Detection of Terriers Dogs Using ResNet on the StanfordDogs Dataset

Master in Data Science
Course: Computer Vision and Deep Learning

Authors:
Samuel Metin
Maxime Aparicio

Date: June 16, 2025

Contents

1 Motivation and Rationale

Detecting and classifying dog breeds is a significant challenge in computer vision, especially when dealing with morphologically similar breeds such as various subcategories of Terriers.

The problem addressed is the detection and correct classification of dog breeds, with a focus on Terrier breeds. This issue is relevant in veterinary domains, breeding, or for consumer applications such as mobile breed recognition apps.

2 State of the Art

Currently, fine-grained object recognition heavily relies on deep convolutional neural network (CNN) architectures, with ResNet18 and ResNet50 being standard models.

Pre-trained models on ImageNet have demonstrated excellent performance on general image classification tasks but may struggle on fine-grained tasks distinguishing between highly similar classes.

Fine-tuning techniques allow these pre-trained models to be adapted to specific datasets by re-training certain deep layers. However, current methods may face difficulties in distinguishing morphologically similar breeds (overfitting, lack of specific data, etc.).

The objective is to leverage these models while performing full retraining after fine-tuning to improve performance on closely related sub-classes.

3 Objectives

The main objectives of this project are the following:

- **Compare ResNet18 and ResNet50:** Evaluate and compare the performance of the two architectures, both using pre-trained weights and training from scratch, in the context of fine-grained dog breed detection.
- **Analyze the benefit of pre-training:** Measure how transfer learning and fine-tuning improve the model's performance compared to training a model entirely from scratch.
- **Evaluate the specialization effect:** Verify whether training the model specifically on Terrier breeds leads to better performance when detecting Terriers, compared to detecting similar but unseen breeds such as Staffordshire.
- **Study the impact of dataset size:** Assess how increasing the number of training images (from 700 to 3600) affects the performance of the models, especially when training from scratch.

4 Methodology

4.1 Data

The dataset used is Stanford Dogs, consisting of approximately 20,000 images across 120 breeds. A subset focused on breeds close to Terriers will be constructed. We made 3 train datasets, one with only Terrier dogs, compose of 200 images, the second with 624, and the last with 3679 images, of a large set of race, all close to terrier. For the test dataset, we took 10 images of Terrier (so a dataset close to the train one), and 10 images of Staffordshire(so a dataset far from the train one), in order to verify that the test result are better for terrier dog than on Staffordshire.

4.2 Models

Two ResNet architectures will be employed: ResNet18 and ResNet50, pre-trained on Coco.

4.3 Training

For both model, we used Fine-tuning and Full training of the entire network. And then, resnet 18, from scratch, with 3600 images.

4.4 Loss Function

We use the Loss Function return by the model Resnet, COCO mAP (Mean Average Precision), So it is a Loss Function based on IoU.

5 Experiments and Results

5.1 Training

- Faster RCNN_Resnet 18, from scratch with 700 images
- Faster RCNN_Resnet and then fine tune with 700 images
- Faster RCNN_Resnet 18, and then fine tune with 3600 images

5.2 700 Images

This table summarizes the results, we use the mAP [IoU=0.50:0.95]

Model	**Terrier**	**Staff**	**Average**
Resnet 50 pretrain	0.855	0.799	0.827
Resnet 50 from scratch	0.090	0.089	0.090
Resnet 18 pretrain	0.587	0.411	0.499
Resnet 18 from scratch	0.146	0.204	0.175
Average	0.4195	0.40475	X

Table 1: result with 700 Images

As expected the pretrain model did very well, their result for the Terrier test are quite superior than the ones with the Staff, this show that the training was effective and that the fine tuning worked. The model from scratch had equivalent results for Terrier and Staff, so they did not learn and they result may come from lucky guess. The worst model is the Resnet 50, from scratch. Because this model is larger than Resnet 18, and therefore need more images to get to its full pentitial.

The pretrain model are already so in the following, we focus more on upgrading the from scratch model. Futhermore, Resnet 50 will need more images that Resnet 18, so we focused only on resnet 18.

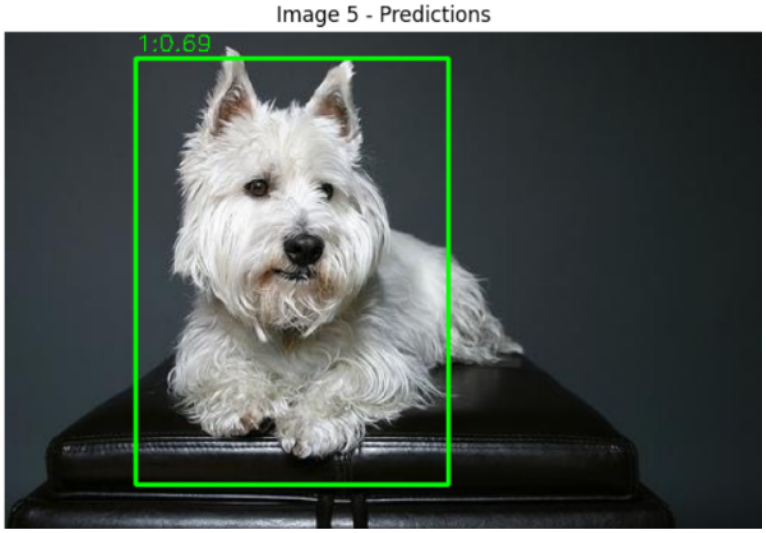

Figure 1: Terrier - RES NET 18 - from scratch

5.2.1 Qualitative result

Figure 2: Staff - RES NET 18 - from scratch

The model we have fully train still get some nice result, like to identifing ths Staff (as we only focus the training on terrier dogs) and getting the Bonding Box correct for the Terrier.

5.3 3600 Images

In this section, we fully trained the model Resnet 18 on 3600 images. For the Staff, we got an average precision of 0.313, and 0.443 for the Terrier. The difference show that the model indeed learn from the Terrier dataset. It is also a surprise that this model got better result on Staff than the previous from scratch model even if it has not seen a Staff in its training. This show that the model may have learn what the shape of a dog is. For the reminder, the pretrained Resnet 18 scored 0.411 on Terrier. So training a model from scratch is as good as getting a pretrain one, and fine tuning it on a unprecise dataset.

5.3.1 Qualitative result

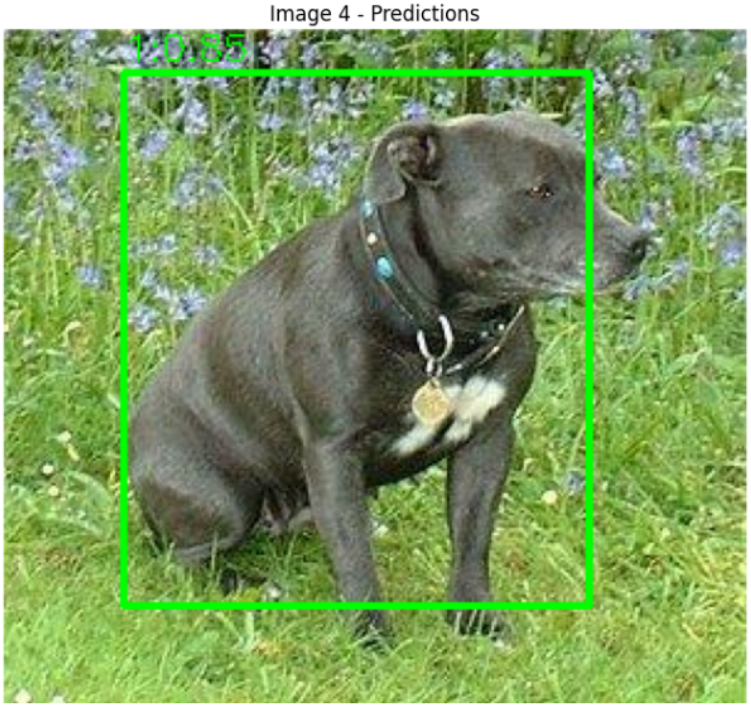

Figure 3: Staff - RES NET 18 - from scratch

On this image, the model has improved.

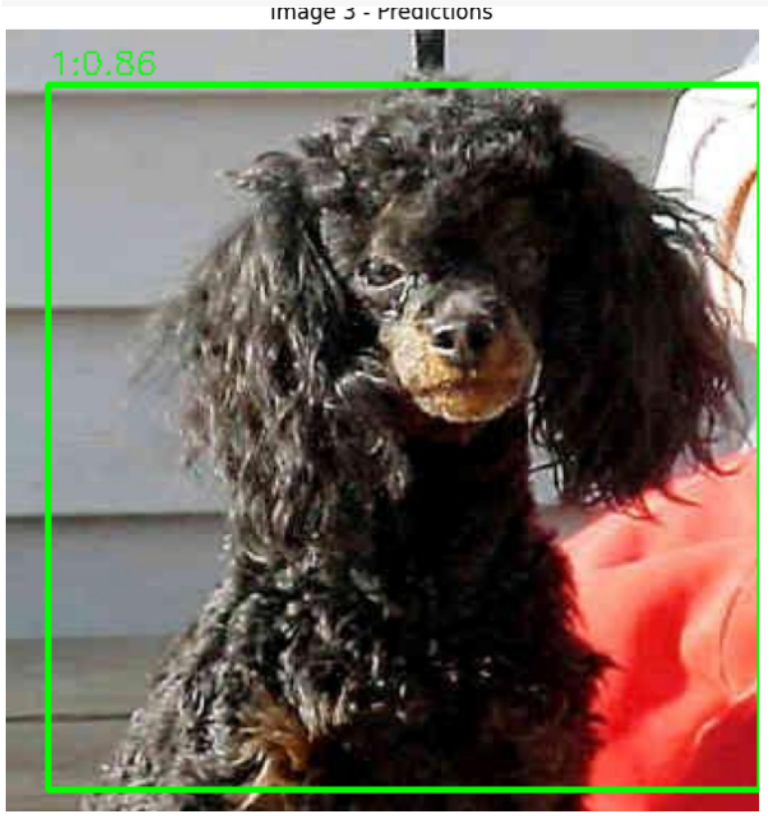

Figure 4: terrier 1- RES NET 18 - from scratch.

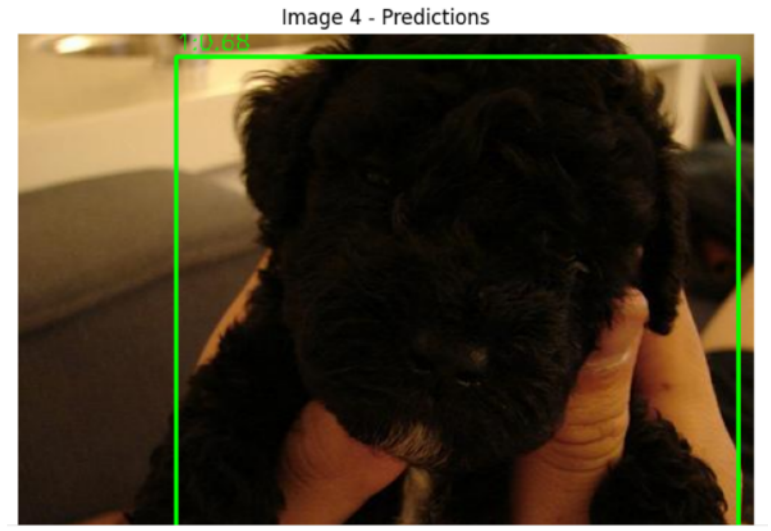

Figure 5: terrier 2- RES NET 18 - from scratch

Futhermore, the model still reconise the dogs even with diffents luminosity levels and shades

6 Conclusions

In this project, we addressed several objectives. First, we compared the performances of ResNet18 and ResNet50. As expected, ResNet50 showed better performance due to its larger capacity. However, the ResNet50 model trained from scratch was far from reaching its full potential, mainly because it requires a much larger amount of training data to fully exploit its capacity.

Secondly, we observed that a model specifically trained on Terrier breeds performs better on Terrier images compared to Staffordshire images, confirming that the model specializes in the data distribution it has been exposed to during training.

Finally, we evaluated the impact of increasing the size of the training dataset. Increasing the number of images from 700 to 3600 significantly improved the model's performance, demonstrating the importance of having a sufficient amount of data for effective training, especially when training models from scratch.

Future work may explore training with even larger datasets, using more advanced architectures such as Vision Transformers, or integrating domain adaptation techniques to improve generalization across similar breeds.

7 code

https://github.com/Ratiio13/Projet-Computer-Vision-deep-learning