1Q-07

高速な Swin Transformer についての考察

加古 遼太郎　　松井 勇佑

東京大学

1　序論

Transformer モデル [1] は，自然言語処理，画像処理などの分野でその有効性を認められ，注目を集めている．Transformer モデルは Self-attention（図 1）をネットワーク内に組み込むことで，様々なタスクの精度向上を可能とした．しかし，Self-attention の計算量は，入力系列長 n に対し空間計算量と時間計算量ともに $\mathcal{O}(n^2)$ であることが知られている [2]．この計算の効率化および高速化が求められており，そのために様々な手法が提案されている．

Swin Transformer [3] は，画像を対称とした Transformer の 1 つである．Swin-Transformer は，画像を分割してそれぞれで Self-attention を計算することにより，計算の高速化を図っている．一方で，画像の分割の方法を固定せず可変にすることにより，画像分類などのタスクの精度も高く維持している．しかし，画像の解像度を大きくする場合は，分割する回数も大きくしないと時間計算量は結局 n の二乗に比例して増加する．さらに，分割を増加させたことによる精度への影響は明らかでない．

そこで本論文では，Swin Transformer の利点を活かしつつさらに高速化する手法について議論，提案する．

2　関連研究

2.1　Self-attention

Self-attention は，Transformer を特徴づける重要な要素技術である [1]．Self-attention $\mathbf{A} \in \mathbb{R}^{n \times c}$ は，以下のように定義される．

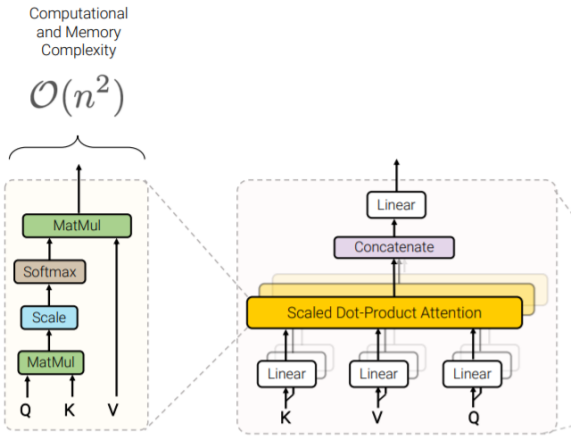

図 1: Self-attention のブロック図 [2]

$$\mathbf{A} = \mathbf{softmax}(\mathbf{Q}\mathbf{K}^\top)\mathbf{V} \tag{1}$$

ここで，$\mathbf{Q}, \mathbf{K}, \mathbf{V} \in \mathbb{R}^{n \times c}$ は，学習される行列，$n \in \mathbb{N}$ は入力系列長，$c \in \mathbb{N}$ は一つの入力系列の特徴量の次元数である．また，**softmax** 関数は，ベクトルに対する softmax 関数を行列のそれぞれの行に適用する関数である．入力系列は，ここでは画像を正方形のグリッドに分割したものである．この式に加え，正規化項などを導入することもある．この定義式より，式 (1) の時間計算量は $\mathcal{O}(n^2 c)$ であることがわかる．

2.2　Swin Transformer

Swin Transformer [3] は，Transformer の計算量を改善したネットワークである．図 2 に示すように，画像をいくつかの単位（Window）に分割し，それぞれに対して Self-attention を計算する．Window の数を w とすると，1 つの Window に含

図 2: Swin Transformer は画像をいくつかの単位（Window）に分割する [3]

まれる系列の数を s とし

$$n = ws \tag{2}$$

が成立する．またこのときの Self-attention の計算の時間計算量は $\mathcal{O}(s^2c)$ であり，これを w 個の Window に対して計算するので合計で $\mathcal{O}(w(s^2c))$ となる．$w \gg s$ のとき，時間計算量は w について一次間数的に増加すると近似することができる．

3 提案手法

提案手法では，$\mathbf{Q}, \mathbf{K}$ の次元を圧縮し，精度を保ちつつ計算量を下げる．つまり，Self-attention を次のように定義し直す．

$$\widehat{\mathbf{A}} = \mathbf{softmax}((\mathbf{QR})(\mathbf{KS})^\top)\mathbf{V} \tag{3}$$

ここで，$\mathbf{Q}, \mathbf{K}, \mathbf{V} \in \mathbb{R}^{n\times c}$ および $\mathbf{R}.\mathbf{S} \in \mathbb{R}^{c\times d}$ は，学習される行列，$n \in \mathbb{N}$ は入力系列長，$c \in \mathbb{N}$ は一つの入力系列の特徴量の次元数，d はハイパーパラメータで $d < c$ である．このとき，$\mathbf{QR}$ および $\mathbf{KS}$ の計算量は $\mathcal{O}(scd)$, $((\mathbf{QR})(\mathbf{KS})^\top)$ の計算量は $\mathcal{O}(s^2d)$ である．したがって, softmax 関数の引数の部分に着目すると，Swin Transformer の計算量は $\mathcal{O}(w(s^2c))$，提案手法の計算量は $\mathcal{O}(w(scd + s^2d))$ である．提案手法では，計算の一部について s について線形時間となる．

4 今後の課題

オリジナルの Swin Transformer モデルと提案手法のモデルで Self-attention の計算時間を計測したものを表 1 に示す．この表に示すとおり，時間計算量の議論上では高速化するが，実際には高速になっていないことがわかる．今後の課題として，この点を考察する必要がある．

表 1: それぞれのネットワークにおける Self-attention を 1 回計算するのにかかる時間．いずれのネットワークも $c = 32, s = 49$

ネットワーク	時間
オリジナル [1]	50.541ms
提案手法 ($d = 32$)	70.194ms
提案手法 ($d = 16$)	66.304ms
提案手法 ($d = 8$)	60.311ms

原因としては，s が小さいことが考えられる．今回の実験では $s = 49$ であったため，計算量における定数倍が無視できず，このような結果になったと考えられる．今後，s を増加させた場合についても検証が必要である．対称とする画像の解像度を高くする場合，一定の精度を得るには s を増加させる必要があるため，s を増加させた場合について考察することは非常に重要である．

参考文献

[1] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pp. 5998–6008, 2017.

[2] Yi Tay, Mostafa Dehghani, Dara Bahri, and Donald Metzler. Efficient transformers: A survey. September 2020.

[3] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. March 2021.