Birds classification Challenge : Object recognition and computer visions

Youssef DAOUD
ENS Paris-Saclay
youssef.daoud@ens-paris-saclay.fr

Abstract

Most machine learning applications often involve classification tasks with many classes. Such tasks have become increasingly important within the research field in recent years. This report will present different methods and computer vision techniques used in order to perform a multi class classification task.

1. Introduction

The data challenge we tackled is a multi class classification problem that aims to predict the right class among the 20 existing classes. The data available in this work is a subset of *Caltech-UCSD Birds-200-2011* dataset. The data in is divided between 1082 training images, 103 images for validation and 517 test images.

2. Preprocessing

While visualizing some images, I observed that the bird can occupy just a little space in the whole image (either in the center or in the sides of the image). Then, instead of doing a random crop, I used a pre-trained YOLO model on MS COCO dataset in order to detect the bird in the image and yield the corresponding bounding box. Once, I have the bounding box of the detected bird, I crop the image and store it in a new dataset. During this process, I have encountered some challenges such as :

- False detection : To tackle this problem, I have used a higher threshold confidence 0.4.
- Detection of other objects : For this case, I specified the id of the wanted object (in our case $\text{id}_{\text{bird}} = 14$).
- Detect a bird twice : Here, I specified for the model to have just one output for each image.

3. Approaches

In this work, I have tried several approaches to tackle this classification problem. in the following sections I will talk about two main approaches that yield best performances.

3.1. A pre-trained CNN model

Since the data is too small, training a new model from scratch is not going to be a great idea. Then, seeing that we have several pre-trained models on ImageNet dataset, it could be smart to use one of them and fine-tune on our data. In fact, I have tried different pre-trained models but the one that yields better performance is called *MobileNet*. So, I have extracted the last layer of the CNN model and add two linear layers. After each one, I added a Dropout layer with a rate 0.5. Additionally, I have tried to do some data augmentation on our dataset. To do so, I performed the following transformations : **RandomRotation**, **VerticalFlip**, **HorizontalFlip** and **ColorJitter**.

3.2. Vision Transformer

Recently, Vision Transformers (ViT) achieved very competitive performance on benchmarks for several computer vision applications, including image classification, object detection, and semantic image segmentation. For this purpose, I have used a pre-trained ViT model on *ImageNet-21k* dataset with a resolution of 224×224. Then, I have trained the model on the cropped dataset (The output of YOLO on our original data). I have also tried to use data augmentation for this model, but it does not boost the performance.

4. Results

Approach	Val score	Test score
YOLO + DA + MobileNet	76%	69%
ViT	90%	74%
YOLO + ViT	**94%**	**88.3%**

Table 1. The accuracy of different implemented approaches

5. Conclusion

This challenge gave us the opportunity to put our Computer Vision knowledge into practice. In this work, we could see the strength of using Transfer Learning, Object Detection models and Transformers in such classification tasks.