Applying Fractal Structure to Vision Transformer (ViT)

Jeong, Sehun (School: North London Collegiate School Jeju)

Current ResNet structure uses convolution kernels to preserve local information of an image. In contrast, ViT divides an image into patches to calculate their global attention. With enough amount of training data, ViT can outperform CNN in image classification. However, the lack of local information considered in ViT results in ViT’s lower performance with a small set of data. In this research, to increase the local information considered by ViT, an input image will be patched into tokens of different levels that are designed to form a fractal pattern. The research improves the accuracy of original ViT while decreasing the amount of parameter being used. Using this, Fractal ViT and ViT are trained with satellite images.

Awards Won:

King Abdulaziz &amp
his Companions Foundation for Giftedness and Creativity: Full Scholarship from King Fahd University of Petroleum and Minerals(KFUPM) (and a $400 cash prize)