Token 合併と後処理量子化に基づく Vision Transformer 推論高速化の実装と検証

s2410155 Feng, Mingyu

近年，人工知能技術の急速な発展に伴い，深層ニューラルネットワークはコンピュータビジョン，自然言語処理，音声認識などの幅広い分野で実用化が進んでいる．深いネットワーク構造や大規模なパラメータ数により，従来の手動設計による特徴量では捉えられなかった高次かつ抽象的な表現を学習できる点が，高い性能を実現している要因である．一方で，このような高性能モデルは計算量およびメモリ消費が大きく，学習時には高性能・高消費電力な計算資源を必要とするだけでなく，推論時においても高い計算負荷，GPU メモリ使用量の増大，および推論遅延を引き起こす．これらの要因は，エッジデバイスやリアルタイム処理，計算資源が制約された環境への展開を困難にしている．

この課題に対し，近年では深層ニューラルネットワークの実運用性を高めるため，モデル圧縮および推論高速化に関する研究が盛んに行われている．モデルプルーニングでは，重みやニューロンを単位とした細粒度プルーニングから，チャネル，層，ブロック単位で構造を簡略化する構造化プルーニングまで，さまざまな手法が提案されており，パラメータ数削減や実行効率向上が報告されている．また，モデル量子化（Quantization）では，重みや活性値を浮動小数点数から低ビット幅の整数表現（例：INT8）へ変換することで，計算量およびメモリ帯域の削減を実現し，GPU Tensor Core などの低精度演算資源を有効活用できる点が注目されている．

しかしながら，これらの手法には依然としていくつかの課題が残されている．高い精度を維持するため，多くのプルーニング手法や量子化対応学習（QAT: Quantization-Aware Training）はファインチューニング（fine-tuning）を前提としており，追加の学習データ，計算時間，計算資源を必要とする．そのため，学習済みモデルのみが利用可能な場合や，迅速なモデル展開が求められる実運用環境においては，これらの学習依存手法が導入上の負担となる．さらに，アルゴリズム上の計算削減効果が，必ずしも実際のハードウェア推論における高速化に直結しない点も課題である．これは，カーネル実装の差異，メモリアクセスのボトルネック，カーネル融合の有無，および推論フレームワークにおける低精度演算対応状況などに起因する．特に Vision Transformer（ViT）では，Self-Attention の計算量が長いトークン系列の二乗に比例するため，入力トークン数の増加が計算量とメモリ帯域の双方を急激に増大させ，高解像度入力や長い系列処理において顕著な性能ボトルネックとなる．したがって，精度を維持しつつトークン数および低精度計算負荷を削減し，それらを実際の推論エンジンおよび GPU ハードウェア上で有効に機能させることが，重要な研究課題である．

以上の背景を踏まえ，本研究では学習済み Vision Transformer を対象とし，再学習を必要としない推論高速化手法の検討を行う．具体的には，モデルレベルでの計算削減手法として Token Merging を用い，推論過程において類似トークンを動的に統合することで長い系列を削減し，Self-Attention 計算量の低減を図る．さらに，学習後量子化（PTQ: Post-Training Quantization）を適用し，モデルを FP32 から INT8 などの低精度表現へ変換することで，演算量およびメモリ使用量の削減を行う．

そのため本論文では，これらの手法を統合した推論パイプラインを構築し，Vision Transformer の

代表的モデルである ViT-B/16 を用いて評価を行った．その結果，分類精度の低下を 1.2% 以内に抑えたまま，FP32 推論と比較して 3.09〜3.55 倍の推論高速化を実現した．また，NVIDIA が提供する Transformer 向け汎用 INT8 量子化手法と比較しても，同等の精度条件下において最大 12.8% 高速な推論性能を達成した．

本研究は，Token Merging と PTQ を統合することで，精度低下を最小限に抑えつつ推論計算量を大幅に削減し，GPU 実機推論における有効性を実証することを目的とする．また，モデル圧縮および低精度推論において生じるボトルネックを整理し，実装レベルでの検討を通じて，Vision Transformer（Vision Encoder）の推論最適化に関する実践的かつ再現性のある知見を示す．