Bird classification on the Caltech-UCSD Birds-200-2011 dataset

Gabriel Fiastre
Computer vision and object recognition
gabriel.fiastre@dauphine.eu

Abstract

For this assignment the goal was to build a deep neural network model for bird classification on the Caltech-USCD Birds-200-2011 dataset. I managed to build a model with 90.9% *accuracy on the test data.*

1. Introduction

The problem focuses on 20 classes (precise bird species). The data is divided into a train set (1082 images) and a validation set (103 images) with both known label, and a test set with unknown label (517 images). The different classes are very close to each other : some bird species differentiate only by small details such as wings shape, slight color variations (and so on).

2. Data exploration and preprocessing

Data augmentation techniques were considered at first for the training set in order to increase data and avoid overfitting. After a first phase of data exploration, it appeared clear that the birds could be randomly oriented and were presented with different enlightenment conditions. Therefore the random tranforms of rotation horizontal flip, as well as colorJitter could be applied. Their respective intensity have been chosen by fine-tuning : a rotation up to 50 degrees appeared to be effective, whereas the color variation were more effective when small (up to 0.1 for both contrast and brightness). The hue parameter of the colorJitter was set to 0 as color was a very important information for identifying the class of a bird. A normalization was also applied to all images.

3. Transfer learning

For this problem it was of great benefit to make profit of the various models which were already trained on ImageNet. The accuracy metric on 30% of the test set was used to choose the best model.

The first models implemented were of the ResNet family : The ResNet50 model with first 6 pretrained layers modules freezed and learning on the 4 remaining achieved 77.4% accuracy while being computationally cheap.

The second models to be considered were the VIT models (Vision transformers) as they can specialize very well with an impressive accuracy on imageNet. The vit_l_16 model converged quickly to results as good as more expensive models such as vit_h_14. The SWAG_E2E_V1 pretrained weights were chosen on the validation set. The whole architecture was frozen, with a small learning MLP network added on top (1, 2, 3 or 4 layers). Dropout before each added layer except the last one was also used to avoid overfitting. $p = 0.25\%$ was the most efficient. The models with 1, 2,3 and 4 layers MLP got respectively 88.3%, 90.3%, 90.3%, and 88.3% accuracy.

For the training convergence, I used SGD fine-tuned with a momentum of 0.9, a starting learning rate of 0.01. A ReduceLROnPlateau scheduler was added to reduce the learning rate ($\times 0.1$) every 4 iterations without finding a new validation loss minimum (cross-entropy), and stopping at 10^-5 (convergence reached). A batch size of 64 was used for the ResNet models and 28 for the VIT models.

4. Final model and leads for improvement

Horizontal ensemble-voting was implemented between models on the ouptut probability per class. The best combination was obtained by the two best VIT models (with 2 and 3 layers MLP on top), when an accuracy of 90.9% was reached. It is a slight improvement, but the model should be more robust that way as it retains the prediction from the most confident model.

There are some leads for improvement : ensemble voting should be implemnted with more different models to make the detected features vary and avoid both models doing the same mistake at the same time. Implementing a mask r-CNN to localize the birds could also add more accuracy.

5. Conclusion

The 2 final models submitted were : the ensemble voting of vit_l_16 with 2 and 3 layers MLP, and the second one alone. They achieved 90.9% and 90.3% accuracy.