Vision Transformer

論文ソース

- AN IMAGE IS WORTH 16X16 WORDS: TRANSFORMERS FOR IMAGE RECOGNITION AT SCALE

概要

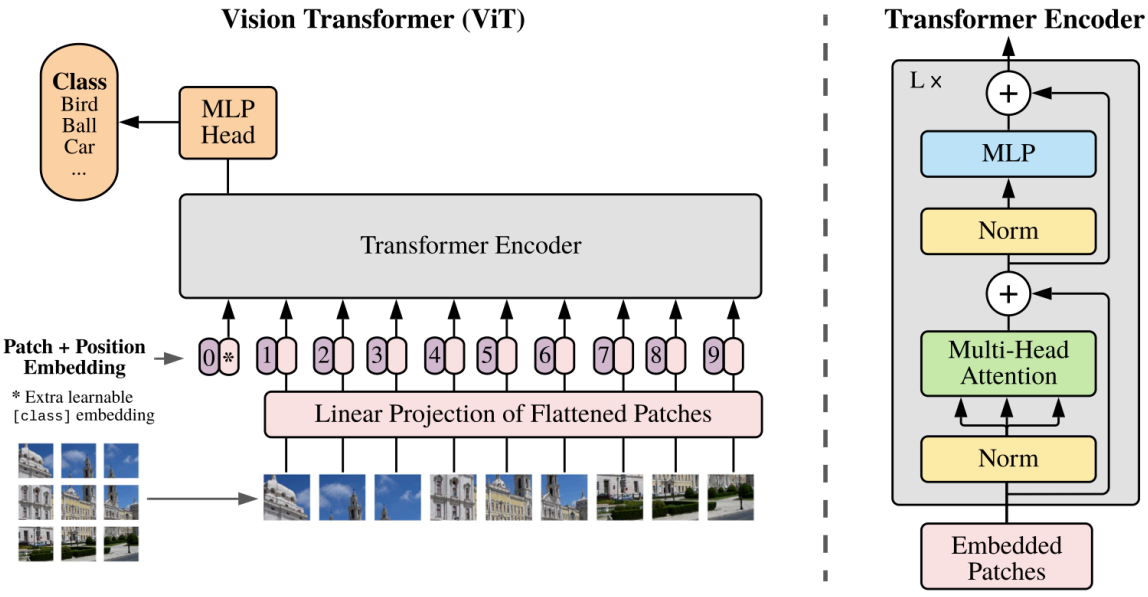
- Transformerを画像でも使うという発想
  - 画像をパッチに分割して横に並べて時系列データにする
- Convolution使わずにSOTA達成

パッチ

- 1辺pの正方形領域
- 画像をパッチに分割することで$N = \frac{HW}{P^2}$個のパッチが得られる
- 各パッチを単語のように扱うためベクトルに変換(Flatten)
  - 本画像 : $x \in \mathbb{R}^{H \times W \times C}$
  - パッチに分割した後 : $\mathbb{R}^{N \times P \times P \times C}$
  - flattenした後 : $x_p \in \mathbb{R}^{N \times P^2 C}$

ViT

- 全体図は以下の左図
- Transformer Encoderの中身が右図×L層



- TransformerのEncoderをほぼそのまま使っていて以下の2点のみ異なる
  - Layer Normalization(図のNorm,数式のLN)がAttentionの前にある
  - MLPの活性化関数がGELU(TransformerはReLU)
- 全体の数式は以下
- i. $z_0 = [x_{class}; x_p^1 E; x_p^2 E; \cdots; x_p^N E] + E_{pos}$
- ii. $z_l' = \mathrm{MSA}\left(\mathrm{LN}\left(\mathrm{z}_{\mathrm{l}-1}\right)\right) + z_{l-1}$
- iii. $z_l = \mathrm{MLP}\left(\mathrm{LN}\left(\mathrm{z}_{\mathrm{l}}'\right)\right) + z_l'$
- iv. $y = \mathrm{LN}\left(\mathrm{z}_{\mathrm{L}}^0\right)$
- 以下が詳細
  - 1が埋め込み及び位置埋め込みの式([;;;;]はconcatenateという意味)
  - 2がMulti head Self Attention
  - 3が2層のネットワークで以下の式
    - $\mathrm{MLP}\left(x\right) = \mathrm{GELU}(xW_1 + b_1)W_2 + b_2$
    - $\mathrm{GELU}(\mathrm{x}) = \mathrm{x}\Phi(\mathrm{x}) = \frac{\mathrm{x}}{2}\left(1 + \mathrm{erf}\left(\frac{\mathrm{x}}{\sqrt{2}}\right)\right)$
  - 2と3は$l = 1, \cdots, L$回繰り返す
  - 4がMLP Head
    - タスクによってyを入れるこの後のMLPが変わる
    - z_L^0はL層目の出力の0番目の要素、つまりCLSトークンのこと
  - $E \in \mathbb{R}^{(P^2C)\times D}, E_{pos} \in \mathbb{R}^{(N+1)\times D}$