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
{jiadeng, wdong, rsocher, jial, li, feifeili}@cs.princeton.edu

Abstract

The explosion of image data on the Internet has the potential to foster more sophisticated and robust models and algorithms to index, retrieve, organize and interact with images and multimedia data. But exactly how such data can be harnessed and organized remains a critical problem. We introduce here a new database called "ImageNet", a large-scale ontology of images built upon the backbone of the WordNet structure. ImageNet aims to populate the majority of the 80,000 synsets of WordNet with an average of 500-1000 clean and full resolution images. This will result in tens of millions of annotated images organized by the semantic hierarchy of WordNet. This paper offers a detailed analysis of ImageNet in its current state: 12 subtrees with 5247 synsets and 3.2 million images in total. We show that ImageNet is much larger in scale and diversity and much more accurate than the current image datasets. Constructing such a large-scale database is a challenging task. We describe the data collection scheme with Amazon Mechanical Turk. Lastly, we illustrate the usefulness of ImageNet through three simple applications in object recognition, image classification and automatic object clustering. We hope that the scale, accuracy, diversity and hierarchical structure of ImageNet can offer unparalleled opportunities to researchers in the computer vision community and beyond.

1. Introduction

The digital era has brought with it an enormous explosion of data. The latest estimations put a number of more than 3 billion photos on Flickr, a similar number of video clips on YouTube and an even larger number for images in the Google Image Search database. More sophisticated and robust models and algorithms can be proposed by exploiting these images, resulting in better applications for users to index, retrieve, organize and interact with these data. But exactly how such data can be utilized and organized is a problem yet to be solved. In this paper, we introduce a new image database called "ImageNet", a large-scale ontology of images. 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.*

ImageNet uses the hierarchical structure of WordNet [9]. Each meaningful concept in WordNet, possibly described by multiple words or word phrases, is called a "synonym set" or "synset". There are around $80,000$ noun synsets in WordNet. In ImageNet, we aim to provide on average 500-1000 images to illustrate each synset. Images of each concept are quality-controlled and human-annotated as described in Sec. 3.2. ImageNet, therefore, will offer tens of millions of cleanly sorted images. In this paper, we report the current version of ImageNet, consisting of 12 "subtrees": *mammal, bird, fish, reptile, amphibian, vehicle, furniture, musical instrument, geological formation, tool, flower, fruit*. These subtrees contain 5247 synsets and 3.2 million images. Fig. 1 shows a snapshot of two branches of the mammal and vehicle subtrees. The database is publicly available at `http://www.image-net.org`.

The rest of the paper is organized as follows: *We first show that ImageNet is a large-scale, accurate and diverse image database* (Section 2). In Section 4, we present a few simple application examples by exploiting the current ImageNet, mostly the mammal and vehicle subtrees. Our goal is to show that *ImageNet can serve as a useful resource for visual recognition applications such as object recognition, image classification and object localization.* In addition, the construction of such a large-scale and high-quality database can no longer rely on traditional data collection methods. Sec. 3 describes how ImageNet is constructed by leveraging Amazon Mechanical Turk.

2. Properties of ImageNet

ImageNet is built upon the hierarchical structure provided by WordNet. In its completion, ImageNet aims to contain in the order of 50 million cleanly labeled full resolution images (500-1000 per synset). At the time this paper is written, ImageNet consists of 12 subtrees. Most analysis will be based on the mammal and vehicle subtrees.

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

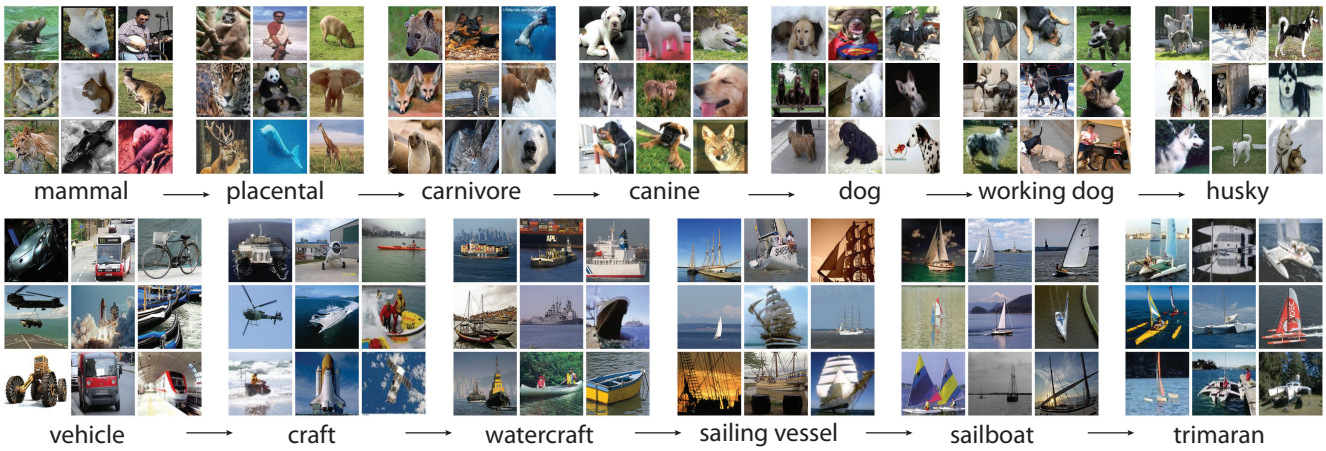
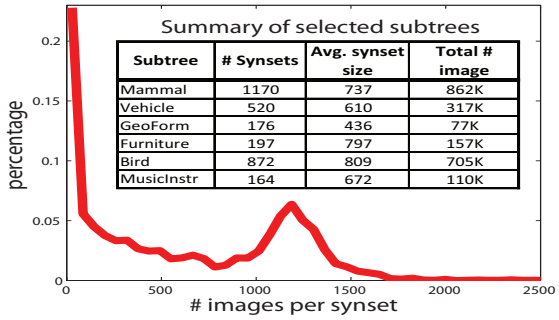


Figure 1: A snapshot of two root-to-leaf branches of ImageNet: the **top** row is from the mammal subtree; the **bottom** row is from the vehicle subtree. For each synset, 9 randomly sampled images are presented.



Subtree	# Synsets	Avg. synset size	Total # image
Mammal	1170	737	862K
Vehicle	520	610	317K
GeoForm	176	436	77K
Furniture	197	797	157K
Bird	872	809	705K
MusicInstr	164	672	110K

Figure 2: Scale of ImageNet. **Red curve**: Histogram of number of images per synset. About 20% of the synsets have very few images. Over 50% synsets have more than 500 images. **Table**: Summary of selected subtrees. For complete and up-to-date statistics visit `http://www.image-net.org/about-stats`.

images spread over 5247 categories (Fig. 2). On average over 600 images are collected for each synset. Fig. 2 shows the distributions of the number of images per synset for the current ImageNet [1]. To our knowledge 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 [2].

Hierarchy ImageNet organizes the different classes of images in a *densely populated* semantic hierarchy. The main asset of WordNet [9] lies in its semantic structure, i.e. its ontology of concepts. Similarly to WordNet, synsets of images in ImageNet are interlinked by several types of relations, the "IS-A" relation being the most comprehensive and useful. Although one can map any dataset with category labels into a semantic hierarchy by using WordNet, the density of ImageNet is unmatched by others. For example, to our knowledge no existing vision dataset offers images of 147 dog categories. Fig. 3 compares the "cat" and "cattle" subtrees of ImageNet and the ESP dataset [25]. We observe that ImageNet offers much denser and larger trees.

[1] About 20% of the synsets have very few images, because either there are very few web images available, e.g. "vespertilian bat", or the synset by definition is difficult to be illustrated by images, e.g. "two-year-old horse".

[2] It is claimed that the ESP game [25] has labeled a very large number of images, but only a subset of 60K images are publicly available.

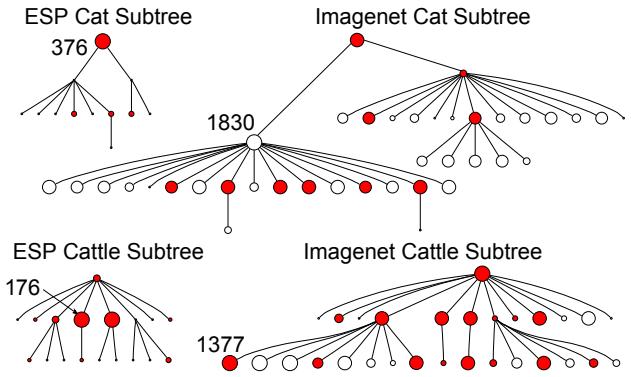


Figure 3: Comparison of the "cat" and "cattle" subtrees between ESP [25] and ImageNet. Within each tree, the size of a node is proportional to the number of images it contains. The number of images for the largest node is shown for each tree. Shared nodes between an ESP tree and an ImageNet tree are colored in red.

Accuracy We would like to offer a clean dataset at all levels of the WordNet hierarchy. Fig. 4 demonstrates the labeling precision on a total of 80 synsets randomly sampled at different tree depths. An average of 99.7% precision is achieved on average. 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.

Diversity ImageNet is constructed with the goal that objects in images should have variable appearances, positions,