC
[Shotton et al. 2006]

Fergus, Perona, Zisserman, CVPR 2003

Holub, et al. ICCV 2005; Sivic et al. ICCV 2005

Fei-Fei et al. CVPR 2004; Grauman et al. ICCV 2005; Lazebnik et al. CVPR 2006
Zhang & Malik, 2006; Varma & Sizzerman 2008; Wang et al. 2006; [....]
Fei-Fei et al. CVPR 2004; Grauman et al. ICCV 2005; Lazebnik et al. CVPR 2006; Zhang & Malik, 2006; Varma & Sizzerman 2008; Wang et al. 2006; [....]

Biederman 1987

Object Recognition

ESP
[Ahn et al, 2006]

LabelMe
[Russell et al, 2005]

TinyImage
Torralba et al. 2007

Lotus Hill
[Yao et al, 2007]
Comparison among free datasets

1. Excluding the Caltech101 datasets from PASCAL
2. No image in this dataset is human annotated. The # of clean images per category is a rough estimation.
Basic evaluation setup

• **IMAGENET**
  - 10,000 categories
  - 9 million images
  - 50%-50% train test split

• Multi-class classification in 1-vs-all framework
  - **GIST+NN**: filter banks; nearest neighbor (Oliva & Torralba, 2001)
  - **BOW+NN**: SIFT, 1000 codewords, BOW; nearest neighbor
  - **BOW+SVM**: SIFT, 1000 codewords, BOW; linear SVM
  - **SPM+SVM**: SIFT, 1000 codewords, Spatial Pyramid; intersection kernel SVM (Lazebnik et al. 2006)

Deng, Berg, Li, & Fei-Fei, *submitted*
Computation issues first

- **BOW+SVM**
  - Train one 1-vs-all with LIBLINEAR $\Rightarrow$ 1 CPU hour
  - 10,000 categories $\Rightarrow$ 1 CPU year

- **SPM + SVM**
  - Maji & Berg 2009, LIBLINEAR with piece-wise linear encoding
  - Memory bottleneck. Modification required.
  - 10,000 categories $\Rightarrow$ 6 CPU year

- **Parallelized on a cluster**
  - Weeks for a single run of experiments

Deng, Berg, Li, & Fei-Fei, *submitted*
Size matters

- 6.5% for 10K categories
- Better than we expected (instead of dropping at the rate of 10x; it’s roughly at about 2x)
- An ordering switch between SVM and NN methods when the # of categories becomes large

Some unpublished results omitted.

Deng, Berg, Li, & Fei-Fei, *submitted*
Size matters

• 6.5% for 10K categories
• Better than we expected (instead of dropping at the rate of 10x; it’s roughly at about 2x)
• An ordering switch between SVM and NN methods when the # of categories becomes large
• When dataset size varies, conclusion we can draw about different categories varies

Some unpublished results omitted.

Deng, Berg, Li, & Fei-Fei, *submitted*
Size matters

- 6.5% for 10K categories
- Better than we expected (instead of dropping at the rate of 10x; it’s roughly at about 2x)
- An ordering switch between SVM and NN methods when the # of categories becomes large
- When dataset size varies, conclusion we can draw about different categories varies
- Purely semantic organization of concepts (by WordNet) exhibits meaningful visual structure (ordered by DFS)

Some unpublished results omitted.

Deng, Berg, Li, & Fei-Fei, submitted
Density matters

• Datasets have very different “density” or “sparcity”
Density matters

• Datasets have very different “density” or “sparcity”
• there is a significant difference in difficulty between different datasets, independent of feature and classier choice.

Some unpublished results omitted.

Deng, Berg, Li, & Fei-Fei, submitted
Hierarchy matters

• Classifying a “dog” as “cat” is probably not as bad as classifying it as “microwave”
• A simple way to incorporate classification cost

Deng, Berg, Li, & Fei-Fei, submitted
Hierarchy matters

• Classifying a “dog” as “cat” is probably not as bad as classifying it as “microwave”
• A simple way to incorporate hierarchical classification cost

Deng, Berg, Li, & Fei-Fei, submitted
IMAGENET is team work!

WordNet friends
- Christiane Fellbaum  
  Princeton U.
- Dan Osherson  
  Princeton U.

co-PI
- Kai Li  
  Princeton U.

Research collaborator; ImageNet Challenge boss
- Alex Berg  
  Columbia U.

Graduate students
- Jia Deng  
  Princeton/Stanford
- Hao Su  
  Stanford U.

Other contributors
- Princeton graduate students
  - Wei Dong
  - Zhe Wang
- Stanford graduate students
  - John Le
  - Pao Siangliulue
- AMT partner
  - Dolores Lab
outline

• Construction of ImageNet
  – 2-step process
  – Crowdsourcing: Amazon Mechanical Turk (AMT)
  – Properties of ImageNet

• Benchmarking: what does classifying 10k+ image categories tell us?
  – Computation matters
  – Size matters
  – Density matters
  – Hierarchy matters

• A “semanticvisual” hierarchy for personal albums
  – Building it from Flickr images and user tags
  – Using the hierarchy for visual recognition tasks
Semantic hierarchy
(purely) visual hierarchy

Sivic, Russell, Zisserman, Freeman, Efros, CVPR 2008

Bart, Porteous, Perona, Welling, CVPR 2008
A “semantivisual” hierarchy of images

Li, Wang, Lim, Blei & Fei-Fei, CVPR, 2010
“Towards total scene understanding”

A “semantivisual” hierarchy of images

Li, Wang, Lim, Blei & Fei-Fei, CVPR, 2010
A “semantivisual” hierarchy of images

\[ N^* = \arg\max (N_{\{R\}}) \]

\( N_{\{R\}} \): No. Regions assigned to the node

\[ p(N_{ik}|R_i, W_k) \propto p(R_i|R_{\text{rest}}, T) \]
\[ p(W_k|W_{\text{rest}}, T, R_i) \]
\[ p(N_{ik}|T, N_{ik_{\text{rest}}}) \]

\[ R: \text{ Region Appearance} \]
\[ W: \text{ Words} \]
\[ N: \text{ Node in the tree} \]
\[ T: \text{ Tree} \]

Li, Wang, Lim, Blei & Fei-Fei, CVPR, 2010
A “semantivisual” hierarchy of images
animal, bride, building, cake, child, christmas, church, city, clouds, dessert, dinner, flower, spring, friends, fruit, green, high-school, calcio, italy, europe, london, love, nature, landscape, macro, paris, party, present, sea, sun, sky, seagull, soccer, reflection, sushi, vacation, trip, water, silhouette, and wife.

Li, Wang, Lim, Blei & Fei-Fei, CVPR, 2010
Li, Wang, Lim, Blei & Fei-Fei, CVPR, 2010

4000 images
4000 images

Li, Wang, Lim, Blei & Fei-Fei, CVPR, 2010
Evaluating and using the hierarchy

Evaluate the quality of image concept clustering by path

Evaluate the quality of hierarchy given a path of the tree

Li, Wang, Lim, Blei & Fei-Fei, CVPR, 2010
Evaluating and using the hierarchy

- Hierarchical annotation

<table>
<thead>
<tr>
<th>Our Result</th>
<th>Flickr</th>
</tr>
</thead>
<tbody>
<tr>
<td>Photo</td>
<td>Photo</td>
</tr>
<tr>
<td>Zoo</td>
<td>Bird</td>
</tr>
<tr>
<td>Flamingo</td>
<td>Animal</td>
</tr>
<tr>
<td>Head</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Our Result</th>
<th>Flickr</th>
</tr>
</thead>
<tbody>
<tr>
<td>Photo</td>
<td>Photo</td>
</tr>
<tr>
<td>Event</td>
<td>Wedding</td>
</tr>
<tr>
<td>Wedding</td>
<td>Bride</td>
</tr>
<tr>
<td>Dress</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>method</th>
<th>accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Our Model</td>
<td>46%</td>
</tr>
<tr>
<td>nCRP</td>
<td>16%</td>
</tr>
</tbody>
</table>

Li, Wang, Lim, Blei & Fei-Fei, CVPR, 2010
Evaluating and using the hierarchy

- Hierarchical annotation
- Image labeling (annotation)

<table>
<thead>
<tr>
<th>Method</th>
<th>Alipr</th>
<th>Corr-LDA</th>
<th>Ours</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>building photo landscape sky people</td>
<td>cake dress garden architecture flower</td>
<td>photo wedding gown bride flower</td>
<td>38%</td>
</tr>
<tr>
<td></td>
<td>card people female fashion cloth</td>
<td>photo birthday bird architecture portrait</td>
<td>photo birthday kid cake human</td>
<td>44%</td>
</tr>
<tr>
<td></td>
<td>people ocean water landscape snow</td>
<td>light cloud photo city human</td>
<td>photo cloud sky architecture building</td>
<td>74%</td>
</tr>
</tbody>
</table>

Li, Wang, Lim, Blei & Fei-Fei, *CVPR*, 2010
Evaluating and using the hierarchy

• Hierarchical annotation
• Image labeling (annotation)
• Image classification

Li, Wang, Lim, Blei & Fei-Fei, CVPR, 2010
Thank you!

Jia Deng
4th year PhD
Princeton;
“ImageNet”

Hao Su
1st year PhD
Stanford;
“ImageNet”

Li-Jia Li
4th year PhD
Stanford;
“Total scene understanding”; “Semantivisual hierarchy”

Chris Baldassano
Juan Carlos Niebles
Bangpeng Yao