Confidence is All You Need

Yacine Mohamed Bouaouni
mohamed.bouaouni@ens-paris-saclay.fr

Abstract

The goal of this assignment is to build a classification model for 20 species of birds from Caltech-UCSD Birds dataset. The model is evaluated based on its accuracy on the test set. In this paper, I propose a solution based on well studied data augmentation techniques and an ensemble learning model relying on classifiers confidence score.

1. Introduction

A first exploratory data analysis shows that the training set is balanced with 52-58 samples for each species. In contrast, the validation set is unbalanced and has classes with only 2 samples, which is quite a small number to evaluate the performance of the model on these classes and the validation performance metrics won't be a good approximation of the test performances. Dealing with this problem requires calling for data augmentation techniques [2]. In this paper, I describe first, the data augmentation techniques used. Then, I describe the model and the maximum confidence ensemble prediction.

2. Data Augmentation.

The size of the dataset is quite small (1082 training sample and 103 validation sample), this may lead the CNN to overfit, unless we augment the dataset using geometric and color space transformations. The transformations used in this solution are: horizontal flipping, Changing the perspective, sharpening, gaussian bluring with a kernel $k = 7$ and $\sigma = 2$, resizing to 386 and random crop a 336x336 and change the brightness and slightly the hue of the image.

3. Model.

3.1. Transfert Learning.

The dataset provided is not sufficient to train a deep learning model on, this is a case where transfer learning comes into the picture. After testing many CNN architectures pre-trained on ImageNet, from small ones like ResNet-18 to more deep networks (e.g. ResNet-152), I build my solution on top of two different versions of ResNet models [1] (Resnet 152 and ResNext 101 32x8d). I froze all the layers of the ResNet except the last block of convolutions (layer4) and the fully connected layer. The reason is that the last layers of the pre-trained models learn high level features of the data, so it's better suited if these layers are trained on the data of our application. The training is performed using Adam with a step learning rate scheduler, starting at a value of $lr = 10^{-4}$ and decreasing each 5 epochs with a factor $\alpha = 0.75$. The models are trained independently for 10 epochs on a batch of 64 images of shape 300x300. The two models reached an accuracy on the validation set 91% for ResNet152 and 90% ResNet 101 32x8d.

3.2. Duplicate Predictions.

The test set is duplicated using the training set augmentation transforms. The purpose behind this technique is to predict many times the label of a single test image (multiple transforms). However, I considered only flip, perspective and crop transforms, because predictions for the other transforms are very different from the original test images predictions (Similarity of only 60%).

Each test image, will have 4 attributed classes (one for each transform) with 4 confidence scores, the selection of the predicted label will either be a max voting or a maximum confidence selection (MCS). The MCS leads to better accuracy (2% on test set). The same procedure is applied for both two models and then the final class prediction will be based on the confidence score of each classifier for that test sample. In simple, words if for a certain test image the first model is more confident then the output will be its prediction. The final test prediction is $\mathbf{83.87}\%$.I tested the classification on the segmented and the cropped images using Mask-RCNN and Faster-RCNN, but it did not improve the results.

References

[1] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition, 2015.

[2] C. Shorten and T. M. Khoshgoftaar. A survey on image data augmentation for deep learning. *Journal of Big Data*, 6(1), July 2019.