Assignment 3: Fine-Grained bird classification

Siwar Mhadhbi
Télécom Paris - Master M2 MVA
siwar.mhadhbi@telecom-paris.fr

Abstract

In this project, our aim is to develop a scheme for Fine-Grained classification of bird species on a subset of the Caltech-UCSD Birds-200-2011 and produce a model that gives the highest possible accuracy as part of a kaggle In-class competition.

1. Introduction

Deep Convolutional Neural Networks (CNNs) have shown exemplary performance on several Computer Vision related tasks. CNNs have shown prevalent performance on the ImageNet[2] dataset. Thenceforth, they have been used via transfer learning[1] in multiple tasks with small datasets. This is where **our main methodology** comes in.

2. Data preparation

2.1 Validation set

We have in all 1702 images divided as 1082 for training, 103 for validation and 517 for test. The training images are almost equally distributed in 20 species. However, the validation set presents 8 for some species and just 2 for others which is not enough to represent a category of birds. Thus, we thought about redistributing the samples so that we obtain 20% of the training samples for each category.

2.2 Crop bird regions

Birds in our dataset are not all well recognized, some are too small, some are blurry, some present occlusions and some appear in background instead of foreground. Thus, to help the algorithm recognize better the birds, we opt for cropping. We apply **faster R-CNN**[6] to predict bounding boxes around birds and we apply cropping when the prediction confidence on birds is higher than 0.7.

2.3 Data augmentation

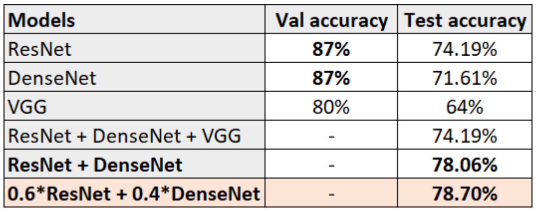
To compensate the small amount of the training set, we apply data augmentation techniques such as *horizontal flips* and *rotations* which will help increase the variability of the inputs. We also resize all images to the highest same size (300, 300) constrained by memory usage on Google Colab.

3. Modeling

We use different pre-trained models on ImageNet: **Resnet-50**[3], **Densenet-161**[7], and **VGG-16**[4]. As early layers of deep neural networks learn general aspects of objects and last ones extract more specific features depending on the underlying task, we chose to freeze the early layers of each pre-trained model and only fine-tune the last ones. We set for training *SGD as optimizer*, with a *learning rate of 0.005*, and a *batch size of 64*. Everytime the validation accuracy stops increasing for 5 successive epochs, we multiply the learning rate by a *factor of 0.8*. We train each model for 100 epochs then select the best model for testing. First, we test **each model** apart. Then, we try ***Ensemble technique*** which combines these base models for a better predictive model. Based on each model performance, we try ***Weighted ensemble technique*** with empirically fixed weights.

4. Results

ResNet50 provided the best accuracy 87% on validation set and 74% on test set. With ensemble learning on ResNet and DenseNet, we increased the accuracy by 4% and we reached 78.06% on test set. The Weighted Ensemble technique improved further the accuracy to **78.70%**.



Models	Val accuracy	Test accuracy
ResNet	**87%**	74.19%
DenseNet	**87%**	71.61%
VGG	80%	64%
ResNet + DenseNet + VGG	-	74.19%
ResNet + DenseNet	-	**78.06%**
0.6*ResNet + 0.4*DenseNet	-	**78.70%**

5. Conclusion

My approach combines data preparation, transfer learning, and weighted ensemble technique to obtain finally a Kaggle *public* accuracy of 78.70%. This can be further enhanced not only by testing a combination with other different models that may lead to a better understanding of the birds characteristics, but also by testing a **Vision Transformer (ViT)** which has been shown to provide excellent results [5] compared to CNNs.

References

[1] S Bozinovski and A Fulgosi. The influence of pattern similarity and transfer learning upon training of a base perceptron b2. In *Proceedings of Symposium Informatica*, pages 3–121, 1976. 1

[2] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 1

[3] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 1

[4] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017. 1

[5] Maithra Raghu, Thomas Unterthiner, Simon Kornblith, Chiyuan Zhang, and Alexey Dosovitskiy. Do vision transformers see like convolutional neural networks? In *Thirty-Fifth Conference on Neural Information Processing Systems*, 2021. 1

[6] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28:91–99, 2015. 1

[7] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 1