Object Recognition and Computer Vision: Image Classification

Matias Etcheverry
Ecole Normale Supérieure
matias.etcheverry9@gmail.com

Abstract

This report aims at describing the full pipeline used to perform species classification on a subset of the Caltech-UCSD Birds-200-2011. First, we will review the dataset and its specifities. We will then study the model and the frameworks used for the training. Finally, we will provide the performances of the model.

1. Dataset

1.1. Decomposition

The studied dataset is a subset of `Caltech-UCSD Birds-200-2011` [3] divided in 3 sets: the training, the validation and the test dataset, made of images containing birds of 20 different species. The objective of the Kaggle challenge is to produce a model which best classify a bird species lying within an image. Fig. 1a shows a batch of images in the training set. The difficulty of this dataset is that it is very small, with only 1,702 images with 517 test images. I noticed that the validation set was very small, and some classes had only 2 validation images. I firstly restructured the datasets, so that each class has 8 validation images. I also cropped every image so that it is focused on the bird, using a a pretrained model on ImageNet [1].

1.2. Data Augmentation

A key to counter the lack of data is to use data augmentation. This process aims to modifying the training images so that the model generalizes better. We create a pipeline which performs random rotations (with $\pm 20°$), crops keeping between at least 40% of the original image, horizontal and vertical flipping and erasing of less than 70% of the original image. Fig. 1b shows a batch of augmented images in the training set.

2. Architecture of the model

The resources available to me are limited in time and computation. Thus, we start with model pretrained on `ImageNet`. The choice of the pretrained model is very important. There is a trade-off between choosing a model with high performances, like top-tier transformers and having an easily trainable model with few parameters like Resnet. In the end, I chose a Vision Transformer of 86M parameters based on SWAG weights [2]. We need to modify the head of the pretrained model so that it fits our tasks. To do so, we add a fully connected block of 2 dense layers, a tanh activation, and a dropout layer.

(a) Original batch. (b) Augmented batch.

Figure 1. Data augmentation on a whole batch.

To perform the training, we use the Adam optimizer coupled with a learning rate scheduler and the cross entropy loss. The whole architecture is firstly frozen except the fully connected block to train it. The convergence is done within 5 epochs. We then unfreeze the whole architecture and train it again on 20 epochs. Finally, we perform a simple grid search on the hyper-parameters and We save the model reaching the best validation accuracy.

3. Results and conclusion

Using the above model, we reach a validation accuracy of 91.88% and a validation loss of 0.357 which results in a public test accuracy of 85.16% in `Kaggle`. Firstly, we notice that there is still a gap between the validation accuracy and the test accuracy, despite the restructuring of the dataset. Thanks to the confusion matrix, we also notice that the model struggles a bit on fined grained classification, ie on birds that look alike. Moreover, with more time, it would have been interesting to study few shots learning classification as it may suit the objective better.

References

[1] N Carion, F Massa, G Synnaeve, N Usunier, A Kirillov, and S Zagoruyko. End-to-end object detection with transformers, 2020. 1

[2] M Singh, L Gustafson, A Adcock, V Reis, B Gedik, RP Kosaraju, D Mahajan, R Girshick, P Dollár, and L van der Maaten. Revisiting weakly supervised pre-training of visual perception models, 2022. 1

[3] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. Caltech-ucsd birds-200-2011. Technical Report CNS-TR-2011-001, California Institute of Technology, 2011. 1