Learning to classify images using Vision Transformers

Aakansha Singh
singh107

Sharvi Tomar
stomar2

Neha Bhardwaj
nehab3

Abstract

Deep learning methods provide state-of-the-art performance in most computer vision problems. Vaswani et al.'s work on Transformers gave rise to implementing transformers (Dosovitskiy et al., 2020) for Image-based data called Vision Transformers (ViT). In this work, we implement transformers for image classification from scratch by utilising a tweaked version of PyTorch implementation of Wightman's code and analyse the training performance on multiple data sets for a variety of settings. Additionally, we perform a comparison of the model performance of ViTB/16 pretrained model with our ViT Model implemented from scratch on a few selected datasets.

1 Introduction

Transformers have long enjoyed great usability in various problems of Natural Language Processing however, their extension to images is fairly limited. The application of transformers in the domain of images and videos is relatively new. The variant of transformers adapted for computer vision tasks is known as Vision Transformer (ViT) (Dosovitskiy et al., 2020), and in the recent works on ViT a major interest area has been image classification. In the case of Natural language processing the transformer measures, a relationship between a pair of input tokens and the cost of this process is exponential with the number of tokens. Vision Transformers overcome this issue by computing the relationship for small sections of an image which considerably reduces the cost associated with it.

The key idea of our work is to implement the transformers architecture for the task of image classification by training the module from scratch. Since training on large datasets is computationally challenging hence, we focused on developing a model capable of performing classification on images from three datasets. In addition, we performed experiments to analyse the training performance in various settings.

2 Related Work

The main idea of the vision transformers in the area of object detection was to replace the existing application of the convolution neural networks with a stack of embedding and encoder layers to predict the class of an image. Tokens from different stages of a vision transformer are assembled into image-like representations at various resolutions and then are progressively combined to form a high-resolution dense prediction using a convolutional decoder. Vision transformers can provide more fine-grained and high-resolution dense predictions as compared to the fully convolutional network (Dosovitskiy et al., 2020). While supervised ViT performs well, it is also observed that self-supervised ViT contains explicit information about the semantic segmentation of an image, which makes it stand out as compared to ConvNets which do not hold this information as clearly as a self-supervised ViT. There is however a drawback to using ViT as opposed to convnets since they are computationally more expensive and require more time to train (Yuan et al., 2021).

Models like Residual Networks (ResNet, and other State-of-the-art CNN models) trained on ImageNet dataset perform better and have higher accuracy compared to transformer based models like ViT. However, when trained on large datasets (14M-300M images) like ImageNet-21k and JFT-300M dataset, ViT matches or even outperforms SOTA models for various benchmarks. In (Caron et al., 2021), the authors demonstrated how with enhanced data augmentation it is possible to achieve improved performance with resulting models be-

ing competitive with CNN on ImageNet. The authors resorted to a two-way approach, training with increased data through augmentation and hard label clustering. DeIt (Data Efficient Image Transformers) are efficient pre-trained models available for use. Another attempt to improve ViT's performance on mid-sized dataset ImageNet when trained from scratch was through the introduction of T2T module (Wang et al., 2022). The paper indicates the use of a layer-wise transformation with soft-splitting i.e. overlapping patches to gather surrounding context along with architecture inspired by CNN to reduce the parameters of vanilla ViT by half.

3 Data & Methodology

3.1 Dataset

We performed analyses on three datasets, namely Fashion-MNIST (Xiao et al., 2017), Cassava (Mwebaze et al., 2019) and CIFAR10 (Krizhevsky et al.). The train and test set split for the Cassava leaf dataset was obtained by keeping 10% of samples of each class in the test set using the stratified sampling strategy. Table 1 summarizes the details for all three datasets.

Dataset	Train	Test	Image-Size
Fashion-MNIST	60000	10000	1∗28∗28
CIFAR10	50000	10000	3∗32∗32
Cassava-Leaf	4242	1414	3∗224∗224

Table 1: Dataset Details

3.2 Methodology: Vision Transformers

Transformer models rely on the attention module to understand the importance of each feature in the input.A transformer econder comprises of a multi-head self-attention layer (MSA) and a Multi-Layer Perceptron (MLP) block,both of which have a layernorm before them. ViT (Fang et al., 2021) proves that a pure Transformer architecture can also attain state-of-the-art performance on image classification.

3.2.1 Network Architecture

The architecture for a Vision transformer can be broadly divided into four parts, wherein the first step looks into the patch and positional embedding of the Image before it can be passed onto the transformer for encoding in the second stage. The third stage involves applying a multi-layer perceptron on the output of the transformer to get to the final classified label for that image. This architecture can be seen in Figure 1 and is discussed more in detail below.

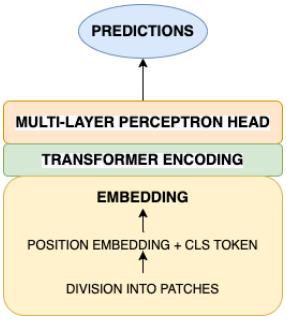

Figure 1: Architecture of Vision Transformer

3.2.2 Embedding

We split each image into fixed-sized patches (in our case the images are being split into 16 X 16 patches) shown in Figure 2. Each patch is then linearly embedded and then assigned with positional encoding vectors so that the architecture can easily identify each patch in case the patches are re-positioned or puzzled as in Figure 3.

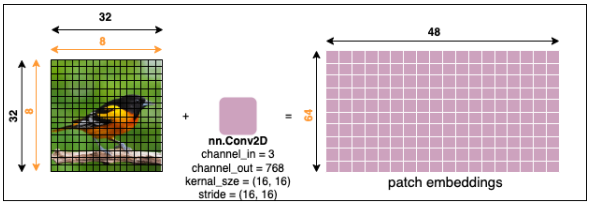

Figure 2: Patch Embeddings

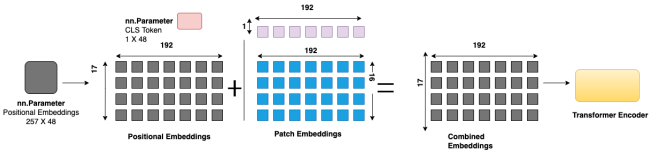

Figure 3: CLS token & Position Embeddings

3.2.3 Transformer Encoding

After successfully adding the linear and position encoding, all these patches are then passed through a transformer encoder, along with an added classification token. So for n patches, $n + 1$ vectors are being passed to the transformer encoder. The transformer encoder is a series of multi-head attention layers, and dense layers stacked sequentially. The multi-head self-attention layers, use three key characteristics of query, key and value to assign

an attention score to each vector to understand the level of attention assigned, which further helps to focus on a few classes and ignore the ones with a very low score. Figure 4 provides the process overview.

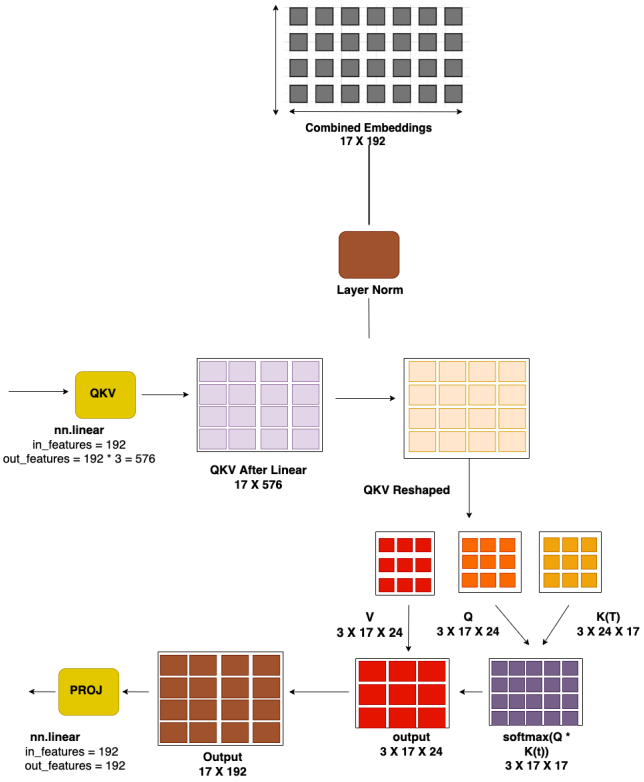

Figure 4: Multi-Head Self-Attention

3.2.4 Classification Layer

Once the patches pass through the transformer encoder, the output would be $n + 1$ encoded vectors. We will extract the classification token for each image and pass them through a softmax classifier. This returns a list of probabilities for each class, where the maximum returned probability classifies the image.

4 Experiments & Results

4.1 Training ViT from scratch

Taking inspiration from the PyTorch implementation of ViT provided by Wightman and (Pufler, n.d), we implemented a Vision transformer model from scratch and tested it on the three datasets. In the case of Fashion-MNIST, we had 28 X 28 images that were split into smaller patches of 4 X 4 and passed into the transformer encoder for prediction. Since this data set was smaller in size with only grayscale pixel values for images, it was computationally the cheapest and also was able to achieve the highest training and testing accuracy. For 40 epochs, the testing accuracy obtained was over 93%.

Cassava data set images resembled the ImageNet dataset in terms of dimensions with a size of 224 X 224 and were split into larger patches of size 16 X 16. The number of samples in the dataset was 5656. This set of samples was split into training and validation sets. So overall, the transformer had few samples accessible in order to train and predict class values. Transformers are data-hungry architectures and with limited data access transformers are known not to perform well. We observed that the training and testing accuracy, in this case, remained stagnant at around 62%.

CIFAR10 was computationally the heaviest data set to train on with colored 50000 images. The image size was 32 X 32 and we divided them into patches of 8 X 8 each. We observed a low testing accuracy of around 43% while training it from scratch for about 40 epochs. But with an increase in the epochs, the accuracy did show a slight upward curve.

The exhaustive model settings for the process of training ViT from scratch for the three datasets are summarized in Table 2.

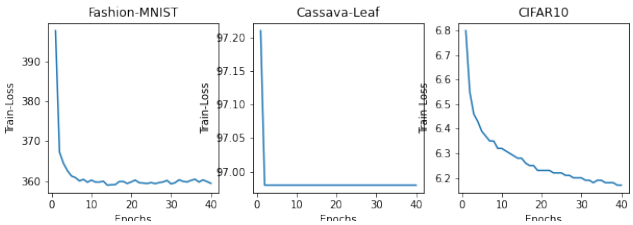

Figure 6: Train Losses per Epoch

All three data sets were trained on the model for around 40 epochs. We set a learning rate of 0.01 for CIFAR10 and Fashion MNIST, and a learning rate of 0.001 for Cassava. It was observed that the training losses seem to converge around 30 epochs for Cassava and Fashion-MNIST datasets and while CIFAR10 train losses take slightly longer to converge as depicted in Figure 6.

Overall looking at the performance of the vision transformer when implemented from scratch we can say that it performs decently well given the number of samples and computational limitations. However, the performance can be further improved by supplying a large amount of data and training for a longer duration of time and epochs.

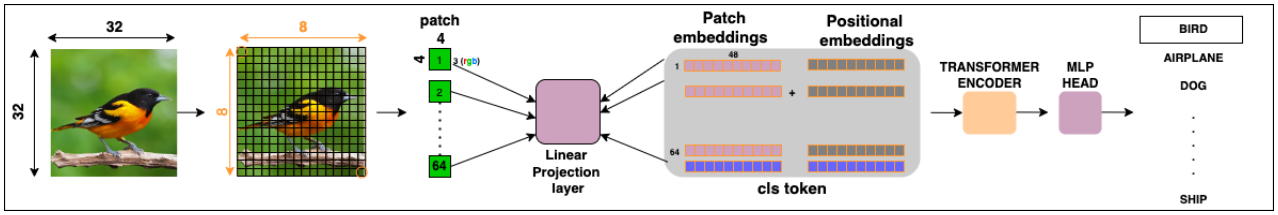

Figure 5: Simplified Model Overview

Datasets	Classes	Patch-size	Batch-size	LR	Epochs	Test Accuracy
Fashion-MNIST	10	4∗4	16	0.01	40	93.09%
CIFAR10	10	8∗8	128	0.01	40	43.26%
Cassava	5	16∗16	16	0.001	40	61.50%

Table 2: Training from Scratch: Model Setting

4.2 Pre-trained ViT vs ViT trained from scratch

For our experiments, we used the ViT-B/16 model which is pretrained on ImageNet-21k and then fine-tuned on ImageNet at 224x224 resolution.(Wu et al., 2020)

In order to analyse the performance of the vision transformer which is trained from scratch with a pretrained ViT, we compared the obtained test accuracies. For a fair comparison, we kept the batch-size, learning-rate, epochs and train and test split same for both scenarios. The details of the model parameters are summarised in Table 3.

Dataset	Epochs	Batch-size	LR
Fashion-MNIST	5	16	0.01
CIFAR10	3	16	0.01
Cassava	7	16	0.001

Table 3: Comparison: Model Setting

While employing the pretrained model to CIFAR10 and Fashion-MNIST dataset, we adapted the image size to be appropriate for processing. The images from these two datasets were resized to a size of 224 pixels in height as the pretrained model as such. In addition to this, pixel values of the images from the Fashion-MNIST dataset were stacked one onto the other to increase the channels from one to three.

We observed that the pre-trained ViT model trained is able to provide higher test accuracy in the case of the Cassava dataset while it is outperformed by ViT which is trained from scratch for Fashion-MNIST and CIFAR10 datasets as evidenced in Figure 7.

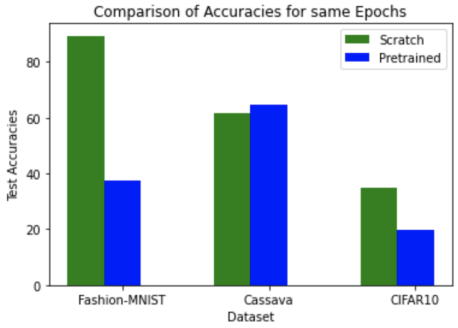

Figure 7: Accuracy Comparison

5 Conclusion

From our experiments, we can infer that a ViT model which is trained from scratch is able to give reasonable performance (test accuracy) even when neither the data provided is large nor the training is performed for large iterations. The accuracy achieved for grayscale Fashion-MNIST is over 93% in just 40 epochs. For Cassava and CIFAR10 colored images, the testing accuracy of over 61% and 43% is also promising with the given number samples and epochs.

However, the performance could be improved even more by expanding the capacity of training samples and training time.

In addition to analysing performance of a ViT trained from scratch, we also compared its performance with that of a pretrained Vit. Both the variants of ViT were trained under similar parameter settings for a fair comparison for the same number of epochs. From the results obtained in terms of testing data accuracy, we can conclude that the ViT trained from scratch outperfoms greatly on grayscale images. However, for colored images

the difference in testing accuracies by both ViT variants is not very significant. It can be derived that the pre-trained ViT is able to perform to its full potential when the training/fine-tuning is carried out for increased number of epochs.

6 Future Work

Due to the limitation of our computational resources, we were not able to compare the performance of a pretrained ViT with a ViT trained from scratch for higher number of epochs and with different settings of hyperparameters such as batch size, learning rate, optimizers etc. Hence, wider exploration of parameter values would help to provide a better picture of comparison. It would also be interesting to see the trade-off of model training time with accuracy in both variants of ViTs.

Additionally, the ViT trained from scratch could also be tweaked to contain overlapping patches, intermediate layer transformations, normalizations and other enhancements to compete with the pre-trained ViT.

7 Contributions

All authors contributed equally to the project and worked together in gaining understanding of the architecture of Vision Transformers and its implementation. The project report is a result of equal efforts by all three project members.

The experiments on three datasets were divided amongst the three authors. Aakansha worked on the Cassava dataset, Sharvi worked on CIFAR10 dataset and Neha worked on Fashion-MNIST dataset.

References

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. 2021. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9650–9660.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Yuxin Fang, Bencheng Liao, Xinggang Wang, Jiemin Fang, Jiyang Qi, Rui Wu, Jianwei Niu, and Wenyu Liu. 2021. You only look at one sequence: Rethinking transformer in vision through object detection. *Advances in Neural Information Processing Systems*, 34.

Alex Krizhevsky, Vinod Nair, and Geoffrey Hinton. Cifar-10 (canadian institute for advanced research).

Ernest Mwebaze, Timnit Gebru, Andrea Frome, Solomon Nsumba, and Jeremy Tusubira. 2019. icassava 2019fine-grained visual categorization challenge.

Brian Pufler. n.d. Papers reimplimentations. https://github.com/BrianPulfer/PapersReimplementations.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. 2022. Pvt v2: Improved baselines with pyramid vision transformer. *Computational Visual Media*, pages 1–10.

Ross Wightman. 2019. Pytorch image models. https://github.com/rwightman/pytorch-image-models.

Bichen Wu, Chenfeng Xu, Xiaoliang Dai, Alvin Wan, Peizhao Zhang, Zhicheng Yan, Masayoshi Tomizuka, Joseph Gonzalez, Kurt Keutzer, and Peter Vajda. 2020. Visual transformers: Token-based image representation and processing for computer vision.

Han Xiao, Kashif Rasul, and Roland Vollgraf. 2017. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms.

Li Yuan, Yunpeng Chen, Tao Wang, Weihao Yu, Yujun Shi, Zi-Hang Jiang, Francis EH Tay, Jiashi Feng, and Shuicheng Yan. 2021. Tokens-to-token vit: Training vision transformers from scratch on imagenet. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 558–567.