From Large Scale Image Categorization to Entry-Level Categories

Vicente Ordonez, Jia Deng, Yejin Choi, Alexander C. Berg, Tamara L. Berg

Stony Brook University

The University of North Carolina at Chapel Hill

Stanford University
What would you call this?

Grampus griseus

Dolphin
What would you call this?

Object
Organism
Animal
Chordate
Vertebrate
Bird
Aquatic bird
Swan
Whistling swan
Cygnus Colombianus
Naming Image Content

Input Image

Vision

Grampus griseus
American black bear
Grizzly bear
King penguin
Cormorant
Homing pigeon
Ball-peen hammer
Spigot
Diskette, floppy
Steel arch bridge
Farmhouse
Soapweed
Brazilian rosewood
Bristlecone pine
Clifdiving
Crabapple

What Should I Call It?

Thousands of Noisy Category Predictions
Entry-Level Category

The category that people are likely to name when presented with a depiction of an object.

*Rosch et al, 1976*
*Jolicoeur, Gluck & Kosslyn, 1984*

**Superordinates:** animal, vertebrate

**Entry Level:** bird

**Subordinates:** Black-capped chickadee
Entry-Level Category

Superordinates: animal, bird
Entry Level: penguin
Subordinates: Chinstrap penguin

The category that people are likely to name when presented with a depiction of an object.

Rosch et al, 1976
Jolicoeur, Gluck & Kosslyn, 1984
Is this hard?

*wordnet hierarchy*
How will we do it?

Linguistic resources

Wordnet

Computer Vision

Google Web 1T

 lots of text

Imagenet

 Labeled Images

 Lots of images with text

SBU Captioned Dataset

Man sits in a rusted car buried in the sand on Waiarere beach

Little girl and her dog in northern Thailand. They both seemed.

The Egyptian cat statue by the floor clock and perpetual motion

Emma in her hat looking super cute

Our dog Zoe in her bed
Scaling Naming Tasks!

48 categories

> 7000 categories
1. Goal: Category Translation

**Detailed Category**

Grampus griseus

What should I Call It?
(Entry-Level Category)

dolphin

e

2. Goal: Content Naming

*Input Image*

What should I Call It?
(Entry-Level Category)

dolphin

e
1. Goal: Category Translation

**Detailed Category**

Grampus griseus

**What should I Call It?**
(Entry-Level Category)

dolphin

2. Goal: Content Naming

**Input Image**

What should I Call It?
(Entry-Level Category)

dolphin
Category Translation by Humans

Friesian, Holstein, Holstein-Friesian

COW
cattle
pasture
fence
1.1 Category Translation: Text-based

\[ \tau(d, \lambda) = \arg\max_w [\phi(e) - \lambda \psi(d, e)] \]
1.2 Category Translation: Image-based

Friesian, Holstein, Holstein-Friesian

(1.9071) cow
(1.1851) orange_tree
(0.6136) stall
(0.5630) mushroom
(0.3825) pasture
(0.3156) sheep
(0.3321) black_bear
(0.3015) puppy
(0.2409) pedestrian_bridge
(0.2353) nest

Vision System
## Category Translation: Examples

<table>
<thead>
<tr>
<th>Common Name</th>
<th>HUMANS</th>
<th>TEXT BASED</th>
<th>IMAGE BASED</th>
</tr>
</thead>
<tbody>
<tr>
<td>cactus wren</td>
<td>bird</td>
<td>bird</td>
<td>bird</td>
</tr>
<tr>
<td>buzzard, Buteo buteo</td>
<td>hawk</td>
<td>hawk</td>
<td>bird</td>
</tr>
<tr>
<td>whinchat, Saxicola rubetra</td>
<td>bird</td>
<td>chat</td>
<td>bird</td>
</tr>
<tr>
<td>Weimaraner</td>
<td>dog</td>
<td>dog</td>
<td>dog</td>
</tr>
<tr>
<td>numbat, banded anteater, anteater</td>
<td>anteater</td>
<td>anteater</td>
<td>cat</td>
</tr>
<tr>
<td>rhea, Rhea americana</td>
<td>ostrich</td>
<td>bird</td>
<td>grass</td>
</tr>
<tr>
<td>Europ. black grouse, heathfowl</td>
<td>bird</td>
<td>bird</td>
<td>duck</td>
</tr>
<tr>
<td>yellowbelly marmot, rockchuck</td>
<td>Squirrel</td>
<td>marmot</td>
<td>rock</td>
</tr>
</tbody>
</table>
1. Goal: Category Translation

Detailed Category

Grampus griseus

What should I Call It?
(Entry-Level Category)

dolphin

d
e

2. Goal: Content Naming

Input Image

What should I Call It?
(Entry-Level Category)

dolphin

e
Large Scale Categorization

Selective Search Windows. van De Sande et al. ICCV 2011
Local descriptors Coding (LLC), Wang et al. CVPR 2010
Spatial pooling

Flat Classifiers

Grampus griseus (0.80)
American black bear (0.41)
Grizzly bear (0.16)
King penguin (0.25)
Cormorant (0.11)
Homing pigeon (0.56)
Ball-peen hammer (0.26)
Spigot (0.06)
Diskette, floppy (0.07)
Steel arch bridge (0.06)
Farmhouse (0.16)
Soapweed (0.03)
Brazilian rosewood (0.12)
Bristlecone pine (0.13)
Cliffdiving (0.04)
Crabapple (0.19)
2.1 Propagated Visual Estimates

\[ f(v, I) - \bar{\psi}(v) \phi(v) \]

Our work

\[ f_{nat}(v, I, \tilde{\lambda}) = f(v, I) \left[ \phi(v) - \tilde{\lambda}\bar{\psi}(v) \right] \]
2.2 Supervised Learning

\[ f_{svm}(\mathbf{v}_i, I, \Theta) = \frac{1}{1 - \exp(a\Theta^T X + b)} \]
Extracting Meaning from Data

Weights learned to recognize images with “tree” in caption

Mammals  Birds  Instruments  Structures  Plants  Other

-0.5 -0.4 -0.3 -0.2 -0.1 0.0 0.1 0.2 0.3 0.4 0.5

0 000 2000 4000 6000 8000 10000

-0.5 -0.4 -0.3 -0.2 -0.1 0.0 0.1 0.2 0.3 0.4 0.5

0 000 2000 4000 6000 8000 10000

snag
shade tree
bracket fungus, shelf fungus
bristlecone pine, Rocky Mountain bristlecone pine, Pinus aristata
Brazilian rosewood, caviuna wood, jacaranda, Dalbergia nigra
redheaded woodpecker, redhead, Melanerpes erythrophythalus
redbud, Cercis canadensis
mangrove, Rhizophora mangle
chiton, coat-of-mail shell, sea cradle, polyplacophore
crab apple, crabapple
papaya, papaia, pawpaw, papaya tree, melon tree, Carica papaya
frogmouth
Extracting Meaning from Data

Weights learned to recognize images with “water” in caption

- water dog
- surfing, surfboarding, surfriding
- manatee, Trichechus manatus
- punt
- dip, plunge
- cliff diving
- fly-fishing
- sockeye, sockeye salmon, red salmon, blueback salmon, Oncorhynchus nerka
- sea otter, Enhydra lutris
- American coot, marsh hen, mud hen, water hen, Fulica americana
- booby
- canal boat, narrow boat, narrowboat
Results: Content Naming

<table>
<thead>
<tr>
<th>Human Labels</th>
<th>Flat Classifier</th>
<th>Deng et al. CVPR’12</th>
<th>Propagated Visual Estimates</th>
<th>Supervised Learning</th>
<th>Joint</th>
</tr>
</thead>
<tbody>
<tr>
<td>farm, fence</td>
<td>gelding</td>
<td>horse</td>
<td>horse</td>
<td>horse</td>
<td>horse</td>
</tr>
<tr>
<td>field</td>
<td>yearling</td>
<td>equine</td>
<td>pasture</td>
<td>pasture</td>
<td>pasture</td>
</tr>
<tr>
<td>horse, mule</td>
<td>shire</td>
<td>perissodactyl</td>
<td>tree</td>
<td>field</td>
<td>field</td>
</tr>
<tr>
<td>kite, dirt</td>
<td>yearling</td>
<td>ungulate</td>
<td>equine</td>
<td>cow</td>
<td>cow</td>
</tr>
<tr>
<td>people</td>
<td>draft</td>
<td>male</td>
<td>male</td>
<td>fence</td>
<td>fence</td>
</tr>
<tr>
<td>tree, zoo</td>
<td></td>
<td>gelding</td>
<td>gelding</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
# Results: Content Naming

<table>
<thead>
<tr>
<th>Human Labels</th>
<th>Flat Classifier</th>
<th>Deng et al. CVPR'12</th>
<th>Propagated Visual Estimates</th>
<th>Supervised Learning</th>
<th>Joint</th>
</tr>
</thead>
<tbody>
<tr>
<td>fence, junk</td>
<td>feeder</td>
<td>woody</td>
<td>tree</td>
<td>logo</td>
<td></td>
</tr>
<tr>
<td>sign</td>
<td>Hyla</td>
<td>tree</td>
<td>structure</td>
<td>street</td>
<td></td>
</tr>
<tr>
<td>stop sign</td>
<td>cleaner</td>
<td>structure</td>
<td>building</td>
<td>neighborhood</td>
<td></td>
</tr>
<tr>
<td>street sign</td>
<td>box</td>
<td>plant</td>
<td>plant</td>
<td>building</td>
<td></td>
</tr>
<tr>
<td>trash can</td>
<td>large</td>
<td>vascular</td>
<td>area</td>
<td>office building</td>
<td></td>
</tr>
<tr>
<td>tree</td>
<td></td>
<td></td>
<td></td>
<td>office</td>
<td></td>
</tr>
</tbody>
</table>

Deng et al. CVPR'12: 
Propagated Visual Estimates: 
Supervised Learning: 
Joint: 

- logo
- street
- neighborhood
- building
- office building
- office
Evaluation: Content Naming

Test Set A – Random Images

Test Set B – High Confidence Prediction Scores
Conclusions/Future Work

• We explored different models for content naming in images.

• Results can be used to improve the larger goal of generating human-like image descriptions.

• Go beyond nouns and infer other type of abstractions on action and attribute words.
Questions?