Object Recognition and Computer Vision
Assignment 3 Report

Borachhun YOU
École Normale Supérieure Paris-Saclay
borachhun.you@ens-paris-saclay.fr

Abstract

The objective of this assignment is to produce a model that achieves the highest possible accuracy in an image classification task. The dataset that is used consists of a total of 20000 sketch images of 250 different object categories. The performance of the model is evaluated on a test set in a Kaggle challenge.

1. Dataset

The original dataset given for the image classification task in this assignment contains 20000 unique sketch images of 250 different object categories, each with a size of 1111×1111 pixels. The dataset is divided into 3 different sets: a training set and a validation set of 12000 and 2250 images respectively for building models, and a test set of 5750 images for evaluating the final accuracy in a Kaggle challenge.

2. Data preprocessing

The given dataset goes through multiple steps of preparation before being used for training and evaluating models. Firstly, each image is downscaled to 224×224 pixels due to memory constraint and so as to improve the training speed. Moreover, all the data are normalized to zero mean and one standard deviation based on statistics collected from the ImageNet dataset. In addition, the training set is augmented with additional images transformed from the original images. This step is crucial as it helps reduce overfitting and improve generalization of trained models which leads to better overall performance. For this assignment, the horizontal reflection of images in the training set are augmented to the set itself. Additional transformations can also be applied to the original images, for instance, rotating by a certain degree, in order to obtain even more data. However, the doubled training set with the reflection alone already takes significantly more training time that with just the original dataset, which is why only one transformation is used for the data augmentation.

3. Architecture of the models

Various models are tested in this assignment to compare for the best performance. Furthermore, transfer learning technique is applied on all models due to computational power constraint and also to reduce the training time and improve generalization of the models. Specifically, each model was previously pretrained on Imagenet dataset.

The first model that is used is ResNet50, which is chosen to see how the residual architecture of the 50-layer deep network performs with the given dataset. The next model is the transformer-based Vision Transformer, more specifically, the ViT Large variant which consists of 24 encoder layers. Another model that is tested is a more recent CNN named ConvNeXt. Proposed in 2022 by Liu et al., the CNN was designed to compete with transformer-based models by making adjustments to the ResNet architecture such that it bears a resemblance to transformers. The adjustments include increasing kernel sizes, using GELU activation functions, having fewer activation functions and normalization layers, and more.

As for the training procedure, the stochastic gradient descent optimizer is used with 0.1 learning rate. Furthermore, the classification head of all models are replaced to match the number of classes in the dataset. Additionally, only parameters in the classification head are optimized during the training process of 10 epochs while the rest of the parameters remain frozen.

4. Result

Based on the table below, ResNet got the highest training accuracy but the lowest validating and testing accuracy, while it is the opposite for ConvNeXt, which achieves the highest testing accuracy of 66%.

Model	Train	Validate	Test
ResNet50	0.7431	0.3275	0.3489
ViT Large	0.7354	0.5417	0.5530
ConvNeXt Large	0.7220	0.6520	0.6666