Modeling Semantic Correlation and Hierarchy for Real-world Wildlife Recognition

Dong-Jin Kim
Hanyang University
djdkim@hanyang.ac.kr

Zhongqi Miao
UC Berkeley / ICSI
zhongqi.miao@berkeley.edu

Yunhui Guo
University of Texas at Dallas
Yunhui.Guo@UTDallas.edu

Stella X. Yu
University of Michigan / UC Berkeley / ICSI
stellayu@umich.edu

Kyle Landolt
United States Geological Survey
klandolt@usgs.gov

Mark Koneff
US Fish and Wildlife Service
mark_koneff@fws.gov

Travis Harrison
United States Geological Survey
tharrison@usgs.gov

Abstract

We explore the challenges of human-in-the-loop frameworks to label wildlife recognition datasets with a neural network. In wildlife imagery, the main challenges for a model to assist human annotation are two-fold: (1) the training dataset is usually imbalanced, which makes the model's suggestion biased, and (2) there are complex taxonomies in the classes. We establish a simple and efficient baseline, including the debiasing loss function and the hyperbolic network architecture, to address these issues. Moreover, we propose leveraging the semantic correlation to train the model more effectively by adding a co-occurrence layer to our model during training. We demonstrate the efficacy of our method in both our real-world wildlife areal survey recognition dataset and the public image classification dataset, CIFAR100-LT and CIFAR10-LT.

1 Introduction

Aerial remote sensing technologies [14] are being increasingly used to monitor and survey wildlife populations safely [20, 22, 24]. However, manually processing real-world data is time-consuming and expensive for researchers and natural resource agencies [2, 7, 9]. As a result, we explore a human-in-the-loop approach [3, 6, 21] to utilize a deep neural network to collaborate with human annotators to efficiently process our real-world wildlife dataset. In particular, our goal is to train a neural network as an image classification model to actively assist human annotators in labeling wildlife recognition datasets [17] by suggesting the class of unlabeled images.

When training a network to suggest image classes for real-world digital aerial imagery applications correctly, we observe two major distinctive points in our multi-species dataset. **(1) Imbalanced data distribution.** Like most real-world datasets [23], our dataset also shows extremely imbalanced data distributions in the animal species [19]. Several dominant species are often observed, along with many infrequent species that are sparsely represented in datasets. In our dataset, the largest class

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

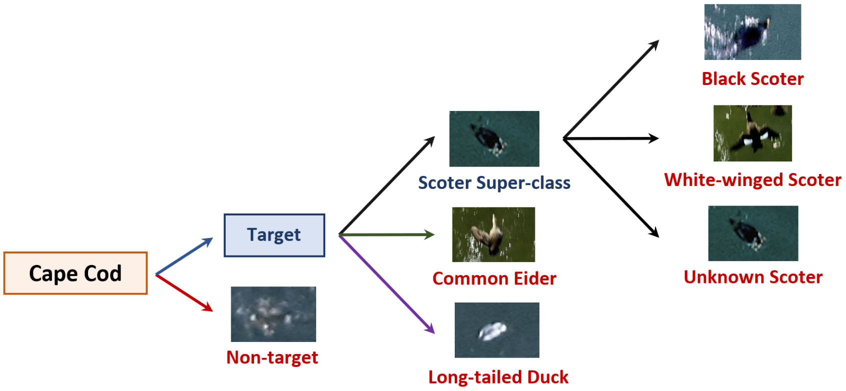

Figure 1: The six classes in our Cape Cod dataset have a hierarchical relationship described in [17]. `Target` consists of three super-classes, `Scoter Super-class`, `Common Eider`, `Long-tailed Duck`. `Scoter Super-class` could be further divided to `Black Scoter`, `White-winged Scoter`, and `Unknown Scoter`.

had 6,246 training images, while the smallest class only had 17 training images. This issue could harm the recognition accuracy as the model prediction can be easily biased towards the abundant class [18]. **(2) Class hierarchy.** In addition, as illustrated in Figure 1, the labels in our dataset have a clear hierarchy. For example, mutually exclusive `Target` and `Non-target` categories have the highest hierarchy. The `Target` category consists of three second-level categories (super-classes): `Scoter`, `Common Eider`, and `Long-tailed Duck`. Finally, the `Scoter` super-class consists of three fine-grained classes such as `Black Scoter`, `White-winged Scoter`, and `Unknown Scoter`. As the categories in our class list have different levels of hierarchy, it is necessary to treat each category differently according to the corresponding hierarchy level.

As a baseline to consider the aforementioned characteristics, we first utilize a simple yet effective debiasing loss [15] by leveraging the prior class distribution. In addition, to leverage the hierarchical nature of the class labels, we utilize hyperbolic neural networks [5], to learn the hierarchy in the labels effectively. To further boost the performance, we propose to learn the semantic correlation by modeling the co-occurrence among the classes. In particular, we add a learnable co-occurrence matrix on top of the model's final layer to refine the probability of the class prediction.

We use a case study of a real-world digital aerial survey dataset of waterbird species collected from the Atlantic Ocean near Cape Cod, Massachusetts, USA [17]. We show the efficacy of our method with both our wildlife dataset and the public image classification datasets, the CIFAR100-LT and CIFAR10-LT dataset [1, 4], where our approach shows favorable performance compared to the baseline methods. In summary, our contributions are three-fold: (1) We explore a new problem to actively label our new wildlife recognition dataset, (2) as a model to actively assist the human annotation process, we establish a simple yet effective baseline model to address the class imbalance problem and exploit label hierarchy, and (3) we propose a method to model the semantic correlation in the class labels by adding a co-occurrence layer, which further improves the performance.

2 Modeling Semantic Correlation and Hierarchy

Given an input image X and the label Y, our goal is to learn a classifier that minimizes the classification error. As mentioned in Sec. 1, our dataset is imbalanced in categories and has a label hierarchy. Therefore, we apply several methods to address these two challenges, (1) debiasing loss, (2) hyperbolic model, and (3) semantic correlation based learning.

Debiasing Loss Function. We leverage one of the popular long-tailed image classification methods named logit adjustment [15] for imbalanced data distribution. Logit adjustment encourages a large relative margin between logits of rare positive versus dominant negative labels to balanced the learning. In particular, we apply the logit adjustment [15] as a debiasing loss function during training. Other concurrent debiasing loss functions, such as margin-based approaches (LDAM [1]), uniformly increase the margin between a rare positive and *all negatives* and have relatively poor generalizability in most scenarios where the negative classes have heavily biased distribution. In contrast, logit

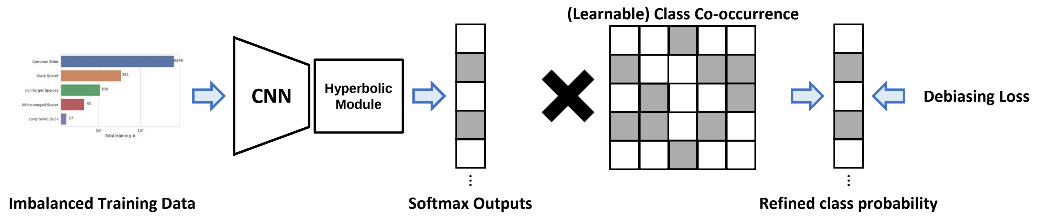

Figure 2: Our training framework including a hyperbolic module, debiasing loss function, and our co-occurrence layer to leverage the label correlations.

adjustment loss increases the margin between a rare positive and a *dominant* negative in the output probability, which is more suitable for realistic datasets like our dataset.

Hyperbolic Models. To effectively learn the hierarchy in the class labels, we add a hyperbolic module [5] in our classifier to create an hyperbolic neural network (HNN) that utilizes the hyperbolic space to embed data with hierarchical structures. In particular, hyperbolic neural networks lift Euclidean features into hyperbolic space for classification. Hyperbolic neural networks have shown to be helpful, especially on datasets with known semantic hierarchies. Specifically, we leverage the clipped hyperbolic classifier [5] method, which is generally effective on standard image classification benchmarks. This method clips the Euclidean feature magnitude in the hybrid architecture connecting Euclidean features to a hyperbolic classifier while training HNNs. Hyperbolic features capture the hierarchical structure of the dataset and improve the model's image classification performance.

Learning with Semantic Correlations. Moreover, we propose taking advantage of the natural *correlations* in the bird species. In particular, if a model gives high likelihood of the black scoter (non-rare class), the model should give a similar likelihood of the white-winged scoter (rare class) as well. In addition, the detection of the Non-target species should preclude other bird classes.

In particular, we define a learnable co-occurrence matrix [10, 11] in the form of $L \times L$ followed by a softmax nonlinearity. Each entry of the matrix represents the conditional probability $p(Y = i|Y' = j)$ that the correct class of the given image is $Y = j$ when a primary model prediction Y' is j. By adding the co-occurrence matrix on top of the model's final layer as a *co-occurrence layer*, the probability of the refined class probability is predicted according to the law of total probability:

$$P(Y = i|X) = \sum_{j \in \mathcal{Y}} p(Y = i|Y' = j) * p(Y' = j|X). \tag{1}$$

This process is illustrated in Fig. 2. Since the co-occurrence layer is learned end-to-end via the back-propagation from the loss function to the model input, it works as prior knowledge [8, 13, 26] that can be automatically learned from the target dataset.

3 Experiments

The main dataset we use is an aerial imagery dataset collected by the U.S. Fish and Wildlife Service at Nantucket Shoals (Cape Cod), Massachusetts, in February 2017. After data collection, wildlife experts manually cropped images of individual birds (*i.e.*, targets) and annotated the images into six different classes, illustrated in Figure 1: The number of samples for each class is shown in Table 3.

Species	Train #	Test #
`Common Eider`	6,246	3,172
`Unknown Scoter`	466	114
`Black Scoter`	341	108
`White-winged Scoter`	45	21
`Long-tailed Duck`	17	5
`Non-target Species`	108	38

Table 1: The number of samples in our dataset for training and testing split. The imbalance ratio of our dataset is 367:1.

Table 2 shows the quantitative results on our dataset. The baseline (Vanilla) method shows relatively poor performance, especially for species with limited training samples such as `Long-tailed Duck` (17 training images) and `White-winged Scoter` (45), which have 0.0% and 38.1% test accuracy, respectively. On the other hand, `Common Eider`, the class with the largest number of training samples (6,246), shows 99.1% test accuracy.

	Test accuracy (%)					
Species	**Vanilla**	**+LDAM**	**+LA**	**+LA+H**	**+LA+Corr**	**+LA+H+Corr (Ours)**
Common Eider	**99.1**	93.8	92.0	86.8	91.5	91.8
Black Scoter	95.4	92.6	97.2	87.0	97.2	**98.0**
White-winged Scoter	38.1	71.4	85.7	61.9	85.7	**85.9**
Long-tailed Duck	0.0	100.0	**100.0**	**100.0**	**100.0**	**100.0**
Non-target Species	39.5	68.4	78.9	**84.1**	81.6	81.9
Average accuracy (%)	54.1	85.3	90.8	82.9	91.2	**91.5**

Table 2: Quantitative results on our Wildlife dataset. Our final model with logit adjustment loss (+LA), hyperbolic module (+H), and our semantic co-occurrence aware learning method (+Corr) shows the best performance is most of the classes.

Methods	CIFAR100-LT	CIFAR10-LT
Vanilla	38.3	70.4
LDAM [1]	42.0	77.0
LA [16]	43.9	77.7
Ours	**45.3**	**79.0**

Table 3: Top-1 accuracy on CIFAR10-LT and CIFAR100-LT dataset [1, 4] with the imbalance ratio of 100. Our final model consistently outperforms the baseline methods across different imbalance factors.

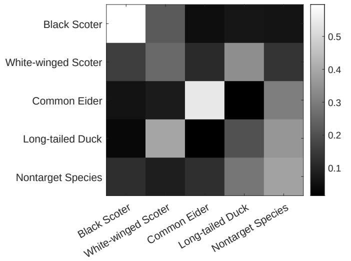

Table 4: The visualization of the learned co-occurrence matrix. Along the Y-axis is the given action, and the X-axis enumerates conditional actions.

While the classification model with a traditional debiasing loss function (LDAM [1]) already improves the recognition performance over the vanilla model, our model with the logit adjustment loss function (LA [16]) shows even more performance improvement in all the classes (+31.2% average classification accuracy improvement from LDAM and +36.7% from LA). The largest gain comes from the two tail classes, Long-tailed Duck and White-winged Scoter, from 0.0% to 100.0% and 38.1% to 85.7%, respectively. Despite the improvements in the less abundant classes, the performance of Common Eider dropped by 7.1%, which is a common phenomenon of imbalanced methods where the performance of large classes is sacrificed [12, 18, 25].

Upon the logit adjustment loss, adding both the hyperbolic module or our co-occurrence based learning method gives noticeable performance improvements in the Non-target Species class by better learning the correlation among the categories. However, the hyperbolic module shows performance drop in most of the other classes. We conjecture that it is because the number of class in our Wildlife dataset is too small (which might have relatively weak hierarchy among the labels) and the hyperbolic module requires large number of parameters (which causes overfitting problem). Note that combining both the hyperbolic module and our co-occurrence based learning method leads to the best performance by compensating the weakness of each other.

We also evaluate our method on standard image classification datasets with class imbalance, CIFAR100-LT and CIFAR10-LT [1, 4] with the imbalance ratio of 100, in order to validate the efficacy of our method to alleviate imbalance problems. The result is shown in Table 3. As shown in the table, our method our method shows the best performance compared to the recent powerful debiasing methods, LDAM [1] or logit adjustment [16]. This signifies that our method is also effective to alleviate class imbalance on popular benchmarks.

Finally, we visualize the learned co-occurrence matrix from our method in Fig. 4. The correlation between Black Scoter and White-winged Scoter and the correlation between Long-tailed Duck and Non-target Species can be found in the co-occurrence matrix. An interesting observation is that there is a correlation between White-winged Scoter and the Long-tailed Duck, which is not intuitive.

4 Conclusion

We explore a human-in-the-loop approach to utilize a network to collaborate with human annotators to process our real-world wildlife dataset. We also present solutions, including the debiasing loss function, hyperbolic model, and our semantic correlation learning. We hope our challenges and solutions inspire future researchers in wildlife surveys.

5 Acknowledgement

The findings and conclusions in this article are those of the author(s) and do not necessarily represent the views of the U.S. Fish and Wildlife Service.

References

[1] Kaidi Cao, Colin Wei, Adrien Gaidon, Nikos Arechiga, and Tengyu Ma. Learning imbalanced datasets with label-distribution-aware margin loss. *Advances in Neural Information Processing Systems (NIPS)*, 2019.

[2] Dominique Chabot and Charles M Francis. Computer-automated bird detection and counts in high-resolution aerial images: a review. *Journal of Field Ornithology*, 87(4):343–359, 2016.

[3] Jae Won Cho, Dong-Jin Kim, Yunjae Jung, and In So Kweon. Mcdal: Maximum classifier discrepancy for active learning. *IEEE transactions on neural networks and learning systems*, 2022.

[4] Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. Class-balanced loss based on effective number of samples. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[5] Yunhui Guo, Xudong Wang, Yubei Chen, and Stella X Yu. Clipped hyperbolic classifiers are super-hyperbolic classifiers. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[6] Dong-Jin Kim, Jae Won Cho, Jinsoo Choi, Yunjae Jung, and In So Kweon. Single-modal entropy based active learning for visual question answering. In *British Machine Vision Conference (BMVC)*, 2021.

[7] Dong-Jin Kim, Jinsoo Choi, Tae-Hyun Oh, and In So Kweon. Dense relational captioning: Triple-stream networks for relationship-based captioning. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[8] Dong-Jin Kim, Jinsoo Choi, Tae-Hyun Oh, and In So Kweon. Image captioning with very scarce supervised data: Adversarial semi-supervised learning approach. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2019.

[9] Dong-Jin Kim, Tae-Hyun Oh, Jinsoo Choi, and In So Kweon. Dense relational image captioning via multi-task triple-stream networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

[10] Dong-Jin Kim, Xiao Sun, Jinsoo Choi, Stephen Lin, and In So Kweon. Detecting human-object interactions with action co-occurrence priors. In *European Conference on Computer Vision (ECCV)*, 2020.

[11] Dong-Jin Kim, Xiao Sun, Jinsoo Choi, Stephen Lin, and In So Kweon. Acp++: Action co-occurrence priors for human-object interaction detection. *IEEE Transactions on Image Processing (TIP)*, 30:9150–9163, 2021.

[12] Ziwei Liu, Zhongqi Miao, Xiaohang Zhan, Jiayun Wang, Boqing Gong, and Stella X Yu. Large-scale long-tailed recognition in an open world. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[13] Cewu Lu, Ranjay Krishna, Michael Bernstein, and Li Fei-Fei. Visual relationship detection with language priors. In *European Conference on Computer Vision (ECCV)*, 2016.

[14] John F McEvoy, Graham P Hall, and Paul G McDonald. Evaluation of unmanned aerial vehicle shape, flight path and camera type for waterfowl surveys: disturbance effects and species recognition. *PeerJ*, 4:e1831, 2016.

[15] Aditya Krishna Menon, Sadeep Jayasumana, Ankit Singh Rawat, Himanshu Jain, Andreas Veit, and Sanjiv Kumar. Long-tail learning via logit adjustment. In *International Conference on Learning Representations (ICLR)*, 2021.

[16] Aditya Krishna Menon, Sadeep Jayasumana, Ankit Singh Rawat, Himanshu Jain, Andreas Veit, and Sanjiv Kumar. Long-tail learning via logit adjustment. In *International Conference on Learning Representations (ICLR)*, 2021.

[17] Zhongqi Miao, Stella X. Yu, Kyle L. Landolt, Mark D. Koneff, Timothy P. White, Luke J. Fara, Enrika J. Hlavacek, Bradley A. Pickens, Travis J. Harrison, and Wayne M. Getz. Challenges and solutions for automated avian recognition in aerial imagery. *Remote Sensing in Ecology and Conservation*, 2022.

[18] Youngtaek Oh, Dong-Jin Kim, and In So Kweon. Daso: Distribution-aware semantics-oriented pseudo-label for imbalanced semi-supervised learning. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[19] Stuart L Pimm, Clinton N Jenkins, Robin Abell, Thomas M Brooks, John L Gittleman, Lucas N Joppa, Peter H Raven, Callum M Roberts, and Joseph O Sexton. The biodiversity of species and their rates of extinction, distribution, and protection. *science*, 344(6187), 2014.

[20] D Blake Sasse. Job-related mortality of wildlife workers in the united states, 1937-2000. *Wildlife society bulletin*, pages 1015–1020, 2003.

[21] Inkyu Shin, Dong-Jin Kim, Jae Won Cho, Sanghyun Woo, KwanYong Park, and In So Kweon. Labor: Labeling only if required for domain adaptive semantic segmentation. In *IEEE International Conference on Computer Vision (ICCV)*, 2021.

[22] Devis Tuia, Benjamin Kellenberger, Sara Beery, Blair R. Costelloe, Silvia Zuffi, Benjamin Risse, Alexander Mathis, Mackenzie W. Mathis, Frank van Langevelde, Tilo Burghardt, Roland Kays, Holger Klinck, Martin Wikelski, Iain D. Couzin, Grant van Horn, Margaret C. Crofoot, Charles V. Stewart, and Tanya Berger-Wolf. Perspectives in machine learning for wildlife conservation. *Nature Communications*, 13(1):792, 2022.

[23] Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alex Shepard, Hartwig Adam, Pietro Perona, and Serge Belongie. The inaturalist species classification and detection dataset. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[24] Dongliang Wang, Quanqin Shao, and Huanyin Yue. Surveying wild animals from satellites, manned aircraft and unmanned aerial systems (uass): a review. *Remote Sensing*, 11(11):1308, 2019.

[25] Xudong Wang, Long Lian, Zhongqi Miao, Ziwei Liu, and Stella X Yu. Long-tailed recognition by routing diverse distribution-aware experts. In *International Conference on Learning Representations (ICLR)*, 2021.

[26] Ruichi Yu, Ang Li, Vlad I Morariu, and Larry S Davis. Visual relationship detection with internal and external linguistic knowledge distillation. In *IEEE International Conference on Computer Vision (ICCV)*, 2017.