Object recognition and computer vision - Assignment 3 - BLANCHARD Théau

A. Introduction

The goal of this challenge was to obtain good accuracy on the classification of bird from a dataset containing 200 different classes of birds adapted from the CUB-200-2011 dataset. This amounted to a transfer learning and data augmentation challenge. I choose to apply an augmentation pipeline to an RCNN followed by a ResNet50.

B. The dataset

The proposed dataset is very small. There is only 1082 train images, equally distributed amongst the 20 classes but only 103 validation images with only up to 2 images for some classes. The final result are computed using the 517 test images.

The main issue lies in the fact that these set are not of equal difficulty. Most of the training/validation images are clear pictures of a single bird, in good resolution and with little to no occlusion. Unfortunately, many test images are of a very different quality. Sometimes the bird is barely visible, either because of poor resolution, capturing condition, or occlusions.

These remarks on the dataset highlight the necessity to take into account these distribution shifts in the training process.

C. Data augmentation

Given the wide variety of image shape I wanted to first find where the bird was in the image before doing any more processing on the image. For this purpose I used the Faster R-CNN [3] network in inference mode. It allows to find the bird in the image and give a bounding box around it. Given the bounding box I am able to crop a square fitting this box so that the image is not weirdly stretched when passed further down the processing pipeline.

In order to account for the small number of available training images we will generate new images whilst also trying to perturb their distributions so as to mitigate the difference between the testing and training distributions.

First we use classic data augmentation tricks such as random rotations, flips, or crops with random Gaussian blurring. Yet these transformations weren't sufficient to achieve good accuracy. In this perspective I decided to use a more heavy and complex data augmentation pipeline based on AugMix [2]. This pipeline involves a serie of random complex transformations affecting the spatial structure of the image but with a limited impact on the color profile of the image so as not to have an augmented image switch class.

D. Classification architecture

Given the small quantity of data and the accuracy standard I wanted to reach, I resolved to transfer learning. But also given that I had time and computational capacity constraints I chose an relatively light architecture based on a ResNet [1]. This specific architecture (*resnet50*) was trained on the ImageNet dataset and reached a 80.85 % accuracy. I adapted the network so as to switch the last fully connected layer with an another fully connected layer leading to the 20 classes of the dataset. During the training all layers were unfrozen and fine-tuned.

Finally the network was trained on 40 epochs until convergence, with a learning rate $\gamma = 0.00001$ using batches of size 64.

E. Conclusion

Finally thanks to this pipeline I managed to achieve a 81.935% accuracy on the public testing leaderboard.

These results could probably be improved by using a more complex network (eg. a Vision Transformer) preceded by an even more complex data augmentation pipeline and with some fine hyper parameter tuning.

References

[1] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015. 1

[2] Dan Hendrycks, Norman Mu, Ekin D. Cubuk, Barret Zoph, Justin Gilmer, and Balaji Lakshminarayanan. Augmix: A simple data processing method to improve robustness and uncertainty. 2019. 1

[3] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks, 2015. 1