

ImageNet: A Large-Scale Hierarchical Image Database

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li and Li Fei-Fei
Dept. of Computer Science, Princeton University, USA
CVPR 2009

Presented by:
Nazanin Mehrasa

Large-scale ontology of images is a critical resource for developing content-based image search and image understanding algorithms

Why we need:

More sophisticated and robust models can be proposed by exploiting these images

Resulting in better application for user to index, retrieve, organize, and interact with these data

We believe that a large-scale ontology of images is a critical resource for developing advanced, large-scale content-based image search and image understanding algorithms, as well as for providing critical training and benchmarking data for such algorithms.

Problem:

How such data can be utilized and organized₂

- A large lexical database of english.
- Nouns, verbs, adj and adverbs are grouped into sets of cognitive synonyms
- Each meaningful concept in WordNet, possibly described by multiple words, is called synonym (synset)

WordNet Search - 3.1

- [WordNet home page](#) - [Glossary](#) - [Help](#)

Word to search for:

Display Options:

Key: "S:" = Show Synset (semantic) relations, "W:" = Show Word (lexical) relations

Display options for sense: (gloss) "an example sentence"

Noun

- [S: \(n\)](#) **German shepherd**, [German shepherd dog](#), [German police dog](#), [alsatian](#)
(breed of large shepherd dogs used in police work and as a guide for the blind)

- A large-scale image dataset
- An image database organized according to the **WordNet** hierarchy
- Each node of hierarchy is depicted by hundreds and thousand of images

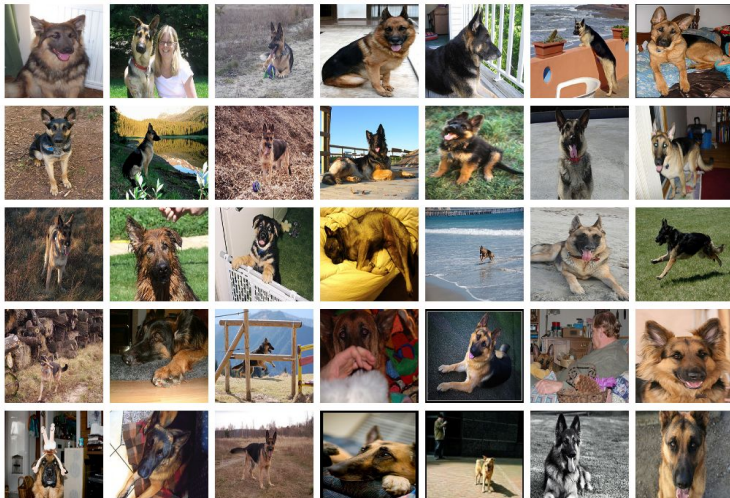
German shepherd, German shepherd dog, German police dog, alsatian 1741 pictures 61.18% Popularity Percentile Wordnet IDs

Breed of large shepherd dogs used in police work and as a guide for the blind

Numbers in brackets: (the number of synsets in the subtree)

- ImageNet 2011 Fall Release (32326)
 - plant, flora, plant life (4486)
 - geological formation, formation (17)
 - natural object (1112)
 - sport, athletics (176)
 - artifact, artefact (10504)
 - fungus (308)
 - person, individual, someone, some1
 - animal, animate being, beast, brute
 - invertebrate (766)
 - homeotherm, homiotherm, hon
 - work animal (4)
 - darter (0)
 - survivor (0)
 - range animal (0)
 - creepy-crawly (0)
 - domestic animal, domesticated
 - domestic cat, house cat, Fell
 - dog, domestic dog, Canis far
 - pooch, doggie, doggy, ba
 - hunting dog (101)
 - dalmatian, coach dog, ca
 - cur, mongrel, mutt (2)
 - corgi, Welsh corgi (2)
 - Mexican hairless (0)
 - lapdog (0)

Treemap Visualization Images of the Synset Downloads



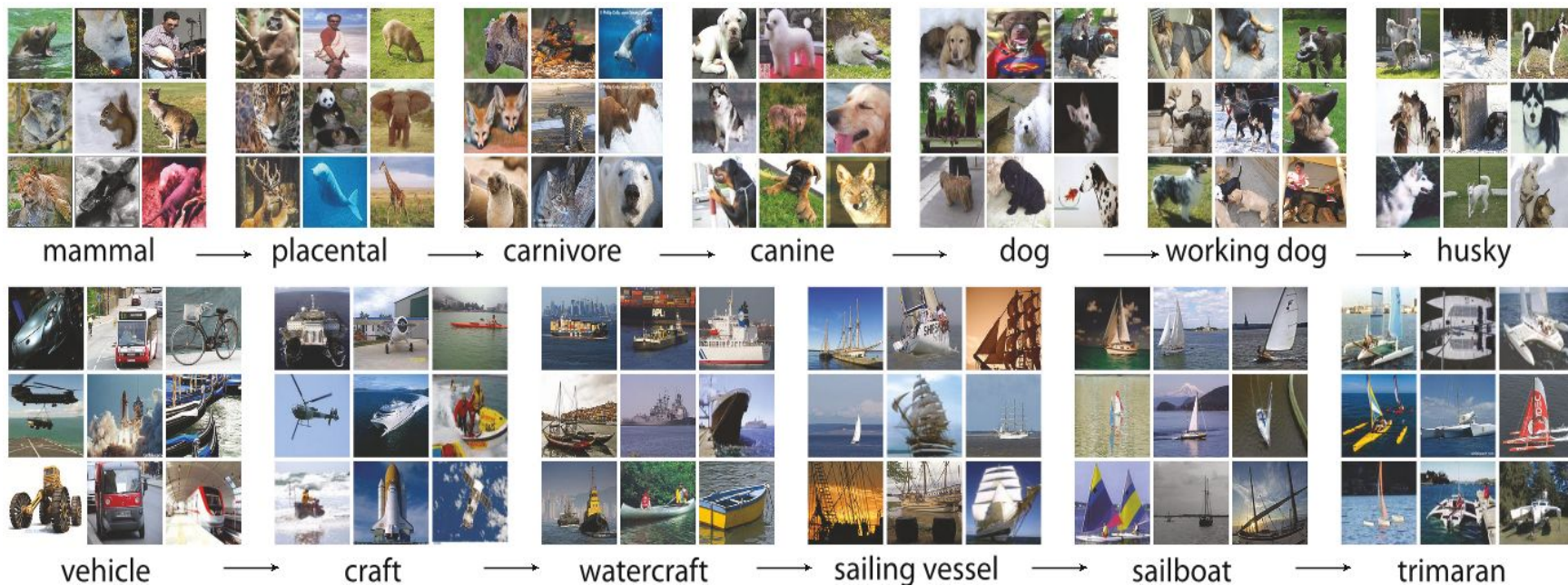
Current version of ImageNet

- Consisting of 12 “subtree”:
mammal, bird, fish, reptile, amphibian, vehicle, furniture, musical instrument, geological formation, tool, flower, fruit.
- These subtrees contain 5247 synsets and 3.2 million images

Goal :

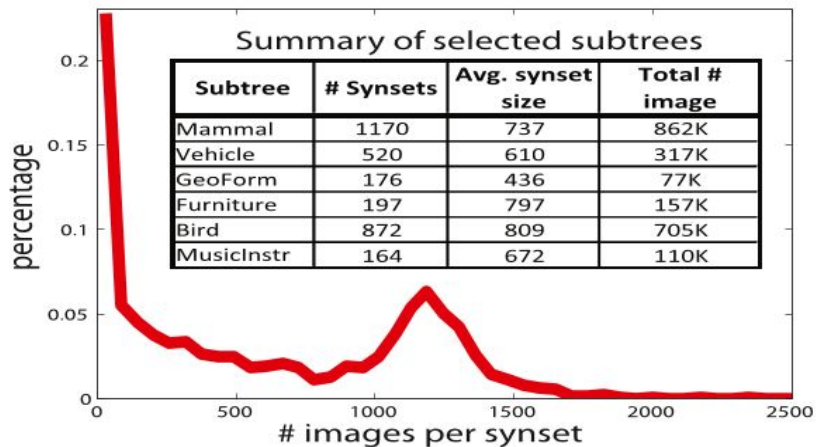
To provide an average of 500-1000 images to illustrate each synset

A snapshot of two root-to-leaf branches of ImageNet



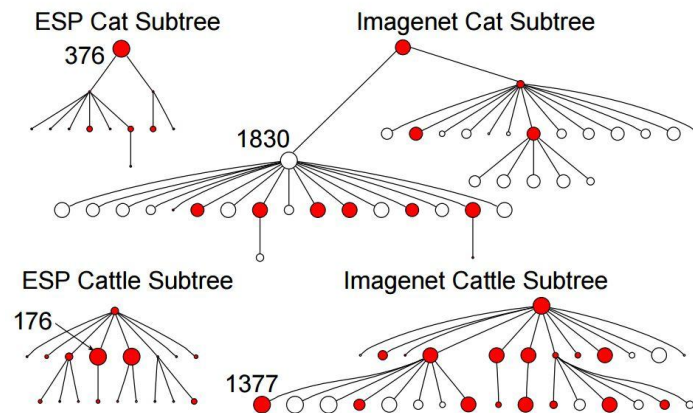
Scale

- ImageNet aims to provide the most comprehensive and diverse coverage of the image world.
- The current 12 subtrees consist of a total of 3.2 million cleanly annotated images spread over 5247 categories
- This is already the largest clean image dataset available to the vision research community, in terms of the total number of images, number of images per category as well as the number of categories



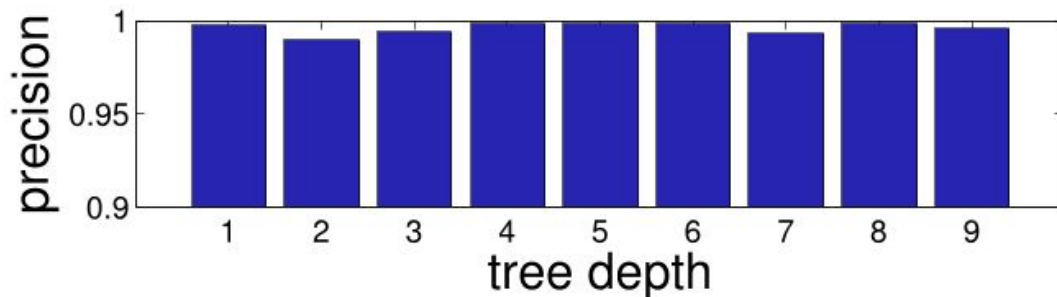
Hierarchy

- Organizing the different classes of images in a densely populated semantic hierarchy
- Synsets of images are interlined by several type of relation like **IS-A**
- Comparing the “cat” and “cattle” subtrees of ImageNet and the ESP dataset, We observe that ImageNet offers much denser and larger trees.



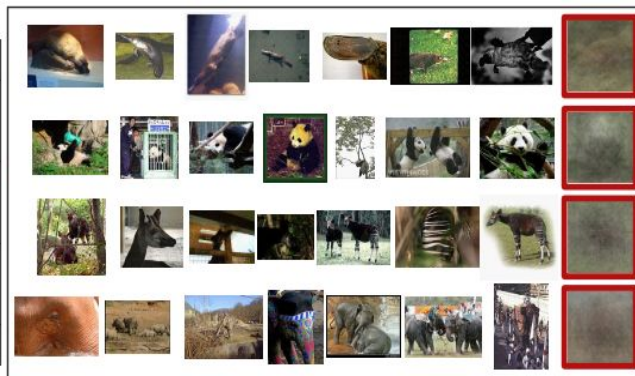
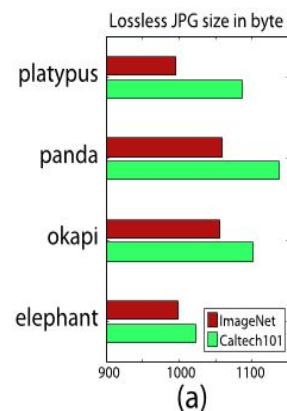
Accuracy:

- We would like to offer a clean dataset at all levels of the WordNet hierarchy
- Achieving a high precision for all depths of the ImageNet tree is challenging because the lower in the hierarchy a synset is, the harder it is to classify, e.g. Siamese cat versus Burmese cat.
- This figure demonstrates the labeling precision on a total of 80 synsets randomly sampled at different tree depths.
- An average of 99.7% precision is achieved for each synset.

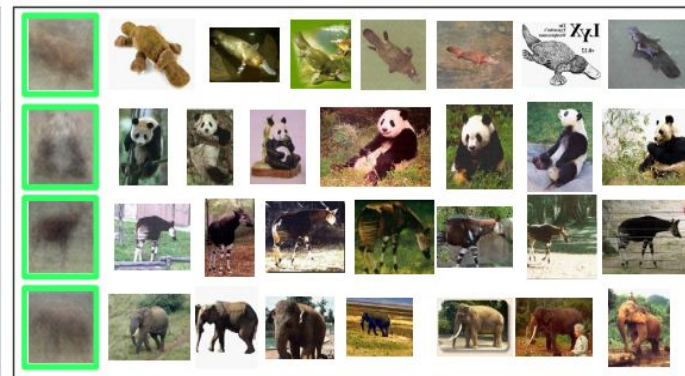


Diversity

- Goal of ImageNet → ‘images should have variable appearances, positions, viewpoints, poses as well as background clutter and occlusions’
- Algorithm of quantifying image diversity:
compute the average image of each synset
measure lossless JPG file size



(b)



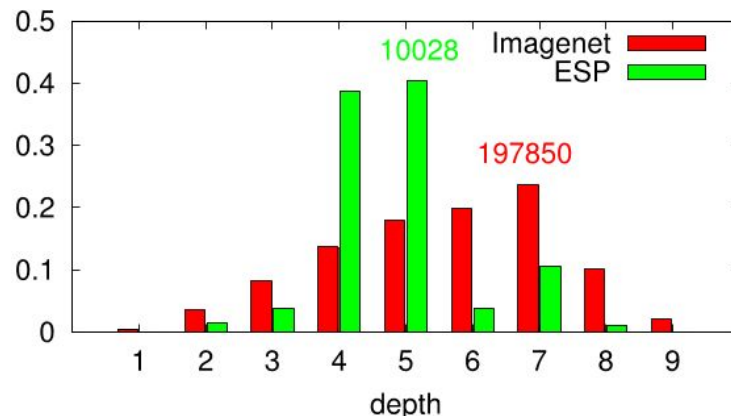
(c)

A synset containing high diverse images will result in a

- Small image datasets: Caltech101/256, MSRC, PASCAL
- TinyImage: 80 million 32*32 low resolution images
- ESP dataset: acquired through an online game
- LabelMe and Lotus Hill dataset: 30K and 50K labeled and segmented images

	ImageNet	TinyImage	LabelMe	ESP	LHill
LabelDisam	Y	Y	N	N	Y
Clean	Y	N	Y	Y	Y
DenseHie	Y	Y	N	N	N
FullRes	Y	N	Y	Y	Y
PublicAvail	Y	Y	Y	N	N
Segmented	N	N	Y	N	Y

Comparison of some of the properties of ImageNet versus other existing datasets



Comparison of the distribution of 'mammal' labels over tree depth levels between ImageNet and ESP games.

Collecting candidate images

- Collect candidate image from the Internet by querying several image search engines
- Queries are the set of WordNet synonyms for each synset
- Expand queries by appending the queries with the word from parent synsets
- translate the queries into other languages (Chinese, Spanish, Dutch, Italian))

Querying “whippet”
according to WordNet ‘s
gloss a “small slender dog
of greyhound type
developed in england”



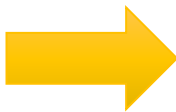
We also search **whippet**
dog and also **whippet**
greyhound

Cleaning candidate images

- Rely on humans to verify each candidate image collected in the previous ()
- **Amazon Mechanical Turk(AMT)**
- We present the users with a set of candidate images and the definition of the target synset (include a link to Wikipedia)
- Ask whether each image contains objects of synset

Two issues:




1. Human makes mistakes
2. Users do not always agree with each



**having multiple users
independently label the
same image**

We developed a simple algorithm to dynamically determine the number of agreements needed for different categories of images.

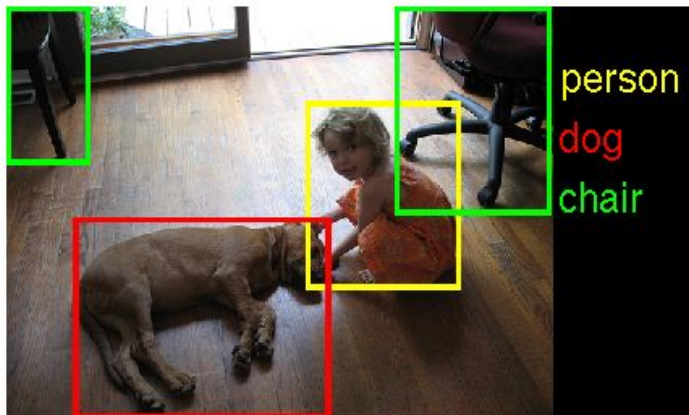
- For each synset, we first randomly sample an initial subset of images.
- At least 10 users are asked to vote on each of these images.
- We then obtain a confidence score table, indicating the probability of an image being a good image given the user votes

				
User 1	Y	Y	Y	#Y
User 2	N	Y	Y	#N
User 3	N	Y	Y	Conf Cat
User 4	Y	N	Y	Conf BCat
User 5	Y	Y	Y	0
User 6	N	N	Y	1
				0.07
				0.23
				0.85
				0.69
				0.46
				0.49
				0.97
				0.83
				0.02
				0.12
				0.99
				0.90
				0.85
				0.68

Confidence score table for 'Cat' and 'Burmese Cat'

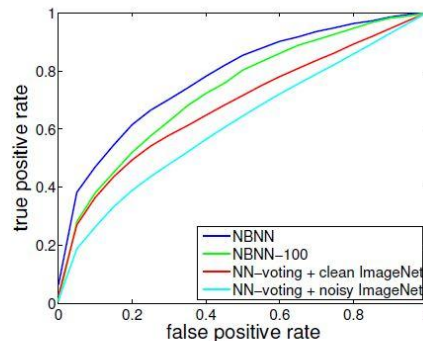
Non parametric object recognition

- Recognizing object class by quering similar images in ImageNet
- Given a large number of images, simple nearest neighbor methods can achieve reasobale reasonable performance despite a high level of noise.

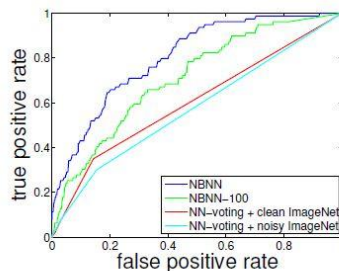


Non parametric object recognition

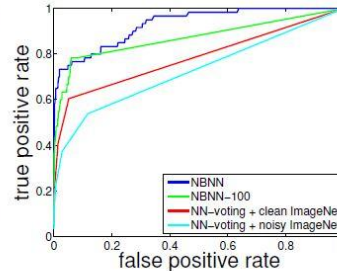
1. NN-voting + noisy ImageNet
2. NN-voting + clear ImageNet:
3. NBNN
4. NBNN-100:



(a) average ROC



(b) elk

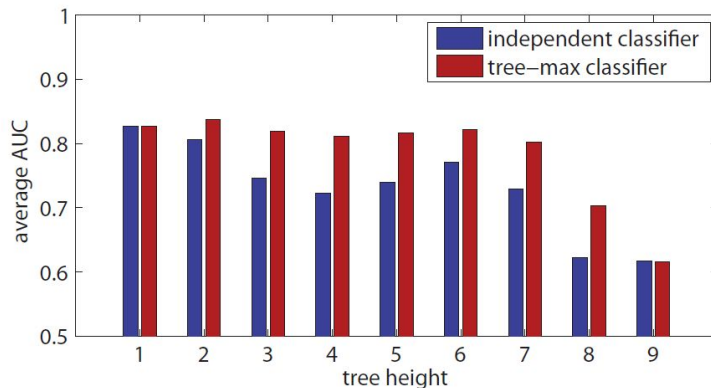


(c) killer-whale

Object recognition experiment results plotted in ROC curves

Tree based image classification

- This experiment illustrates the usefulness of the ImageNet hierarchy
- A classifier at each synset node of the tree
- We want to decide whether an image contains an obj of synset or not
- The maximum of all the classifier responses in this subtree becomes the classification score of the query image.
- Using AdaBoost-based classifier proposed by Collins



Automatic Object Localization

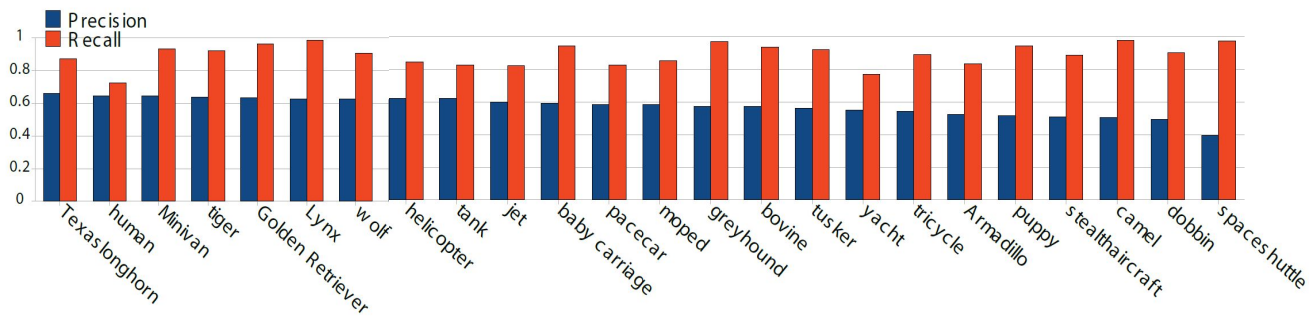
By adding spatial extent of the object in each image two application comes:

1. Training a robust object detection algorithm often need localized objects in different poses
2. For object localization

We used non-parametric graphical model to learn visual representation of object against background

Every input image is represented as a “bag of words”

The output is the probability for each image patch to belong to the topics z_i of a given category





Average images and image samples of the detected bounding boxes from the 'tusker' and 'stealth aircraft' categories



Samples of detected bounding boxes around different objects

- Completing imagenet

Current version= 10% of the wordNet synset

- Speed up construction process
- to have roughly 50 milion clean diverse full resolution images spread over approximatly 50K synsets
- Deliver ImageNet to research communicate by making it publicly available
- Extend to include more information such as localization
- Foster an ImageNet community and develop an online platform where every one can contribute to and benebif from imageNet

We hope ImageNet will become a central resource for a broad of range of vision related research.

We envision these application:

- A training resource
- A benchmark dataset
- Introducing new semantic relation for visual modeling
- Human vision research

Thank you

