JAIST Repository
https://dspace.jaist.ac.jp/

Title	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 [課題研究報告書]
Author(s)	馮, 名宇
Citation	
Issue Date	2026-03
Type	Thesis or Dissertation
Text version	author
URL	https://hdl.handle.net/10119/20560
Rights	
Description	Supervisor:井口　寧, 先端科学技術研究科, 修士(情報科学)

JAIST
JAPAN
ADVANCED INSTITUTE OF
SCIENCE AND TECHNOLOGY

Japan Advanced Institute of Science and Technology

Token Merging and Post-Training Quantization for Implementation and Verification of Vision Transformer Inference Acceleration

s2410155 Feng, Mingyu

In recent years, with the rapid advancement of artificial intelligence technologies, deep neural networks have been increasingly deployed in a wide range of fields, including computer vision, natural language processing, and speech recognition. Owing to their deep network architectures and large numbers of parameters, these models are able to learn high-level and abstract representations that could not be captured by conventional manually designed features, which is a key factor behind their high performance. On the other hand, such high-performance models require large amounts of computation and memory. They not only demand high-performance and high-power-consumption computational resources during training, but also incur substantial computational load, increased GPU memory usage, and inference latency during deployment. These factors make it difficult to apply such models to edge devices, real-time processing, and environments with limited computational resources.

To address these challenges, extensive research has been conducted on model compression and inference acceleration to improve the practical usability of deep neural networks. In model pruning, a variety of methods have been proposed, ranging from fine-grained pruning at the level of individual weights or neurons to structured pruning that simplifies network architectures at the channel, layer, or block level, achieving reductions in parameter count and improvements in execution efficiency. In addition, model quantization converts weights and activations from floating-point representations to low-bit-width integer formats (e.g., INT8), thereby reducing computational cost and memory bandwidth requirements, and enabling effective utilization of low-precision computing resources such as GPU Tensor Cores.

However, several challenges still remain with these approaches. To maintain high accuracy, many pruning methods and quantization-aware training (QAT) techniques assume fine-tuning, which requires additional training data, computational time, and computational resources. As a result, in practical deployment scenarios where only pretrained models are available or rapid model deployment is required, such training-dependent methods impose a significant burden. Furthermore, reductions in computational complexity at the algorithmic level do not necessarily translate directly into actual inference speedups on real hardware. This is due to differences in kernel implementations, memory access bottlenecks, the presence or absence of kernel fusion, and the degree of support for low-precision computation in inference frameworks. In particular, for Vision Transformers (ViTs), the computational cost of self-attention scales quadratically with the length of the token sequence. As a result, an increase in the number of input tokens leads to a rapid growth in both computational cost and memory bandwidth requirements, becoming a significant performance bottleneck for high-resolution inputs and long-sequence processing. Therefore, reducing the number of tokens and the load of low-precision computation while maintaining accuracy, and ensuring that these reductions function effectively on actual inference engines and GPU hardware, is an important research challenge.

Based on the above background, this study investigates retraining-free inference acceleration

methods for pretrained Vision Transformers. Specifically, Token Merging is employed as a model-level computation reduction technique, in which similar tokens are dynamically merged during the inference process to shorten long token sequences and reduce the computational cost of self-attention. In addition, post-training quantization (PTQ) is applied to convert the model from FP32 to low-precision representations such as INT8, thereby reducing computational cost and memory usage.

Accordingly, this thesis constructs an inference pipeline that integrates these techniques and evaluates its performance using ViT-B/16, a representative Vision Transformer model. Experimental results show that, while keeping the classification accuracy degradation within 1.2%, the proposed approach achieves a 3.09–3.55× inference speedup compared with FP32 inference. Furthermore, when compared with NVIDIA's general-purpose INT8 quantization method for Transformer models, the proposed approach achieves up to 12.8% faster inference performance under equivalent accuracy conditions.

This study aims to demonstrate the effectiveness of integrating Token Merging and PTQ in significantly reducing inference computational cost while minimizing accuracy degradation through GPU-based inference experiments. In addition, this work organizes the bottlenecks that arise in model compression and low-precision inference, and through implementation-level investigations, presents practical and reproducible insights into inference optimization for Vision Transformers (Vision Encoders).