Taxonomic Curriculum Learning for Fine-Grained Object Recognition

Joseph Marino

California Institute of Technology

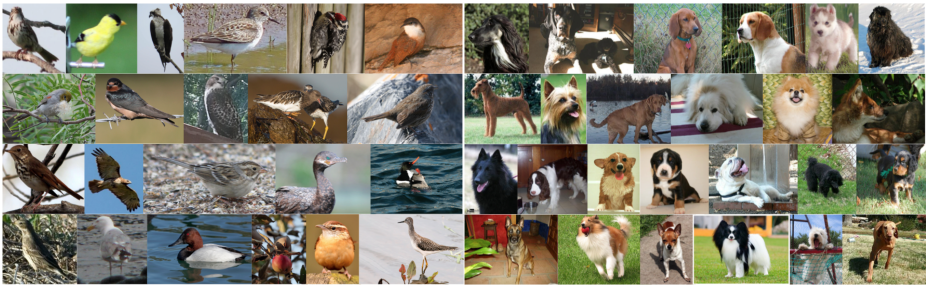
Abstract. Drawing inspiration from taxonomic methods of learning and curriculum learning, we propose a new learning method: taxonomic curriculum learning. With this method, given a visual taxonomy of fine-grained classes, a classifier is trained in a sequential coarse to fine manner. We compare our method with two baseline methods: the standard approach of purely fine-grained learning and multi-task learning. We present results on the Stanford Dogs dataset as well as a collection of birds datasets. In the case of training with labels that are non-uniformly distributed across the taxonomy, taxonomic curriculum learning outperforms both baseline methods in terms of accuracy. Additionally, even with uniformly distributed labels, taxonomic curriculum learning results in lower average taxonomic distance, making more reasonable predictions.

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.

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. With these shared visual features, we can group fine-grained classes into coarse clusters. Particularly in biological domains, which are the focus of many fine-grained object recognition tasks, these clusters tend to form a visual taxonomy of classes, for the most part mirroring the underlying phylogenetic taxonomy. For instance, all birds share a certain set of 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. The relevant visual features for distinguishing between classes in the taxonomy become more detailed with depth.

Humans tend to learn object classes following this coarse to fine ordering. A child might start by first learning a base object class, such as a dog. As the child sees and interacts with dogs, it might learn to distinguish between big dogs



(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. Visual similarity between classes in these domains tends to follow a taxonomic structure.

and small dogs, then between classes of breeds, such as poodles, spaniels, and bulldogs. This stands in contrast to current methods of training deep networks, which involve training on all fine-grained classes simultaneously. Being a non-convex optimization problem, there is evidence to suggest that the order in which learning occurs can play a significant role in determining the final performance of the model [20]. We pose the following question: does ordering training into a series of coarse- to fine-grained tasks help in training a deep network? Here, we define 'help' to be either resulting in better fine-grained accuracy or more interpretable predictions, as defined by the taxonomy.

Working in the domains of both birds and dogs, we construct visual taxonomies in order to transform the fine-grained tasks into series of coarse- to fine-grained tasks, ranging in difficulty. We train deep networks on these tasks in sequential order, a learning technique that we refer to as taxonomic curriculum learning. Comparing with baselines of purely fine-grained learning and multi-task learning, we show that taxonomic curriculum learning outperforms these methods under certain circumstances. The remaining sections are outlined as follows: Section 2 reviews previous work in this area, Section 3 details our method of learning, Section 4 presents our experimental results, and Section 5 provides a brief conclusion.

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 one could formulate taxonomic curriculum learning as a global approach.

More 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, however these classifiers are learned separately.

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 compare our method with this approach in sections 3 and 4.

2.3 Curriculum Learning

The inspiration for our technique stems from curriculum learning, in which a model is trained on progressively more difficult examples, defined by some metric, as a non-convex optimization technique [20, 21]. Instead of training on a series of increasingly difficult examples, our approach trains on a series of increasingly difficult tasks, as defined by the depth in the taxonomy.

3 Method

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 a human. In contrast, a learned taxonomy is not guaranteed to be visually interpretable. 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.

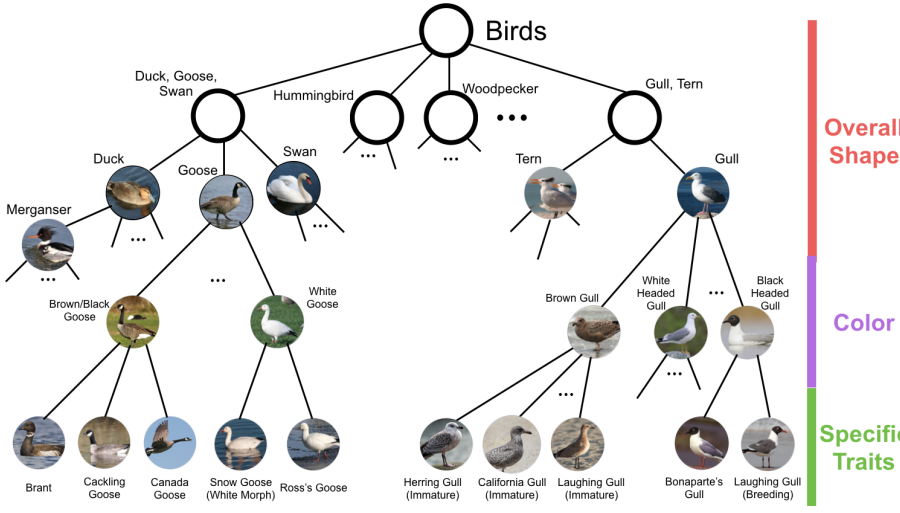


Fig. 2: Visual taxonomy of North American birds. Fine-grained classes are grouped first by overall shape and form, then by color, and finally by specific traits, such as patterns. The visual taxonomy roughly follows the phylogenetic taxonomy of birds.

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 by color, and finally by specific traits. This is shown in Figure 2. 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.). Coarse-grained classes should only be formed up to the point at which they are visually cohesive, i.e. the constituent classes share a large number of visual features. Grouping classes smaller and smaller until the root node may be detrimental to training. Phylogenetic taxonomies can aid in construction, but they have their limitations. 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). Likewise, House

sparrows are phylogenetically distant from other sparrows, however they have a similar appearance.

There is always the concern that manually constructing a visual taxonomy is too subjective. In the Appendix, we outline a method for evaluating the quality of a taxonomy. One final point is that we are not constrained to taxonomic structures. We prefer taxonomies because fine-grained labels automatically provide coarse-grained labels. They also represent biological domains well. Nevertheless, alternative coarse clusterings of varying difficulty may work as well.

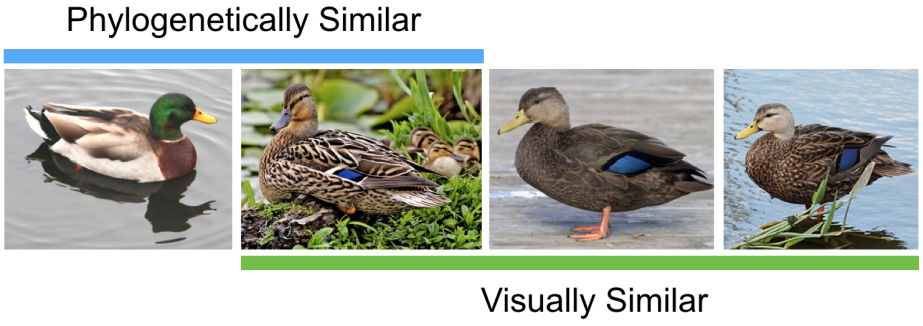


Fig. 3: Phylogenetically similar classes are typically also visually similar. Certain cases go against this rule, such as Mallard ducks. Female Mallard ducks are much more visually similar to American Black ducks and Mottled ducks than male Mallard ducks.

3.2 Training

Since we are incorporating a class taxonomy, we must modify our network to handle taxonomic inputs and outputs. For example, 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.

In taxonomic curriculum learning, we start by training a model on the first level of the taxonomy, with each input example having a coarse-grained label. Once the network converges, we remove the classifier for the first level, replace

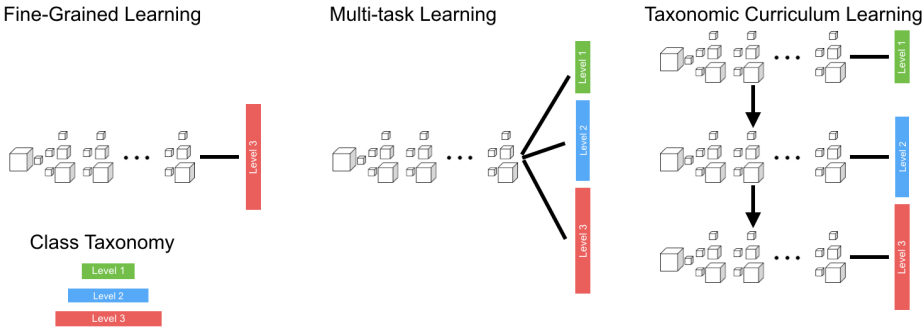


Fig. 4: A class taxonomy comprised of three levels is shown on the bottom left. Fine-grained learning represents the standard approach of training purely on the fine-grained class labels. During multi-task learning, labels from all levels of the taxonomy are trained on simultaneously. In taxonomic curriculum learning, labels from the various levels of the taxonomy are trained on 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.

it with a classifier for the second level, and, again, train until convergence. This process is repeated until we reach the end of the taxonomy. It should be noted that examples need not all have fine-grained labels: only examples for which we have corresponding labels are trained on at each level.

We compare taxonomic curriculum learning with two baseline methods. The first is the standard approach of ignoring the taxonomy and training purely on labels at the fine-grained level. The second is multi-task learning, in which labels from all levels of the taxonomy are trained on 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. If an example does not have a label at a particular level, then its gradient contribution for that level is zero. By comparing the performance of these two baselines with taxonomic curriculum learning, we can determine (1) whether or not the taxonomy assists in learning and (2) whether or not training on the levels of the taxonomy sequentially has any added benefit. Further details for all three learning methods are given in the Appendix.

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 second 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. Example images from both of these datasets are shown in Figure 1.

Dogs are a single species, so there is no phylogenetic taxonomy to follow. 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 the respective levels. 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 matching coarse classes for species that are phylogenetically highly unique. The visual taxonomy for birds has four levels, with 17, 130, 241, and 555 classes at the respective levels.

4.2 Accuracy-Based Comparison

We first compare taxonomic curriculum learning with the baseline methods on the basis of classification accuracy at each level of the taxonomy. For each trial, we re-initialized the loss layers in a pre-trained GoogLeNet [22] network, then fine-tuned the entire network. With the standard approach to fine-grained learning, we trained only on the final level of the taxonomy and propagated predictions up to internal classes. With this approach, the accuracy a_ℓ at taxonomic level ℓ in a taxonomy of L layers is given by

$$a_\ell = \frac{1}{N} \sum_{i=1}^{N} 1_{[\hat{y}_{i,L} \in \mathcal{L}(y_{i,\ell})]}, \tag{1}$$

where $\mathcal{L}(y_{i,\ell})$ is the set of leaf classes in the sub-tree rooted at the internal class $y_{i,\ell}$, the ground truth label at level ℓ for example i, and $\hat{y}_{i,L}$ is the fine-grained prediction of the network for example i. With taxonomic curriculum learning and multi-task learning, accuracy at each level adheres to the normal definition, since we have predictions and labels at each level. Note that we do not enforce consistency in these predictions.

We compared all three learning methods on both datasets for two cases. In the first case, all training examples have labels at the fine-grained level of the taxonomy, meaning there is an equal number of labels for each level: 52,701 for birds and 12,000 for dogs. In the second case, the labels for the training examples are skewed toward the coarse levels of the taxonomy. For both datasets, we selected subsets of images for each class (9 per class for birds and 5 per class

[1] http://birdsnap.com

for dogs) to have fine-grained labels, with all other examples having only coarse-grained labels. Figure 5 contains the test accuracy results of training with each method in each case.

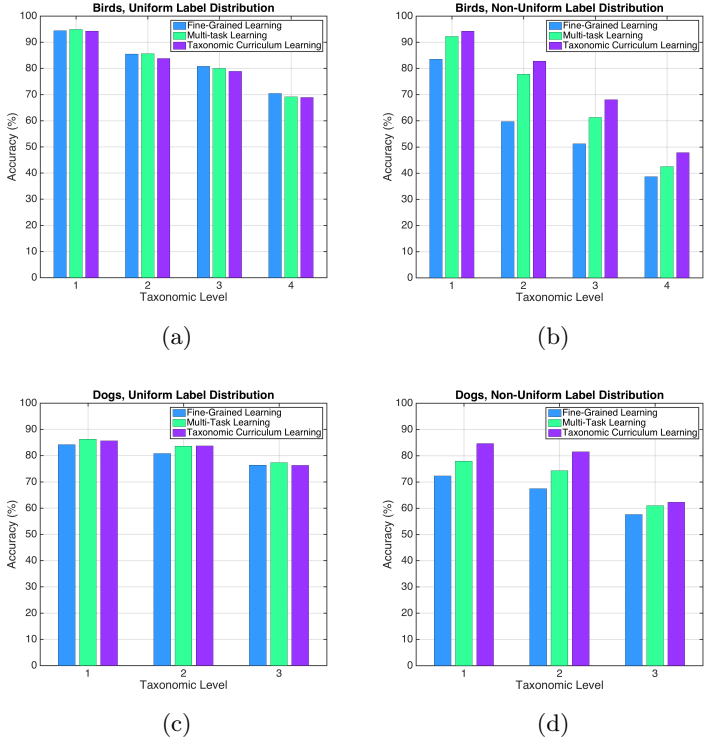


Fig. 5: Accuracy at each level of the taxonomy for all three training methods. For both the birds and dogs datasets, taxonomic curriculum learning outperforms both baseline methods when the labels are non-uniformly distributed across the levels of the taxonomy.

In the case of uniformly distributed labels at all levels of the taxonomy, taxonomic curriculum learning performs roughly as well as both baseline methods at all levels of the taxonomy. However, when the number of coarse-labeled examples is much greater than the number of examples with fine-grained labels, taxonomic curriculum learning surpasses the baselines. In the case of fine-grained learning, this is not surprising: the network has fewer training examples. With multi-task learning, it is more difficult to assess the reason for the model's poor performance. There appears to be adverse competition between the levels of the taxonomy resulting from the difference in the number of data examples at each level, however further experiments will be needed to explore this in more detail.

4.3 Taxonomic Distance-Based Comparison

Another metric with which to evaluate the performance of taxonomic curriculum learning is through taxonomic distance. We define our taxonomic distance measure as follows:

$$D_{\text{tax}} = L - \max_{\ell}(\ell \cdot 1_{\hat{y}_\ell = y_\ell}). \tag{2}$$

Thus, for a given example, the taxonomic distance is the number of levels above the leaf level at which the prediction first matches the label. A smaller taxonomic distance represents a more reasonable prediction, quantifying the fact that in a taxonomic setting, not all errors are equal.

We trained all three learning methods on the birds dataset by again re-initializing the final loss layer and training with the subset of 9 images per class. We chose this setting because (1) the birds taxonomy is larger, allowing for more variation in taxonomic distance, and (2) the subset of data should help differentiate the methods, demonstrating which methods work well in the case of few training examples. Comparing the methods first on the basis of test accuracy, Figure 6 shows that the results are, for the most part, consistent with those from the full dataset in Section 4.2, with fine-grained learning performing slightly worse at coarse levels of the taxonomy. The average taxonomic distances for each method, evaluated on the test set, are shown in Table 1. Taxonomic curriculum learning outperforms both baseline methods slightly, despite having lower accuracy at the fine-grained level. This suggests that taxonomic curriculum learning results in more reasonable predictions, compared with the baseline methods.

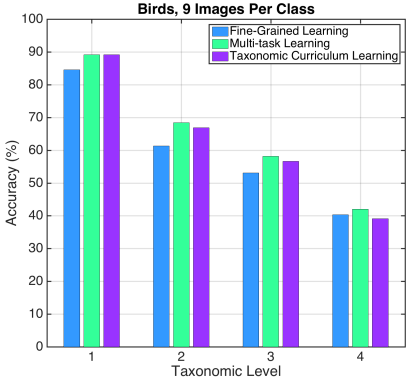


Fig. 6: Comparison of taxonomic curriculum learning with both baseline methods according to accuracy for the subset of 9 images per class on the birds dataset.

Table 1: Comparison of taxonomic curriculum learning with both baseline methods according to average taxonomic distance.

	Ave. D_{tax}
Fine-Grained	1.608
Multi-task	1.289
Taxonomic Curriculum	**1.280**

Since we expect that taxonomic curriculum learning may help during training as a non-convex optimization technique, we increased the difficulty of the optimization problem by re-initializing the network further back by three inception modules (see Appendix for details). We again trained with each method on the subset of the birds dataset. The results are shown in Figure 7 and Table 2. As with the previous results, taxonomic curriculum learning does not result in a substantial improvement in accuracy over multi-task learning. However, we see that with a deeper re-initialization, taxonomic curriculum learning results in a noticeably lower average taxonomic distance.

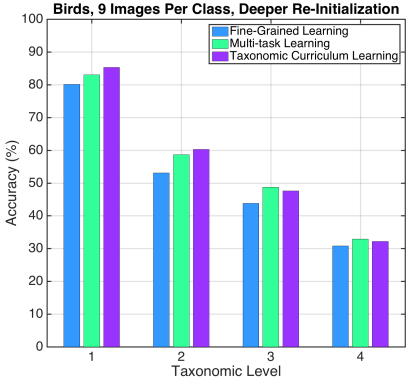


Fig. 7: Comparison of taxonomic curriculum learning with both baseline methods according to accuracy for the subset of 9 images per class on the birds dataset using a deeper re-initialization.

Table 2: Comparison of taxonomic curriculum learning with both baseline methods according to average taxonomic distance using a deeper re-initialization.

	Ave. D_{tax}
Fine-Grained	1.922
Multi-task	1.609
Taxonomic Curriculum	**1.541**

5 Conclusion

We have proposed a new learning method, taxonomic curriculum learning, in which we sequentially train on a class taxonomy from coarse- to fine-grained classes. In the case of non-uniformly distributed labels across the levels of the taxonomy, taxonomic curriculum learning outperforms baseline methods of standard fine-grained learning and multi-task learning. With uniformly distributed labels, it results in similar accuracy, but still outperforms the baseline methods in terms of average taxonomic distance. Additional work will be needed to confirm these results with other datasets and with other data formats. Also, more principled methods of taxonomy construction should be explored, with an eye toward further improving the gains from taxonomic curriculum learning.

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. 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

6 Appendix

6.1 Learning Methods

With the standard approach to fine-grained learning, the network is typically trained using the softmax cross-entropy loss function:

$$\mathscr{L} = -\frac{1}{N}\sum_{i=1}^{N}\log p_{i,y_i}, \tag{3}$$

where

$$p_{i,y_i} = \frac{e^{f_{i,y_i}}}{\sum_k e^{f_{i,k}}} \tag{4}$$

is the probability(-like) output, given the activations, f, at the output layer. y_i is the ground truth class for example i. If we take $\mathscr{L}$ to be the loss for a mini-batch of stochastic gradient descent (SGD), then N is the number of examples in the mini-batch. In the case of multi-task learning, we have a loss for each level of the taxonomy:

$$\mathscr{L} = -\frac{1}{N}\sum_{\ell=1}^{L}\sum_{i=1}^{N}\log p_{i,y_{i,\ell}}. \tag{5}$$

$p_{i,y_{i,\ell}}$ is now the probability of example i having label $y_{i,\ell}$, the label at level ℓ of the taxonomy. At each level, the denominator in (4) is a sum over all classes in that level. Since the levels have different total numbers of classes, the losses at smaller, coarser levels will end up dominating losses at larger, finer levels. To account for this, we weight the losses appropriately. Assuming the fine-grained level has loss weight $w_L = 1$, then the loss weight at each level will be

$$w_\ell = \frac{|M_\ell|}{|M_L|}, \tag{6}$$

where $|M_\ell|$ is the total number of classes at level ℓ. In practice, properly weighting the losses for each level affects the network only minimally. With the Stanford Dogs dataset, the difference in accuracy between weighting and not weighting the losses was about 1%.

Up through this point, we have assumed that each example has labels at all levels of the taxonomy. Relaxing this assumption, we must properly account for the loss contributions for each level. A dataset with labels that are heavily skewed toward the coarse levels of the taxonomy will tend to focus more on those levels if the examples are sampled uniformly. To encourage learning at all levels of the taxonomy equally, we would rather sample uniformly according to level. However, since fine-grained labels automatically provide coarse-grained labels, we cannot do so without ignoring coarsely labeled examples. In our implementation of multi-task learning, we present fine-grained labels without their corresponding coarse grained labels to account for this.

For both taxonomic curriculum learning and multi-task learning, there are network architecture and hyper-parameter decisions that must be made. For both approaches, we must decide whether to include what are referred to as 'private hidden layers' [19], shown in Figure 8. These are layers of the network to which a particular task has private access. With taxonomic curriculum learning, this is equivalent to re-initializing the final layers of the network (or adding/removing layers) when switching between levels of the taxonomy. We leave investigation of these architectures for future work.

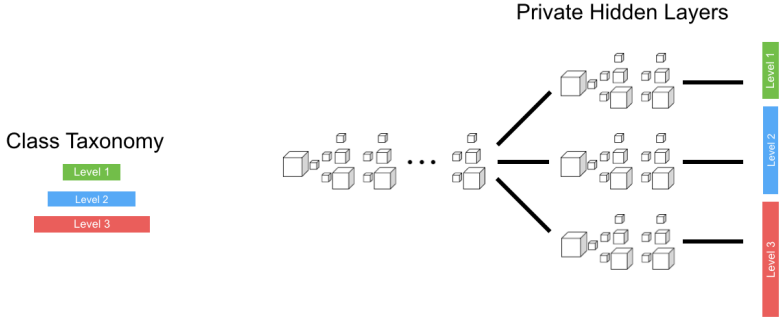


Fig. 8: Diagram of possible network architecture using private hidden layers.

6.2 Re-Initialization

In comparing the learning methods on the basis of taxonomic distance, we tried two different network re-initializations. The first was our approach from Section 4.2, in which we re-initialized only the final loss layers of the network. To help determine whether taxonomic curriculum learning acts as a non-convex optimization technique, we also trained using a larger re-initialization. This included the loss layers, as well as the final three inception modules in the GoogLeNet network architecture. The two re-initializations are shown in Figure 9.

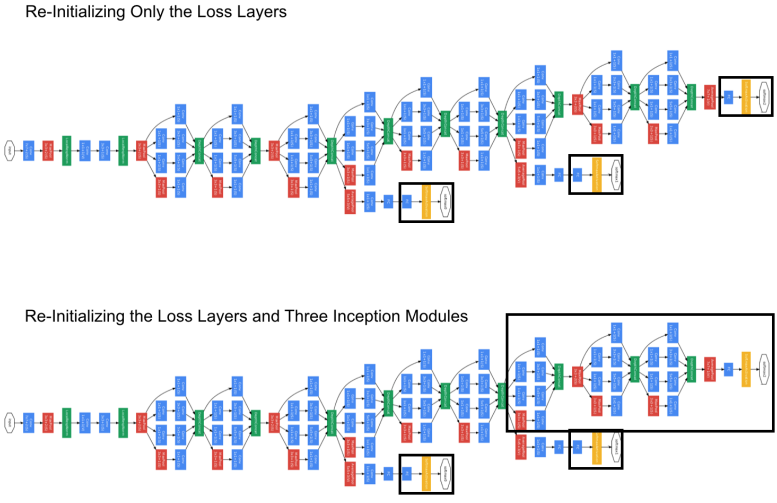


Fig. 9: Diagram showing the two re-initializations used in comparing the learning methods on the basis of taxonomic distance.

6.3 Analyzing the Taxonomy

We can assess the accuracy of coarse-grained classes to get a broad sense of where the model is making mistakes and look for potential flaws in the visual taxonomy. Figure 10 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.

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 11 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).

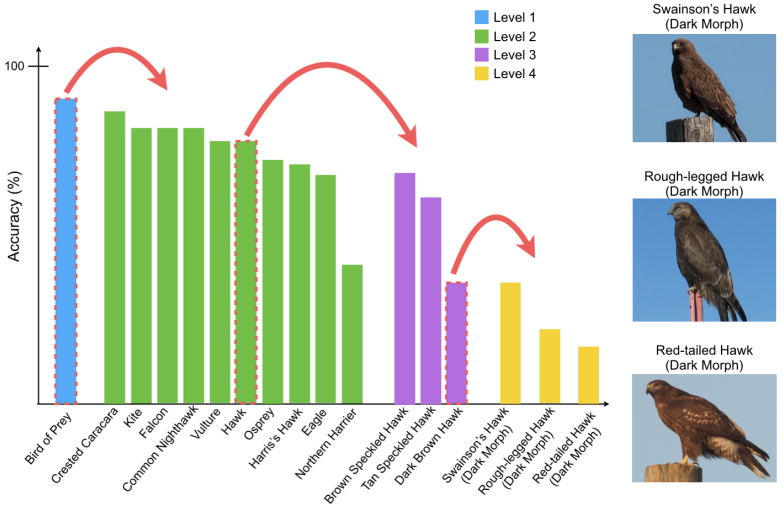


Fig. 10: Example analysis of coarse class accuracies. We start with the coarse class 'Bird of Prey' at the first level of the taxonomy. By splitting this class into its constituent classes, we can see which classes are dragging down performance. We repeat this at each taxonomic level. The coarse class 'Dark Brown Hawk' is underperforming significantly with respect to its siblings. This may suggest a flaw in the visual taxonomy.

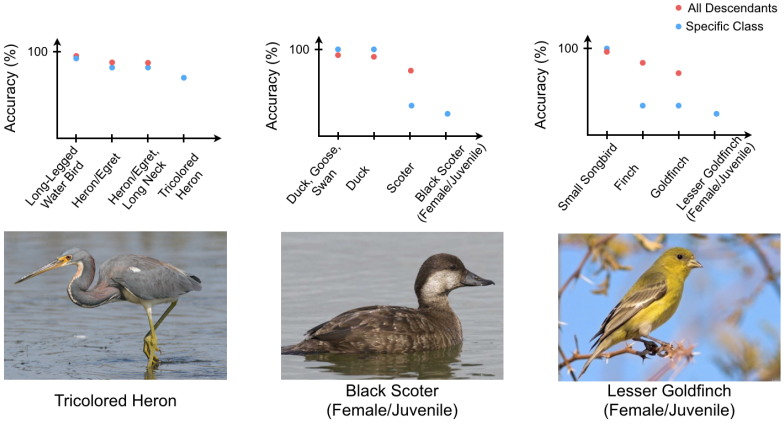


Fig. 11: Accuracies throughout the visual taxonomy for three representative fine-grained classes. Red dots denote the accuracy of all descendants from the corresponding coarse class. Blue dots denote accuracy from only the fine-grained class. For example, the red dot for finch indicates what percentage of finches were correctly classified as such. The blue dot indicates what percentage of Lesser Goldfinch Females/Juveniles were correctly classified as finches.