What Does Classifying More Than 10,000 Image Categories Tell Us?

Jia Deng[1,3], Alexander C. Berg[2], Kai Li[1], and Li Fei-Fei[3]

[1] Princeton University
[2] Columbia University
[3] Stanford University

Abstract. Image classification is a critical task for both humans and computers. One of the challenges lies in the large scale of the semantic space. In particular, humans can recognize tens of thousands of object classes and scenes. No computer vision algorithm today has been tested at this scale. This paper presents a study of large scale categorization including a series of challenging experiments on classification with more than $10,000$ image classes. We find that a) computational issues become crucial in algorithm design; b) conventional wisdom from a couple of hundred image categories on relative performance of different classifiers does not necessarily hold when the number of categories increases; c) there is a surprisingly strong relationship between the structure of WordNet (developed for studying language) and the difficulty of visual categorization; d) classification can be improved by exploiting the semantic hierarchy. Toward the future goal of developing automatic vision algorithms to recognize tens of thousands or even millions of image categories, we make a series of observations and arguments about dataset scale, category density, and image hierarchy.

1 Introduction

Recognizing categories of objects and scenes is a fundamental human ability and an important, yet elusive, goal for computer vision research. One of the major challenges is the sheer scale of the problem, both in terms of the very high dimensional physical space of images, and the large semantic space humans use to describe visual stimuli. In particular, psychologists have postulated that humans are able to categorize at least tens of thousands of objects and scenes [1].

The breadth of the semantic space has important implications. For many real world vision applications, the ability to handle a large number of object classes becomes a minimum requirement, *e.g.* an image search engine or an automatic photo annotator is significantly less useful if it is unable to cover a wide range of object classes. Even for tasks in restricted domains, *e.g.* car detection, to be effective in the real world, an algorithm needs to discriminate against a large number of distractor object categories.

Recent progress on image categorization has been impressive and has introduced a range of features, models, classifiers, and frameworks [2,3,4,5,6,7,8,9,10]. In this paper we explore scaling up the number of categories considered in recognition experiments from hundreds to over 10 thousand in order to move toward

K. Daniilidis, P. Maragos, N. Paragios (Eds.): ECCV 2010, Part V, LNCS 6315, pp. 71–84, 2010.

reducing the gap between machine performance and human abilities. Note that this is not simply a matter of training more and more classifiers (although that is a challenging task on its own). With such large numbers of categories there is a concomitant shift in the difficulty of discriminating between them as the categories sample the semantic space more densely. The previously unexplored scale of the experiments in this paper allow this effect to be measured.

Recognition encompasses a wide range of specific tasks, including classification, detection, viewpoint understanding, segmentation, verification and more. In this paper we focus on category recognition, in particular the task of assigning a single category label to an image that contains one or more instances of a category of object following the work of [11,12,13,14].

We conduct the first empirical study of image categorization at near human scale. Some results are intuitive – discriminating between thousands of categories is in fact more difficult that discriminating between hundreds – but other results reveal structure in the difficulty of recognition that was previously unknown. Our key contributions include:

- The first in-depth study of image classification at such a large scale. Such experiments are technically challenging, and we present a series of techniques to overcome the difficulty. (Sec. 5)
- We show that conventional wisdom obtained from current datasets does not necessarily hold in some cases at a larger scale. For example, the ordering by performance of techniques on hundreds of categories is not preserved on thousands of categories. Thus, we cannot solely rely on experiments on the Caltech [13,14] and PASCAL [12] datasets to predict performance on large classification problems. (Sec. 6)
- We propose a measure of similarity between categories based on WordNet[15] – a hierarchy of concepts developed for studying language. Experiments show a surprisingly strong correlation between this purely linguistic metric and the performance of visual classification algorithms. We also show that the categories used in previous object recognition experiments are relatively sparse – distinguishing them from each other is significantly less difficult than distinguishing many other denser subsets of the 10,000 categories. (Sec. 7)
- Object categories are naturally hierarchical. We propose and evaluate a technique to perform hierarchy aware classification, and show that more informative classification results can be obtained. (Sec. 8)

2 Related Work

Much recent work on image classification has converged on bag of visual word models (BoW) [16] based on quantized local descriptors [17,3,4] and support vector machines [3,2] as basic techniques. These are enhanced by multi-scale spatial pyramids (SPM) [4] on BoW or histogram of oriented gradient (HOG) [18,4] features. In the current state-of-the-art, multiple descriptors and kernels are combined using either ad hoc or multiple kernel learning approaches [19,5,20,21]. Work in machine learning supports using winner-takes-all between 1-vs-all

classifiers for the final multi-class classification decision [22]. We choose SPM using BoW because it is a key component of many of the best recognition results [19,5,20,21] and is relatively efficient. Recent work allows fast approximation of the histogram intersection kernel SVM, used for SPM, by a linear SVM on specially encoded SPM features [23]. See Appendix for the modifications necessary to allow even that very efficient solution to scale to very large problems.

There are very few multi-class image datasets with many images for more than 200 categories. One is Tiny Images [6], 32x32 pixel versions of images collected by performing web queries for the nouns in the WordNet [15] hierarchy, without verification of content. The other is ImageNet [24], also collected from web searches for the nouns in WordNet, but containing full images verified by human labelers. To date there have been no recognition results on large numbers of categories published for either dataset[1]. Fergus *et al*. explore semi-supervised learning on 126 hand labeled Tiny Images categories [25] and Wang *et al*. show classification on a maximum of 315 categories ($< 5\%$) [26].

Recent work considering hierarchies for image recognition or categorization [27,28,29,30] has shown impressive improvements in accuracy and efficiency, but has not studied classification minimizing hierarchical cost. Related to classification is the problem of detection, often treated as repeated 1-vs-all classification in sliding windows. In many cases such localization of objects might be useful for improving classification, but even the most efficient of the state of the art techniques [7,20,31] take orders of magnitude more time per image than the ones we consider in this study, and thus cannot be utilized given the scale of our experiments.

3 Datasets

The goals of this paper are to study categorization performance on a significantly larger number of categories than the current state of the art, and furthermore to delve deeper toward understanding the factors that affect performance. In order to achieve this, a dataset with a large number of categories spanning a wide range of concepts and containing many images is required. The recently released ImageNet dataset consists of more than 10,000,000 images in over 10,000 categories organized by the WordNet hierarchy [24]. The size and breadth of this data allow us to perform multiple longitudinal probes of the classification problem. Specifically we consider the following datasets:

- **ImageNet10K.** 10184 categories from the Fall 2009 release of ImageNet [32], including both internal and leaf nodes with more than 200 images each (a total of 9 million images).
- **ImageNet7K.** 7404 leaf categories from ImageNet10K. Internal nodes may overlap with their descendants, so we also consider this leaf only subset.
- **ImageNet1K.** 1000 leaf categories randomly sampled from ImageNet7K.

[1] Previous work on Tiny Images [6] and ImageNet [24] shows only proof of concept classification on fewer than 50 categories.

- **Rand200{a,b,c}.** Three datasets, each containing 200 randomly selected leaf categories. The categories in Rand200a are sampled from ImageNet1K while Rand200b and Rand200c are sampled directly from ImageNet7K.
- **Ungulate183, Fungus134, Vehicle262.** Three datasets containing all the leaf nodes that are descendants of particular parent nodes in ImageNet10K (named by the parent node and number of leaves).
- **CalNet200**. This dataset serves as a surrogate for the Caltech256 dataset – containing the 200 image categories from Caltech256 that exist in ImageNet.

Note that all datasets have non-overlapping categories except ImageNet10K. Following the convention of the PASCAL VOC Challenge, each category is randomly split 50%-50% into a set of training and test images, with a total of 4.5 million images for training and 4.5 million images for testing. All results are averaged over two runs by swapping training and test, except for ImageNet7K and ImageNet10K due to extremely heavy computational cost. In all cases we provide statistical estimates of the expected variation. The number of training images per category ranges from 200 to 1500, with an average of 450.

4 Procedure

The main thrust of this paper is image classification: given an image and K classes, the task is to select one class label. We employ two evaluation measures:

Mean accuracy. The accuracy of each class is the percentage of correct predictions, *i.e.* predictions identical to the ground truth class labels. The mean accuracy is the average accuracy across all classes.

Mean misclassification cost. To exploit the hierarchical organization of object classes, we also consider the scenario where it is desirable to have non-uniform misclassification cost. For example, misclassifying "dog" as "cat" might not be penalized as much as misclassifying "dog" as "microwave". Specifically, for each image $x_i^{(k)} \in X, i = 1, \ldots, m$ from class k, we consider predictions $f(x_i^{(k)}) : X \rightarrow \{1, \ldots, K\}$, where K is the number of classes (*e.g.* $K = 1000$ for ImageNet1K) and evaluate the cost for class k as $L_k = \frac{1}{m} \sum_{i=1}^{m} C_{f(x_i^{(k)}),k}$, where C is a $K \times K$ cost matrix and $C_{i,j}$ is the cost of classifying the true class j as class i. The mean cost is the average cost across all classes. Evaluation using a cost based on the ImageNet hierarchy is discussed in Sec. 8.

We use the following four algorithms in our evaluation experiments as samples of some major techniques used in object recognition:

- `GIST+NN` Represent each image by a single GIST [33] descriptor (a commonly accepted baseline descriptor for scene classification) and classify using *k-nearest-neighbors* (kNN) on L2 distance.
- `BOW+NN` Represent each image by a histogram of SIFT [17] codewords and classify using kNN on L1 distance, as a baseline for BoW NN-based methods.
- `BOW+SVM` Represent each image by a histogram of SIFT codewords, and train and classify using linear SVMs. Each SVM is trained to distinguish one class from the rest. Images are classified by the class with largest score (a 1-vs-all framework). This serves as a baseline for classifier-based algorithms.

– `SPM+SVM` Represent each image by a spatial pyramid of histograms of SIFT codewords [4]. Again a 1-vs-all framework is used, but with approximate histogram intersection kernel SVMs [23,3,4]. This represents a significant component of many state of the art classifiers [19,5,20,21].

5 Computation Matters

Working at the scale of 10,000 categories and 9 million images moves computational considerations to the forefront. Many common approaches become computationally infeasible at such large scale.

As a reference, for this data it takes 1 hour on a 2.66GHz Intel Xeon CPU to train *one* binary linear SVM on bag of visual words histograms (including a minimum amount of parameter search using cross validation), using the extremely efficient LIBLINEAR [34]. In order to perform multi-class classification, one common approach is 1-vs-all, which entails training 10,000 such classifiers – requiring more than 1 CPU year for training and 16 hours for testing. Another approach is 1-vs-1, requiring 50 million pairwise classifiers. Training takes a similar amount of time, but testing takes about 8 years due to the huge number of classifiers. A third alternative is the "single machine" approach, *e.g.* Crammer & Singer [35], which is comparable in training time but is not readily parallelizable. We choose 1-vs-all as it is the only affordable option.

Training `SPM+SVM` is even more challenging. Directly running intersection kernel SVM is impractical because it is at least 100× slower (100+ years) than linear SVM [23]. We use the approximate encoding proposed by Maji & Berg [23] that allows fast training with LIBLINEAR. This reduces the total training time to 6 years. However, even this very efficient approach must be modified because memory becomes a bottleneck [2] – a direct application of the efficient encoding of [23] requires 75GB memory, far exceeding our memory limit (16GB). We reduce it to 12G through a combination of techniques detailed in Appendix A.

For NN based methods, we use brute force linear scan. It takes 1 year to run through all testing examples for GIST or BOW features. It is possible to use approximation techniques such as locality sensitive hashing [36], but due to the high feature dimensionality (*e.g.* 960 for GIST), we have found relatively small speed-up. Thus we choose linear scan to avoid unnecessary approximation.

In practice, all algorithms are parallelized on a computer cluster of 66 multicore machines, but it still takes weeks for a single run of all our experiments. Our experience demonstrates that computational issues need to be confronted at the outset of algorithm design when we move toward large scale image classification, otherwise even a baseline evaluation would be infeasible. Our experiments suggest that to tackle massive amount of data, distributed computing and efficient learning will need to be integrated into any vision algorithm or system geared toward real-world large scale image classification.

[2] While it is possible to use online methods, *e.g.* stochastic subgradient descent, they can be slower to converge [34].

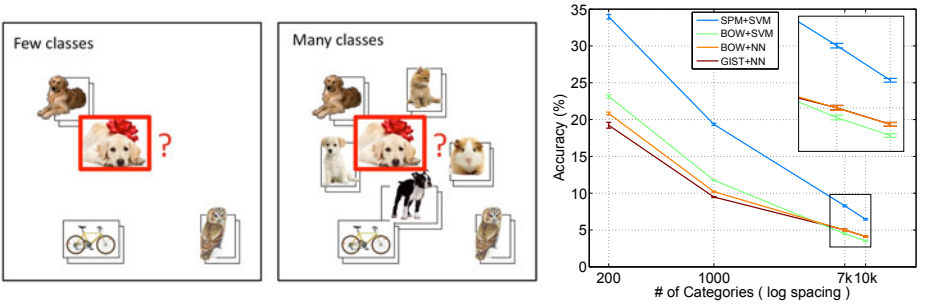

Fig. 1. Given a query image, the task of "image classification" is to assign it to one of the classes (represented by a stack of images) that the algorithm has learned. **Left:** Most traditional vision algorithms have been tested on a small number of somewhat distinct categories. **Middle:** Real world image classification problems may involve a much larger number of categories – so large that the categories can no longer be easily separated. **Right:** Mean classification accuracy of various methods on Rand200$\{a, b, c\}$, ImageNet1K, ImageNet7K and ImageNet10K.

6 Size Matters

We first investigate the broad effects on performance and computation of scaling to ten-thousand categories. As the number of categories in a dataset increases, the accuracy of classification algorithms decreases, from a maximum of 34% for Rand200{a,b,c} to 6.4% for ImageNet10K (Fig. 1 *right*). While the performance drop comes at no surprise, the speed of decrease is slower than might be expected – roughly a 2× decrease in accuracy with 10× increase in the number of classes, significantly better than the 10× decrease of a random baseline.

There is a surprise from *k-nearest-neighbor* (*kNN*) classifiers, either using GIST features or BoW features. For Rand200{a,b,c}, these techniques are significantly worse than linear classifiers using BoW features, around 10% lower in accuracy. This is consistent with the experience of the field – methods that do use *kNN* must be augmented in order to provide competitive performance [2,37]. But the picture is different for ImageNet7K or ImageNet10K categories, where simple *kNN* actually outperforms linear SVMs on BoW features (`BOW+SVM`), with 11-16% *higher* accuracy. The small absolute gain in mean accuracy, around 0.5%, is made significant by the very small expected standard deviation of the means 0.1% [3]. **A technique that significantly outperforms others on small datasets may actually underperform them on large numbers of categories**.

This apparent breakdown for 1-vs-all with linear classifiers comes despite a consistent line of work promoting this general strategy for multi-class classification [22]. It seems to reveal issues with calibration between classifiers, as the majority of categories have comparable discriminative power on ImageNet7K and Rand200a (Fig 2 *left*), but multi-way classification is quite poor for ImageNet7K

[3] Stdev for ImageNet7K and ImageNet10K are estimated using the individual category variances, but are very small *cf* standard error and the central limit theorem.

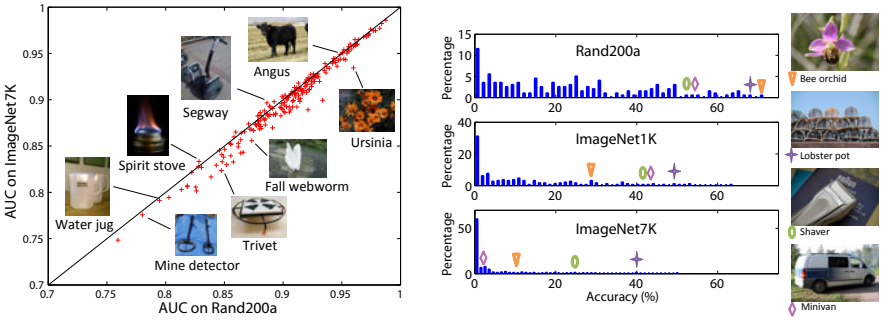

Fig. 2. Left: Scatter plot comparing the area under ROC curve (AUC) of `BOW+SVM` for the 200 categories in Rand200a when trained and evaluated against themselves(x-axis) and when trained and evaluated against ImageNet7K(*y*-axis). **Right:** Histograms of accuracies for the same 200 categories in Rand200a, ImageNet1K, and ImageNet7K, example categories indicated with colored markers.

(Fig 2 *right*). One explanation is that for the one-against-all approach, a correct prediction would require that the true classifier be more confident than any other classifiers, which becomes more difficult with a larger number of classes as the chance of false alarms from others greatly increases. Then the behavior starts to resemble kNN methods, which are only confident about close neighbors.

Looking in more detail at the confusion between the categories in ImageNet7K reveals additional structure (Fig. 3). Most notable is the generally block diagonal structure, **indicating a correlation between the structure of the semantic hierarchy (by WordNet) and visual confusion between the categories**. The two most apparent blocks roughly align with "artifacts" and "animals", two very high level nodes in WordNet, suggesting the least amount of confusion between these two classes with more confusion within. This is consistent with both computational studies on smaller numbers of classes [30] and some human abilities [38]. Sections of the confusion matrix are further expanded in Fig. 3. These also show roughly block diagonal structures at increasingly finer levels not available in other datasets. The pattern is roughly block diagonal, but by no means exact. There is a great deal of noise and a fainter "plaid", oscillating pattern of stripes, indicating that the ordering of categories in WordNet is not completely in agreement with the visual confusion between them.

The block patterns indicate that it is possible to speed up the classification by using a sublinear number of classifiers in a hierarchy, as Griffin & Perona have demonstrated on Caltech256 [30]. They built a hierarchy of classifiers directly from the confusion matrix. Here we confirm their findings by observing a much stronger pattern on a large number of classes. Moreover we note that such a grouping may actually be directly obtained from WordNet, in which case, the output of an internal classifier in the hierarchy would be semantically meaningful.

Also of note is that in scaling to many classes, only a small subset of the distractor classes are truly distracting, possibly explaining the smaller than

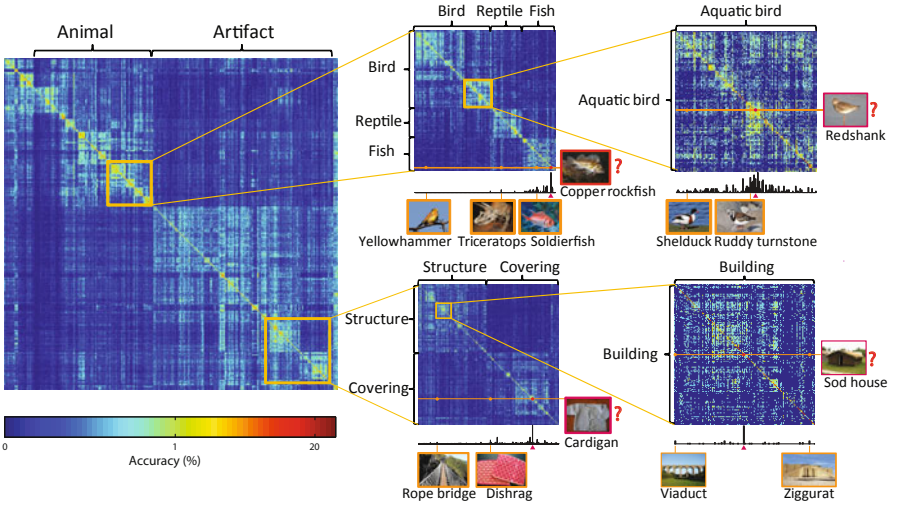

Fig. 3. Confusion matrix and sub-matrices of classifying the 7404 leaf categories in ImageNet7K, ordered by a depth first traversal of the WordNet hierarchy, using `SPM+SVM`. **Left:** Downsampled 7404 × 7404 confusion matrix, each pixel representing max confusion over 4 × 4 entries. **Middle:** Zoom-in to two sub-matrices (top: 949 × 949; bottom: 1368 × 1368), each pixel 2 × 2 entries. One row of the matrix is plotted below each matrix (corresponding to red outlined images). The correct class is indicated by a red triangle. Examples of other classes are also shown. **Right:** Further zoom-in (top: 188 × 188; bottom: 145 × 145), each pixel representing the confusion between two individual categories.

expected performance drop. For example, to classify "German shepherd", most of the distractor classes are "easy" ones like "dishrag", while only a few semantically related classes like "husky" add to the difficulty. It suggests that one key to improving large scale classification is to focus on those classes, whose difficulty correlates with semantic relatedness. We quantify this correlation in Sec. 7.

7 Density Matters

Our discussion so far has focused on the challenges arising from the sheer number of categories. Figure 3 reveals that the difficulty of recognition varies significantly over different parts of the semantic space. Some classifiers must tackle more semantically related, and possibly visually similar, categories. Accurate classification of such categories leads to useful applications, *e.g.* classifying groceries for assisting the visually impaired, classifying home appliances for housekeeping robots, or classifying insect species for environmental monitoring [39]. We refer to sets of such categories as *dense* and study the effect of density on classification.

We begin by comparing mean classification accuracy for classifiers trained and tested on each of the small datasets – Fungus134, Ungulate183, Vehicle262,

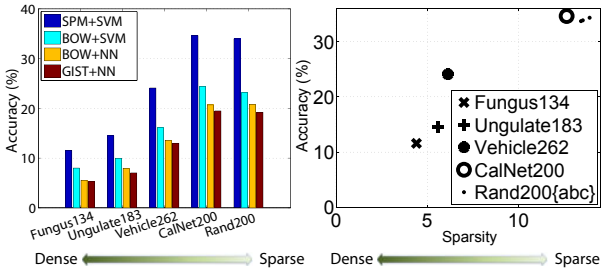

Fig. 4. Left: Accuracy on datasets of varying density. Note that CalNet200 (the Caltech 256 categories in ImageNet) has low density and difficulty on par with a set of 200 randomly sampled categories. **Right**: Accuracy (using `SPM+SVM`) versus dataset density measured by mean distance in WordNet (see Sec. 7).

CalNet200, Rand200 – across descriptors and classifiers in Fig. 4. Note that while `SPM+SVM` produces consistently higher accuracies than the other approaches, the ordering of datasets by performance is exactly the same for each approach[4]. This indicates that **there is a significant difference in difficulty between different *datasets*, independent of feature and classifier choice**.

Next we try to predict the difficulty of a particular dataset by measuring the density of the categories, based on the hierarchical graph structure of WordNet. We define the distance, $h(i, j)$, between categories i and j, as the height of their lowest common ancestor. The height of a node is the length of the longest path down to a leaf node (leaf nodes have height 0). We measure the density of a dataset as the mean $h(i, j)$ between all pairs of categories – smaller implies denser. See Fig. 5 for an illustration and for examples of pairs of categories from each dataset that have distance closest to the mean for that dataset. There is a very **clear correlation between the density in WordNet and accuracy of visual classification; *denser* datasets predict lower accuracy** (Fig. 4). This is despite the fact that WordNet was not created as a visual hierarchy!

Classification accuracy on 200 randomly chosen categories (Rand200{a,b,c}) is more than 3 times higher than on the 134 categories from Fungus134. The large gap suggests that the methods studied here are not well equipped for classifying dense sets of categories. In fact, there have been relatively few efforts on "dense classification" with some notable exceptions, *e.g.* [40,41,39]. The results seem to call for perhaps more specialized features and models, since it is one key to improving large scale classification performance as discussed in Sec. 6

Also of note is that the Caltech256 categories that occur in ImageNet (CalNet200) have very low density and relatively high accuracy – in almost exactly the same range as random sets of categories. **The Caltech categories are very sparse and do not exhibit the difficulty of dense sets of categories,**

[4] Ordering of datasets is consistent, but ordering of methods may change between datasets as noted in Sec. 6 where `BOW+SVM` and the kNN approaches switch order.

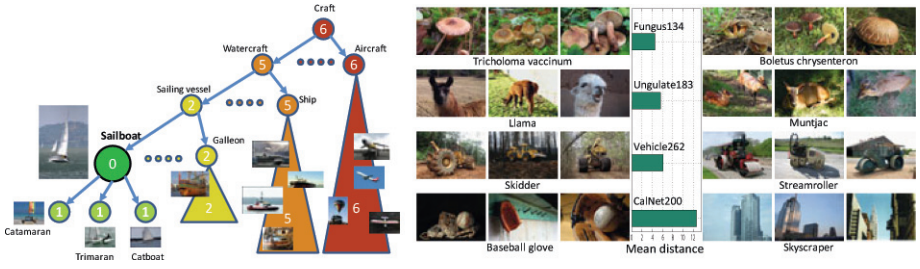

Fig. 5. Left: Illustration of the inter-class distance (indicated by the numbers) between "sailboat" and other classes, as defined in Sec. 7. Any descendant of ship is further from sailboat than gallon but closer than those in aircraft. Note that one step up the hierarchy may increase the distance by more than one as the tree height is the length of the longest path to a leaf node. **Right:** Each row shows a pair of example categories from the dataset indicated in the center column. The pairs are chosen to have distance near the mean distance in WordNet (Sec. 7) between categories in the dataset, indicated by the bars in the center column.

making Caltech-like datasets incomplete as an evaluation resource towards some of the real-world image classification problems.

Finally we note that our WordNet based measure is not without limitations, *e.g.* "food tuna" and "fish tuna" are semantically related but belong to "food" and "fish" subtrees respectively, so are far away from each other. Nonetheless as a starting point for quantifying semantic density, the results are encouraging.

8 Hierarchy Matters

For recognition at the scale of human ability, categories will necessarily overlap and display a hierarchical structure [11]. For example, a human may label "redshank" as "shorebird", "bird", or "animal", all of which are correct but with a decreasing amount of information. Humans make mistakes too, but to different degrees at different levels – a "redshank" might be mistaken as a "red-backed sandpiper", but almost never as anything under "home appliance".

The implications for real world object classification algorithms are two fold. First a learning algorithm needs to exploit real world data that inevitably has labels at different semantic levels. Second, it is desirable to output labels as informative as possible while minimizing mistakes at higher semantic levels.

Consider an automatic photo annotator. If it cannot classify "redshank" reliably, an answer of "bird" still carries much more information than "microwave". However, our classifiers so far, trained to minimize the 0-1 loss [5], have no incentive to do so – predicting "microwave" costs the same as predicting "bird".

[5] The standard loss function for classification, where a correct classification costs zero and any incorrect classification costs 1.

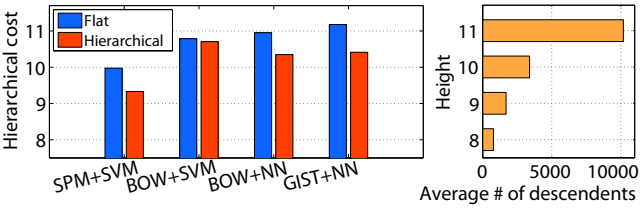

Fig. 6. Left: Hierarchical cost of flat classification and hierarchical classification on ImageNet10K across different methods. **Right**: Mean number of descendants for nodes at each height, indicating the effective log-scale for hierarchical cost.

Here we explore ways to make classifiers more informative. We define a hierarchical cost $C_{i,j}$ for classifying an image of class j as class i as $C_{i,j} = 0$ when $i = j$ or when i is a descendant of j, and $C_{i,j} = h(i, j)$, the height of their lowest common ancestor in WordNet, otherwise. This cost definition directly measures the semantic level at which a misclassification occurs – a more informative classifier, one able to discriminate finer details, would have lower cost. It also takes care of the overlapping categories – there is penalty for classifying an image in an internal node as its (more general) ancestor but no cost for classifying it as any of its (more specific) descendants. As an example, in Fig. 5 *left*, for an image labeled as "sailboat", classifying it as "catamaran" or any other descendant incurs no cost [6] while classifying as any descendant of "aircraft" incurs cost 6.

We can make various classification approaches cost sensitive by obtaining probability estimates (Appendix). For a query image x, given posterior probability estimates $\hat{p}_j(x)$ for class j, $j \in \{1, \ldots K\}$, according to Bayesian decision theory, the optimal prediction is obtained by predicting the label that minimizes the expected cost $f(x) = \arg\min_{i=1,\ldots,K} \sum_{j=1}^{K} C_{i,j}\hat{p}_j(x)$.

Comparing the mean hierarchical cost for the original (flat) classifier with the mean cost for the cost sensitive (hierarchical) classifier, we find a consistent reduction in cost on ImageNet10K(Fig. 6). It shows that the hierarchical classifier can discriminate at more informative semantic levels. While these reductions may seem small, the cost is effectively on a log scale. It is measured by the height in the hierarchy of the lowest common ancestor, and moving up a level can more than double the number of descendants (Fig. 6 *right*).

The reduction of mean cost on its own would not be interesting without a clear benefit to the results of classification. The examples in Fig. 7 show query images and their assigned class for flat classification and for classification using hierarchical cost. While a whipsnake is misclassified as ribbon snake, it is still correct at the "snake" level, thus giving a more useful answer than "sundial". It demonstrates that **classification based on hierarchical cost can be significantly more informative.**

[6] The image can in fact be a "trimaran", in which case it is not entirely correct to predict "catamaran". This is a limitation of intermediate level ground truth labels.

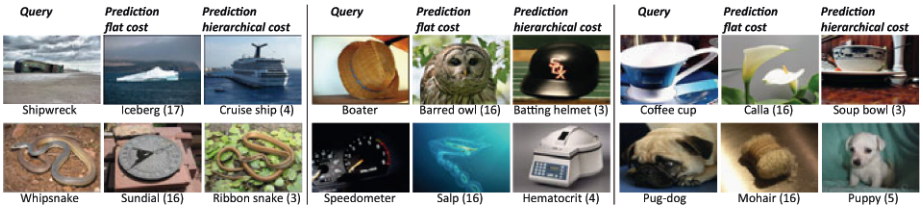

Fig. 7. Example errors using a flat vs. hierarchical classifier with `SPM+SVM` on ImageNet10K, shown in horizontal groups of three: a query, prediction by a flat classifier (minimizing 0-1 loss), and by a hierarchical classifier (minimizing hierarchical cost). Numbers indicate the hierarchical cost of that misclassification.

9 Conclusion

We have presented the first large scale recognition experiments on 10,000+ categories and 9+ million images. We show that challenges arise from the size and density of the semantic space. Surprisingly the ordering of NN and Linear classification approaches swap from previous datasets to our very large scale experiments – we cannot always rely on experiments on small datasets to predict performance at large scale. We produce a measure of category distance based on the WordNet hierarchy and show that it is well correlated with the difficulty of various datasets. We present a hierarchy aware cost function for classification and show that it produces more informative classification results. These experiments point to future research directions for large scale image classification, as well as critical dataset and benchmarking issues for evaluating different algorithms.

Acknowledgments. We thank Chris Baldassano, Jia Li, Olga Russakovsky, Hao Su, Bangpeng Yao and anonymous reviewers for their helpful comments. This work is partially supported by an NSF CAREER grant and a Google Award to L.F-F, and by NSF grant 0849512, Intel Research Council and Gigascale Systems Research Center.

References

1. Biederman, I.: Recognition by components: A theory of human image understanding. PsychR 94, 115–147 (1987)
2. Zhang, H., Berg, A.C., Maire, M., Malik, J.: SVM-KNN: Discriminative Nearest Neighbor Classification for Visual Category Recognition. In: CVPR 2006 (2006)
3. Grauman, K., Darrell, T.: The pyramid match kernel: Discriminative classification with sets of image features. In: ICCV (2005)
4. Lazebnik, S., Schmid, C., Ponce, J.: Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In: CVPR 2006 (2006)
5. Varma, M., Ray, D.: Learning the discriminative power-invariance trade-off. In: ICCV (2007)
6. Torralba, A., Fergus, R., Freeman, W.: 80 million tiny images: A large data set for nonparametric object and scene recognition. PAMI 30, 1958–1970 (2008)

7. Felzenszwalb, P., Mcallester, D., Ramanan, D.: A discriminatively trained, multiscale, deformable part model. In: CVPR 2008 (2008)
8. Tuytelaars, T., Mikolajczyk, K.: Local Invariant Feature Detectors: A Survey. Foundations and Trends in Computer Graphics and Vision 3, 177–820 (2008)
9. Fei-Fei, L., Fergus, R., Torralba, A.: Recognizing and learning object categories. CVPR Short Course (2007)
10. Fei-Fei, L., Fergus, R., Torralba, A.: Recognizing and learning object categories. ICCV Short Course (2009)
11. Rosch, E., Mervis, C.B., Gray, W.D., Johnson, D.M., Braem, P.B.: Basic objects in natural categories. Cognitive Psychology 8, 382–439 (1976)
12. Everingham, M., Zisserman, A., Williams, C.K.I., van Gool, L., et al.: The 2005 pascal visual object classes challenge. In: Quiñonero-Candela, J., Dagan, I., Magnini, B., d'Alché-Buc, F. (eds.) MLCW 2005. LNCS (LNAI), vol. 3944, pp. 117–176. Springer, Heidelberg (2006)
13. Fei-Fei, L., Fergus, R., Perona, P.: One-shot learning of object categories. PAMI 28, 594–611 (2006)
14. Griffin, G., Holub, A., Perona, P.: Caltech-256 object category dataset. Technical Report 7694, Caltech (2007)
15. Fellbaum, C.: WordNet: An Electronic Lexical Database. MIT Press, Cambridge (1998)
16. Dance, C., Willamowski, J., Fan, L., Bray, C., Csurka, G.: Visual categorization with bags of keypoints. In: ECCV International Workshop on Statistical Learning in Computer Vision (2004)
17. Lowe, D.: Distinctive image features from scale-invariant keypoints. IJCV 60, 91–110 (2004)
18. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: CVPR 2005, pp. 886–893 (2005)
19. Chum, O., Zisserman, A.: An exemplar model for learning object classes. In: CVPR 2007, pp. 1–8 (2007)
20. Vedaldi, A., Gulshan, V., Varma, M., Zisserman, A.: Multiple kernels for object detection. In: ICCV (2009)
21. Gehler, P.V., Nowozin, S.: On feature combination for multiclass object classification. In: ICCV (2009)
22. Rifkin, R., Klautau, A.: In defense of one-vs-all classification. JMLR 5, 101–141 (2004)
23. Maji, S., Berg, A.C.: Max-margin additive models for detection. In: ICCV (2009)
24. Deng, J., Dong, W., Socher, R., Li, L., Li, K., Fei-Fei, L.: ImageNet: A large-scale hierarchical image database. In: CVPR 2009 (2009)
25. Fergus, R., Weiss, Y., Torralba, A.: Semi-supervised learning in gigantic image collections. In: NIPS (2009)
26. Wang, C., Yan, S., Zhang, H.J.: Large scale natural image classification by sparsity exploration. ICASP (2009)
27. Nister, D., Stewenius, H.: Scalable recognition with a vocabulary tree. In: CVPR 2006, pp. II: 2161–2168 (2006)
28. Marszalek, M., Schmid, C.: Semantic hierarchies for visual object recognition. In: CVPR 2007, pp. 1–7 (2007)
29. Zweig, A., Weinshall, D.: Exploiting object hierarchy: Combining models from different category levels. In: ICCV 2007, pp. 1–8 (2007)
30. Griffin, G., Perona, P.: Learning and using taxonomies for fast visual categorization. In: CVPR 2008 (2008)

31. Everingham, M., Van Gool, L., Williams, C.K.I., Winn, J., Zisserman, A.: The PASCAL Visual Object Classes Challenge, VOC 2009 Results (2009), http://www.pascal-network.org/challenges/VOC/voc2009/workshop/
32. http://www.image-net.org
33. Oliva, A., Torralba, A.: Modeling the shape of the scene: A holistic representation of the spatial envelope. IJCV 42, 145–175 (2001)
34. Fan, R.E., Chang, K.W., Hsieh, C.J., Wang, X.R., Lin, C.J.: LIBLINEAR: A library for large linear classification. JMLR 9, 1871–1874 (2008)
35. Crammer, K., Singer, Y., Cristianini, N., Shawe-Taylor, J., Williamson, B.: On the algorithmic implementation of multiclass kernel-based vector machines. JMLR 2 (2001)
36. Andoni, A., Indyk, P.: Near-optimal hashing algorithms for approximate nearest neighbor in high dimensions. In: FOCS 2006, pp. 459–468 (2006)
37. Boiman, O., Shechtman, E., Irani, M.: In defense of nearest-neighbor based image classification. In: CVPR 2008 (2008)
38. Thorpe, S., Fize, D., Marlot, C.: Speed of processing in the human visual system. Nature 381, 520–522 (1996)
39. Martinez-Munoz, G., Larios, N., Mortensen, E., Zhang, W., Yamamuro, A., Paasch, R., Payet, N., Lytle, D., Shapiro, L., Todorovic, S., Moldenke, A., Dietterich, T.: Dictionary-free categorization of very similar objects via stacked evidence trees. In: CVPR 2009 (2009)
40. Nilsback, M.E., Zisserman, A.: A visual vocabulary for flower classification. In: CVPR 2006, pp. 1447–1454 (2006)
41. Ferencz, A., Learned-Miller, E.G., Malik, J.: Building a classification cascade for visual identification from one example. In: ICCV 2005, pp. 286–293 (2005)
42. Vedaldi, A., Fulkerson, B.: VLFeat: An open and portable library of computer vision algorithms (2008), http://www.vlfeat.org
43. Lin, H.T., Lin, C.J., Weng, R.C.: A note on platt's probabilistic outputs for support vector machines. Mach. Learn. 68, 267–276 (2007)

A Experimental Details

We obtain BoW histograms (L1-normalized) using dense SIFT [42] on 20x20 overlapping patches with a spacing of 10 pixels at 3 scales on images resized to a max side length of 300, and a 1000 codebook from KMeans on 10 million SIFT vectors. We use the same codewords to obtain spatial pyramid histograms (3 levels), ϕ_2 encoded [23] to approximate the intersection kernel with linear SVMs. Due to high dimensionality (21k), we only encode nonzeros (but add a bias term). This preserves the approximation for our, non-negative, data, but with slightly different regularization. We found no empirical performance difference testing up to 1K categories. To save memory, we use only two bytes for each entry of encoded vectors (sparse) by delta-coding its index (1 byte) and quantizing its value to 256 levels (1 byte). We further reduce memory by only storing every other entry, exploiting redundancy in consecutive entries. We use LIBLINEAR [34] to train linear SVMs, parameter C determined by searching over 3 values (0.01, 0.1, 1 for ImageNet10K) with 2-fold cross validation. We use smaller weight for negative examples(100× smaller for ImageNet10K) than positives. We obtain posterior probability estimates by fitting a sigmoid function to the outputs of SVMs [43], or by taking the percent of neighbors from a class for NN.