MVA RecVis 2020 Assignment 3: Bird image classification competition

Vincent LIU
ENS Paris-Saclay
liuvincent25@gmail.com

Abstract

The objective of this competition is to produce a model that gives the highest accuracy at recognizing different species of birds from a subset of the Caltech-UCSD Birds-200-2011 bird dataset. This report presents my solutions.

1. Methods and results

1.1. Baseline on original images

First, I merged the provided training and validation data into a single folder, so I could split the merged data to carry out a cross validation (CV) kfold strategy, in order to have robust measurement of the performances. I changed the vanilla convnet to an Imagenet pretrained convnet. What worked best was finetuning EfficientNet[9] architecture. The architecture as well as the input size had quite an impact: changing from 224×224 to 300×300, from ResNet18 to EfficientNet was better on both my CV and public leaderboard (LB).

Input	Model	CV	Public LB
224×224	ResNet18	80.58 ± 3.3	63.87
224×224	Efficient $B1$	84.2 ± 3.1	71.61
300×300	Efficient $B3$	92 ± 1.6	74.19

Table 1. Preliminary Results on original images (CV 5folds)

1.2. Bird detection and cropped image classification

The testing data is significantly different from the training/val data, which explains the gap between LB and CV. I decided to crop the birds using pytorch Mask R-CNN [1] trained on COCO dataset. Using these images pushed my LB score to 78.71, but val score was lower. The reason could be that the model trained on cropped images generalizes better on the difficult test images.

Methods	Validation	Public LB
B4	89.58 ± 1.1	78.71
B4 with PL	91.39 ± 0.5	80.64
B4 + SeResNeXt + TTA	93 ± 0.5	81.9

Table 2. Results on cropped images with input size 300×300. Validation is one fold, average over best epochs.

1.3. Pseudo Label and Unlabeled external data

Afterwards, I used external unlabeled data from NABirds dataset. I performed the pseudo label (PL) [7] strategy which consists of training with both training and unlabeled data. I had a small increment of the LB to 80.64.

1.4. Final submission: Ensemble models with TTA

Last model is a bit more sophisticated. We generate bounding box coordinates with Mask R-CNN for each training image. Then at training time, the bounding boxes are increased by a random small factor at each load, then we apply data augmentation such as random erasing [12], rotation, horizontal flipping and color jitter on the cropped image. At test time, we crop the test images with both Mask R-CNN and RetinaNet [8] to obtain two representations of each image. We predict the probability for each representation with an ensemble of checkpoints of EfficientNet B4 and SE-ResNeXt [6]. Finally, we aggregate the probability distribution corresponding to each image to obtain the final label. My final public CV is 93 ± 0.5 and public LB is 81.9.

1.5. What did not work during the competition

I tried lot of things that unfortunately did not work. I tried Central Loss [10] to improve the neural network ability to find fine-granularity between categories that are visually close. Also, I tried knowledge distillation [2] to leverage the unlabeled external data but it resulted to noisy labels. I tried a fancy data augmentation approach that is Mixup [11] but results couldn't increase my LB score. I used other people's implementation for mixup [3], central loss [4] and label smoothing [5] mentioned in the file *lossfunction.py*. The rest is an adjustement from provided starter code.

References

[1] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn, 2018.

[2] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network, 2015.

[3] https://github.com/facebookresearch/mixup cifar10/blob/master/train.py. Code for mixup.

[4] https://github.com/KaiyangZhou/pytorch-center loss/blob/master/center_loss.py. Code for center loss.

[5] https://stackoverflow.com/questions/55681502/label-smoothing-in pytorch. Code for label smoothing.

[6] Jie Hu, Li Shen, Samuel Albanie, Gang Sun, and Enhua Wu. Squeeze-and-excitation networks, 2019.

[7] Dong-Hyun Lee. Pseudo-label : The simple and efficient semi-supervised learning method for deep neural networks. 2013.

[8] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection, 2018.

[9] Mingxing Tan and Quoc V. Le. Efficientnet: Rethinking model scaling for convolutional neural networks, 2020.

[10] Yandong Wen, Kaipeng Zhang, Zhifeng Li, and Yu Qiao. A discriminative feature learning approach for deep face recognition. volume 9911, pages 499–515, 10 2016.

[11] Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization, 2018.

[12] Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang. Random erasing data augmentation, 2017.