Figure 4: Percent of clean images at different tree depth levels in ImageNet. A total of 80 synsets are randomly sampled at every tree depth of the mammal and vehicle subtrees. An independent group of subjects verified the correctness of each of the images. An average of 99.7% precision is achieved for each synset.

Table 1: Comparison of some of the properties of ImageNet versus other existing datasets. ImageNet offers disambiguated labels (LabelDisam), clean annotations (Clean), a dense hierarchy (DenseHie), full resolution images (FullRes) and is publicly available (PublicAvail). ImageNet currently does not provide segmentation annotations.

<table>
<thead>
<tr>
<th></th>
<th>ImageNet</th>
<th>TinyImage</th>
<th>LabelMe</th>
<th>ESP</th>
<th>LHill</th>
</tr>
</thead>
<tbody>
<tr>
<td>LabelDisam</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
<td>N</td>
<td>Y</td>
</tr>
<tr>
<td>Clean</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
<td>N</td>
<td>Y</td>
</tr>
<tr>
<td>DenseHie</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
<td>N</td>
<td>N</td>
</tr>
<tr>
<td>FullRes</td>
<td>Y</td>
<td>N</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
</tr>
<tr>
<td>PublicAvail</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
<td>N</td>
</tr>
<tr>
<td>Segmented</td>
<td>N</td>
<td>Y</td>
<td>N</td>
<td>N</td>
<td>Y</td>
</tr>
</tbody>
</table>

ESP dataset The ESP dataset is acquired through an online game [25]. Two players independently propose labels to one image with the goal of matching as many words as possible in a certain time limit. Millions of images are labeled through this game, but its speeded nature also poses a major drawback. Rosch and Lloyd [20] have demonstrated that humans tend to label visual objects at an easily accessible semantic level termed as “basic level” (e.g. bird), as opposed to more specific level (“sub-ordinate level”, e.g. sparrow), or more general level (“super-ordinate level”, e.g. vertebrate). Labels collected from the ESP game largely concentrate at the “basic level” of the semantic hierarchy as illustrated by the color bars in Fig. 6. ImageNet, however, demonstrates a much more balanced distribution of images across the semantic hierarchy. Another critical difference between ESP and ImageNet is sense disambiguation. When human players input the word “bank”, it is unclear whether it means “a river bank” or a “financial institution”. At this large scale, disambiguation becomes a non-trivial task. Without it, the accuracy and usefulness of the ESP data could be affected. ImageNet, on the other hand, does not have this problem by construction. See section 3.2 for more details. Lastly, most of the ESP dataset is not publicly available. Only 60K images and their labels can be accessed [1].

LabelMe and Lotus Hill datasets LabelMe [21] and the Lotus Hill dataset [27] provide 30k and 50k labeled and segmented images, respectively. These two datasets provide complementary resources for the vision community compared to ImageNet. Both only have around 200 categories, but the outlines and locations of objects are provided. ImageNet in its current form does not provide detailed object outlines (see potential extensions in Sec. 5.1), but the number of categories and the number of images per category

---

3 We also compare with Caltech256 [12]. The result indicates the diversity of ImageNet is comparable, which is reassuring since Caltech256 was specifically designed to be more diverse.

4 We focus our comparisons on datasets of generic objects. Special purpose datasets, such as FERET faces [19], Labeled faces in the Wild [13] and the Mammal Benchmark by Fink and Ullman [11] are not included.

5 All statistics are from [21, 27]. In addition to the 50K images, the Lotus Hill dataset also includes 587k video frames.
Figure 5: ImageNet provides diversified images. (a) Comparison of the lossless JPG file sizes of average images for four different synsets in ImageNet (the mammal subtree) and Caltech101. Average images are downsampled to $32 \times 32$ and sizes are measured in byte. A more diverse set of images results in a smaller lossless JPG file size. (b) Example images from ImageNet and average images for each synset indicated by (a). (c) Examples images from Caltech101 and average images. For each category shown, the average image is computed using all images from Caltech101 and an equal number of randomly sampled images from ImageNet.

Figure 6: Comparison of the distribution of “mammal” labels over tree depth levels between ImageNet and ESP game. The y-axis indicates the percentage of the labels of the corresponding dataset. ImageNet demonstrates a much more balanced distribution, offering substantially more labels at deeper tree depth levels. The actual number of images corresponding to the highest bar is also given for each dataset.

already far exceeds these two datasets. In addition, images in these two datasets are largely uploaded or provided by users or researchers of the dataset, whereas ImageNet contains images crawled from the entire Internet. The Lotus Hill dataset is only available through purchase.

3. Constructing ImageNet

ImageNet is an ambitious project. Thus far, we have constructed 12 subtrees containing 3.2 million images. Our goal is to complete the construction of around 50 million images in the next two years. We describe here the method we use to construct ImageNet, shedding light on how properties of Sec. 2 can be ensured in this process.

3.1. Collecting Candidate Images

The first stage of the construction of ImageNet involves collecting candidate images for each synset. The average accuracy of image search results from the Internet is around 10% [24]. ImageNet aims to eventually offer 500-1000 clean images per synset. We therefore collect a large set of candidate images. After intra-synset duplicate removal, each synset has over $10^4$ images on average.

We collect candidate images from the Internet by querying several image search engines. For each synset, the queries are the set of WordNet synonyms. Search engines typically limit the number of images retrievable (in the order of a few hundred to a thousand). To obtain as many images as possible, we expand the query set by appending the queries with the word from parent synsets, if the same word appears in the gloss of the target synset. For example, when querying “whippet”, according to WordNet’s gloss a “small slender dog of greyhound type developed in England”, we also use “whippet dog” and “whippet greyhound”.

To further enlarge and diversify the candidate pool, we translate the queries into other languages [10], including Chinese, Spanish, Dutch and Italian. We obtain accurate translations by WordNets in those languages [3, 2, 4, 26].

3.2. Cleaning Candidate Images

To collect a highly accurate dataset, we rely on humans to verify each candidate image collected in the previous step for a given synset. This is achieved by using the service of Amazon Mechanical Turk (AMT), an online platform on which one can put up tasks for users to complete and to get paid. AMT has been used for labeling vision data [23]. With a global user base, AMT is particularly suitable for large scale labeling.

In each of our labeling tasks, we present the users with a set of candidate images and the definition of the target synset (including a link to Wikipedia). We then ask the users to verify whether each image contains objects of the synset. We encourage users to select images regardless of occlusions, number of objects and clutter in the scene to...
While users are instructed to make accurate judgment, we need to set up a quality control system to ensure this accuracy. There are two issues to consider. First, human users make mistakes and not all users follow the instructions. Second, users do not always agree with each other, especially for more subtle or confusing synsets, typically at the deeper levels of the tree. Fig. 7(left) shows an example of how users’ judgments differ for “Burmese cat” images.

The solution to these issues is to have multiple users independently label the same image. An image is considered positive only if it gets a convincing majority of the votes. We observe, however, that different categories require different levels of consensus among users. For example, while five users might be necessary for obtaining a good consensus on “Burmese cat” images, a much smaller number is needed for “cat” images. We develop 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 (Fig. 7(right) shows examples for “Burmese cat” and “cat”). For each of remaining candidate images in this synset, we proceed with the AMT user labeling until a pre-determined confidence score threshold is reached. It is worth noting that the confidence table gives a natural measure of the “semantic difficulty” of the synset. For some synsets, users fail to reach a majority vote for any image, indicating that the synset cannot be easily illustrated by images. Fig. 4 shows that our algorithm successfully filters the candidate images, resulting in a high percentage of clean images per synset.

![Table of User Votes](image)

<table>
<thead>
<tr>
<th>User</th>
<th>Y</th>
<th>Y</th>
<th>Y</th>
</tr>
</thead>
<tbody>
<tr>
<td>User 1</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
</tr>
<tr>
<td>User 2</td>
<td>N</td>
<td>Y</td>
<td>Y</td>
</tr>
<tr>
<td>User 3</td>
<td>N</td>
<td>Y</td>
<td>Y</td>
</tr>
<tr>
<td>User 4</td>
<td>Y</td>
<td>N</td>
<td>Y</td>
</tr>
<tr>
<td>User 5</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
</tr>
<tr>
<td>User 6</td>
<td>N</td>
<td>N</td>
<td>Y</td>
</tr>
</tbody>
</table>

**Figure 7:** Left: Is there a Burmese cat in the images? Six randomly sampled users have different answers. Right: The confidence score table for “Cat” and “Burmese cat”. More votes are needed to reach the same degree of confidence for “Burmese cat” images.

### 4. ImageNet Applications

In this section, we show three applications of ImageNet. The first set of experiments underline the advantages of having clean, full resolution images. The second experiment exploits the tree structure of ImageNet, whereas the last experiment outlines a possible extension and gives more insights into the data.

#### 4.1. Non-parametric Object Recognition

Given an image containing an unknown object, we would like to recognize its object class by querying similar images in ImageNet. Torralba et al. [24] has demonstrated that, given a large number of images, simple nearest neighbor methods can achieve reasonable performances despite a high level of noise. We show that with a clean set of full resolution images, object recognition can be more accurate, especially by exploiting more feature level information.

We run four different object recognition experiments. In all experiments, we test on images from the 16 common categories between Caltech256 and the mammal subtree. We measure classification performance on each category in the form of an ROC curve. For each category, the negative set consists of all images from the other 15 categories. We now describe in detail our experiments and results(Fig. 8).

1. **NN-voting + noisy ImageNet** First we replicate one of the experiments described in [24], which we refer to as “NN-voting” hereafter. To imitate the TinyImage dataset (i.e. images collected from search engines without human cleaning), we use the original candidate images for each synset (Section 3.1) and downsample them to $32 \times 32$. Given a query image, we retrieve 100 of the nearest neighbor images by SSD pixel distance from the mammal subtree. Then we perform classification by aggregating votes (number of nearest neighbors) inside the tree of the target category.

2. **NN-voting + clean ImageNet** Next we run the same NN-voting experiment described above on the clean ImageNet dataset. This result shows that having more accurate data improves classification performance.

3. **NBBN** We also implement the Naive Bayesian Nearest Neighbor (NBBN) method proposed in [5] to underline the usefulness of full resolution images. NBBN employs a bag-of-features representation of images. SIFT [15] descriptors are used in this experiment. Given a query image $Q$ with descriptors $\{d_i\}, i = 1, \ldots, M$, for each object class $C$, we compute the query-class distance $D_C = \sum_{i=1}^{M} \delta(d_i, C)$.

---

6 An alternative explanation is that we did not obtain enough suitable candidate images. Given the extensiveness of our crawling scheme, this is a rare scenario.

7 The categories are bat, bear, camel, chimp, dog, elk, giraffe, goat, gorilla, greyhound, horse, killer-whale, porcupine, raccoon, skunk, zebra. Duplicates (~ 20 per category) with ImageNet are removed from the test set.
The ROC curve for "elk" and "killer-whale". Caltech256 images serve as testing images.

The mammal subtree. Note that our algorithm is agnostic to any method used to learn image classifiers for each synset. In this case, we use an AdaBoost-based classifier proposed by [6]. For each synset, we randomly sample 90% of the images to form the positive training image set, leaving the rest of the 10% as testing images. We form a common negative image set by aggregating 10 images randomly sampled from each synset. When training an image classifier for a particular synset, we use the positive set from this synset as well as the common negative image set excluding the images drawn from this synset, and its child and parent synsets.

We evaluate the classification results by AUC (the area under ROC curve). Fig. 9 shows the results of AUC for synsets at different levels of the hierarchy, compared with an independent classifier that does not exploit the tree structure of ImageNet. The plot indicates that images are easier to classify at the bottom of the tree (e.g. star-nosed mole, minivan, polar bear) as opposed to the top of the tree (e.g. vehicles, mammal, artifact, etc.). This is most likely due to stronger visual coherence near the leaf nodes of the tree.

At nearly all levels, the performance of the tree-max classifier is consistently higher than the independent classifier. This result shows that a simple way of exploiting the ImageNet hierarchy can already provide substantial improvement for the image classification task without additional training or model learning.
4.3. Automatic Object Localization

ImageNet can be extended to provide additional information about each image. One such information is the spatial extent of the objects in each image. Two application areas come to mind. First, for training a robust object detection algorithm one often needs localized objects in different poses and under different viewpoints. Second, having localized objects in cluttered scenes enables users to use ImageNet as a benchmark dataset for object localization algorithms. In this section we present results of localization on 22 categories from different depths of the WordNet hierarchy. The results also throw light on the diversity of images in each of these categories.

We use the non-parametric graphical model described in [14] to learn the visual representation of objects against a global background class. In this model, every input image is represented as a “bag of words”. The output is a probability for each image patch to belong to the topics $z_i$ of a given category (see [14] for details). In order to annotate images with a bounding box we calculate the likelihood of each image patch given a category $c$: $p(x|c) = \sum p(x|z_i, c) p(z_i|c)$. Finally, one bounding box is put around the region which accumulates the highest likelihood.

We annotated 100 images in 22 different categories of the mammal and vehicle subtrees with bounding boxes around the objects of that category. Fig. 10 shows precision and recall values. Note that precision is low due to extreme variability of the objects and because of small objects which have hardly any salient regions.

Fig. 11 shows sampled bounding boxes on different classes. The colored region is the detected bounding box, while the original image is in light gray.

In order to illustrate the diversity of ImageNet inside each category, Fig. 12 shows results on running k-means clustering on the detected bounding boxes after converting them to grayscale and rescaling them to $32 \times 32$. All average images, including those for the entire cluster, are created with approximately 40 images. While it is hard to identify the object in the average image of all bounding boxes (shown in the center) due to the diversity of ImageNet, the average images of the single clusters consistently discover viewpoints or common poses.

5. Discussion and Future Work

Our future work has two goals:

5.1. Completing ImageNet

The current ImageNet constitutes ~ 10% of the WordNet synsets. To further speed up the construction process, we will continue to explore more effective methods to evaluate the AMT user labels and optimize the number of repetitions needed to accurately verify each image. At the completion of ImageNet, we aim to (i) have roughly 50 million clean, diverse and full resolution images spread over approximately 50K synsets; (ii) deliver ImageNet to research communities by making it publicly available and readily ac-
cessible online. We plan to use cloud storage to enable efficient distribution of ImageNet data; (iii) extend ImageNet to include more information such as localization as described in Sec. 4.3, segmentation, cross-synset referencing of images, as well as expert annotation for difficult synsets and (iv) foster an ImageNet community and develop an online platform where everyone can contribute to and benefit from ImageNet resources.

5.2. Exploiting ImageNet

We hope ImageNet will become a central resource for a broad range of vision related research. For the computer vision community in particular, we envision the following possible applications.

A training resource. Most of today’s object recognition algorithms have focused on a small number of common objects, such as pedestrians, cars and faces. This is mainly due to the high availability of images for these categories. Fig. 6 has shown that even the largest datasets today have a strong bias in their coverage of different types of objects. ImageNet, on the other hand, contains a large number of images for nearly all object classes including rare ones. One interesting research direction could be to transfer knowledge of common objects to learn rare object models.

A benchmark dataset. The current benchmark datasets in computer vision such as Caltech101/256 and PASCAL have played a critical role in advancing object recognition and scene classification research. We believe that the high quality, diversity and large scale of ImageNet will enable it to become a new and challenging benchmark dataset for future research.

Introducing new semantic relations for visual modeling. Because ImageNet is uniquely linked to all concrete nouns of WordNet whose synsets are richly interconnected, one could also exploit different semantic relations for instance to learn part models. To move towards total scene understanding, it is also helpful to consider different depths of the semantic hierarchy.

Human vision research. ImageNet’s rich structure and dense coverage of the image world may help advance the understanding of the human visual system. For example, the question of whether a concept can be illustrated by images is much more complex than one would expect at first. Aligning the cognitive hierarchy with the “visual” hierarchy also remains an unexplored area.

Acknowledgment

The authors would like to thank Bangpeng Yao, Hao Su, Barry Chai and anonymous reviewers for their helpful comments. WD is supported by Gordon Wu fellowship. RS is supported by the ERP and Upton fellowships. KL is funded by NSF grant CNS-0509447 and by research grants from Google, Intel, Microsoft and Yahoo!. LFF is funded by research grants from Microsoft and Google.

References