Bird Classification – Kaggle Challenge

Megi Dervishi
École Normale Supérieure
45 Rue d'Ulm
megi.dervishi@ens.fr

Abstract

The challenge aims to classify 20 different bird species which were extracted from the Caltech-UCSD Birds-200-2011 dataset. The objective is to develop a model/ method that gives the highest possible accuracy on the test dataset which contains the same categories.

1. Introduction

The first thing we did to solve this challenge is pre-process the data by cropping the images around the most prominent bird. Second, we use data-augmentation and then try to optimize the hyperparameter of our model. Lastly, we trained a classification model on the new cropped dataset.

2. Dataset

The images in the validation dataset are centered and well-focused with a blurred background, whereas in the test set birds are overwhelmed by their background (i.e., hidden by trees or have a small size). This hinders the model to learn the features of the bird and instead focuses more on the background. To avoid that we thus crop the dataset.

Furthermore, we notice that the valid. set is unbalanced. We regroup and then randomly split the train and valid. dataset with ratio 0.9:0.10. In section 3 we train and evaluate on both datasets (balanced, unbalanced).

2.1. Cropped dataset

We use the Detectron2 [1] library from Facebook AI to detect birds in the dataset. First, we use the Mask-R-CNN model [1] and do instance segmentation on the dataset. Among the train, valid. and test data there are respectively 3,0 and 5 undetected bird images. We then do instance segmentation with PointRend [3] and panoptic segmentation [1] on the remaining images and get the most prominent bird. Figure 1 was the only bird not detected, so we cropped it manually. We crop the images according to the detected bounding box. If there is more than one bird detected, we take the one with the highest score.

Figure 1: Not-detected bird from test set.

2.2. Data Augmentation

In order, to avoid overfitting, the train images are randomly flipped horizontally with probability 0.4, and random sized sections are erased from the image with probability 0.2. The first transformation is to make the model not affected by changes in orientation and the second is to allow the model to learn bird features even though the bird may be hidden by other objects e.g. trees.

3. Method

We use the base 16x16 patch size and 224x224 vision transformer [2] pre-trained on ImageNet21K and finetuned on ImageNet1K. The model is then concatenated with a classifier (hidden size in table1) and is trained on the new cropped datasets (balanced and unbalanced, but we do not observe a significant performance difference).

First, we optimize the hyperparameters of the model via the python package ax-platform. The Bayesian optimization has 20 steps, and each step does 3 epochs of training. Due to GPU constraints, we couldn't optimize the batch size, so it is set to 16. The final hyperparameters are:

	LR	Weight decay	Momentum	Hidden size
Balanced	0.085	3.67e-05	0.193	604
Unbalanced	0.065	3.53e-05	0.207	512

Table1: Optimized hyperparameters for (un)balanced dataset

We train for 30 epochs with CrossEntropy loss and SGD optimizer and get the following results:

	Acc on valid	Avg. loss	Acc. kaggle
Balanced	94%	0.0133	87.096%
Unbalanced	93%	0.0154	89.677 %

Table2: Results. Accuracy (Acc.) and Average (Avg.) loss

4. Conclusion

Although we are happy with the results, we believe that there could have been other improvements. Since the biggest bottleneck of this challenge is the low data, one could use self-supervised learning on a larger unlabeled set of the NABirds/iNaturalist. One could also use more advanced augmentation e.g., deep augmentation.

References

[1] Yuxin Wu, Alexander Kirillov, Francisco Massa and Wan-Yen Lo, & Ross Girshick. (2019). Detectron2. https://github.com/facebookresearch/detectron2

[2] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *Int. Conf. Learn. Represent., 2021*

[3] Alexander Kirillov and Yuxin Wu and Kaiming He and Ross Girshick. PointRend: Image Segmentation as Rendering. (2019). *ArXiv:1912.08193*