Future Computing Technologies Lab: Creative Inquiry

TRANSFORMERS IN COMPUTER VISION: A BASIC IMPLEMENTATION AND REVIEW

May 4, 2023

Adam Pickeral
Clemson University
Department of Electrical and Computer Engineering
apicker@g.clemson.edu

1 Introduction

Computer vision tasks, such as object detection and image classification, have traditionally relied on convolutional neural networks (CNNs) as the go-to architecture. However, a recent paper titled ”An Image is Worth 16x16 Words” introduced a new architecture called the Vision Transformer (ViT), which uses a self-attention mechanism to capture global relationships between image patches. This approach shows promising results in image classification tasks and has been gaining popularity in the computer vision community.

In this report, I present a project that creates a basic implementation of the Vision Transformer architecture described in the paper and applies it to the MNIST Handwritten Digits dataset. The MNIST dataset is a classic benchmark dataset used to test and evaluate image classification models, and its small size and simplicity make it an ideal dataset to experiment with new architectures.

My project aims to investigate the performance of the Vision Transformer on this dataset and compare it with the performance of traditional CNNs. I first implemented the architecture and then trained and evaluated the model on the MNIST dataset. I present my results and analyze the performance of the Vision Transformer in comparison with traditional CNNs.

Overall, this project serves as an introduction to the Vision Transformer architecture and its application in computer vision tasks, specifically for image classification. It also provides insights into the performance of the ViT architecture on the MNIST dataset and serves as a starting point for future research in this area.

2 Materials and Methods

Dataset

The MNIST Handwritten Digits dataset was used as the dataset for this project. The dataset consists of 60,000 training images and 10,000 test images of handwritten digits from 0 to 9. The images are grayscale and have a resolution of 28x28 pixels.

Figure 1: Example Images from the MNIST Dataset

Vision Transformer Architecture

The Vision Transformer architecture described in the paper ”An Image is Worth 16x16 Words” was implemented. The architecture consists of a series of self-attention layers and feedforward layers. The self-attention layers are used to capture global relationships between image patches, and the feedforward layers are used to generate the final output.

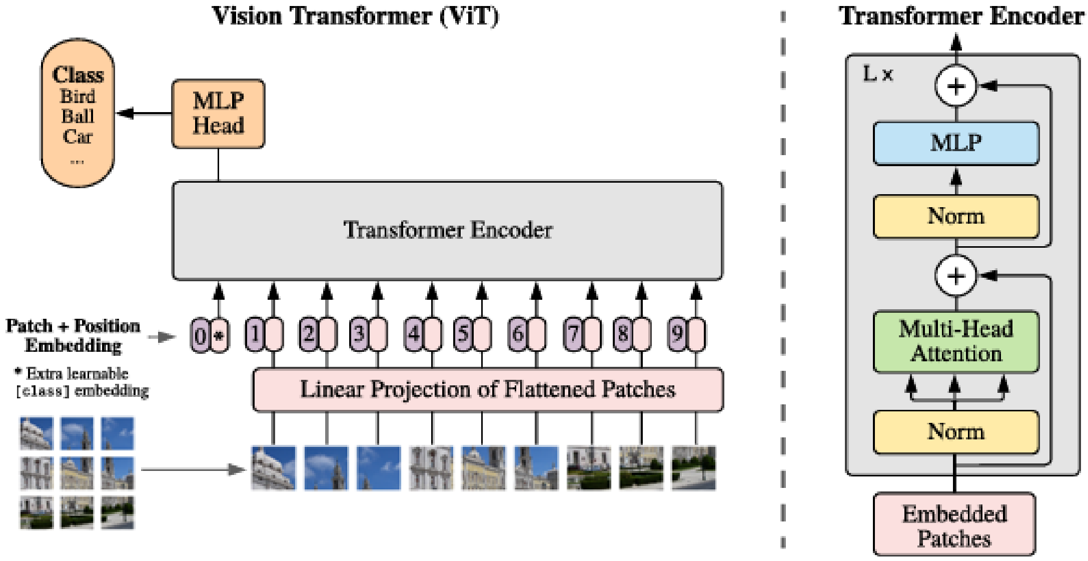

Figure 2: The ViT Architecture (Left) and Singular Transformer Architecture (Right)

Preprocessing

The images were preprocessed by normalizing the pixel values between 0 and 1.

Training

The Vision Transformer model was trained using the Adam optimizer with a learning rate of 3e-4 for 5 epochs. A batch size of 100 was used and the model was trained on a single NVIDIA A100 GPU on the Palmetto Supercomputing Cluster.

Evaluation

The performance of the model was evaluated on the test set by computing the accuracy, precision, and recall. The performance of the Vision Transformer was compared with a traditional CNN architecture (in this case, a simple 3-layer CNN)..

Code

The Vision Transformer was implemented and the model was trained and evaluated using the PyTorch deep learning framework.

Hardware

All the experiments were conducted on a machine with a single NVIDIA A100 GPU.

Metrics

Accuracy and training time were used to evaluate the performance of the model. Accuracy is the proportion of correctly classified images, and training time is the time elapsed when training the model.

3 Results

Model	**Accuracy**	**Training Time**
ViT	97.19%	129.05 sec
CNN	98.52%	57.45 sec

Table 1: Results of Training and Testing Both Models on MNIST

The results from this experiment show that the Vision Transformer is an extremely viable model, learning deep connections in this dataset after only 5 epochs. However, in this example when compared to a CNN, it isn't as valuable. The CNN was able to have an improved accuracy in half the training time.

4 Discussion

As the cited paper explains, ViT's have a lot of potential in large datasets where global attention could be beneficial. In fact, in the paper's results, their ViT model outperformed ResNet 18, the current SOTA CNN for image detection, and was training in a fourth of the time. I believe that with more complex, color image datasets, a ViT will outperform the current CNN architectures. In this case with a simple dataset, the complexity of the transformer was unnecessary.

5 Experience

I have enjoyed this Creative Inquiry. It has allowed me to explore new topics in areas I wouldn't have known about had I not been given the freedom.

6 Conclusions and Future Work

In the future, I would love to improve this model to use with a more complex dataset. I would like to change it to work with the CIFAR-10 dataset, an introductory RGB set. With the increased complexity of these images, given color and various classes, I believe the ViT could show significant improvement in its comparison with CNN architectures.

This study showed that ViTs are certainly a viable model in the field of Computer Vision.

7 Works Cited

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ... & Houlsby, N. (2021). An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 3156-3164). [link]

Arora, A. (2021, April 29). Visual Transformers: A New Computer Vision Paradigm. *Medium.* Retrieved from https://medium.com/swlh/visual-transformers-a-new-computer-vision-paradigm-aa78c2a2ccf2