MaxViT: Multi-Axis Vision Transformer

ソース

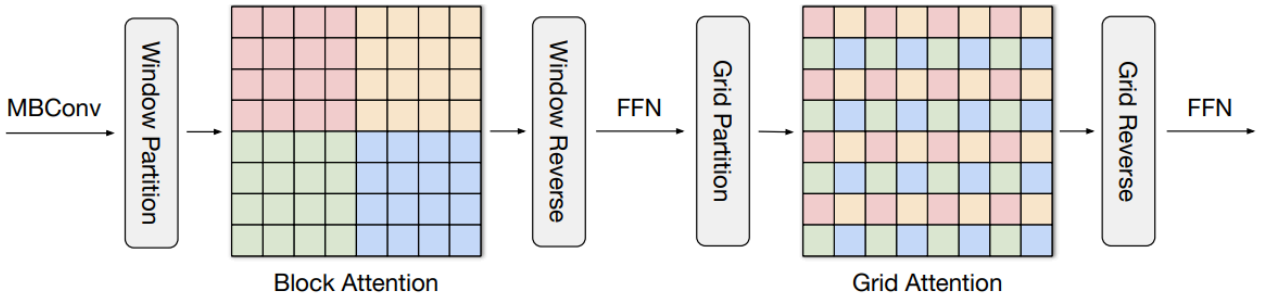
- https://arxiv.org/pdf/2204.01697

概要

- ViTの派生
- Grid Attentionというものを提案
- Swin Transformer, CoAtNetなどに勝ちSota達成

Grid Attention

- 図を見ればわかる
- 図の同じ色になっているピクセルでself attentionを計算する
- Block Attentionでは入力をnon-overlapping windowsに分割して$\left(\frac{H}{P} \times \frac{W}{P}, P \times P, C\right)$にreshapeして各ブロックでself attentionを計算する
- Grid Attentionでは$G \times G$のグリッドに分割して$\left(G \times G, \frac{H}{G} \times \frac{W}{G}, C\right)$
- つまり, Block Attentionでは$\frac{HW}{P^2}$ 個の (P, P)のサイズのwindowでattentionを計算
- Grid AttentionではG^2個の$\left(\frac{H}{G}, \frac{W}{G}\right)$のサイズでattentionを計算
- また、図の通りgridは飛んでいる
- よって、block attentionはlocal featureを取得してgrid attentionはglobal featureを取得する



MaxViT

- アーキテクチャは以下の通り
- 4層のMaxViT Blockから構成

- MaxViT BlockはMBConv, Block Attention, Grid Attentinから構成