ISSN 2186-5647

P-40

ブロックウィンドウシフトを用いたECViT

日大生産工　○佐藤　旭広　　日大生産工　山内　ゆかり

1. まえがき

画像認識の分野ではCNNが主流であったが、Dosovitskiy らが提案したVision Transformer (ViT)[1]の登場により画像認識分野の大きな転換点となった。ViTは画像内の長期距離依存関係を捉えることが可能になり、大規模なデータセットでの学習においてCNNを上回る性能を発揮することができた。しかし、小規模なデータセットでの学習はCNNに匹敵できるほどの性能に至っておらず、ViTとCNNを組み合わせるハイブリッドモデルの研究が活発に行われている。Zhoujie QianらはViTにピラミッド構造を統合することで、トークン数および特徴次元を段階的に調整する設計を採用し、性能を損なうことなくリソースの最適利用を目指したEfficient Convolutional Vision Transformer (ECViT)[2]を提案した。これによりCIFAR-10などの小規模なデータセットにおいても他のCNNベースモデルと同等な精度を発揮し、他のViTモデルよりも高い精度を発揮したことが報告がされている。しかし、問題点としてトークンをブロックに分割し、ブロックごとにMulti-head Self-Attentionを行うことで画像全体の関係性を直接計算する能力に制限がある。

本研究では，その問題に対して、ブロックウィンドウをシフトすることで、広い範囲の関係性が計算できるような構造を提案する。

2. 従来研究

2.1 ECViT[2]

ECViTは3つのステージで構成されている。第1ステージでは画像から特徴マップを生成し、その後に従来のViTと同様にトークン化を行う。第2、3ステージは同じ構造を持ち、Transformer Encoder層とマージング層から構成されている。Encoder層にはP-MSAとI-FFNの2つのサブレイヤを持ち、このEncoder層を複数回繰り返して学習を行う。

2.1.1低ランク特徴抽出による画像トークン化

従来のViTでは画像をそのままパッチに分割し、トークン化を行うが、ECViTでは畳み込み処理により得られた特徴マップからトークンを生成する。具体的には2つの畳み込み層と1つのMaxPooling層からなる。

初めに、水平方向の特徴の抽出し、出力$f_h(x)$を得る。式(1)に出力$f_h(x)$を示す。

$$f_h(x) = GeLU(BN(Conv2d(x))) \tag{1}$$

ここでxは入力である。

次に出力$f_h(x)$に対して垂直方向の特徴の抽出を行い、出力$f_v(x)$を得る。

その後、MaxPoolingを適用して重要な特徴を抽出し、出力x_lを得る。式(2)は出力x_lを示す。

$$x_l = MaxPooling(f_v(f_h(x))) \tag{2}$$

これにより特徴マップが得られ、従来のViT同様にパッチトークンを生成し、空間的な位置情報を加え、学習可能なクラストークンをパッチトークンの先頭に追加する。このトークンが第2ステージのTransformer Encoderの入力となる。

2.1.2.Partitioned Multi-head Self-Attention

標準的なTransformerアーキテクチャではグローバルなSelf-Attentionを用いるが、この方法ではトークン数に対して計算量が2次的に増加する。そのためPartitioned Multi-head Self-Attention(P-MSA)が提案された。

P-MSAではトークンをクラストークンx_c^lとパッチトークンx_p^lに分解する。パッチトークンをB個のブロックに分割し、各ブロックの先頭にクラストークンを追加する。この動作を式(3)(4)に示す。

$$X_p = \text{Partition}(x_p^l) = [x_p^1, \cdots, x_p^b, \cdots, x_p^B] \tag{3}$$

$$x_t^b = Concat(x_c^l, x_p^b) \tag{4}$$

x_t^bは各ブロックに分けた後にクラストークンを追加したものを表しており、それに対してMulti-head Self-Attention(MSA)を適用する。その出力$\hat{x}_t^b$を式(5)に示す。

$$\hat{x}_t^b = MSA(x_t^b) \tag{5}$$

最後に全てのブロックの結果を結合し、各ブロックのクラストークンに線形層を用いて1つのクラストークンにまとめる。この様にすること

ECViT with Block Window Shift

Asahiro SATO and Yukari YAMAUCHI

で全ブロックのローカル情報が統合され、画像の表現力が向上する。

2.1.3 Interactive Feed-forward Network

CNNの持つ局所情報抽出の強みと、Transformerが持つ長距離依存関係のモデリング能力を活かし、異なる非重複ブロック間の相互作用を促進するためにInteractive Feed-forward Network(I-FFN)が提案された。

動作としてはまず、クラストークンx_c^lとパッチトークンx_p^lに分解する。そして、パッチトークンx_p^lを元画像における相対位置に基づき2次元の特徴マップに展開し、特徴マップに対し、カーネルサイズ分の水平方向と垂直方向の2種類のdepthwise Separable Convolutionを適用する。これらの操作を式(6)(7)に示す。

$$x_p^e = ExpandTo2d(x_p^l) \tag{6}$$

$$x_p^c = GELU(BN(DepthConv(x_p^e))) \tag{7}$$

x_p^e, x_p^cはこれらの操作の出力を示す。これにより局所的な特徴が強化され、異なる空間方向の情報が統合される。畳み込み後のパッチトークンを再び1次元に変換してクラストークンと結合し、元と同じ形状のトークン列を得る。

2.1.4トークンマージ

各ステージの最後にマージ層を配置し、トークン特徴の次元数とトークン数を調整することでTransformerにピラミッド構造を組み込む。トークンマージによってトークン数を段階的に削減しながら特徴次元を拡張する。これにより、トークンはより広い空間領域にわたる複雑な視覚パターンを捉えることができる。

手順としてはまず、クラストークンとパッチトークンを分解する。パッチトークンを2次元の特徴マップに展開し、MaxPoolingを適用してダウンサンプリングを行う。これにより、空間次元が縮小されつつ、指定ウィンドウ内の最大値を選択することで最も重要な特徴が保持される。ダウンサンプリングされたパッチトークンとクラストークンを再結合し、線形層を適用してトークンの特徴次元を拡張するこれにより、縮小されたトークンをより高次元の空間へ写像し、より複雑なパターンや関係性を捉える表現能力が強化される。

3. 提案手法

P-MSAの各ブロック内のみにMSAを適用することにより画像全体のグローバルな依存関係が得られない問題に対して、ブロックウィンドウを半分シフトすることで広い範囲の関係性を求めることができる構造を提案する。

具体的には2つのTransformer Blockを1つのセットとして考える。１つ目では今まで同様にブロック内のみにMSAを適用する。そして、2つ目ではブロックウィンドウを半分ずらしてMSAを適用する。このセットを繰り返していくことで全体の関係性を求めることが可能にする。

4. 実験および検討

実験では小規模なデータセットとして知られているCIFAR-10を用いて、従来のECVITとブロックウィンドウシフトを導入したECViTの分類精度で評価を行う。結果を表1に示す。

Table 1　精度の比較

Model	acc[%]
ECViT	88.65
ECViT with Block Window Shift	88.53

結果としては、ブロックウィンドウシフトを導入しても精度に大きく変化はなかった。原因としては従来のECViTでブロックにクラストークンを先頭に追加しており、クラストークンにより全体の情報を既に共有できていると考えられる。

5. まとめ

本研究では、各ブロック内のみにMSAを適用することにより画像全体のグローバルな依存関係が得られないという問題に対し、ブロックウィンドウをシフトすることで画像全体の依存関係求める手法を提案した。しかし、精度に変化なかったので新たな手法を考案する必要がある。

参考文献

[1]A.Dosovitskiy, L.Beyer,A.Kolesnikov, D.Weissenborn, X.Zhai, T.Unterthiner, M.Dehghani, M.Mindere, G.Heigold, S.Gelly, J.Uszkoreit and N.Houlsby,"An Image is Worth 16×16 Words: Transformers for Image Recognition at Scale" arXiv:2010.11929v2(2020)pp.1-21

[2]Z.Qian,"ECViT：Efficient Convolutional Vision Transformer with Local-Attention and Multi-scale Stages" arXiv:2504.14825v1(2025)