Non-expert Labels Improve Fine-Grained Object Recognition

Joseph Marino

California Institute of Technology

Abstract. Fine-grained object recognition typically requires domain experts to provide class label annotations, making these labels difficult to collect when experts are rare. Often, fine-grained classes can be organized into a visual class taxonomy according to shared visual features. In this case, non-experts can easily distinguish between coarse classes, allowing them to annotate examples at a coarse level. We show that by supplementing a small set of expert labeled examples with a larger set of non-expert labeled examples, one can significantly boost performance. We also investigate methods of training using a visual taxonomy, show the effectiveness of taxonomic learning in the context of self-learning, and demonstrate methods of analyzing the model's performance with regard to the taxonomy.

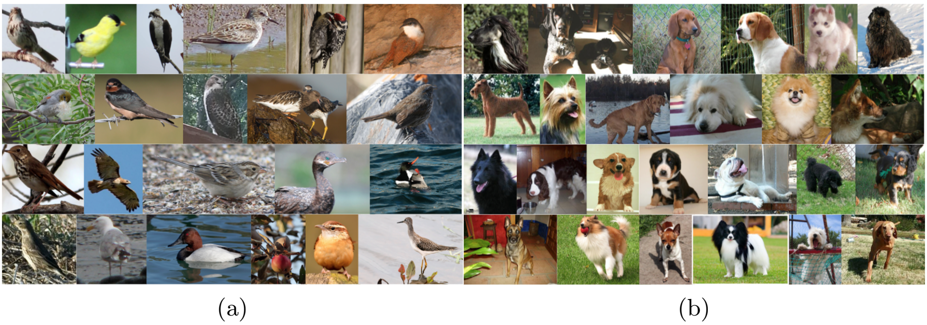
Keywords: Fine-Grained Object Recognition, Class Taxonomy, Multi-task Learning, Transfer Learning, Curriculum Learning

1 Introduction

The successes of deep convolutional neural networks have led to their widespread adoption in the area of object recognition. With record-breaking performance on large object recognition datasets, such as ImageNet, there has been a push to develop more difficult object recognition tasks: more object categories, with more similarity between categories [1]. A number of fine-grained object recognition datasets have been introduced, in both natural [2–5] and non-natural [6–8] object domains.

Collecting a fine-grained object dataset presents a unique challenge. Data examples, while often still plentiful, are typically too difficult for Amazon Mechanical Turk annotators to accurately label. Untrained annotators are unfamiliar with class names and lack the expertise to distinguish between fine-grained classes. Likewise, training annotators for a task of significant size, with many fine-grained classes, can be prohibitively time-consuming and costly. One is then left to rely on individuals with domain knowledge, such as field experts and enthusiasts, to provide class label annotations [2]. This is not only costly, but is in many cases infeasible due to the relatively small number of such individuals. As the object recognition task becomes more fine-grained, collecting an accurately labeled dataset becomes increasingly difficult. The result is a smaller dataset, thereby limiting the performance of a model that is trained to perform the desired task.

Fine-grained object classes, by definition, share many visual features: people have torsos, limbs, and faces; buildings have walls, windows, and doors; and cars have wheels, windshields, and lights. These shared visual features allow us to group the fine-grained classes into coarse clusters. Particularly in biological domains, which are the focus of many fine-grained object recognition tasks, these coarse clusterings will tend to form a visual taxonomy of classes, for the most part mirroring any underlying phylogenetic taxonomy. For instance, all birds share certain visual features, all birds of prey share a more specific set of visual features, and all hawks share an even more specific set of visual features.



(a) (b)

Fig. 1: Fine-grained object recognition attempts to distinguish between highly visually similar object classes in a particular domain. Example domains include (a) birds and (b) dogs. Labeling these examples requires extensive domain knowledge, making it difficult to collect large annotated datasets for many domains.

By constructing a visual taxonomy, one can transform the fine-grained recognition task into a series of coarse- to fine-grained recognition tasks, ranging in difficulty. Coarse splittings near the root of the taxonomy can be easily distinguished by non-experts, whereas correctly distinguishing between branches near the leaf level requires extensive domain knowledge. Depending on an annotator's level of expertise, he or she can now provide a class label annotation at whichever level of the taxonomy that he or she feels most confident making a classification. While not as informative as fine-grained labels, non-expert labels nevertheless contain visually relevant information and in some cases, may be far easier to collect. In this way, one can collect a fine-grained object recognition dataset in which non-experts provide labels for a majority of the examples, with experts labeling only a small subset. By training on both expert and non-expert labels, one can substantially boost performance over using expert labels alone.

The main contributions of this work are as follows:

1. We demonstrate the performance gains of using non-expert labeled examples with a small set of expert labeled examples.
2. We compare two methods of taxonomic training: taxonomic curriculum learning and multi-task learning.
3. We show the effectiveness of taxonomic classifiers in the context of self-learning in the case of very weakly labeled data.

We study the use of non-expert labels on the tasks of fine-grained recognition of North American bird species and dog breeds. Assuming a visual taxonomy of the classes exists, our overall method is independent of the specific fine-grained task and should generalize to other fine-grained biological domains and many non-biological domains as well.

2 Related Work

2.1 Taxonomic Classifiers

There has been considerable work in incorporating class taxonomies into multi-class classification. This work started in the area of document classification, where taxonomic versions of learning algorithms and loss functions were combined with document taxonomies to aid in classification [9, 10]. The taxonomic classifier review in [11] categorizes these approaches as either local or global. Local approaches learn a classifier for each node or level of the taxonomy, whereas global approaches build a classifier that operates across the entire taxonomy. The method presented here is a local approach, though the technique of using non-expert data is independent of the approach taken.

Recently, taxonomies have been applied to object recognition. Griffin, *et al.* [12] introduced a method for learning taxonomies of object classes. When faced with an image of an object for which the classifier is uncertain, the approach taken in [13] trades off specificity for accuracy using a class taxonomy. Sfar, *et al.* [14] present a method for using a taxonomy to return a set of fine-grained classes that contain the correct class with high probability. In addition to learning taxonomies, the approach in [15] uses a class taxonomy to impose a prior on weights as a means for learning from few examples. Wang, *et al.* [16] take a similar approach to our method, learning a classifier per level of the taxonomy. Our method compliments theirs, as we incorporate further data at the coarse-grained levels of the taxonomy. However, we note the importance of using a visual taxonomy over a semantic or phylogenetic taxonomy.

2.2 Transfer Learning

When multiple tasks share some level of representation, the tasks can assist each other through transfer learning. Since many natural images share low level features with other natural images, transfer learning can be particularly helpful in object recognition [17]. For instance, one can boost object recognition performance substantially by training a network on ImageNet and fine-tuning the network for a related task with a smaller dataset [18]. This technique of using pre-trained models has facilitated dramatic performance improvements on benchmark datasets that would otherwise be far too small to train an entire deep convolutional network. During transfer learning, if the tasks are trained concurrently, it is referred to as multi-task learning. As was pointed out in [19], given a taxonomy of class labels, one can perform multi-task learning, with each task being a classification at a different level in the taxonomy. We explore this training approach in Sections 3 and 4.

2.3 Curriculum Learning

In curriculum learning, a model is trained on progressively more difficult examples, defined by some metric, as a non-convex optimization technique [20, 21]. We take inspiration from this technique, and formulate a taxonomic version of curriculum learning. We train on progressively deeper levels of the taxonomy. Instead of training first on easy examples, the model first trains on easy tasks, separating coarse classes of objects. The motivation is that, perhaps by ordering training in this sequential manner, a better final solution will be found.

2.4 Self-Learning

When faced with many unlabeled examples, one approach is to train a model on a set of labeled examples, use the model to label the unlabeled examples, combine the datasets, and train on the entire dataset. This form of semi-supervised learning is referred to as self-learning [22]. In fine-grained object recognition, where labels are relatively difficult to collect, harnessing unlabeled or weakly-labeled examples from, for instance, Flickr, can have large potential gains. This is the approach taken in [23].

3 Method

Fine-grained object classes, especially those found in the natural world, can often be characterized by a visual taxonomic tree structure. The examples that we use in the following sections are dogs and North American birds, but the overall method is independent of the particular taxonomy. That is, we only assume that a measure of visual similarity exists in the domain and that some fine-grained classes are more visually similar than others. Our objectives are to (1) construct a visual taxonomy and (2) train a model such that it can handle data and make predictions at any level of the taxonomy.

3.1 Constructing a Visual Taxonomy

While methods exist for learning class taxonomies directly from data examples [12, 15], we have taken the approach of constructing visual taxonomies manually. Although more labor intensive, this approach guarantees that internal classes in the taxonomy are, at some level, visually–and often semantically–interpretable to an annotator who is unfamiliar with the domain. This is a key consideration, because in order to provide useful annotations, non-experts need to be able to interpret and classify coarse-level classes. In contrast, a learned taxonomy is not guaranteed to be visually interpretable to a non-expert, potentially making annotation more difficult. The outputs of a model trained with a non-interpretable taxonomy will also be similarly difficult to interpret. Additionally, in biological domains, phylogenetic taxonomies provide a convenient starting point for manual taxonomy construction.

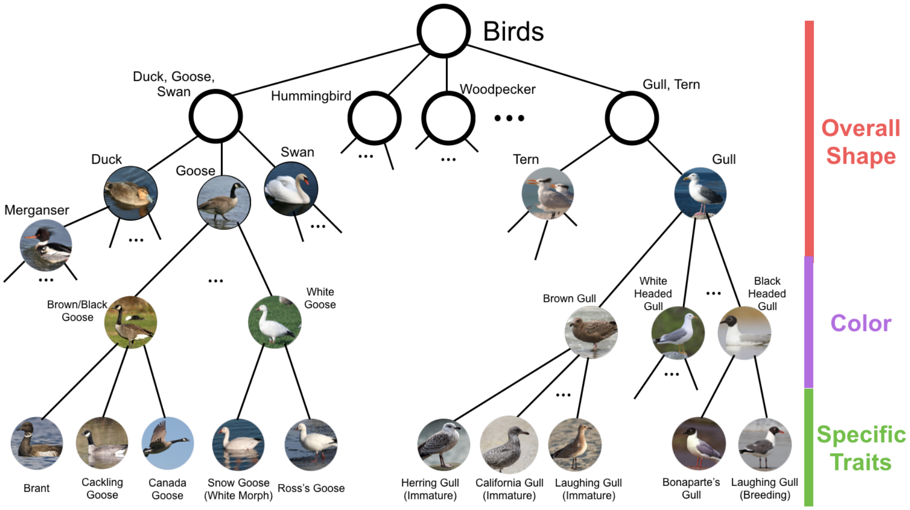


Fig. 2: Visual taxonomy of North American birds. Fine-grained classes are grouped first by overall shape and form, then by color, and finally by specific traits, such as patterns. The visual taxonomy roughly follows the phylogenetic taxonomy of birds. The purpose of the taxonomy is to aid non-expert annotators in labeling examples.

Given a phylogenetic taxonomy or even just a set of fine-grained classes, how does one construct a visual taxonomy? Our approach consists of grouping object classes first by overall shape and form, then color, and finally by specific traits. This is shown in Figure 2. Though open to interpretation, for the dog and bird taxonomies that we tested, this method resulted in coarse-level classes that we believe roughly mirror how non-experts tend to classify these domains. For instance, in classifying an unfamiliar bird, a non-expert might first describe the overall type (duck, hummingbird, owl, etc.), then the color, and finally any other peculiarities of the bird (patterns, eye color, bill shape, etc.). We make no claims as to whether a taxonomy constructed in this manner is optimal, or even desirable, for training a model. For our purposes, the objective of constructing the taxonomy is to provide a means for non-experts to easily distinguish between groupings of fine-grained classes, allowing them to contribute their knowledge toward the overall task.

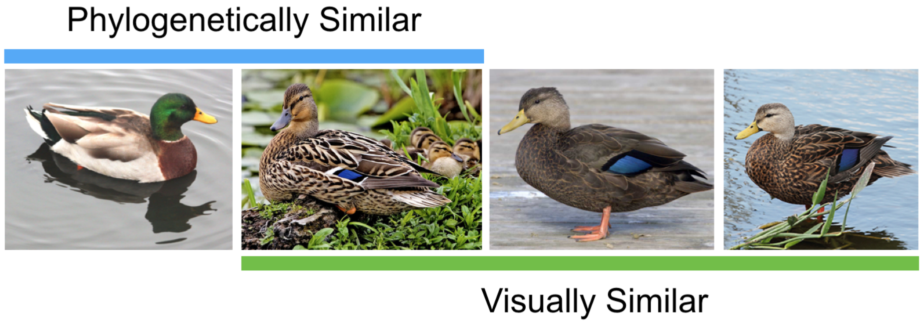


Fig. 3: Phylogenetically similar classes are typically also visually similar. Certain cases go against this rule, such as Mallard ducks. Female Mallard ducks are much more visually similar to American Black ducks and Mottled ducks than male Mallard ducks.

We offer three final thoughts for taxonomy construction. First, coarse-grained classes should only be formed up to the point at which they are helpful to the non-expert annotator. Grouping classes smaller and smaller until the root node does not necessarily aid a non-expert. In addition, having coarse groups that do not share a high degree of visual similarity may be detrimental to the training process. The second point is that phylogenetic taxonomies should be followed loosely. For instance, female Mallard ducks, although phylogenetically identical to male Mallard ducks, are much more visually similar to American black ducks and Mottled ducks (see Figure 3). It is often helpful to look at the common names, as humans tend to give visually similar classes similar names. The House sparrow, for example, is phylogenetically quite distant from all other sparrows, however they have a similar appearance. The final point is that we are not constrained to tree taxonomic structures. We prefer trees for the fact that fine-grained labels automatically provide coarse-grained labels. They also represent biological domains well.

3.2 Training

By incorporating a class taxonomy, our classifier is no longer constrained to predict and train with labels at only the fine-grained level. A variety of classifier modifications can enable these capabilities. For instance, one can learn individual classifiers for each internal node in the taxonomy, or learn a classifier over all classes in the taxonomy by defining some loss function using a taxonomic distance measure between classes [11]. We have taken the approach of making distinct cuts progressively through the taxonomy, which we refer to as levels. Some classes may appear in multiple levels, while other classes may not appear in any. We then learn a classifier for each level. In the context of convolutional networks, this has two main advantages: it supports batch processing and it affords us the flexibility to separate out the different classifiers, allowing us to learn different features that are relevant for classification at each level of the taxonomy. With the levels of the taxonomy defining a set of similar, yet distinct learning tasks, we can employ some form of transfer learning between them. This is the power behind this approach; we can make up for a lack of data on the fine-grained task by increasing the amount of data on the coarse-grained tasks.

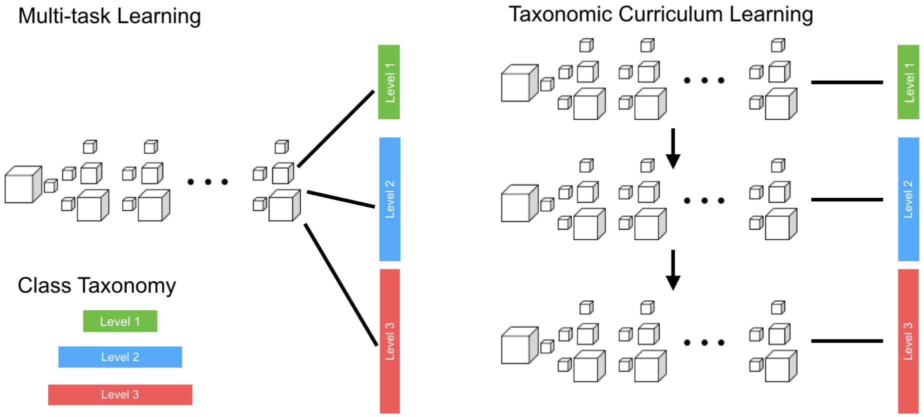


Fig. 4: A class taxonomy comprised of three levels is shown on the bottom left. All levels of the taxonomy are learned simultaneously during multi-task learning. During taxonomic curriculum learning, the levels of the taxonomy are learned sequentially. The model is trained on the first level of the taxonomy until convergence, then trained on the second level and so on until the final level of the taxonomy.

We are still left to decide how to train a model using this taxonomically-defined set of tasks. The two approaches that we decided to investigate are multi-task learning and what we refer to as taxonomic curriculum learning. In multi-task learning, the network is trained on all of the tasks simultaneously. We achieve this by adding multiple output layers to the network, one for each task, so that, given an input example, the network produces a prediction at each level of the taxonomy. During training, if an example does not have a label at a particular level in the taxonomy, then there is no gradient contribution from the corresponding output. In taxonomic curriculum learning, we train on each level in sequential order: we train the network on the first level of the taxonomy until convergence, then replace the output with a classifier for the second level and repeat until we have reached the final, fine-grained level of the taxonomy. To produce predictions throughout the taxonomy, one can use the converged model from each level. Both of these transfer learning methods require certain decisions about hyper-parameters and network structure. For example, each task may have a certain number of branching 'private hidden layers' [19], and loss contributions from each task may be weighted differently. We did not thoroughly explore the implications of these parameters, but hypothesize that they may further improve these methods.

Neither of the previously mentioned learning methods enforce taxonomic consistency between levels, but this can be added. This may be a desirable property, especially when the classifiers at the fine-grained level are fairly weak. Assuming that the tasks are progressively more difficult, when the fine-grained classifier is faced with uncertainty over many different classes, it can then rely on the output of the previous level to help guide its output. One can also set thresholds for each level of the network based on some desired error rate. When highly uncertain, this gives the model the opportunity to forgo a fine-grained prediction in favor of a more coarse-grained prediction. This property is one of the main advantages of taxonomic classifiers in general.

4 Experimental Results

4.1 Datasets and Taxonomies

We demonstrate our methods using two fine-grained datasets. The first is Stanford Dogs [5], which contains 20,580 images of 120 classes of dogs. Each class contains 100 training images of dogs at various ages and colors. The other dataset is a collection of 76,613 images of 555 classes of North American birds compiled from NABirds [2], CUB-200-2011 [3], and Birdsnap.[1] There is an average of 95 images per class in the birds dataset.

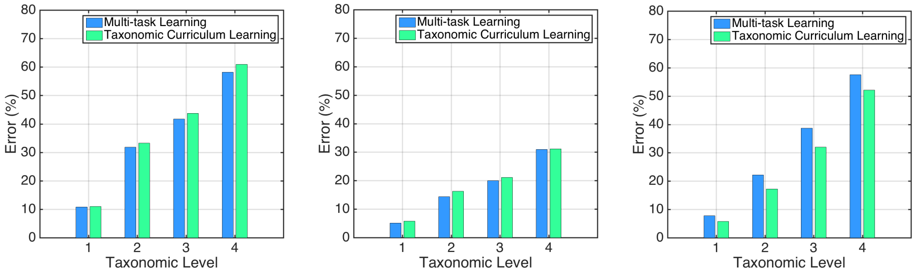
For the dog and bird domains, defining their respective visual taxonomies was fairly straightforward. Dogs are a single species, so we did not have a phylogenetic taxonomy to follow during visual taxonomy construction. Instead, we grouped fine-grained classes with highly similar appearances (e.g. Collie and Shetland Sheepdog) that a non-expert could easily confuse. These classes were then grouped into somewhat visually similar groups (e.g. Doberman Pinscher and Miniature Pinscher) that are distinguishable to most non-experts. Some of these classes are semantically interpretable (e.g. Poodle), while others are based on visual traits (e.g. White Wolf-Like). The resulting taxonomy has three levels, with 35, 72, and 120 classes at each level. The birds domain has a phylogenetic taxonomy, which, for the most part, follows the visual taxonomy well. A majority of the effort in creating the visual taxonomy was spent separating out fine-grained classes (e.g. juveniles from adults, males from females, etc.) and finding coarse classes for species that are phylogenetically highly unique. For example, Belted Kingfishers are the only bird from their phylogenetic order in the dataset, but to a non-expert, they are somewhat visually similar to other

[1] http://birdsnap.com

small birds. The visual taxonomy for birds has four levels, with 17, 130, 241, and 555 classes at each level.

4.2 Comparison of Training Methods

We compared the performance of multi-task learning and taxonomic curriculum learning using the birds dataset. The size of this dataset is atypical of the amount of data available for many real world fine-grained domains. A researcher creating a vision system for an obscure domain may only have access to a handful of expert labeled images per class. For this reason, we selected a subset containing 4,995 images (9 images per class) from our 52,701 training images as our more 'realistic' dataset. We tested both learning approaches in three different settings: 4,995 expert labeled examples, 52,701 expert labeled examples, and a partitioning of our dataset into 47,706 non-expert labeled images and 4,995 expert labeled images. Non-expert labeled images have coarse labels at the second level of the taxonomy, whereas expert labeled images have fine-grained labels at the fourth taxonomic level. Each model was trained by re-initializing the loss layers in a pre-trained GoogLeNet [24] network and fine-tuning the entire network. The results are shown in Figure 5.



(a) (b) (c)

Fig. 5: Comparison of multi-task learning and taxonomic curriculum learning using (a) 4,995 expert labeled images, (b) 52,701 expert labeled images, and (c) 47,706 non-expert labeled images and 4,995 expert labeled images. Multi-task learning barely beats taxonomic curriculum learning when labels are entirely at the fine-grained level. When there are many more coarse-grained examples than fine-grained examples, taxonomic curriculum learning outperforms multi-task learning.

Multi-task learning slightly outperformed taxonomic curriculum learning on all levels of the taxonomy when training with only expert labeled data. When training with many more non-expert labels than expert labels, multi-task learning performed noticeably worse than taxonomic curriculum learning. Although not shown, when the data sampling rate for multi-task learning was set uniformly, it performed substantially worse. The data shown in Figure 5c is from using expert-heavy sampling. Without extensively adjusting the network architecture or hyper-parameters, it is difficult to assess the generality of these findings. For our purposes, the important point is that taxonomic curriculum learning is more stable when dealing with labels distributed unequally across the taxonomy.

4.3 Performance Gains from Non-expert Data

Figures 5 (a) and (c) illustrate our main point: supplementing a small set of expert labeled data with a much larger set of non-expert labeled data can result in a substantial improvement in performance, measured in accuracy at the fine-grained level. The exact values are given in Table 1.

Table 1: Accuracy at the fine-grained level of the birds taxonomy using different amounts of expert and non-expert labeled training data. Supplementing the subset of 4,995 expert labeled images with an additional 47,706 non-expert labeled images results in a 6.1% increase in fine-grained accuracy.

Labels	Accuracy (%)
52,701 Expert	69.1
4,995 Expert	41.8
47,706 Non-Expert and 4,995 Expert	**47.9**

We explored this further using the Stanford Dogs dataset. We selected subsets of 5, 10, and 50 images from the 100 images for each class. These subsets are our expert labeled training sets, with labels at level 3 of the dog taxonomy. We trained models using all 100 images with non-expert labels at either the first or second level of the taxonomy in combination with each of the subsets. For comparison, we also trained models using only the expert labeled subsets. The fine-grained accuracy of these models is plotted in Figure 6. Increasing the ratio of non-expert labels to expert labels leads to a larger gain in performance. The plot also portrays the relative information contained in the labels at each level. For a fixed number of expert labeled examples, training with non-expert labels at level 2 is better than training with the same amount of non-expert labels at level 1, as should be expected. However, training with only 10 expert labels per class outperforms training with 5 expert labels and 95 level 2 non-expert labels per class.

Determining the relative information contained in labels at each level is an important consideration when trying to minimize annotation costs. From analyzing the dog and bird domains, we have found that for a given amount of additional expert labeled data, one typically requires an order of magnitude more non-expert labeled data to achieve the same performance improvement. Of

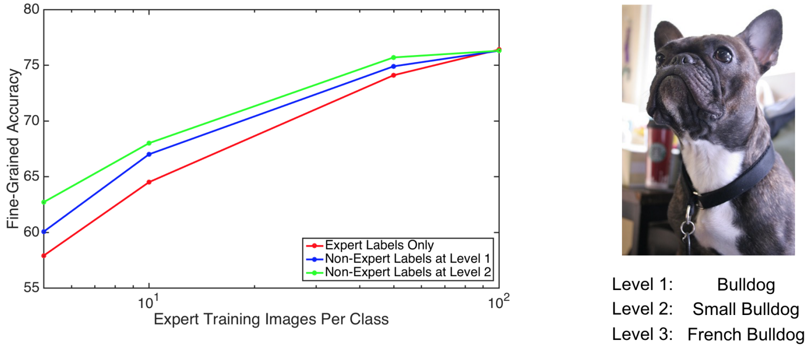


Fig. 6: Left: Fine-grained accuracy on the Stanford Dogs dataset as a function of the number of expert labeled examples used during training. The total number of images is kept fixed at 100 images per class. The three curves represent different types of non-expert labels. Right: Sample labels from each level of the taxonomy.

course, there may be a limit to this improvement. Collecting large amounts of non-expert labeled data can be daunting, but in certain domains, where online communities of non-experts exist, it may be possible to collect these examples at virtually no cost.

4.4 Taxonomic Self-Learning with Weak Labels

One can also increase the number of examples using self-learning. This entails using the trained model to label a set of unlabeled or weakly-labeled examples, then adding them to the training set for further learning. A taxonomic classifier adds an extra element, as we are now able to label examples up to different levels. We use the birds dataset to demonstrate this, training initially on 4,995 expert labeled images.

We gathered an additional 100,000 images from Flickr by querying the scientific names of each bird species in the dataset. Some of these images are mislabeled as other classes, while others do not contain a class from the dataset. Bird enthusiasts, for the most part, though, are quite adept at correctly labeling their images. We estimated at least half of the images had correct fine-grained labels. Our goal is to study self-learning in the context of very noisy labels due to non-experts. We expect that this may be more common in obscure domains. Therefore, we add additional noise to the labels by probabilistically mislabeling the examples according to the taxonomy. In addition to the noise already inherent in the dataset, for levels of the taxonomy, we set respective probabilities of 1%, 5%, 15%, and 75% for being mislabeled, an arbitrary approximation of a non-expert's error rate. Incorrect labels were chosen randomly from the corresponding level and propagated randomly down the taxonomy.

We used our model to generate predictions at each level of the taxonomy. Examples were accepted if (1) the prediction matched the weak label and (2)

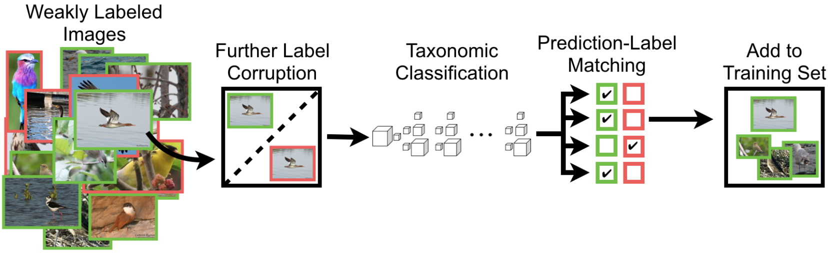


Fig. 7: The taxonomic self-learning pipeline. We further corrupt the labels of weakly labeled images from Flickr. The images are then run through the initial model, generating taxonomic predictions. If the prediction at a particular level is consistent and matches the weak label, the image is added to the training set.

the prediction was consistent with previous levels of the taxonomy. A separate set of examples were collected using only the model's fine-grained prediction. Table 2 contains the number of weakly labeled images added at each level of the taxonomy. Note that the fine-grained training set contains more level 4 images than the taxonomic training set. We then trained two models: one using the taxonomic labels and another using only the fine-grained labels. The taxonomic model outperformed the fine-grained model, 46.5% to 44.6%. These are improvements of 4.7% and 2.8% respectively over the base model's accuracy of 41.8%.

The key point is this: when we assume that weakly labeled data is mislabeled progressively worse at deeper levels of the taxonomy, a taxonomic approach to self-learning can acquire more labels, outperforming a purely fine-grained approach. When the labels are not very noisy, or the noise does not increase through the taxonomy, then standard fine-grained self-learning works just as well.

Table 2: Weakly labeled images added to the training set (cumulatively) at each level of the taxonomy. A total of 100,000 weakly labeled images were tested. More images were added at level 4 to the fine-grained training set than the taxonomic training set due to the consistency constraint.

Level	Number of Images Added
1	78,706
2	50,065
3	30,446
4	5,490
4 (Fine-Grained Only)	7,088

4.5 Coarse-Grained Classes

Organizing fine-grained object classes into a visual taxonomy allows us to work with a set of coarse-grained classes. These classes provide a new level of insight in analyzing the model's performance. We can assess the accuracy of coarse-grained classes to get a broad sense of where the model is making mistakes. Figure 8 gives an example of this analysis. At the first level of the taxonomy, we start with the class for all birds of prey. Splitting this class into its constituent classes, we observe the distribution of accuracies. After repeating this process, we see that the class 'Dark Brown Hawk' is substantially underperforming. Among other factors, this could be the result of a flawed visual taxonomy.

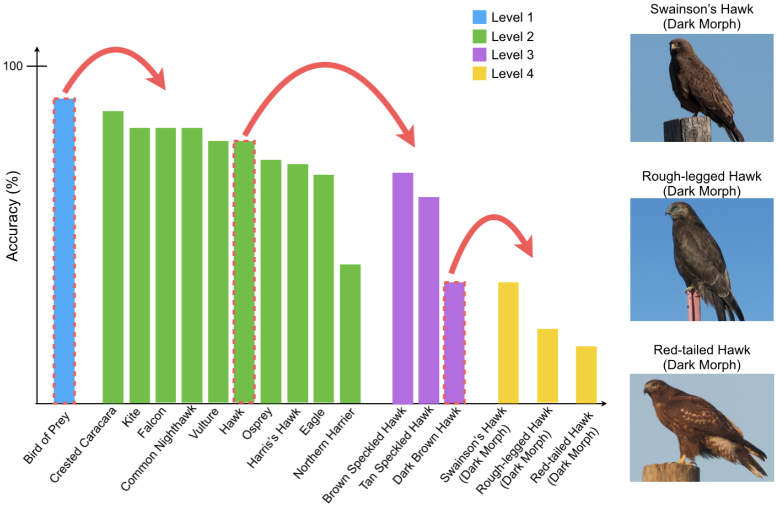


Fig. 8: Example analysis of coarse class accuracies. We start with the coarse class 'Bird of Prey' at the first level of the taxonomy. By splitting this class into its constituent classes, we can see which classes are dragging down performance. We repeat this at each taxonomic level. The coarse class 'Dark Brown Hawk' is underperforming significantly with respect to its siblings. This may suggest a flaw in the visual taxonomy.

To help determine possible flaws in the taxonomy, we can plot coarse class accuracies for a fine-grained class. For a specific fine-grained class, we determine what percentage of examples from that class were correctly labeled at each level of the taxonomy. We then can compare these percentages against the total accuracies for those coarse classes. Figure 9 gives a concrete example. The fine-grained class for Tricolored Heron seems well situated in the visual taxonomy. Its coarse-grained accuracies track well with other members of its coarse classes, and the accuracies remain fairly high throughout. Compare this with the fine-grained class Black Scoter (Female/Juvenile). Examples from this class are correctly classified as 'Duck, Goose, Swan' and 'Duck', but the model has a particularly difficult time classifying them as 'Scoter' as compared with other scoters. This suggests that the class Black Scoter (Female/Juvenile) should be moved elsewhere in the taxonomy, or perhaps the class 'Scoter' should be split. A similar phenomenon takes place with the class Lesser Goldfinch (Female/Juvenile).

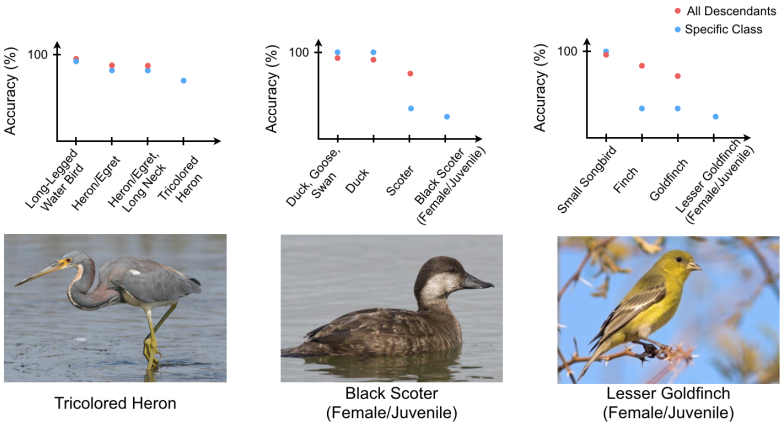


Fig. 9: Accuracies throughout the visual taxonomy for three representative fine-grained classes. Red dots denote the accuracy of all descendants from the corresponding coarse class. Blue dots denote accuracy from only the fine-grained class. For example, the red dot for finch indicates what percentage of finches were correctly classified as such. The blue dot indicates what percentage of Lesser Goldfinch Females/Juveniles were correctly classified as finches.

5 Conclusion

Visual taxonomies allow for the possibility of using non-expert labels for fine-grained object recognition. These labels are relatively easy to collect, and when combined with a small number of expert labels, can result in significant performance gains. We investigated two methods of taxonomic training, but much work remains to be done in this area. Taxonomic training methods work well when labels at different levels are able to assist each other, but if these tasks are directly competing for parameters in the network, performance can suffer. This is the motivation for private hidden layers for each task, which we did not explore. Each task could have a branch in the network, allowing for separate representations of each taxonomic level. Work also remains to be done in defining alternative class structures. Our method is not entirely confined to taxonomic tree structures; other coarse clusterings of classes could work as well. For object domains like food, this may be a more representative structure.

References

1. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on, IEEE (2009) 248–255
2. Van Horn, G., Branson, S., Farrell, R., Haber, S., Barry, J., Ipeirotis, P., Perona, P., Belongie, S.: Building a bird recognition app and large scale dataset with citizen scientists: The fine print in fine-grained dataset collection. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (June 2015)
3. Welinder, P., Branson, S., Mita, T., Wah, C., Schroff, F., Belongie, S., Perona, P.: Caltech-UCSD Birds 200. Technical Report CNS-TR-2010-001, California Institute of Technology (2010)
4. Nilsback, M.E., Zisserman, A.: Automated flower classification over a large number of classes. In: Computer Vision, Graphics & Image Processing, 2008. ICVGIP'08. Sixth Indian Conference on, IEEE (2008) 722–729
5. Khosla, A., Jayadevaprakash, N., Yao, B., Li, F.F.: Novel dataset for fine-grained image categorization: Stanford dogs. In: Proc. CVPR Workshop on Fine-Grained Visual Categorization (FGVC). (2011)
6. Maji, S., Rahtu, E., Kannala, J., Blaschko, M., Vedaldi, A.: Fine-grained visual classification of aircraft. arXiv preprint arXiv:1306.5151 (2013)
7. Krause, J., Deng, J., Stark, M., Fei-Fei, L.: Collecting a large-scale dataset of fine-grained cars. Second Workshop on Fine-Grained Visual Categorization (FGVC2) (2013)
8. Berg, T.L., Berg, A.C., Shih, J.: Automatic attribute discovery and characterization from noisy web data. In: Computer Vision–ECCV 2010. Springer (2010) 663–676
9. Hofmann, T., Cai, L., Ciaramita, M.: Learning with taxonomies: Classifying documents and words. In: NIPS workshop on syntax, semantics, and statistics. (2003)
10. Cesa-Bianchi, N., Gentile, C., Zaniboni, L.: Incremental algorithms for hierarchical classification. The Journal of Machine Learning Research **7** (2006) 31–54
11. Silla Jr, C.N., Freitas, A.A.: A survey of hierarchical classification across different application domains. Data Mining and Knowledge Discovery **22**(1-2) (2011) 31–72
12. Griffin, G., Perona, P.: Learning and using taxonomies for fast visual categorization. In: Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on, IEEE (2008) 1–8
13. Deng, J., Krause, J., Berg, A.C., Fei-Fei, L.: Hedging your bets: Optimizing accuracy-specificity trade-offs in large scale visual recognition. In: Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on, IEEE (2012) 3450–3457
14. Sfar, A.R., Boujemaa, N., Geman, D.: Confidence sets for fine-grained categorization and plant species identification. International Journal of Computer Vision **111**(3) (2015) 255–275
15. Srivastava, N., Salakhutdinov, R.R.: Discriminative transfer learning with tree-based priors. In Burges, C.J.C., Bottou, L., Welling, M., Ghahramani, Z., Weinberger, K.Q., eds.: Advances in Neural Information Processing Systems 26. Curran Associates, Inc. (2013) 2094–2102
16. Wang, D., Shen, Z., Shao, J., Zhang, W., Xue, X., Zhang, Z.: Multiple granularity descriptors for fine-grained categorization. In: Proceedings of the IEEE International Conference on Computer Vision. (2015) 2399–2406

17. Fei-Fei, L.: Knowledge transfer in learning to recognize visual objects classes. In: Proceedings of the Fifth International Conference on Development and Learning. (2006)
18. Donahue, J., Jia, Y., Vinyals, O., Hoffman, J., Zhang, N., Tzeng, E., Darrell, T.: Decaf: A deep convolutional activation feature for generic visual recognition. arXiv preprint arXiv:1310.1531 (2013)
19. Caruana, R.: Multitask learning. Machine Learning **28**(1) 41–75
20. Bengio, Y., Louradour, J., Collobert, R., Weston, J.: Curriculum learning. In: Proceedings of the 26th annual international conference on machine learning, ACM (2009) 41–48
21. Kumar, M.P., Packer, B., Koller, D.: Self-paced learning for latent variable models. In: Advances in Neural Information Processing Systems. (2010) 1189–1197
22. Chapelle, O., Schölkopf, B., Zien, A., et al.: Semi-supervised learning. (2006)
23. Xu, Z., Huang, S., Zhang, Y., Tao, D.: Augmenting strong supervision using web data for fine-grained categorization. In: Proceedings of the IEEE International Conference on Computer Vision. (2015) 2524–2532
24. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A.: Going deeper with convolutions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2015) 1–9