Topic: Max ViT : Multi Axis Vision Transformer 21/16 June 2023

Abstract:

Transformers have recently gained significant attention in the computer vision community. However, the lack of scalability of self attention mechanics with respect to image size has limited their wide adoption in state of the art vision backbones. In this paper we introduce an efficent and scalable attention model we call multi axis attention, which consists of two aspects, blocked local and dilated global attention. These designs choices allows global local spatial interactions on arbitrary input resolutions; with only linear complexity. We also present a new architectural element by effectively blending our proposed attention model with convultions and accordingly propose a simple hierarchical vision backbone, dubbed MaxViT by simply repeating the basic building block over multiple stages. Notably, MaxViT is able to "see" globally throughout the entire network, even in earlier, high resolution stages. We demonstrate the effectiveness of our model on a broad spectrum of vision tasks on image classification, Max ViT achieves state-of the art results under various settings: without extra data, MaxViT attains 86.5% ImageNet 1K top1 accuracy with ImageNet 21K pre training, our model achieves 88.9% top1 accuracy. for downstream tasks, MaxViT as a backbone delivers favorable performance on object detection as well as visual asthetic assessment we also show that our proposed model expresses strong generative modeling capability on ImageNet demonstrating the superior potential of MaxViT blocks as universal vision module.

Conclusion:

While recent tasks in the 2020s have arguably shown that Conv Nets and vision Transformers can achieve similar performance on image recognition, our work presents a unified design that takes advantages of best of both ~~models~~ worlds - efficient, convultion and sparse attention and demonstrates that a model built on top, namely MaxViT, can achieve state of the art performance on a variety of vision tasks, and more importantly, scale extermely well to massive scale data sizes. Even though we present our model in the context of vision tasks, the proposed multi axis approach can easily extend to language modelling to capture both local and global dependencies in linear time. We also look foward to studying other forms of sparse attention in higher dimensional or multi modal signals such as videos, point clouds and vision languages.