
DUAL-TASK LEARNING FOR SPECIES AND PHYSICAL ATTRIBUTES CLASSIFICATION

A PREPRINT

Johanna S. Karras*

Department of Computer Science
California Institute of Technology
Pasadena, CA 91125
jkarras@caltech.edu

ABSTRACT

Automated fine-grain species and physical attributes classification of animals are important, but challenging, problems for ecologists. State-of-the-art methods require large amounts of image labels and natural-language captioning, both of which are labor-intensive and time-consuming tasks. In order to reduce the required amount of labelled data for training neural network classification models, we introduce a jointly-optimized deep-learning model that leverages mutual information between species and physical attributes. Our model simultaneously predicts both the species class and a vector of detailed image attributes from an input image. Instead of full natural-language image captions, our approach utilizes binary feature vectors representing the presence or non-presence of expert pre-defined physical features. This choice of input reduces data-collection labor and time, while still providing valuable details. We test our method on the Caltech-UC San Diego Birds 200-2011 dataset and our key finding is that a dual-task approach improves multi-attribute image labelling F1-score by over 11%. As such, we show that our method is capable of providing more accurate fine-grain details of birds than a traditional single-task model, which is useful for ecological research and educational applications.²

Keywords Computer Vision · Machine Learning · Image Classification · Ecology

1 Introduction

Deep learning has been proven to be extremely useful in recent years for an array of computer vision problems, including fine-grain species classification and captioning in ecology. However, training deep-learning models for these tasks requires large amounts of labelled data. Although there is an abundance of unlabelled ecological data thanks to camera surveillance, citizen science applications, and lidar technology, labelling images is a labor-intensive and expensive task. As such, a persistent challenge in computer vision and ecology is training deep-learning models to classify species and provide fine-grain physical details of animals in images without requiring large amounts of labelled and captioned images.

Two state-of-the-art approaches to this challenge are transfer learning and multi-task classification. Transfer learning is a machine learning technique that utilizes neural network layers already trained for one task as a

^{*}This project was created for Professor Pietro Perona's EE/CNS/CS 148: Selected Topics in Computer Vision course at Caltech.
²

starting point for training another neural network for a different, but similar, task. For image classification, this often consists of using intermediate layers from a neural network previously trained on one task in order to extract deep features with which to train a new classifier for a related task [Pan and Yang, 2010]. Transfer learning has been shown to successfully reduce the amount of training samples needed to achieve high accuracy [Pan and Yang, 2010]. Another approach is to combine input data from multiple domains, such as visual data and natural-language image captions [Gebu et al., 2017] [He and Peng, 2017]. However, natural language captions are often labor-intensive and time-consuming to collect. Moreover, captions gathered from non-expert annotators are not guaranteed to contain useful discriminating and identifying information about the species in the image.

In this paper, we propose a blended transfer learning and multi-task learning model. Our **dual-task learning model** leverages mutual information between species class and physical attributes to simultaneously predict both the species class and a vector of fine-grain attributes of an image. Our intuition is that shared information in the dual-task approach will reduce the required amount of labelled training data required to achieve high accuracy predictions for both tasks. We explore our approach for classifying birds and predicting multi-label attributes in the Caltech UC-San Diego Birds (CUB) 200-2011 dataset. Instead of natural-language image captions, we use CUB’s crowd-sourced, binary attribute vectors representing expert pre-defined, visual features. As animal species contain visually distinct features, such as color, size, and pattern, attributes are easily labelled as present or not-present even by non-expert annotators. We implement and train two single-task models, one for species classification and one for attributes classification. Then, we implement a dual-task model for both species and attribute classification. Finally, we analyze the results of both single-task and dual-task classification networks with varying amounts of labelled training examples.

In the next section, we discuss previous work in transfer learning and multi-task image classification. Then, in section 3 we describe the CUB 200-2011 dataset. In section 4, we outline our single- and dual-task model architectures. In section 5, we discuss our evaluation metrics and experimental results. Finally, in sections 5 and 6, we discuss our findings and conclusions.

2 Related Work

2.1 Transfer Learning for Species Classification

Traditionally, species classification uses transfer learning to transfer the inner layers from a pre-trained convolutional neural network (CNN) trained on one classification task as black-box feature extractor for training a new classification [Pan and Yang, 2010]. However, such approaches do not leverage expert domain knowledge about which features are useful for differentiating and identifying species, risking learning coincidental correlations in the dataset. By fine-tuning the transferred feature-extractor layers for both species and attributes classification, our model weakly supervises the visual features used for species classification. Since the attributes are pre-defined by experts, this approach injects domain-knowledge into the model. Additionally, by outputting attributes along with class, our method provides additional information about the species classified in the image, which is useful for validating the reasonableness of the species prediction and for educational purposes.

2.2 Visual and Textual Inputs for Image Classification

Previous work in He and Peng [2017] explores the benefits of training bird species classifiers using both visual and natural language inputs. By fine-tuning an encoder using the CUB 200-2011 dataset to extract features and object region, they show improvements in object localization and classification accuracy. Similarly, Gebu et al. [2017] show the benefits of incorporating annotations along with visual data for fine-grain classification models in order to reduce the required amount of labelled images required for training.

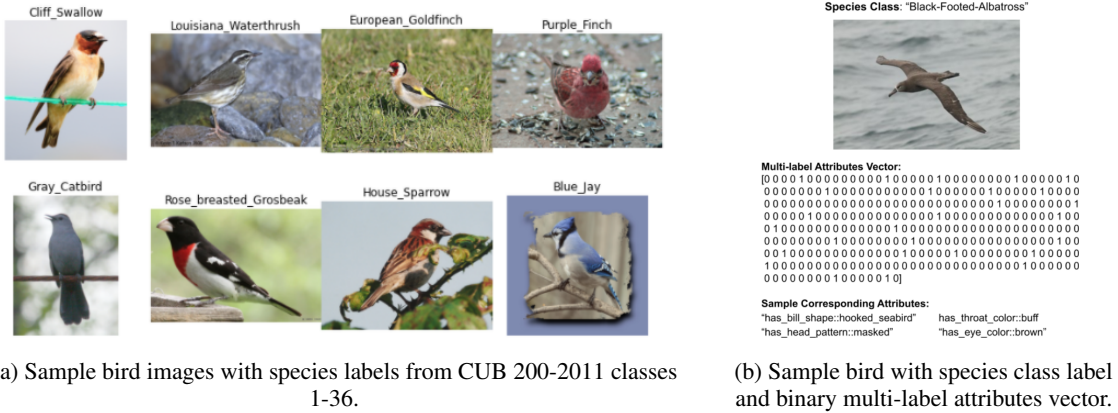


Figure 1: Visualization of the Caltech UC-San Diego (Birds) 200-2011 dataset.

We similarly explore visual and non-visual inputs for classification, but take an alternative approach by pairing the image input with a multi-label binary attribute vector. The benefit of our approach is that the task of attribute labelling is significantly easier and faster than that of writing a full, natural-language caption. Another benefit is that attributes classification is significantly easier to train than natural-language caption generation, while still providing useful information about the image. Finally, by specifying the attributes that annotators should label, we inject expert supervision into the data-gathering process, limiting the scope of possible attributes to the most important ones.

3 Dataset – CUB 200-2011

The Caltech-UC San Diego Birds (CUB) 200-2011 dataset contains 11,788 labeled images of birds in 200 species categories [Wah et al., 2011]. For each image, the dataset includes a 312-element vector of confidence scores corresponding to 312 fine-grain attributes, such as bill shape, wing color, eye color, etc. We provide a visualization of the dataset, species labels, and attribute vectors in Figure 1. These annotations are collected through the Amazon Mechanical Turk platform, where the annotators were not provided additional information about the species. To speed up training for the purposes of this paper, we train using the first 2000 images (36 species classes) of CUB 200-2011.

4 Methods

We propose a model that classifies bird species and bird attributes simultaneously. The intuition behind our model is that image features that correspond to certain visual attributes, like color, size, or shape, also discriminate and identify the overall image class. Additionally, by using pre-defined attribute labels specified by domain experts, we inject domain knowledge into model training. Attribute vectors also allow us to crowd-source only the most relevant image attributes even from non-expert annotators, avoiding irrelevant or non-discriminating information in free-form natural-language captions.

4.1 Binary Multi-label Image Attributes

For each image in the CUB 200-2011 dataset, we construct a 312-element binary vector where each element corresponds to one of 312 fine-grain attributes that are collected from human annotators. Attributes that are labelled “present” for the image are marked with a “1” in the attribute vector, and non-present images are marked with a “0”. The attribute arrays are sparse, with an average of only 29 “1”s in each array.

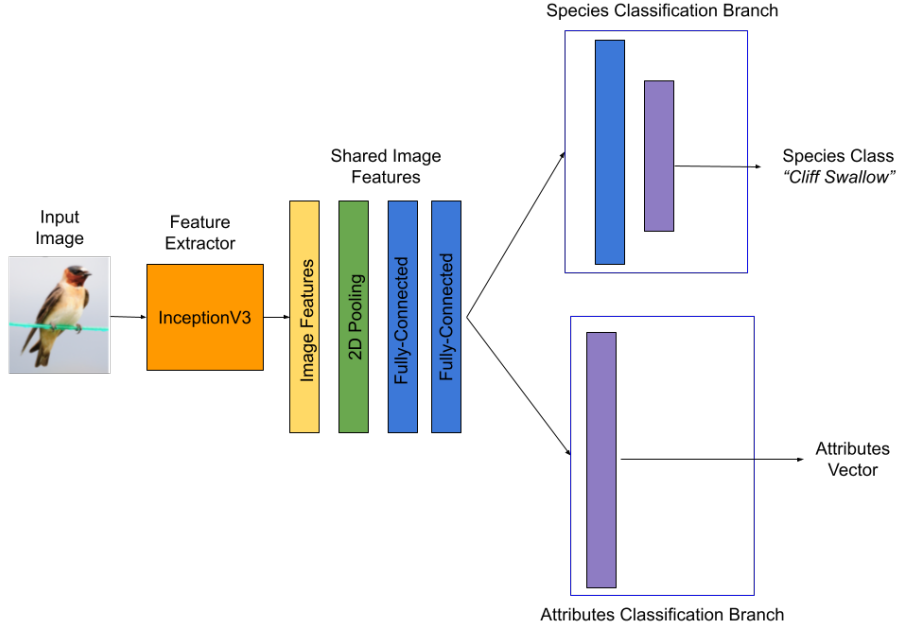


Figure 2: Dual-task model architecture for classifying bird species and binary attributes.

4.2 Single-Task Models

In order to contrast our dual-task approach, we implement state-of-the-art models for species classification and attribute multi-label classification separately. Both models use Google’s InceptionV3 CNN pre-trained on ImageNet as a feature extractor and are fine-tuned using CUB 200-2011 [Szegedy et al., 2015][Deng et al., 2009].

The species classifier’s decoder consists of three fully-connected layers connected to the final prediction layer of 36 nodes corresponding to each species class. The species classifier is trained with a categorical-crossentropy loss function between the true and predicted species label and Adam optimizer for 20 epochs with a learning rate of 0.0001.

The multi-label attributes classifier’s decoder consists of two fully-connected layers connected to the final prediction layer of 312 nodes corresponding to each attribute label. We set the confidence threshold to 0.2, above which labels are set to “1” and below which they are set to “0”. The attributes classifier is trained with binary-crossentropy loss between the true and predicted attribute vectors for 30 epochs with a learning rate that shrinks from 0.01 to 0.0001.

4.3 Dual-Task Model

Our dual species and attribute-classifier model learns to predict both species class and multi-label attribute vector simultaneously. It uses the InceptionV3 CNN pre-trained on ImageNet as a shared feature extractor on the input image [Szegedy et al., 2015][Deng et al., 2009]. We first jointly model the extracted features for both outputs with a pooling and two fully-connected dense layers. In these shared layers, mutual information between species class and image attributes can be shared and combined. Then, we separately train two branches, one for outputting species class and one for outputting attributes class. In order to provide a fair comparison with the single-task models, both branches including the feature-extractor shared layers are identical to their respective single-task architectures. The full architecture can be seen in Fig. 2.



Figure 3: Comparison of species prediction and of select true positive (TP), false positive (FP), and false negative (FN) attribute predictions made by the single-task and dual-task models. Included are also the total scores for the full 312-element vectors.

We use two different loss functions for each output, corresponding to the loss functions used by the single-task models. Namely, we use categorical-cross-entropy loss for the species class output and binary-cross-entropy loss for the attributes output. We train the dual-task model for 30 epochs with a learning rate that shrinks from 0.01 to 0.0001.

5 Experiments

We trained a single-task species classifier, a single-task attributes classifier, and a dual species- and attributes-classifier on 2000 CUB 200-2011 images pertaining to 36 different bird species. We partition the reduced CUB 200-2011 dataset into 80% training, 10% validation, and 10% test images with attribute labels.

5.1 Metrics and Baseline Scores

We evaluate the species classification score based on accuracy and F1-score and we evaluate attribute classification performance based on F1-score only. Since the binary attribute vectors are sparse, a high accuracy score is misleading even when there are lots of false 0’s in the predicted array. This can be seen in Fig. 1, where a baseline accuracy score for a random attribute classifier is 90.5%.

Accuracy and F1-score are given by

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \quad (1)$$

$$F1 - Score = \frac{TP}{TP + \frac{1}{2}(FP + FN)} \quad (2)$$

where TP is the number of true positives, TN is the number of true negatives, FP is the number of false positives, and FN the number of false negatives.

Each image belongs to one of 36 species classes and corresponds to 312 binary attributes. Each 312-element binary attribute vector contains a mean of 29 1’s and 283 0’s. Using a random species classifier and a random attributes classifier that assigns an element to “1” with probability $29/312 = 0.093$ as baseline models, we summarize their expected accuracy and F1-scores in Table 1. We would thus expect our species classifiers to reach accuracy of more than 0.278 and our attributes classifiers to reach F1-score of more than 0.171.

| | Species Classification | Attributes Classification |
|-------------------|------------------------|---------------------------|
| Baseline Accuracy | 0.278 | 0.905 |
| Baseline F1-Score | 0.054 | 0.171 |

Table 1: Baseline accuracy and F1-scores of random species and attributes classifiers.

5.2 Quantitative Results

We compare the results of the single-task transfer models and dual-task model trained on the full dataset in Table 2. We note that both methods significantly outperform baseline random chance scores for both classification tasks. Although the single-task species classifier outperforms the dual-task classifier for species classification by 10.09%, the classification F1-score for attribute classification by the dual-task model exceeds the single-task attributes classifier by 11.82%.

Table 2: Results of single-task and dual-task species and attribute classification for CUB 200-2011 dataset 1, with and without dual-task transfer learning from CUB 200-2011 dataset 2.

| | Species Classification Accuracy | Attributes Classification F1-Score |
|-----------------------------------|------------------------------------|---------------------------------------|
| Single-Task Species Classifier | 0.7995 | N/A |
| Single-Task Attributes Classifier | N/A | 0.4340 |
| Dual-Task Model | 0.6986 | 0.5522 |

Table 3: Comparison of species classification accuracy scores and attributes classification F1-scores for single-task and dual-task models. Scores in bold indicate the highest (best) score for each task.

Next, we evaluated the performance of single- and dual-task models when reducing the number of expert-labelled training examples. The results are plotted in Fig. 4. We observe that the dual-task model continues to dominate the single-task attributes classifier, even with only 40% of labelled training examples available for training.

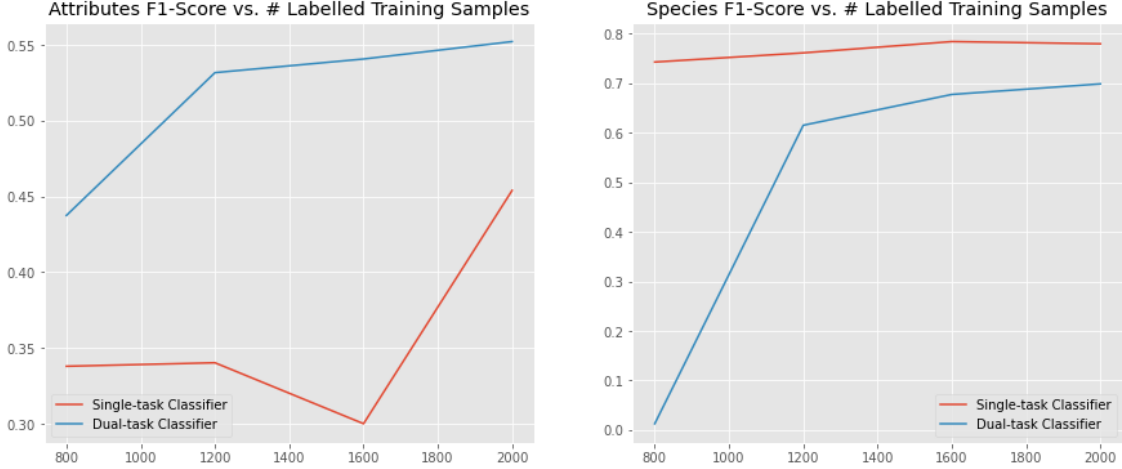


Figure 4: Species classification accuracy versus number of labelled training images for single-task, cross-task, and dual-task transfer learning models.

5.3 Qualitative Results

From our quantitative analysis, we can generalize that for the current implementation, the dual-task model is largely an improvement on the single-task model for attribute identification, but not so for species classification. Below, we summarize our positive and negative findings.

5.3.1 Success Cases

Our main success is showing that the dual-task model largely increases the F1-score for attributes classification. In Fig. 3, we show three examples comparing single-task and dual-task model attribute predictions. In each case, we show that the dual-task model **identifies bird attributes with a greater number of true positives and fewer numbers of false positives and false negatives** than the single-task model. In particular, we find that the dual-task model **better predicts colors, relative sizes of body parts, and fine-grain shapes** than the single-task model. The dual-task model performs especially well for birds that exhibit more common features or very visible features, while still identifying finer-grain details than the single-task model.

5.3.2 Failure Cases

Compared to the single-task species classifier, the dual-task model struggles with three examples of **incorrect false species predictions**. The dual-task model is an improvement on the single-task attributes classifier, but still prone to **false positive and false negative attribute predictions**. We include several examples of failure cases in Fig. 5.

We note that the even “incorrect” attributes that are not included in the ground truth are often still reasonable. For example, “has_shape::pigeon-like” is a true label, but “has_shape::perching-like” is a false label, even though “pigeon” and “perching” are likely indistinguishable to the non-expert annotator. As such, the ground truth annotations may have inconsistencies between similar attribute labels.

Our model **struggles with uncommon attributes**, as shown in the last example of Fig. 5. Our model easily learns to identify common bird features and colors, such as gray color and solid patterns, but more poorly for rare attributes, like green wings and red breast. This may also be a limitation of the selected species included in our training set.

6 Discussion

In our experiments, we demonstrate that dual-task learning can improve multi-label attribute classification and provide insights into physical species properties useful for identification.

Examples of Failure Cases




| Input Image | Species Class | True Positive Attribute Examples | False Positive Attribute Examples | False Negative Attribute Examples |
|---|--|---|--|--|
|  | True: "Shiny Cowbird" Single-Task Prediction: "Gray Crowned Rosy Finch" Dual-Task Prediction: "Black-Footed Albatross" | 'has_wing_color::black' 'has_upperparts_color::black' 'has_underparts_color::black' 'has_breast_pattern::solid' 'has_back_color::black' | 'has_bill_shape::all-purpose' 'has_size::medium(9-16in)' 'has_shape::perching-like' | 'has_bill_shape::cone' 'has_tail_shape::pointed_tail' 'has_shape::pigeon-like' |
|  | True: "Sooty Albatross" Single-Task Prediction: "Laysan Albatross" Dual-Task Prediction: "Black-footed Albatross" | 'has_underparts_color::buff' 'has_breast_color::orange' 'has_throat_color::orange' 'has_forehead_color::orange' 'has_size::very_large(32-72in)' | 'has_bill_shape::spatulate' 'has_wing_color::orange' 'has_back_color::orange' 'has_head_pattern::eyering' 'has_eye_color::orange' | 'has_bill_shape::needle' 'has_wing_color::rufous' 'has_tail_shape::pointed_tail' 'has_head_pattern::masked' 'has_eye_color::rufous' 'has_shape::owl-like' 'has_primary_color::rufous' |
|  | True: "Painted Bunting" Single-Task Prediction: "Gray Crowned Rosy Finch" Dual-Task Prediction: "Bobolink" | 'has_bill_shape::cone' 'has_breast_pattern::solid' 'has_eye_color::black' 'has_forehead_color::blue' 'has_nape_color::blue' 'has_size::small(5-9in)' | 'has_wing_color::blue' 'has_wing_color::black' 'has_upperparts_color::blue' 'has_upperparts_color::black' 'has_underparts_color::blue' 'has_back_color::blue' | 'has_wing_color::green' 'has_underparts_color::red' 'has_back_color::green' 'has_tail_shape::forked_tail' 'has_head_pattern::malar' 'has_breast_color::red' 'has_throat_color::purple' |

Figure 5: Examples of error cases where image species are misclassified by both classifiers. We also select several examples for each input image of true positive, false positive, and false negative attribute labels predicted by the dual-task model.

6.1 Applications

We hope that by providing improved attribute labeling for species, our model can improve ecological research and educational tools. For research, our model can be adapted to a variety of ecological domains to help inform researchers about key identifying features for different species. Additionally, our dual-task model offer. For education, our model can be used to learn about identifying features to look for in birds for those who are interested in learning to identify birds in the wild.

6.2 Limitations

Our work in this paper faces several limitations due to time and resources. First, the model is only trained and tested using the first 2000 images form CUB 200-2011 and that is in order to work within the limited EE/CNS/CS 148 course time-frame and available computing resource constraints. Second, the binary attribute vector is constructed by assigning a "0" or "1" if an attribute is labelled "not present" or "present", without taking into account confidences for these labels. As a result, some attributes that may actually be reasonable for an image with lower confidence are labelled "not present" in the ground truth. Additionally, some attribute labels could be indistinguishably similar to non-experts, such as "*has_nape_color::red*" and "*has_breast_color::red*", resulting in false positives and false negatives in the ground truth attribute vectors. Finally, given time constraints, we have not yet fully examined alternative model architectures that may improve results.

6.3 Future Work

In future work, we aim to improve our model architecture to achieve improved species classification accuracy, in addition to our already-improved attribute classification F1-score. We would also like to evaluate our feature-extractor that was fine-tuned using the dual-model architecture for species and attributes labelling as a transferred feature-extractor for a related species classification task, in order to see whether it would improve accuracy compared to an out-of-box

feature-extractor. Our ultimate goal is to be able to deliver more accurate species classification with the fewest possible number of expert-labelled images.

7 Conclusion

In this paper, we present a new dual-task, deep-learning model architecture for simultaneously classifying bird species and multi-label bird attributes. We demonstrate that a dual-task architecture greatly improves multi-label attribute prediction F1-score, with only a slight compromise to species classification accuracy. We hope that our model can empower educators and researchers to build more descriptive and detailed identification and tracking tools for species in the wild.

References

- Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359, 2010. doi:10.1109/TKDE.2009.191.
- Timnit Gebru, Judy Hoffman, and Li Fei-Fei. Fine-grained recognition in the wild: A multi-task domain adaptation approach, 2017.
- Xiangteng He and Yuxin Peng. Fine-grained image classification via combining vision and language. 2017. doi:10.1109/CVPR.2017.775.
- C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. The Caltech-UCSD Birds-200-2011 Dataset. Technical report, 2011.
- Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. *CoRR*, abs/1512.00567, 2015. URL <http://arxiv.org/abs/1512.00567>.
- Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.