Object Recognition and Computer Vision 2019 - Assignment 3

Antoine Yang
Ecole Polytechnique - ENS Paris-Saclay
antoineyang3@gmail.com

Abstract

The goal of this assignment is to design a model that has maximum test accuracy on Image Classification of a subset of 20 *classes and* 1702 *images of the Birds200 dataset (2).*

1. Introduction

A data analysis shows that the task is hard. This motivates a data augmentation achieved by detecting birds with a pretrained Mask-R-CNN model (3). Then I extract features with a pretrained Inception3 model (1) and train a 2-layer network to perform classification. Notebooks run on Google Colab and clean (unused) python files are provided.

2. Data Augmentation

2.1. Data Analysis

In some images, the bird seems very hard to distinguish; classes are balanced in terms of size in the training set, but not in the validation set, which is also quite small; a t-SNE embedding shows classes are very close to each other.

2.2. Bird Detection and Horizontal Flipping

I use a pretrained (on ImageNet) Mask R-CNN model with ResNet+FPN backbone implemented in the Detectron2 framework (3) to detect birds. Training, validation, and test datasets are augmented with images of highly confident expanded bounding boxes. With a detection threshold of 0.7, it only fails to detect birds in 1.76%, 2.91% and 3.63% cases of respectively the training, validation and test sets. In this case, the image is horizontally flipped. Then, following Test Time Augmentation heuristic, the test set is duplicated with its horizontal flipped version.

3. The classifying bird model

3.1. Feature extraction

I use an Inception3 model pretrained on ImageNet and more importantly iNaturalist2017 (1). From each image resized at 299×299, I extract a 2048-feature vector. At this stage, I have 4096 features for each training and validation image, 8182 for each test image.

3.2. Feature Classification

I use a 2-layer network with 512 hidden neurons. The optimizer is SGD with momentum 0.9, weight decay 3×10^{-4} and gradient clipping 5. The learning rate follows a cosine annealing schedule starting from 0.01 for 50 epochs, with batch size 32. The best model is obtained with early stopping after 21 epochs and achieves 95% validation accuracy. Averaging predictions on the test features and test features obtained with horizontal flipping, I achieve 92.258% accuracy on the public Kaggle leaderboard.

4. Discussion

Code provided also includes feature map visualization script, other models tried (Yolo detector, stacked EfficientNet and Resnet, Resnext), and the implementation of different tricks such as Cutout, Mixup, Pairwise Confusion, Label Smoothing, finetuned Resnext features extraction and merging with Inception3 features, rebalancing the validation set. The superiority of the selected approach inherently comes from a better detector and a feature extractor already pretrained on a fine-grained image classification task.

5. Conclusion

I presented an efficient approach (which can be run in less than two hours on Google Colab) achieving 92.258% accuracy on the public Kaggle leaderboard and the work that led to this approach.

References

[1] Yin Cui and al. Large scale fine-grained categorization and domain-specific transfer learning. https://github.com/richardaecn/cvpr18-inaturalist-transfer, 2018.

[2] Catherine Wah and al. The caltech-ucsd birds-200-2011 dataset. 2011.

[3] Yuxin Wu and al. Detectron2. https://github.com/facebookresearch/detectron2, 2019.