Attribute learning in large-scale datasets

Olga Russakovsky and Li Fei-Fei

Stanford University
{olga,feifeili}@cs.stanford.edu

Abstract. We consider the task of learning visual connections between object categories using the ImageNet dataset, which is a large-scale dataset ontology containing more than 15 thousand object classes. We want to discover *visual* relationships between the classes that are currently missing (such as similar colors or shapes or textures). In this work we learn 20 visual attributes and use them in a zero-shot transfer learning experiment as well as to make visual connections between semantically unrelated object categories.

1 Introduction

Computer vision has traditionally focused on object categories: object classification, object segmentation, object retrieval, and so on. Recently, there has been some interest in transitioning from learning visual nouns (whether object categories, such as cars or pedestrians, or object parts, such as "wheel" or "head") to visual adjectives (such as "red" or "striped" or "long") which can be used to describe a wide range of object categories [1–6]. Learning visual attributes has been shown to be beneficial for improving performance of detectors [3] but especially for transferring learned information between object categories. For example, learning the color "red" or the pattern "striped" from a series of training images can then be used to recognize these attributes in a variety of unseen images and object categories [1, 3].

The term "attribute" is defined in Webster's dictionary as "an inherent characteristic" of an object, and various types of attributes have been explored in the literature: appearance adjectives (such as color, texture, shape) [1–5, 7], presence or absence of parts [1, 4, 6] and similarity to known object categories [1, 5, 6]. Attributes have also been broken up into (1) semantic, i.e., those that can be described using language [1, 4, 7], and (2) non-semantic but discriminative [3] or similarity-based [5, 6]. In this paper, we focus on semantic appearance attributes.

Attributes and parts-based models are particularly important when building large-scale systems, where it is infeasible to train an object classifier independently for each object class. Given a sufficiently rich dataset of learned adjectives, new categories of objects can be recognized simply from a verbal description consisting of a list of the attributes [1, 3] or a verbal description in combination with just a few training examples [3].

In this paper, we consider learning multiple visual attributes on ImageNet [9], which is a large-scale ontology of images built upon WordNet [8]. It contains more

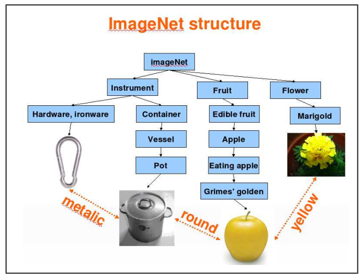


Fig. 1. The goal of our work is to build visual connections between object categories. We focus on the large-scale ImageNet dataset which currently uses WordNet [8] to provide a semantic hierarchy provides a semantic hierarchy of categories. Discovering a visual hierarchy would be useful for a variety of tasks; for example, targeted retrieval.

than 11 million images representing more than 15 thousand concepts. While the dataset already provides useful structure and connections between object classes through the hierarchical semantic ontology of WordNet, we want to learn *visual* relationships or hierarchies between the classes (see Figure 1). We begin by describing the existing connections within the ImageNet dataset in Section 2, and discussing prior work for attribute learning in Section 3. In Section 4 we describe our approach to obtaining ground truth human labeling of attributes. We then learn 20 visual attributes on the ImageNet data and present results on a number of tasks in Section 5. We conclude and discuss future work in Section 6.

2 Learning visual connections in ImageNet

The ImageNet dataset [9] contains representative images for more than 15 thousand image categories, or *synsets* as they are called in WordNet.[1] Recently, bounding box annotations have been released for some of the categories, making it easier to perform object categorization or attribute learning. However, the dataset remains highly challenging, with lots of variety within the synsets, as shown in Figure 2.

Noun hierarchies such as WordNet have been very successfully used in natural language processing. However, the WordNet noun hierarchy is far from visual; for example, human-made objects within ImageNet are organized by their high-level purpose and animals are organized by their evolutionary relation, and as a result the sibling synsets are often very far from each other in appearance (see Figure 2). Evolutionary hierarchies are fundamental in genomics and evolutionary biology, but for computer vision, it would be more useful to be able to derive a hierarchy of (or at least a set of relations between) object categories that's based on visual adjectives or attributes of objects, rather than their evolutionary relation.

Connections based on the visual attribute such as "striped" are missing: striped animals (zebras, raccoons, tigers), striped insects (hairstreak butterfly),

[1] We use the terms "synset" and "object category" interchangeably.

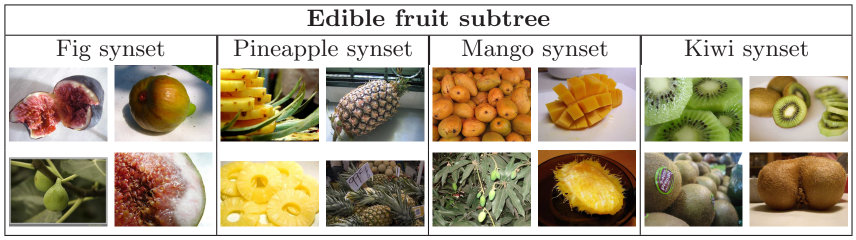


Fig. 2. Example images of synsets that are direct descendants of the edible fruit synset. First, the high variability within each of the four synsets makes classification on this dataset very challenging. Second, the four object classes are sibling synsets in WordNet since they are all children of the "edible fruit" synset; however, visually they are quite different from each other in terms of color, texture and shape.

striped flowers (butterfly orchid, moosewood tree), striped vegetables (cushaw, watermelon), striped fish (black sea bass, lionfish) and inanimate objects such as striped fabric are not related within ImageNet. To the best of our knowledge, previous work on attributes has focused on making connections within a much more narrow set of object categories (such as animals [1, 6], cars [3, 4] or faces [5]). We are interested in discovering visual relations between all categories of ImageNet, from fruits to animals to appliances to fabrics. We show in Section 5.4 that our algorithm indeed manages to do that.

3 Related work

Ferrari and Zisserman [2] proposed learning attributes using segments as the basic building blocks. They distinguish between unary attributes (colors) involving just a single segment and binary attributes (stripes, dots and checkerboards) involving a pattern of alternating segments. Since their method relies on obtaining a near-perfect segmentation of the pattern, in practice it's difficult to apply to challenging natural images – for example, the stripes of a tiger are very difficult to segment out perfectly, and the orange background stripes would often get merged into a single segment, contrary to what their attribute classification algorithm expects.

Yanai and Barnard [7] learned the "visualness" of 150 concepts by performing probabilistic region selection for images labeled as positive and negative examples of a concept, and computing the entropy measure which represents how visual this concept is. They evaluated their algorithm on Google search images, and also considered each image to be a collection of regions obtained from segmentation, but didn't consider the pairwise relationship between the regions.

Recently, Lampert et al. [1] considered the problem of object classification when the test set consists entirely of previously unseen object categories, and the transfer of information from the training to the test phase occurs entirely through attribute text labels. They introduced the Animal with Attributes dataset with 30,000 images annotated with 50 classes. They are interested in performing zero-shot object classification (where the object classes in the training and test sets are disjoint) based on attribute transfer rather than learning the attributes themselves or building an attribute hierarchy. Interestingly, some of their attributes are not even fundamentally "visual" (for example, "strong" or "nocturnal"), but were nevertheless found to be useful for classification [1]. One interesting thing to point out in relation to our work is that ImageNet already has subtrees for some of their adjectives, such as edible, living, predator/prey/scavenger, young, domestic, male/female, even insectivore/omnivore/herbivore. While many of their other attributes are animal-specific, such as "has paws," and thus not as useful in our setting for making connections between a broad range of object categories, we were inspired by their list in creating our own.

Farhadi et al. [3] worked on describing objects by parts, such as "has head," or appearance adjectives, such as "spotty." They wanted to both describe unfamiliar objects (such as "hairy and four-legged") and learn new categories with few visual examples. They distinguished between two types of attributes: semantic ("spotty") and discriminative (dogs have it but cat don't). Similarly, Kumar et al. [5] considered two types of attributes for face recognition: those trained to recognize specific aspects of visual appearance, such as gender or race, and "simile" classifiers which represent the similarity of faces to celebrity faces. We focus on semantic attributes in the current work, but argue that ultimately discriminative and comparative attributes are necessary because language is insufficient to precisely describe, e.g., the typical shape of a car or the texture of a fish.

Rohrbach et al. [6] use semantic relationships mined from language to achieve *unsupervised* knowledge transfer. They found that path length in WordNet is a poor indicator of attribute association (for example, the "tusk" synset is very far from the "elephant" synset in the hierarchy, making it impossible to infer that elephants would have tusks). They show that web search for part-whole relationships is a better way of mining attribute annotations for object categories. In our work, we also explore using WordNet to mine attribute associations, but consider using the WordNet synset definitions rather than path length.

Most recently, Farhadi et al. [4] discussed creating the right level of abstraction for knowledge transfer. They learned part and category detectors of objects, and described objects by spacial arrangement of their attributes and the interaction between them. They focused on finding animal and vehicle categories not seen during training, and inferring attributes such as function and pose. They learn both the parts that are visible and not visible in each image.

4 Building and labeling an attribute dataset

In order to learn and evaluate attribute labels, we first need to obtain ground truth annotations of the images. [6] discusses various data mining strategies; however, it focuses on parts-based attributes, mining for relations such as "leg is a part of dog" or "dog's leg." WordNet provides a definition for every synset it contains; since we are instead interested in appearance-based attributes, we considered two strategies: mining these definitions directly (which is different than the path length discussed in [6]), and manual labeling (which was the approach of [1, 4]).

WordNet synset definitions are not well-suited for mining visual adjectives for several reasons. First, the mined adjectives don't necessarily correspond to visual characteristics of the full object and require understanding of the object parts (e.g., animals with a "striped tail"). Second, the mined adjectives often need to be understood in the context of other adjectives in the definition (e.g., a flower described as "yellow or red or blue"). Also, sometimes the adjectives are extremely difficult to detect visually (e.g., a flag is defined as "rectangular" but usually doesn't look rectangular in the image). However, since ImageNet is a very large-scale dataset, mining for attributes in this very simple way can help restrict attention to just a subset of the ImageNet data which is likely to contain a sufficient amount of positive examples for each attribute. To construct the dataset of 384 synsets that we use for our experiments, for every attribute we searched for all synsets (from among those with available bounding box annotations) which contained this attribute in either the synset name or the synset definition, and included that synset along with all of its siblings in the training set. The motivation for including the siblings was to provide a rich enough set of negative examples that are likely to differ from the positive synsets in only a few characteristics, and specifically in the characteristic corresponding to the mined attribute. For example, if a zebra is characterized as a "striped" equine, it's reasonable to infer that other equines, such as horses, are not striped.

In order to obtain the ground truth data we use workers on Amazon Mechanical Turk (AMT) to label 25 images randomly chosen from each synset. We present each worker with 106 images (25 each from 4 different synsets plus 6 randomly injected quality control images) and one attribute, and ask to make a binary decision of whether or not this attribute applies to the image. For color attributes (black, blue, brown, gray, green, orange, pink, red, violet, white and yellow), we ask whether a significant part of the object (at least 25%) is that color. For all other attributes (furry, long, metallic, rectangular, rough, round, shiny, smooth, spotted, square, striped, vegetation, wet, wooden), we ask if they would describe the object *as a whole* using that attribute.

Each image is labeled by 3 workers, and we consider an image to be positive (negative) if all workers agree that it's positive (negative); otherwise, we consider it ambiguous and don't include it in our training sets. Unfortunately, for 5 of our attributes (blue, violet, pink, square and vegetation) we did not get sufficient positive training data (at least 75 images) to include them in our experiments.

We analyze the overlap between the mined synsets and the human labeling in Table 1. We consider a synset to be labeled positive for an attribute by AMT workers if more than half of its labeled images are unanimously labeled as positive. Interestingly, some obvious annotations such as "green salad" or "striped zebra" were not present in the human labels. This shows that data obtained from AMT can be extremely noisy, and that better quality control and/or more annotators are needed. Currently we are only considering an image to be a positive or negative example if it is labeled unambiguously; while this gives us good precision in our training set, the recall is much lower than we would like, and thus the number of training examples for each attribute is low despite the large dataset size. Overall, we have 384 synsets $\times$ 25 images per synset = 9600 images labeled with 20 attributes, with 4% of all labels being positive, 68% negative, and 28% ambiguous.

Attr.	WN	Both	AMT
Green	salad, sukiyaki, absinthe	green lizard, grass	sunflower, bonsai
Rectang.	flag, sheet, towel	box	bench, blackboard, cabinet
Round	feline, pita, shortcake	ball, button, pot	basketball, drum, Ferris wheel
Spotted	cheetah, giraffe, pinto	jaguar	garden spider, strawberry, echidna
Striped	aardwolf, zebra		garden spider, skunk, basketball
Wooden	cross, popsicle	marimba	cabinet, pool table, ski
Yellow	grizzly, yolk, honey	sunflower	margarine

Table 1. Examples of synsets labeled positive by mining WordNet definitions ("WN"), by both WordNet and AMT labelers ("Both"), and just by AMT labelers ("AMT").

5 Experiments

We have described the procedure for obtaining 384 imageNet synsets, all of which have bounding box annotations released, with 25 images within each labeled as positive, negative or ambiguous for each of 20 attributes. In this section we show classification and retrieval performance of attribute classifiers trained using this data, as well as apply these classifiers to a simple transfer learning task following the framework of Lampert et al. [1]. Finally, we show the visual links that were discovered between distant ImageNet synsets.

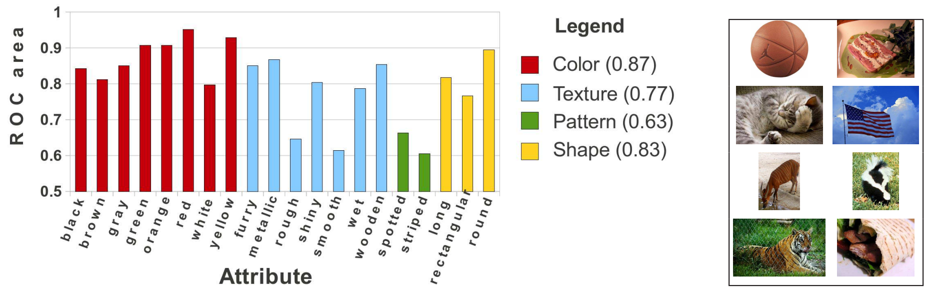
5.1 Implementation

We represent each image using three types of normalized histogram features: (1) color histogram of of quantized RGB pixels using a codebook of size 50, (2) texture histogram of quantized SIFT descriptors at multiple levels using a codebook of size 1000 [10, 11], and (3) shape histogram of quantized shape-context features [12] with edges computed using the *Pb* edge detector [13, 14] using a codebook of size 500. Each of the three feature histograms was normalized independently to have L1 unit length. We use an SVM with a histogram intersection kernel [15, 16], which in our experiments significantly outperforms both the linear and RBF kernels. We use a holdout set to determine the regularization.

5.2 Learning image attributes

First, we train the classifiers to recognize each attribute individually and evaluate the generalization performance. All images in our training set are labeled by 3 AMT workers, and we consider an image to be a positive (negative) example of an attribute if all subjects agree that this is a positive (negative) example. We use 5-fold cross-validation, making sure that no synset appears in multiple folds.

Some classifiers, such as those corresponding to the color attributes, generalize quite well in this setting. We point out the two main challenges we face when



(a) (b)

Fig. 3. (a) Performance of attribute classifiers (as measured by the area under the ROC curve) sorted by attribute type. The average performance of each type is reported in parentheses. (b) Some example images labeled by the human subjects as "striped." This shows the difficultly of learning a good "striped" classifier on this dataset.

training the attribute classifiers. First, the "pattern" classifiers corresponding to "striped" and "spotted" attributes perform poorly as a result of the great *variety* of the exemplars (see Figure 3(b) for examples of "striped" images). There is a lack of training data especially in light of this variety (only 99 images were labeled as "striped" and 146 as "spotted"). As the number of object categories increases, so does the variety of appearances of certain attributes, and thus the amount of training data collected should be sufficient to account for this.

Second, the two texture attributes "rough" and "smooth" suffer from *ambiguity* as evidenced by the lack of labeling consensus. The labelers unanimously agreed on only 66% of the images in the dataset when labeling with the "smooth" attribute, and 72% when labeling with the "rough" attribute. In contrast, for every other attribute the annotators unanimously agreed on more than 79% of the images. As a result, whether an image was labeled as a positive or negative training example for "rough" or "smooth" was largely dependent on the specific set of labelers assigned to it. Such attributes require further refinement and/or better definitions during the labeling process.

In Figures 4-6 we show qualitative retrieval results using the trained classifiers. Note that many of the top correctly retrieved images were not used in the quantitative evaluation because they were not unanimously labeled by the labelers. This further reinforces the need for more rigorous labeling procedures.

5.3 Transfer learning using attributes

We use the learned classifiers in a small-scale transfer learning experiment following the Direct attribute prediction (DAP) model of Lampert et al. [1]. Briefly, we are given L test classes $z_{1,\dots,L}$ not seen during training, and M attributes, where the test classes are annotated with binary labels a^l_m for each class l and attribute m. In our experiments we consider $L = 5$ test classes: chestnut, green lizard, honey badger, zebra, and spitz, and $M = 20$ attributes described above.

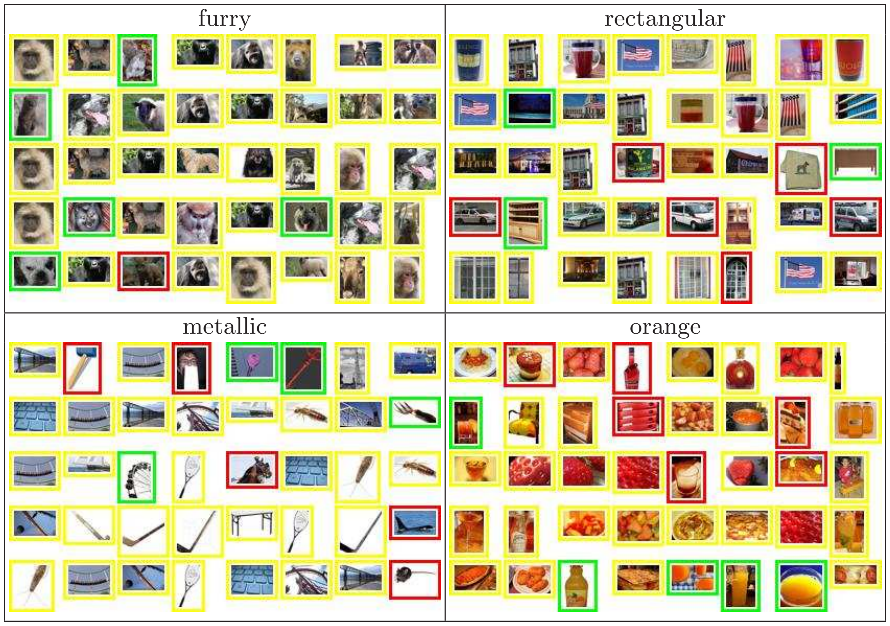


Fig. 4. Visualization of four of the learned attributes (for the other attributes, see Figures 5 and 6). For each attribute, the 5 rows represent the 5 training folds, and each row shows the top 8 images retrieved from among all synsets that didn't appear in that fold's training set. The border around each image corresponds to the human labeler annotation (green is positive, red is negative, yellow is ambiguous).

The synset-level annotations come from AMT human labelers.[2] We use 25 images per object class as above, Given an image x, the DAP model defines the probability of this image belonging to class z as

$$p(z|x) = \sum_{a \in \{0,1\}^M} p(z|a)p(a|x) = \frac{p(z)}{p(a^z)} \prod_{m=1}^{M} p(a^z_m|x)$$

where $p(a^z_m|x)$ is given by the learned attribute model, $p(z)$ is assumed to be a uniform class prior, and $p(a^z)$ is the prior on seeing an example with the same set of attributes as the ground truth for the target class z, computed from training data assuming a factorial distribution over attributes. Image x is assigned to class $c(x)$ using:

$$c(x) = \arg\max_{l=1,\ldots,L} \prod_{m=1}^{M} \frac{p(a^{z_l}_m|x)}{p(a^{z_l}_m)}$$

[2] Out of 100 class-attribute labels, 18 were ambiguous, meaning that less than half the images within that class were unanimously annotated as either positive or negative for that attribute by all 3 workers. We manually disambiguated the annotations.

Fig. 5. Continuation of Figure 4 visualizing the learned attributes.

Fig. 6. Continuation of Figures 4 and 5 visualizing the learned attributes.

We apply this model to our learned classifiers and report our result in Table 2. The main source of errors is the zebra class, which relies on the poorly generalizing "striped" attribute (see results in Figure 3).

						
chestnut: brown,smooth		**0.52**	0.16	0.12	0.12	0.08
green lizard: green, long		0	**0.84**	0	0.12	0.04
honey badger: black, gray, rough, furry		0.32	0	**0.60**	0.04	0.04
zebra: black, white, striped, smooth		0.36	0.08	0.40	**0.08**	0.08
spitz: white, furry		0.08	0	0.36	0.08	**0.48**

Table 2. On the left are the animal classes and the corresponding human attribute annotations, and on the right is the confusion table from the transfer learning experiments. The rows of the confusion table are the ground truth labels and the columns are the classifier outputs.

5.4 Synset-level connections

Given the attribute classifiers we can now consider making synset-level connections within ImageNet, which was the main objective of our work. For each attribute, we have 5 learned classifiers, one for each of the 5 folds. We fit a sigmoid to the output of each classifier to obtain normalized probabilities [15, 17]. We run each classifier on all images that were not part of its training set synsets. For each test synset, we compute the median confidence score of the classifier on images within that synset. Figures 7 and 8 show the top returned synsets.

There are various interesting observations that could be made about the retrieved synsets. "Green," "white" and "round" classifiers discover connections between synsets which are very far apart in the WordNet hierarchy – for example, salad, which is a node 6 levels deep under the "food, nutrient" subtree of ImageNet, green lizard, which is 13 levels deep under the "animal" subtree, and bonsai, which is 9 levels deep under the "tree" subtree. Similarly, the "white" classifier connects various breeds of dogs as well as Persian cats, sails, and sheets. The round classifier connects, e.g., basketball, ramekin, which is "a cheese dish made with egg and bread crumbs that is baked and served in individual fireproof dishes" [8], and egg yolk.

More interesting is to look at attributes such as "long," which are more contextual and relative, and see the kinds of synsets that were learned. It is not immediately clear that the classifier is picking up on the synsets that human would classify as "long," although bottles and forks definitely are.

Finally, "striped" and "wet" discovered some interesting connections – even though it is extremely difficult to learn the high variability of stripes in natural

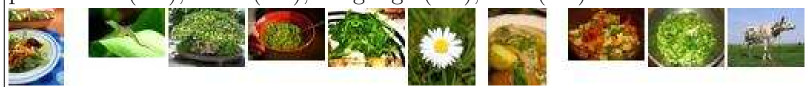
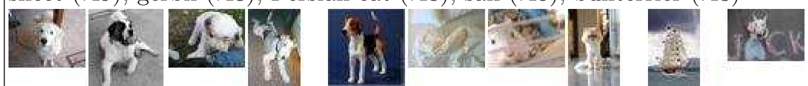
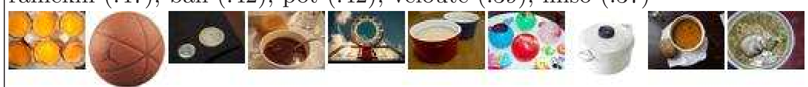
green	salad (.84), green lizard (.73), bonsai (.52), pesto (.43), saute (.37), daisy (.30) pot-au-feu (.12), salsa (.12), roughage (.11), cow (.11) 
white	kuvasz (.70), Saint Bernard (.67), clumber (.65), wirehair (.62), foxhound (.60) sheet (.49), gerbil (.48), Persian cat (.48), sail (.45), bullterrier (.43) 
round	egg yolk (.75), basketball (.68), button (.63), goulash (.56), basket (.49), ramekin (.47), ball (.42), pot (.42), veloute (.39), miso (.37) 
long	kirsch (.83), sail (.77), rorqual (.74), police van (.72), fork (.69), rack (.67), killer whale (.58), window (.54), transporter (.50), pool table (.49) 
striped	barn spider (.36), daisy (.17), zebra (.17), echidna (.16), backboard (.13), drum (.12), coloring (.12), roller coaster (.12), bridge (.11), colobus (.11) 
wet	rorqual (.59), sidecar (.55), orangeade (.53), flan (.52), screwdriver (.47), killer whale (.44), bowhead (.43), maraschino (.41), dugong (.40), porpoise (.40) 

Fig. 7. This figure shows the top 10 synsets that were returned by the algorithm as the most representative for a subset of the attributes (see Figure 8 for the remainder). The number in parenthesis represents the median probability assigned to images within that synset by the attribute classifier.

scenes, zebras and echidnas were retrieved, as well as "garden spiders," which actually often do look striped upon inspection even though it is not a common example that humans would think of as a striped insect. The "wet" classifier especially was able to pick up on some very promising connections: besides just learning that the ocean tends to be wet and thus marine animals are likely wet, it also made the connection to cocktail drinks such as sidecar and screwdriver.

6 Conclusion

In this work we began building a set of visual connections between object categories on a large scale dataset. Our ultimate goal is to automatically discover a large variety of visual connections between thousands of object categories.

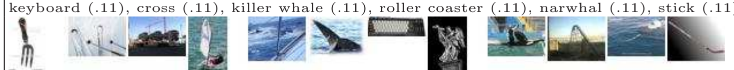
black	colobus (.78), siamang (.75), guereza (.73), groenendael (.71), binturong (.69), chimpanzee (.66), schipperke (.66), silverback (.63), aye-aye (.54), gorilla (.54), skunk (.53), bowhead (.50) 
brown	puku (.82), lechwe (.73), kob (.73), steenbok (.66), sassaby (.65), redbone (.62), bushbuck (.60), ragout (.59), dhole (.57), chestnut (.56), bovid (.54), sambar (.54) 
furry	keeshond (.94), chacma (.93), macaque (.90), grivet (.90), grizzly (.88), gorilla (.88), baboon (.88), mandrill (.88), koala (.86), simian (.86), guenon (.85), kit fox (.85) 
gray	koala (.42), abrocome (.39), gorilla (.38), grivet (.33), keeshond (.29), manul (.29), schnauzer (.29), chacma (.29), viscacha (.28), vervet (.28), hominid (.27), otter (.26) 
metallic	fork (.72), transporter (.56), roller coaster (.49), stick (.41), wheel (.38), police van (.37), keyboard (.34), sail (.31), bridge (.31), building (.28), ski (.25), bowhead (.25) 
orange	orangeade (.73), egg yolk (.58), sunflower (.44), strawberry (.43), fork (.42), maraschino (.42), casserole (.39), screwdriver (.37), pizza (.35), croquette (.30), vermouth (.30), moussaka (.29) 
rectangular	police van (.90), transporter (.84), cabinet (.61), marimba (.50), window (.44), varietal (.42), flag (.38), bridge (.38), kummel (.31), pot (.29), generic (.28), pool table (.26) 
red	shortcake (.70), basketball (.67), catsup (.55), teriyaki (.43), salad (.42), pizza (.37), chili (.30), flan (.26), ragout (.23), slumgullion (.22), bordelaise (.20), police van (.18) 
rough	fork (.11), ski (.11), transporter (.11), sail (.11), rorqual (.11), bowhead (.11), keyboard (.11), cross (.11), killer whale (.11), roller coaster (.11), narwhal (.11), stick (.11) 
shiny	rorqual (.95), bowhead (.82), killer whale (.61), dugong (.54), narwhal (.52), manatee (.44), porpoise (.31), police van (.27), kirsch (.27), flag (.21), stick (.21), ski (.20) 
smooth	sail (.65), kirsch (.64), varietal (.63), champagne (.62), generic (.61), green lizard (.58), bottle (.56), egg yolk (.55), window (.55), mallet (.54), pool table (.53), tower (.53) 
spotted	barn spider (.37), zebra (.26), Ferris wheel (.24), cheetah (.19), insectivore (.16), badger (.15), carnivore (.15), grass (.15), kudu (.14), groundhog (.13), pesto (.12), dik-dik (.12) 
wooden	fork (.75), rack (.66), bridge (.54), police van (.52), pool table (.46), table (.43), kirsch (.42), marimba (.40), squash racket (.36), transporter (.35), cue (.35), slivovitz (.27) 
yellow	egg yolk (1.00), sunflower (.86), omelet (.70), kedgeree (.64), flan (.61), tostada (.48), succotash (.42), pizza (.35), zabaglione (.26), ravigote (.25), curry (.23), casserole (.21) 

Fig. 8. Continuation of Figure 7 showing the visual connections made between synsets.

Discovering semantic attributes can aid in more intelligent image retrieval: for example, the user can specify exactly what he's looking for using a known dictionary of attributes instead of visual training examples. More interestingly, clustering the attributes into categories, such as shape, texture, color, and so on, and working with non-semantic attributes, can potentially lead to at least two major advantages. First, this can allow for new ways of object classification training: instead of showing the algorithm a large variety of cars during training, one can simply inject a bit of prior knowledge that cars can come in all colors but shape is the important characteristic. Second, in retrieval, instead of asking to find an image closest to the query, the user can instead specify that he's looking for something that's close in color to the query image, but round.

Acknowledgements We give warm thanks to Alex Berg, Jia Deng, Bangpeng Yao and Li-Jia Li for their help with this work. Olga Russakovsky was supported by the NSF graduate fellowship.

References

1. Lampert, C., Nickisch, H., Harmeling, S.: Learning to detect unseen object classes by between-class attribute transfer. CVPR (2009)
2. Ferrari, V., Zisserman, A.: Learning Visual Attributes. In: NIPS. (2007)
3. Farhadi, A., Endres, I., Hoiem, D., Forsyth, D.: Describing objects by their attributes. In: CVPR. (2009)
4. Farhadi, A., Endres, I., Hoiem, D.: Attribute-Centric Recognition for Cross-category Generalization. In: CVPR. (2010)
5. Kumar, N., Berg, A.C., Belhumeur, P.N., Nayar, S.K.: Attribute and Simile Classifiers for Face Verification. In: ICCV. (2009)
6. Rohrbach, M., Stark, M., Szarvas, G., Gurevych, I., Schiele, B.: What Helps Where – And Why? Semantic Relatedness for Knowledge Transfer. CVPR (2010)
7. Yanai, K., Barnard, K.: Image Region Entropy: A Measure of Visualness of Web Images Associated with One Concept. In: ACM Multimedia. (2005)
8. Fellbaum, C.: WordNet:An Electronic Lexical Database. Bradford Books (1998)
9. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fe, L.: ImageNet: A Large-Scale Hierarchical Image Database. CVPR (2009)
10. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. IJCV (2004)
11. Berg, A., Deng, J., Fei-Fei, L.: ImageNet Large Scale Visual Recognition Challenge development kit (2010)
12. Belongie, S., Malik, J., Puzicha, J.: Shape matching and object recognition using shape contexts. PAMI (2002)
13. Martin, D., Fowlkes, C., , Malik, J.: Learning to detect natural image boundaries using brightness and texture. NIPS (2002)
14. Catanzaro, B., Su, B.Y., Sundaram, N., Lee, Y., Murphy, M., Keutzer, K.: Efficient, high-quality image contour detection. ICCV (2009)
15. Chang, C.C., Lin, C.J.: LIBSVM: a library for support vector machines (2001)
16. Swain, M.J., Ballard, D.H.: Color indexing. IJCV (1991)
17. Platt, J.: Probabilistic outputs for support vector machines and comparison to regularized likelihood methods. Advances in Large Margin Classifiers (2000)