ISSN 2186-5647

P-39

変形可能なローカルアテンション

日大生産工　○山上　真和　　日大生産工　山内　ゆかり

1. まえがき

畳み込みを用いない手法として、Vision Transformer(ViT)[1]が提案され、画像認識分野で大きな結果を出している。しかし、ViTは画像全体に対して計算を行うため、計算量が非常に大きいという欠点がある。この課題を克服するために、Swin Transformer[2]やDeformable Attention Transformer(DAT)[3]などの階層型トランスフォーマーが提案されてきた。多くの階層型トランスフォーマーでは、ローカルなアテンションとグローバルなアテンションの二回処理を行うため、依然として計算コストが高い。そこで本研究では、ローカルなアテンションに対してDeformable Attentionという変形可能なアテンションを組み合わせることにより、二段階の処理を一回に統合して計算量を削減することを目指す。また、DATと比較実験を行い、その結果を報告する。

2. 従来研究

2.1 Vision Transformer

2.1.1 概要

Vison Transformer(ViT)はVaswaniらによって提案されたTransformerアーキテクチャを画像認識分野に応用したものである。それまで主流であった畳み込みニューラルネットワークを用いないで精度・計算コストともにそれまでの最先端のモデルを上回った。ViTの構造図をFigure1に示す。

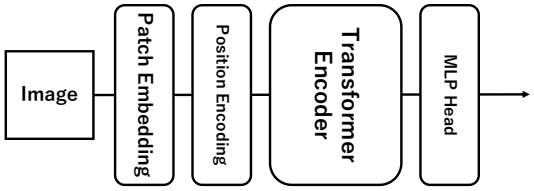

Figure 1　ViT の構造

2.1.2 Patch Embedding

ViTは自然言語処理で使われるTransforme[4]アーキテクチャを画像に応用したものであるが、画像は文章ではないので、Transformerに合う形式に変更する必要がある。そこで、画像をトークンに分割することで、画像を単語の集まり(文章)と同等に扱うことができる。

2.1.3 Position Encoding

画像をパッチに分割したことにより、文字と同等に扱うことができるようになったが、このパッチが画像のどの部分に対応するかの情報を与えるために、位置エンコーディングを行う。それと同時に、クラストークンという分類専用のパッチ(トークン)も追加する。

2.1.4 Multi-Head Attention

Multi-Head Attention(MHA)では、キューとバリュー、クエリの3つの値を用いて計算を行う。

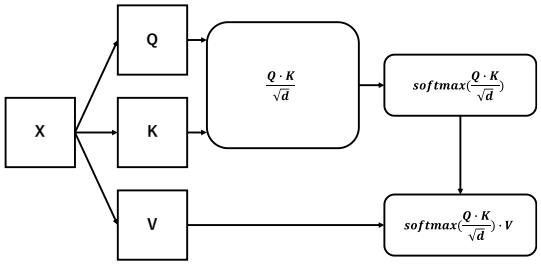

Figure 2　Multi Head Attention の構造

まず、クエリQとキーKを掛け算する。その後$\sqrt{d_k}$(dはキーベクトルの次元)で割ることで正規化し、softmax演算を行い、バリューと掛け算する。つまり、出力は

$$出力 = softmax\left(\frac{QK}{\sqrt{d_k}}\right)V \qquad (1)$$

ということになる。

2.2 Deformable Attention Transformer

2.2.1 概要

Vision Transformerでは、画像全体に対してアテンション計算を行っていたが、アテンション範囲を制限することで計算効率の向上を図る。そこで、画像の内容によってアテンション範囲を決めるDeformable Attention Transformerが提案された。DATの構造図を以下のFigure3に示す。

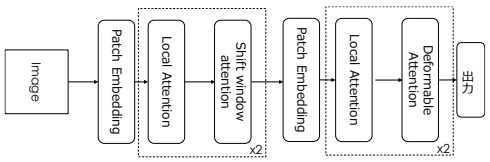

Figure 3　DAT の構造

4ステージによる構成で、前半2ステージはウインドウアテンションとシフトウインドウアテ

Deformable local attention

Masakazu YAMAGAMI and Yukari YAMAUCHI

ンションを使用しているが、後半は、ウインドウアテンションとDeformable Attentionが採用されている。

2.2.2 W-MSAとSW-MSA

DATにおいてもすべてでDeformable Attentionを使用するのではなく、序盤はよりローカルな情報を得るためにウインドウアテンションとシフトウインドウアテンションを利用している。

Figure 4　Window と Shift Window

図の左側は、ウインドウサイズごとに等間隔で分割する方法である。また、右側はウインドウサイズの半分だけシフトした位置から、ウインドウサイズごとに分割する方法である。これにより、一つ前のレイヤーで分割された隣接するウインドウと相互作用することができる。左側の方法をW-MSA、右側の方法をSW-MSAと呼んでいる。

2.2.3 Deformable Attention

Deformable attentionでは、まず、入力画像xを線形変換し、クエリqを計算する。その後サブネットワークθ_{offset}により、格子点pからの$offset\Delta p$を計算している。バイリニア補間された参照点$\tilde{x}$を線形変換して、キーとバリューを求め、それをもとにAttentionを計算している。参照点を移動させることによってアテンションする範囲を決定している。

3.　提案手法

前章までで説明した階層型トランスフォーマーは、局所的な受容野と大局的な受容野の二種類を使っていた。これに対し本研究では、ウインドウ分割されたAttention範囲に対してDATの参照点を使用する方法を参考にAttention範囲を少しだけ大きくする手法を提案する。これにより、局所的な受容野としての機能を大きく失うことなく、また、参照点によって効率的に追加されたAttention範囲によって大局的な受容野を実現している。これによって本来局所的受容野、大局的受容野の2回でAttentionを計算していたところを1回にまとめて処理することができる。以下に提案手法の全体の構造を示す。

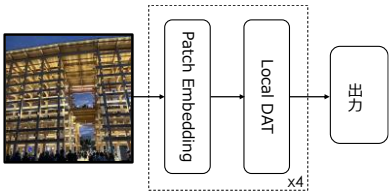

Figure 5　Deformable Attention の構造

4.　実験および検討

実験条件について、データセットはCIFAR-10、学習率0.0005、30エポック、バッチサイズ128で画像分類タスクによる実験を5回行った。5回の平均値を以下の表に示す。

Table 1　実験結果

	DAT	LocalDAT
実行時間[s]	1983.97	2276.80
精度[%]	69.79	68.74

実験結果では、従来手法の方が実行時間が短いという結果になっている。これの原因としてDeformable Attentionの処理が重いということが挙げられる。従来研究の章で説明した通り、DATはクエリに対して参照点をオフセットネットワークを通して移動するという処理を行う為、その操作がボトルネックとなっている化膿性が考えられる。また、従来のDATではstage3,4のみにDeformable Attentionを採用しているが、提案手法ではstage1~4すべてに使用している。これらが実行時間の増加を招いていると考えられる。

5.　まとめ

本研究では、階層型トランスフォーマーのさらなる計算量の削減を目指して実験を行った。画像分類タスクによる実験では精度、実行速度ともに従来手法を上回ることができなかった。この結果により、全ての階層においてLocalDATを採用することは不適切であり、DAT同様低いステージにおいてはウインドウアテンションとシフトウインドウアテンションを採用するなどモデルの構造について再考する必要があることが分かった。

今後は、各ステージにおけるDeformable Attentionの採用不採用にした場合における実験を行い、どの階層でLocalDATを採用するのがよいかを研究する。

参考文献

1)　Alexey Dosovitrkiy, Lucas Beyer, Alexander Kolesnikov, Dirk

Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Miderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, Neil Houlsby, “An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale”, arXiv:2010.11929(2021)

2) Ze Liu, Yutong Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, Baining Guo, “Swin Transformer: Hierarchical Vision Transformer using Shifted Windows”, arXiv:2103.14030(2021)
3) Zhuofan Xia, Xuran Pan, Shiji Song, Li Erran Li, Gao Huang, “Vision Transformer with Deformable Attention”, arXiv2201.00520(2022)
4) Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aiden N. Gomez, Lukasz Kaiser, Illia Polosukhin, “Attention Is All You Need”, arXiv:1706.03762