Object Recognition and Computer Vision: Assignment 3

Gauthier Multari
ENS Paris-Saclay
gauthier.multari@ens-paris-saclay.fr

Abstract

The goal of this report is to describe the process used to train different Deep Learning models in order to find the best suited to an image classification task. The target dataset is the classifysketch dataset. We will first describe the training strategies used to compare the different models before discussing the best model found for this assignement.

1. Introduction

The target dataset is the ClassifySketch dataset [1].It consists of sketchs drawn by humans for given categories. In total there are 20,000 sketches evenly distributed in 250 object categories, making it 80 images for each category. The goal of the Kaggle challenge constituting this assignment is find a model that can accurately classify these sketches. The main issue regarding this dataset is that it has a very low amount of images for each category.

2. Data augmentation

In order to train the models, the dataset is split into train, validation, and test subsets. The train subset is constituted by 60% of the data, the validation by 10% and the test 30%. This leads to a an even lower amount of images for each class. To help alleviate this issue we use data augmentation techniques. The train datasets are augmented by adding random horizontal and vertical flips, as well as random rotations. We chose not to overdo the the data augmentation by adding techniques like random cropping because we felt like it would have a detrimental effect on the data given its simplicity.

3. Model selection

Given the amount of time given for this assignment and the computing power available (a Google Colab T4 GPU, and a laptop RTX 3060 6Gb 130W GPU) as well as the low amount of data available, the choice was made to focus on pretrained models. For the same reasons, we focused on models of small size that have achieved state of the art performance. The chosen models are: mobilenet_v2 (3.4M paramters), ResNet18 (11.7M), ResNet34 (21M), ResNet50 (26M) and ViT-16 (86M).

Two different classifier layers have been tested, a simple linear layer and a more complex constituted of a sequence of six layers. A linear layer followed by a ReLU and a Dropout repeated one and ending with a linear layer.

Various tests were performed with differerent learning rates, momentum and batch size values. Given the time requiered to perform such experiments many did not converge.

The different runs were logged using WandB.

4. Results and discussion

The model that gave the best results is the ViT-16 one. The weigths used to submit for the kaggle competition led to a 70% accuracy on the validation set as well as a 70% accuracy on the Kaggle test set.

A common point with all the models is that they all had a way better accuracy on the training dataset than the test one. This means that the models were overfitting. A better classification head could help solve this issue as well as of course more data. A smarter loss function would definitely help as well.

References

[1] Mathias Eitz, James Hays, and Marc Alexa. How do humans sketch objects? *ACM Trans. Graph. (Proc. SIGGRAPH)*, 31 (4):44:1–44:10, 2012. 1