Assignment 3: Bird Classification Challenge

Michaël Soumm
ENSAE
michael.soumm@ensae.fr

Abstract

Image classification can be a difficult task when classes are such that only a specialist could manually distinguish between them. Moreover, in such specific tasks, training data is often small. To overcome this problem, we propose to use transfer learning, cropping, and self-supervised learning.

1. Introduction

From [1], we download a bird dataset of 1082 images evenly distributed amongst 20 classes. After inspection, the validation set seems not representative of the train set, so we re-generate it. With so little images, and their quality, it is necessary to work on them to enhance their informational content.

2. Bird detection

We use Mask R-CNN [2] pre-trained on COCO dataset to crop the images on the birds. We select only bounding boxes associated with birds, and chose the one with the highest probability, if it is above 85%. After inspection, only 25 birds were not detected.

3. First-stage model

We use a Vision Transformer model [3], pretrained on ImageNet. This choice is motivated by performance, as ResNet and EfficientNet models were also tested with far lesser results. After unfreezing the last 3 blocks (trained at a smaller learning rate), changing the head layer, and adding a final classification layer, we train the model for 10 epochs on the non-cropped dataset, with an Adam optimiser. We continue training of this *hard model* on the cropped images for 10 epochs to get an *easy model*, without refreshing the optimiser. Convergence is reached in 2-8 epochs.

4. Second-stage model

We propose to augment the CUB dataset with the NaBirds dataset [4], by using pseudo-labeling [5]. We predict the classes of 2428 unlabeled images, using the *hard model* on the NaBirds dataset, and the *easy model* on the cropped NaBirds dataset (with provided bounding boxes). In order to have a minimum number of mis-labeling, we only add the new images to our dataset if both *models* give the same output, or if one of them has a high confidence. Our final augmented datset thus contains 2925 images. We note that in second-stage we do have a class inbalance, which we overcome with a weighted sampler. A figure of the full model is available in the appendix (Figure 1).

5. Results

The following table presents our results on cropped and non-cropped images, with and without self-supervised learning. Test images were cropped whenever possible.

Method	Orig. val.	Cropped val.	public test
First-stage	95%	97%	87,7%
Second stage	97%	98%	87,1%

Table 1. Results on original val. dataset, cropped val. dataset, and public test dataset

6. Conclusion

Even if the second stage model seems to have a smaller public test score, we see an improvement in the validation score. As the public test dataset is very small (˜150 images), we will submit test predictions for both first and second stage models.

References

[1] Willow, "Recvis20 a3." Available at https://github.com/willowsierra/recvis20_a3. 1

[2] K. H. et al., "Mask r-cnn," 2018. 1

[3] A. D. et al., "An image is worth 16x16 words: Transformers for image recognition at scale," 2020. 1

[4] C. L. of Ornithology, "Nabirds dataset." Available at https://dl.allaboutbirds.org/nabirds. 1

[5] D.-H. Lee, "Pseudo-label : The simple and efficient semi-supervised learning method for deep neural networks," *ICML 2013 Workshop*, 07 2013. 1

Appendix

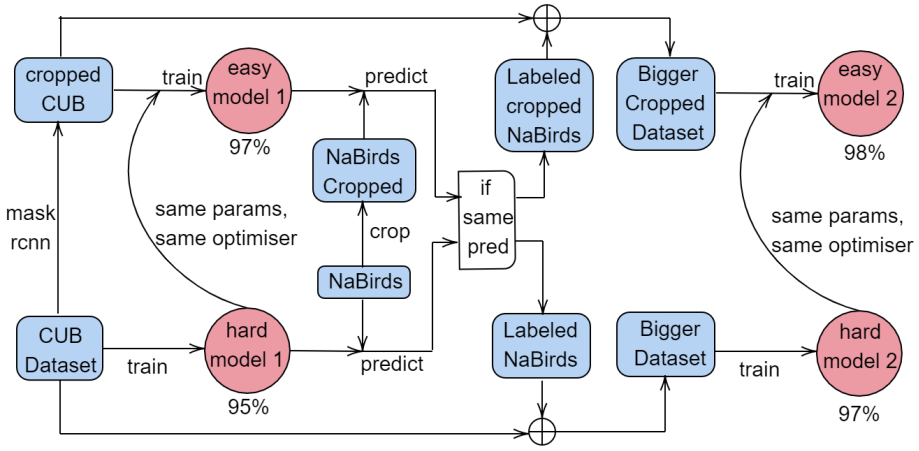

Figure 1. The full model pipeline