Object recognition and computer vision 2024 : Assignment 3

Amer Essakine
ENS Paris Saclay
amer.essakine@ens-paris-saclay.fr

Abstract

The assignment involves a classification task using a dataset comprising 500 distinct classes of sketches, adapted from the ImageNet Sketch dataset. This dataset poses a significant challenge due to the high intra-class variance characteristic of freehand sketches. Sketches of the same object can differ substantially in style, detail, and representation, adding complexity to the task.

This short report outlines the dataset and the preprocessing steps undertaken, followed by a detailed description of the model architecture, the fine-tuning process, and the methodology used to optimize performance.

1. Dataset

The dataset comprises 22,500 sketch images, divided into 20,000 images for training and 2,500 images for validation, and 500 distinct classes is considered.

1.1. Prepocessing

To address memory and computation time constraints, it was necessary to downsize the images. Initially, I attempted resizing the images to 64x64; however, this resolution was incompatible with DinoV2, which requires image dimensions to be multiples of 14. I also experimented with 70x70 images, but this resulted in lower accuracy. Ultimately, cropping the images to 224x224 proved to be the most suitable approach, effectively balancing accuracy and training time.

For augmentation, I tested various transformations. Random cropping yielded worse results compared to center cropping, which is expected since most sketches are drawn in the center of the image on a white background. Rotation also failed to improve accuracy, as the images generally lack orientation issues, given that the sketches are centered. Similarly, random erasing and blur performed poorly, likely because the sketches are drawn on a white background that contains no additional information.

An additional idea was to explore an augmentation technique specifically tailored for sketches, involving the deformation of sketch curves. This approach was investigated in [2], where the Bézier Pivot Deformation method was introduced as a relevant augmentation strategy. I attempted to adapt this method; however, the provided code was available only in MATLAB, and due to time constraints, I was unable to fully implement it and get it to work.

Finally, the best result were obtained by resizing the images to 224x224 and normalizing them.

1.2. Method

I used the pre-trained DinoV2 ViT-B/14 model, loaded through torch.hub, to extract features for the classification task. The input images were first passed through the DinoV2 backbone to generate feature embeddings. These features were then normalized using the model's built-in normalization layer before being passed to the classifier for training.

Initially, I froze all the parameters of the backbone model, which resulted in an accuracy of 81%. Next, I experimented with freezing 50% of the parameters, which improved the results.I conducted a grid search and found that freezing 75% of the backbone parameters yielded the best performance.

In addition, I utilized autocasting during the training phase to accelerate the process and optimize performance by leveraging mixed precision computation. This approach dynamically adjusts the precision of operations, using FP16 where possible to speed up calculations and reduce memory usage, while retaining FP32 for operations requiring higher numerical stabilit. This method improved the accuracy obtained.

Finally, I used a learning rate of $510^{-}3$, which decayed to $510^{-}4$, starting from the seventh epoch. I also experimented with assigning a separate learning rate to the unfrozen parameters of the backbone model and tried the learning rate recommended by the original authors, but neither approach improved performance.

I intended to train additional models (such as SwinV2), and perform weight averaging [1] across these models to improve performance. However, due to time constraints, I was unable to implement this approach.

References

[1] Mitchell Wortsman, Gabriel Ilharco, Samir Ya Gadre, Rebecca Roelofs, Raphael Gontijo-Lopes, Ari S Morcos, Hongseok Namkoong, Ali Farhadi, Yair Carmon, Simon Kornblith, and Ludwig Schmidt. Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 23965–23998. PMLR, 17–23 Jul 2022. 1

[2] Qian Yu, Yongxin Yang, Feng Liu, Yi-Zhe Song, Tao Xiang, and Timothy M Hospedales. Sketch-a-net: A deep neural network that beats humans. *International Journal of Computer Vision*, 122(3):411–425, 2017. 1