Image Classification on CIFAR10

Sruthi Reddy Rekula
August 06 2023
https://github.com/shruthireddyrekula/Image_classification_cifar10_tf

Observations for CIFAR-10 data using different models (MLP, ResNet-50, and Vision Transformer) with the model's performance on the dataset and insights from the results. Here are the observations for each model:

Multi-Layer Perceptron (MLP):

- MLP model used here consists of an input layer, two hidden layers with dropout, and an output layer. The model is designed for a classification task with 10 output classes. The 'relu' activation function is used in the hidden layers to introduce non-linearity, and 'softmax' is used in the output layer for multi-class probability estimation.
- The CIFAR-10 dataset is proprocessed with normalization where Pixel values[0,255] of images are scaled to the range [0, 1], flattened the images and One-Hot Encoding for converting labels to binary vectors
- The hyperparameters used during training are dropout rate of 0.2, batch-size of 128, adam optimizer with categorical crossentropy loss and 100 epochs
- There are slight signs of overfitting since the training accuracy and loss are slightly better than the validation accuracy and loss. However, the differences are not substantial, and the model seems to be learning to some extent from the data.
- Overall Analysis: The MLP model does not perform exceptionally well on the CIFAR-10 test dataset. The test accuracy of approximately 50.49% suggests that the model struggles to generalize to unseen data effectively. This may be due to the limited capacity of the MLP architecture in capturing complex patterns present in the CIFAR-10 dataset, which contains images of various objects and backgrounds.

RESNET50

- The ResNet50 model with the settings(`include_top = False, weights='imagenet',input_shape=(32,32,3))` are used to extract features from images and then added a fully connected layer on top of the base model to perform classification, object detection or other image-related work. The pre-trained model weights from ImageNet will help the model capture meaningful patterns in the images and make it easier to train for new tasks
- The CIFAR-10 dataset is proprocessed with normalization where Pixel values[0,255] of images are scaled to the range [0, 1], flattened the images and One-Hot Encoding for converting labels to binary vectors
- The hyperparameters used during training are learning rate of 0.001, batch-size of 64, adam optimizer with categorical crossentropy loss and 100 epochs
- The model's performance is quite good, as it achieved high accuracy on both the training and validation datasets, and it was able to generalize well to new, unseen data in the test dataset. The test accuracy is in line with the validation accuracy, indicating that the model is robust and capable of making accurate predictions on real-world data.
- Experiment with different learning rates, batch sizes, and optimizers to see if the model's performance can be further improved.
- The ResNet50 model shows significantly higher accuracy on the training set compared to the MLP, indicating that it is better at learning the representations of the training data. However, there is a noticeable drop in accuracy on the validation and test sets, suggesting some degree of overfitting. Despite this, it still outperforms the MLP on the validation and test sets

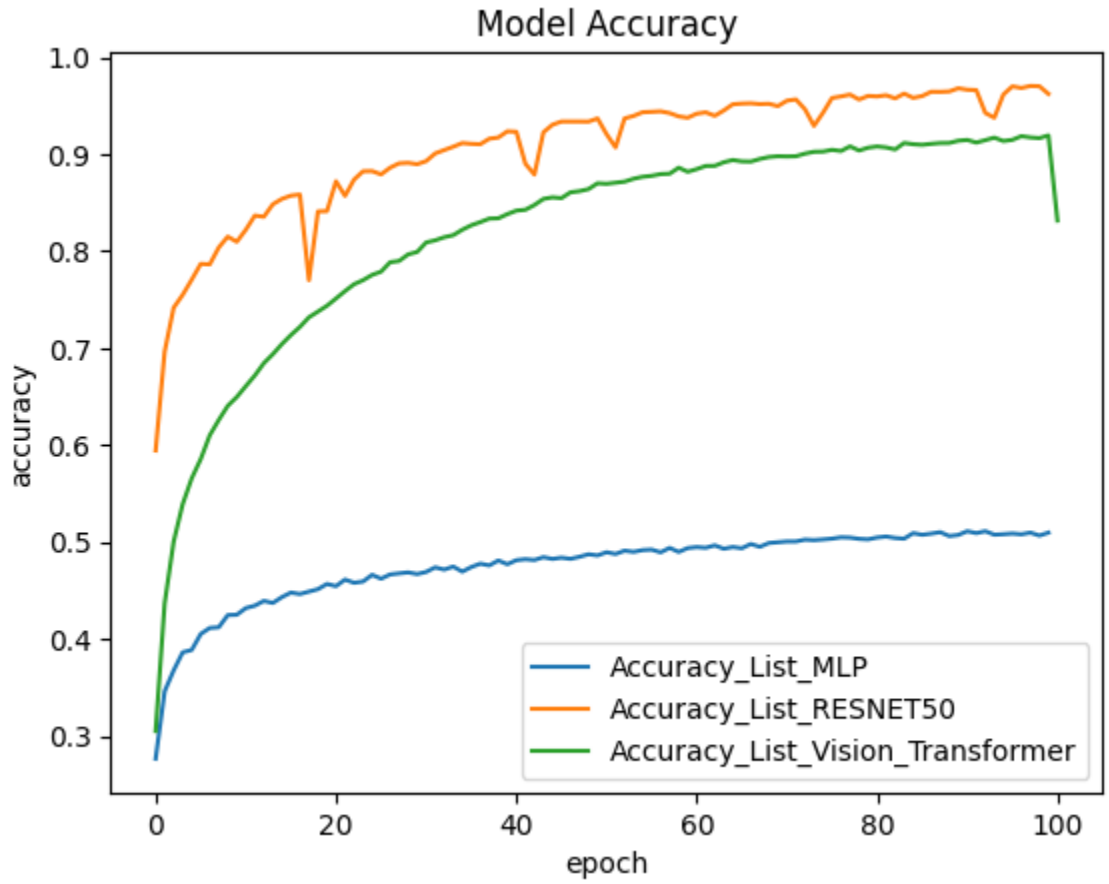
Vision Transformer (ViT):

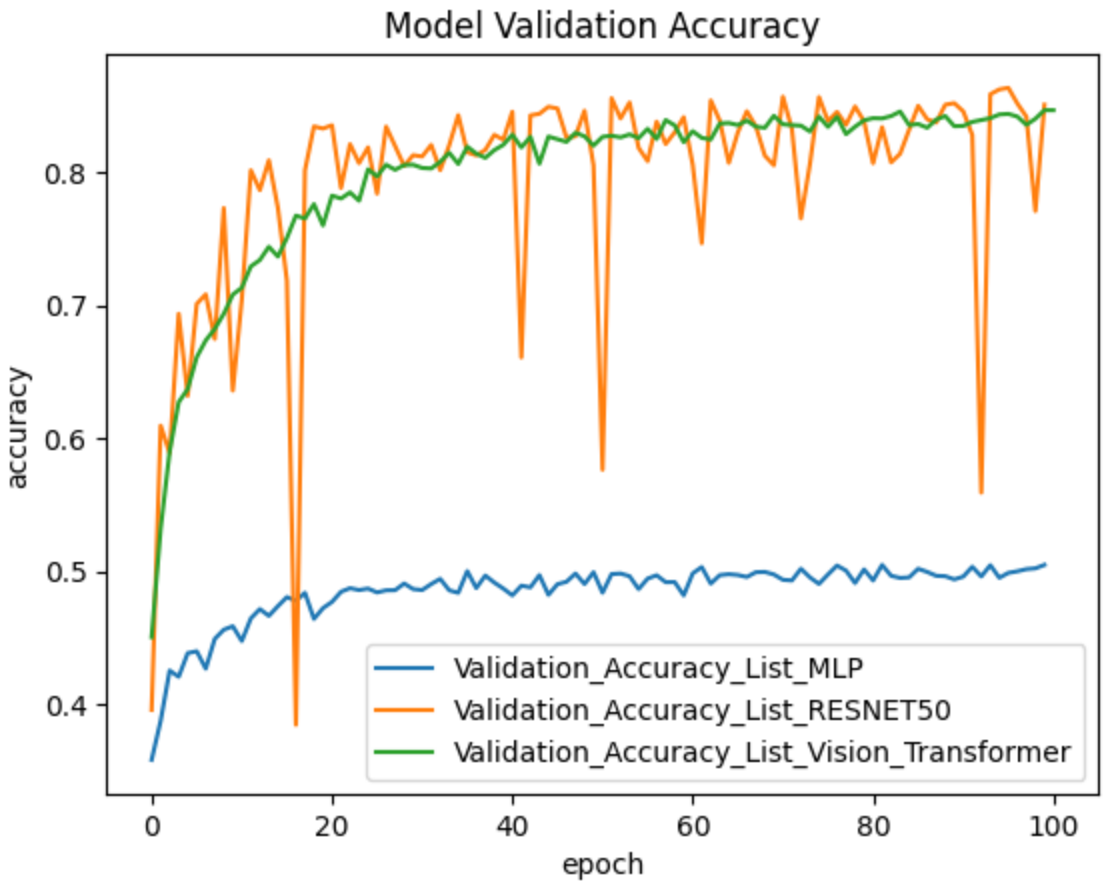
- The ViT is deep learning architecture which is used for image classification problem where an image is transformed into smaller non overlapping patches using patch embedding.These patches are then linearly embedded into fixed-size vector representation.And then we add positional embedding to the patches to provide the location of each patch in image since transformer model does not inherently understand spacial relationships between patches. That patch embeddings along with positional embeddings are then passed to encoder which uses self attention mechanism to capture contextual dependencies and representation of the image. And then a classification head is added, typically a fully connected layer to predict the class label of the input image.
- The hyperparameters used during training are learning rate of 0.001,weight decay of 0.0001, batch-size of 256, AdamW optimizer with categorical crossentropy loss and 100 epochs
- The Vision Transformer (ViT) model demonstrates exceptional accuracy on both the training and validation sets, indicating that it has learned the complex patterns in the data very well. It also achieves high accuracy on the test set, suggesting that it generalizes effectively to unseen data. The ViT model outperforms both the MLP and ResNet50 by a significant margin.

In summary, the MLP model has its strengths in simplicity, fast training, and low resource requirements. However, it may not be the best choice for image classification tasks like CIFAR-10, where spatial relationships and hierarchical features play a crucial role in achieving high accuracy. Deeper models like ResNet50 and advanced models like Vision Transformers have shown superior performance on challenging computer vision tasks, due to their ability to capture intricate patterns and relationships in images.

Model	Train_Accuracy	Validation_Accuracy	Test_Accuracy
MLP	50.96%	50.49%	50.49%
RESNET50	96.21%	85.07%	85.07%
Vision Transformer	**99.89%**	99.26%	99.12%

The graphical representations provide below is a clearer visualization of the performance differences between the models, helping us understand the strengths and weaknesses of each model on the task at hand.




Model Validation Accuracy
accuracy
epoch
0.4
0.5
0.6
0.7
0.8
0
20
40
60
80
100
Validation_Accuracy_List_MLP
Validation_Accuracy_List_RESNET50
Validation_Accuracy_List_Vision_Transformer