SwinIR

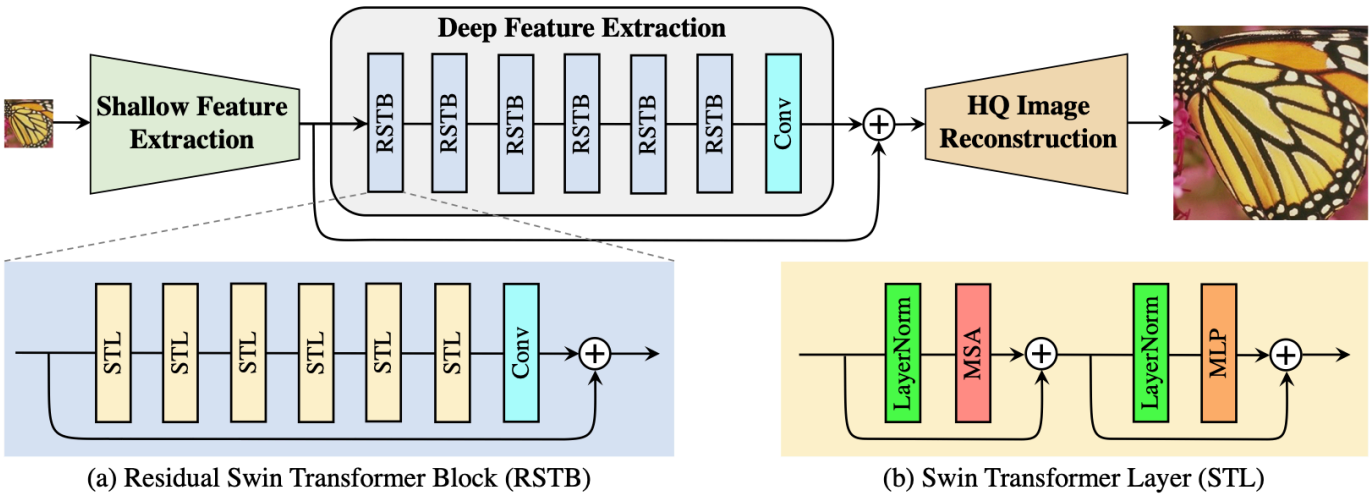
論文ソース

- SwinIR: Image Restoration Using Swin Transformer

概要

- SOTA達成(当時のSOTAはCNNベース)
- Super ResolutionをCNNベースでやってるけどSwinTransformer使ったほうが精度いいよというお話
- 3層からなるアーキテクチャ
  - shallow feature extraction
  - deep feature extraction
  - high-quality(HQ) image reconstruction

SwinIRの構造



(a) Residual Swin Transformer Block (RSTB)

(b) Swin Transformer Layer (STL)

Shallow Feature Extraction

- 1層の3×3convolutinal layerで畳み込みしてチャネル数変更してるだけ
- 入力:low-quality input $I_{LQ} \in \mathbb{R}^{H \times W \times C_{in}}$
- 出力:shallow feature $F_0 \in \mathbb{R}^{H \times W \times C}$

- $F_0 = H_{SF}\left(I_{LQ}\right)$

Deep Feature Extraction

- 入力F_0で出力$F_{DF} \in \mathbb{R}^{H \times W \times C}$で形状不変
  - $F_{DF} = H_{DF}\left(F_0\right)$
- H_{DF}:deep feature extraction module
  - K個のresidual Swin Transformer blocks(RSTB)と最後に1層のconvolutionから成る
  - 数式で書くと以下
  - $F_i = H_{RSTB_i}\left(F_{i-1}\right), \quad i = 1, 2, \cdots, K,$
  - $F_{DF} = H_{CONV}\left(F_K\right)$

RSTB

- L個のSwin Transformer layers(STL)と最後の1層のconvolutionから成る
  - 最初と最後で残差接続
- $F_{i,j} = H_{STL_{i,j}}\left(F_{i,j-1}\right), \quad j = 1, 2, \cdots, L,$
- $F_{i,out} = H_{CONV_i}\left(F_{i,L}\right) + F_{i,0}$

STL

- Swin Transformer layerと同じ
- $M \times M$のウィンドウに分割するので入力が$H \times W \times C$から$\frac{HW}{M^2} \times M^2 \times C$にreshape
- $\frac{HW}{M^2}$個のwindowごとにMulti head Self Attention
  - $Q = XP_Q, K = XP_K, V = XP_V$
  - $\mathrm{Attention}\left(Q, K, V\right) = \mathrm{SoftMax}\left(\frac{QK^T}{\sqrt{d}} + B\right)V$
- 全体としてはSwin Transformerと同じくLayer Normalization -> Multi head Self Attention -> residual -> LN -> MLP(2層GELU) ->residual
- MSAはSW-MSA->W-MSAを交互に繰り返す
  - ソースコード見てないけどL偶数になっているはず

HQ Image Reconstruction

- shallow featureとdeep featureの残差接続が入力
  - $I_{RHQ} = H_{REC}\left(F_0 + F_{DF}\right)$
- sub-pixel convolution layerを使って実装
  - ESPCN
  - https://arxiv.org/pdf/1609.05158.pdf