Object Recognition and Computer Vision: Assignment 3

1. Introduction

In this challenge, I used different approaches to tackle the multi-classification problem of birds. I used transfer learning using pre-trained models on "Imagenet", along with bagging. I also explored a state of the art architecture using an attention approach. I will explain briefly my ideas and present my results.

2. Prepossessing

Since we had 1185 images in the training and validation data set and 517 images in the test-set, I immediately thought of using transfer learning to tackle the challenge. Exploring the validation and test dataset for every class, I found that the images were not equally distributed in the validation set. For that reason, to have a good validation metric, I regrouped the images in a way that we end up with 85 % in the training set and 15 % in the validation set. This representation allows us to have a good local metric and without a leak to the training dataset. I resized the images to 400×400

2.1. Image cropping

Since the images were taken from a variety of angles, there was a lot of background noise. I used FasterRCNN pretrained on the COCO dataset having resent50 as a backbone to detect the birds before feeding them to the model.

2.2. Data augmentation

To deal with the small dataset that we have, I decided to use data augmentation techniques (only for the training dataset). I ended up choosing only random rotation and random horizontal flip.

3. Architectures

3.1. Attention model

I found it interesting to explore a state-of-the-art model that uses the attention mechanism [1].I choose to work on the paper "See Better Before Looking Closer: Weakly Supervised Data Augmentation Network for Fine-Grained Visual Classification". The idea behind this paper, alongside the data augmentation that I applied, is to select some specialized crops in the image since random crops increase the background noise. We represent the parts of the birds or visual patterns by attention maps that are generated by convolutions at a high level to extract sequential local features to solve the fine-grained classification problem. It applied to do that 2 operations, attention cropping, to select one of the attention maps that have been generated in an advanced neuron layer and to do cropping around this attention area to improve the representation of the local features and attention dropping, to encourage the network to focus also on other parts of the model so that it doesn't overfit.

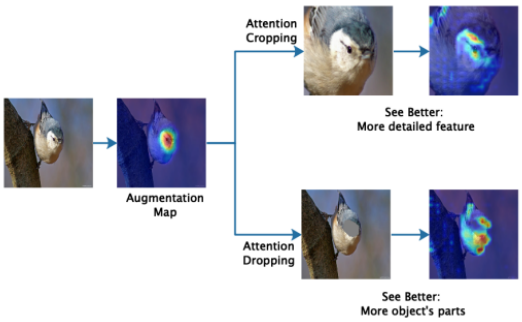

Figure 1. See Better: Attention maps represent discriminative parts of the object.

At the end, we pass the input image, the cropped version and the dropped version to the network in order to do the training. I tested this architecture both using Resnet50 and inception v3 as backbone based on this implementation. [2]

3.2. Transfer learning models

For transfer learning, I trained Resnet50, SEnet50 (Squeeze-and-Excitation) [3],Resnet101,and EfficientNet-b5 all loaded with Imagenet weights (For bagging).To speed up the training, I froze some of the first layers(The number depends on the architecture of the model). For Resnet50 for example, I choose to freeze the first 6 layers. For all these models, I chose to work with SGD (lr=0.005) with momentum(0.8) as an optimizer. Also, to avoid overfitting I used a scheduler (ReduceOnPlateau) and early stopping (20 epochs). I used Batch size = 16.

4. Results

Different results		
Model name	Validation score	Private test score
Resnet 50 (6 layers frozen)	89.44%	80%
Resnet 50 (6 layers frozen without cropped data)	91.07%	76.12%
Bagging model(Resnet50+EfficientNetb5+ SeNet50+Resnet101)	92.74%	80.64%
Attention model	94%	80%
Attention model without cropping	91%	76.77%

For the Public test, I will consider the highlighted models as my final submissions.

References

[1] T. Hu, H. Qi, Q. Huang, and Y. Lu, "See better before looking closer: Weakly supervised data augmentation network for fine-grained visual classification," *arXiv preprint arXiv:1901.09891*, 2019. 1

[2] https://github.com/wvinzh/WS_DAN_PyTorch. 1

[3] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7132–7141, 2018. 1