Tyler Beetle
Second Final Report
Sankalp Jajee Grad Presentation
CSCE 581

This is an article about leveraging transformers for image recognition. In the past, transformers have primarily been used for natural language processing but this is the first time they have been used for image recognition. The authors display that the transformer architecture can be used to achieve effective results on image classification tasks. They do this by applying transformers to a sequence of image patches. This is different from the typical convolutional neural network (CNN) architecture, which typically uses a grid of pixels. The transformer offers multiple benefits over the CNN based model, The authors evaluate their approach on a variety of image classification benchmarks, including CIFAR-10, CIFAR-100, and ImageNet. Their results consistently demonstrate that transformers can achieve competitive or even superior performance compared to CNNs. Overall, the proposed transformer based approach demonstrated the potential of transformers to achieve state of the art image classification.

In this paper, the authors developed a novel architecture called Vision Transformer (ViT) that utilizes transformers for image recognition tasks. Unlike conventional convolutional neural networks (CNNs), ViT directly processes sequences of image patches, resulting in improved performance and reduced computational costs. The authors demonstrated the effectiveness of ViT by achieving state-of-the-art results on various image classification benchmarks, including CIFAR-10, CIFAR-100, and ImageNet. Additionally, ViT requires significantly fewer parameters compared to CNNs, making it more efficient to train and deploy. This transformative approach paves the way for a new era of image recognition using transformers. The overall article provides affirmation that the continuous development of the transformer architecture can lead to methods such as detection and segmentation. The success of ViT underscores the remarkable versatility of transformers in tackling diverse visual tasks, extending their dominance beyond natural language processing and into the realm of computer vision.

My main critique of the paper is that the authors could have expanded more on the background of the convolutional neural network and other methods of image classification. Giving the reader more context in this regard would give a better understanding as to why the transformer architecture is important. Another critique I have is that the authors could have dove deeper into the direct limitations of the transformer and how expanding on these limitations will push the transformer architecture into the future. A final area the authors could have given more information on is diving into the the ethical considerations of the transformer.