Image Processing GNN: Breaking Rigidity in Super-Resolution

ソース

- https://openaccess.thecvf.com/content/CVPR2024/papers/Tian_Image_Processing_GNN_Breaking_Rigidity_in_Super-Resolution_CVPR_2024_paper.pdf
- https://github.com/huawei-noah/Efficient-Computing/tree/master/LowLevel/IPG

概要

- CVPR2024 best paper candidate
- Graph Neural NetworkでSuper Resolution
- PSNRでHAT上回った
- Super Resolutionはunbalanceである
  - 情報量が多い部分(カラフルな物体など)は周辺の多くのピクセルの情報が必要
  - 情報量が少ない部分(青空とか、雲とか)は周辺の数個のピクセルだけでよい
  - このように領域によって周辺の使いたいピクセル数が可変だと嬉しい
  - けど、ConvolutionもAttentionも固定だからよくない
  - そこで、Graph Neural Networkで可変にする
- Vision GNNでは画像のpatchをnodeとするが、IPGでは画像のpixelをnodeとする
  - patchを決めたときに位置ずれのせいで精度落ちるかもしれないから
- 画像からグラフを構成するメジャーな手法はKNN
  - k-nearest neighbors
  - あるノードから距離の小さい順にk個のノードに辺を張ったグラフ
  - これは任意の頂点について次数がkで固定でありSuper Resolutionのunbalanceを無視することになるから却下

提案手法

Degree Flexibility

- Degree Flexibilityを実現するためにdetail-rich indicatorを導入する

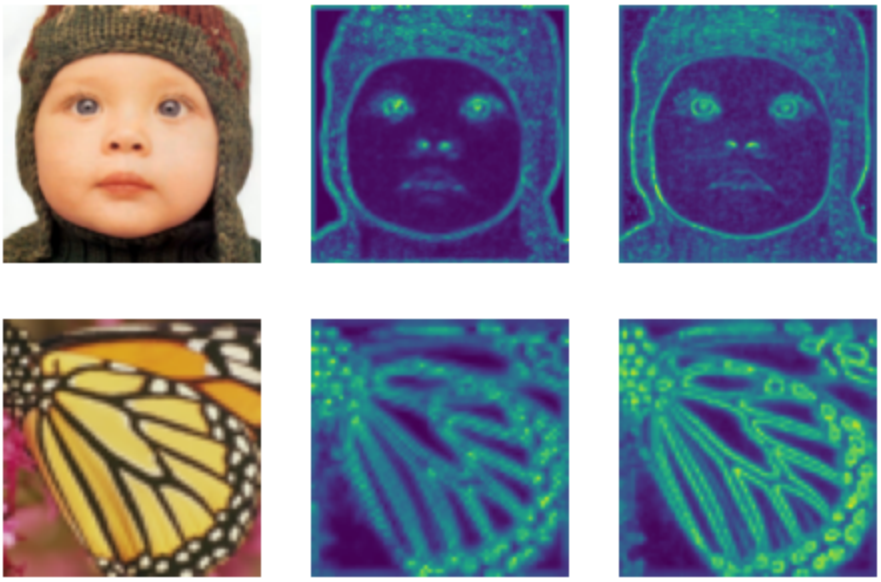
    - detail-rich indicatorは画像のどの領域がdetail-rich(情報量多い,色の変化が激しい,flatでない)かを示すもの
    - この値が大きいほど頂点の次数を増やしたい
- detail-rich indicatorは以下の式で定まる

$$D_F := \sum_C |F - F_{\downarrow s \uparrow s}|$$

- ここでFはfeature mapであり$F \in \mathbb{R}^{H \times H \times C}$
- sはdownsampling ratio(s=2としている)
- $F_{\downarrow s \uparrow s}$は$F$に対してdownsamplingしてから線形補完でupsamplingしたもの
- このときpixel node $v \in F$の次数を以下のようにする

$$\deg(v) \propto D_F(v)$$

- D_Fの出力結果は以下のようになる



(論文より引用)

Sampling Strategy

- グラフを構成するときに、あるノードからどのノードに辺を張るかが問題になる
- local graphとglobal graphを構成するために、local samplingとglobal samplingを行う

(論文より引用)

- 左が対象ノード、真ん中がlocal、右がglobal
- 以上の方法で画像からグラフを構成できた

Graph Aggregation

- Graph Aggregationとしてedge-conditioned aggregationを採用
- edge-conditioned aggregationは以下で定まる

NNのk層目の出力であるnode feature $\boldsymbol{h}^k$のv個目の要素$\boldsymbol{h}_v^k$は$k-1$層目の出力$\boldsymbol{h}^{k-1}$を用いて

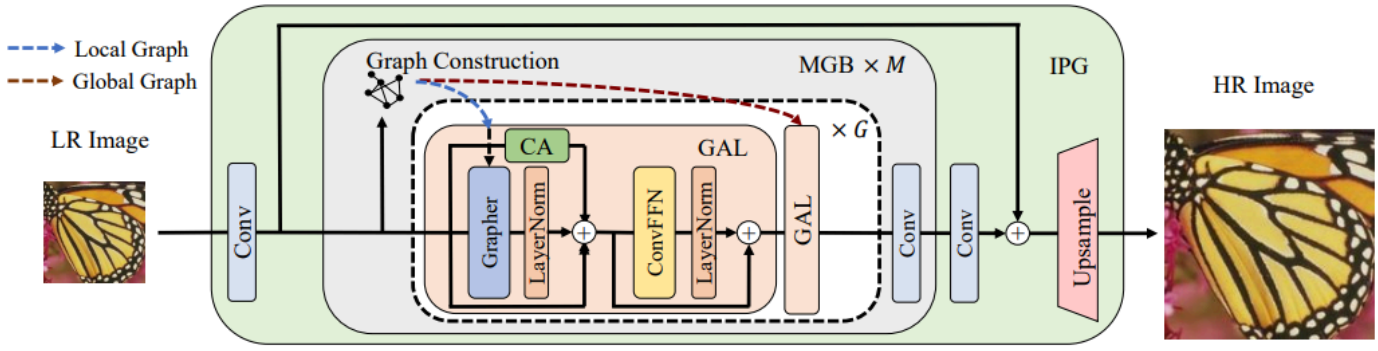
$$\boldsymbol{h}_v^k = \frac{1}{C^k} \sum_{u \in N(v)} \exp\left(f^k\left(u, v\right)\right) \boldsymbol{h}_u^{k-1} \tag{1}$$

$$C^k = \sum_{u \in N(v)} \exp\left(f^k\left(u, v\right)\right) \tag{2}$$

とかける

$f^k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$は2頂点の相関を示す写像で、本論文ではコサイン類似度を使用

Model Architecture



(論文より引用)

- よくある3層構造
  - shallow feature
  - deep feature
  - reconstruction
- shallow featureはこれまでと同じくConvのみ
- reconstructionもこれまでと同じくpixel shuffle upsampler
- deep featureはM個のMGB(Multiscale Graph-aggregation Blocks)から成る
- MGBはG個のGAL(Graph Aggregation Layers)から成る
- GALはよくあるTransformerの構造でGrapher -> LayerNorm -> ConvFFN -> LayerNormの形で途中にChannel Attentionとの残差接続を追加
- GrapherがGraph Aggregationを計算する層
- GALはlocal graph, global graphを交互に計算する