Image Classification of CIFAR-10 using Compact Convolutional Transformers

Ankit Basu [UIN: 934008673]
Texas A&M University
Master of Science in Data Science

Spring 2024

1 Introduction

Compact Transformers (CCT) offer a breakthrough in small-scale learning, outperforming state-of-the-art CNNs on small datasets. With as little as 0.28M parameters, CCT achieves competitive results on CIFAR-10 while being smaller than other transformers and ResNet50. CCT's compact design makes it feasible for researchers with limited computing resources and small datasets, extending research efforts in data-efficient transformers. In this project, we will try to improve the Vision Transformer for low-resolution image datasets.

2 Methodology

In this implementation, we use CNN convolution as a preprocessing embedding layer to the transformers, so that the CNN layers help in extracting more features that can help the transformer, to learn better from the convoluted and pooled images. Also, we use Max Pooling to get the most useful features from the image and feed it to a group of transformer layers.

3 Hyperparameters

In this implementation, we have used the following hyperparameters:
image size: 32px,
input channels: 3,
kernel size: 2,
depth of attention: 7,
number of heads: 4,
embedding dropout: 0.1,
transformer dropout: 0.1,
dropout: 0.1,

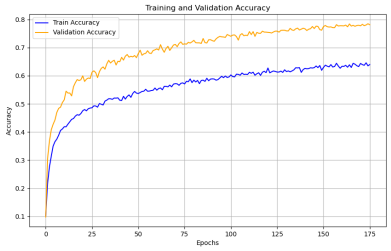
number of classes: 10,
hidden dimemsions: 64,
head dimension: 64,
scaled dimemsions multiplier: 4
number of epochs: 200,
learning rate: $1e-3$,
batch size: 64,
weight decay: $1e-4$

The model uses a AdamW optimizer and uses a custom scheduler that will increase the learning rate from 0 to a certain number of warmup epochs and finally start cosine to reduce the learning rate. As part of data preparation and transformation, random data augmentation, CutMix and MixUp have been used to increase generalization and overall accuracy.

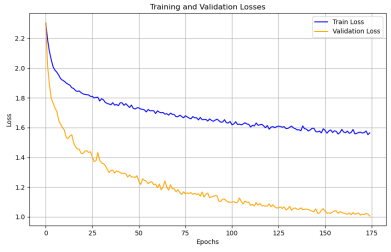
4 Training and Testing

In the limited training we have conducted for the epochs 200, we have seen the following trends as part of training, validation, and testing metrics.

Accuracy Curve:

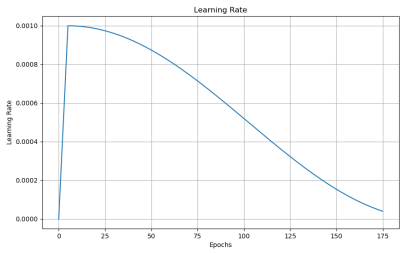


Loss Curve:



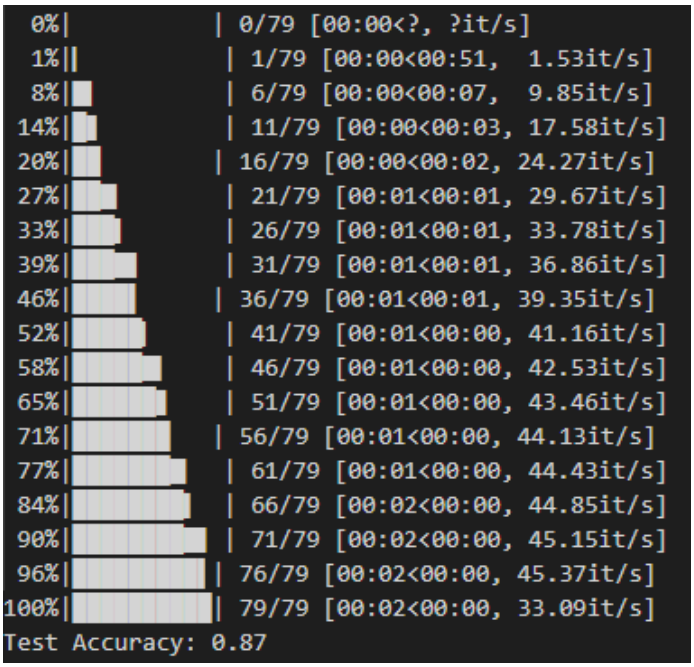
Here, we can see the model can optimize its performance by training for more epochs. Also, the validation scores are better than the training scores because the training data is highly augmented.

Learning Rate Curve:



5 Conclusion

As part of the evaluation, the model has achieved a validation accuracy up to 87%. This is pretty decent considering how small the images and the dataset are and how much data transformers need to accurately understand the patterns in the image. Image below:



As we can see, the model still has room for improvement. The model can be trained for longer epochs with more constraints on the learning rate, or overall using a better optimizer for training. This way the model will be able to properly understand the intricacies of the data and provide better outcomes.

6 References

1. Escaping the Big Data Paradigm with Compact Transformers: https://arxiv.org/pdf/2104.05704
2. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale: https://arxiv.org/abs/2010.11929
3. CSCE636 class lecture notes