Ryan Ott
14862565
ryan.ott@student.uva.nl

Practical Assignment 2
Deep Learning 1

2023-11-27

1 Transfer Learning

1.1 Comparing Models

1.1.a Inference Speed per Image against Accuracy and Number of Parameters

With a Pearson correlation coefficient of 0.85 there seems to be a strong positive linear correlation between the top-1 accuracy and the inference speed. The ViT-B/32 model performs best overall with a higher accuracy for its inference speed than the trend suggests. For inference speed against number of trainable parameters, we see no correlation. The number of trainable parameters should only play a role when the model is being trained, not during inferencing, so this result matches our expectations.

1.1.b Inference Speed per Image with and without torch.no_grad()

With torch.no_grad() the inference speed should be faster because gradients are not computed and stored for the backward pass within the context manager's scope, thus requiring less operations and saving compute. This is confirmed by the results, the inference speed is slightly faster with torch.no_grad() across the models.

1.1.c vRAM Usage with and without torch.no_grad()

Like with the inference speed, the vRAM usage should be lower with torch.no_grad() because gradients are not stored for the backward pass when performing tensor operations, freeing up memory. Much more considerably than the inference speed, the vRAM usage is lower when using torch.no_grad() as seen in the plot.

1.2 Fine-tuning

1.2.a Retraining ResNet-18's Fully Connected Layer

Using the specifications and default hyperparameters from the assignment, the retrained model achieved a test accuracy of 0.5855 in CIFAR-100.

1.2.b Increasing Model Performance using Data Augmentation

By applying `torchvision.transforms.RandomHorizontalFlip(p=0.5)` around half of the images in the training set are flipped horizontally. This should increase the model's performance because it is being trained on more varied data, thus making it more robust to different inputs. The model achieved a test accuracy of 0.5927, a small improvement.

1.2.c Last vs First Convolutional Layer

The first convolutional layers tend to capture lower-level features like edges, corners or textures, while the last layers capture higher-level features like shapes and objects in a larger receptive field. As such, better performance should be achievable by fine-tuning the last layers (along with the classifier layers) because they are more specialized to the downstream task than the first layers.

2 Visual Prompting

2.1 CLIP Baseline

2.1.a Top-1 Accuracy on CIFAR-10 and CIFAR-100 with CLIP ViT-B/32 Backbone

Dataset	Train Accuracy (%)	Test Accuracy (%)
CIFAR-10	88.726	88.940
CIFAR-100	63.564	63.150

Table 1: Zero-shot CLIP Top-1 Accuracy on CIFAR-10 and CIFAR-100