Hardware-Aware Design-Space Exploration for Token Merging in Recursive Vision Transformers

ID: MP-ML-OPTIMISE-AUTO
Level: MSc
Contact: dr.ir. Uraz Odyurt u.odyurt@utwente.nl, dr. Amirreza Yousefzadeh a.yousefzadeh@utwente.nl

Background

Vision Transformers (ViTs) achieve strong performance in computer vision, but their computational and memory requirements can make deployment on resource-constrained hardware challenging. The recent **MergeOver** framework [1] combines recursive weight sharing with post-training Token Merging (ToMe), reducing the number of tokens processed during inference without requiring retraining.

MergeOver evaluates several token-reduction schedules, including constant, linear, stage-wise exponential, and stage-wise single-shot strategies. The original work uses a limited set of manually selected configurations and shows that the trade-offs between accuracy, latency, throughput, and memory consumption depend strongly on the target hardware platform and batch size.

This motivates a natural follow-up question: **can the token-merging strategy and merge rates be automatically adapted to the target platform and optimisation objective?**

Project Objective

The objective of this MSc project is to develop a **design-space exploration (DSE) framework for MergeOver** that automatically searches for suitable token-merging configurations for a given hardware platform and target metric.

Rather than relying on a predefined merge schedule, the framework will explore combinations of merge rates and scheduling strategies and evaluate their impact on metrics such as:

- classification accuracy;
- inference latency;
- throughput;
- peak memory consumption;
- computational cost; and, where supported,
- energy consumption.

The resulting system should allow a user to specify a target platform and optimisation goal—for example, *minimise latency on an ARM-based edge platform while limiting the loss in accuracy to two percentage points*—and return an appropriate MergeOver configuration.

Research Question

The central research question is:

To what extent can automated design-space exploration identify platform- and objective-specific token-merging configurations that improve upon manually selected MergeOver schedules?

Relevant sub-questions include:

1. Which parameters of the token-merging schedule have the greatest impact on accuracy and hardware performance?
2. How do optimal merge configurations differ across GPU, x86 CPU, and embedded ARM platforms?
3. How does the preferred configuration change when optimising for latency, throughput, memory consumption, energy, or combinations of these metrics?
4. Can efficient DSE methods identify near-Pareto-optimal configurations using only a limited number of hardware measurements?
5. Do flexible stage- or block-specific merge policies provide better trade-offs than predefined schedule shapes?

Design Space

The project will extend the configurable token-reduction mechanisms of MergeOver. Candidate design parameters may include:

- merge rate per stage, e.g. ρ_1, ρ_2, ρ_3;
- merge-rate schedule, such as constant, linear, exponential, or single-shot;
- schedule-specific parameters, such as an exponential decay factor;
- placement of merging operations across blocks or recursive iterations;
- different merge strategies for different stages.

A candidate configuration may therefore be represented as

$$\theta = (\text{strategy}, \rho_1, \rho_2, \rho_3, \alpha, ...),$$

with the exact parameters depending on the selected strategy. All generated configurations must respect the architectural constraints imposed by MergeOver and the underlying model.

Design-Space Exploration

The first step will be to establish a reproducible benchmarking interface that takes a MergeOver configuration and measures its accuracy and hardware performance.

Different DSE techniques can then be investigated, including random or grid search, Bayesian optimisation, and multi-objective evolutionary optimisation.

For a single objective, the problem may be formulated as

$$\min_{\theta} L_{\text{target}}(\theta)$$

subject to

$$A_{\text{baseline}} - A(\theta) \leq \Delta A_{\max},$$

where L_{target} represents a deployment metric such as latency and $\Delta A_{\max}$ specifies the maximum acceptable accuracy degradation.

For multi-objective optimisation, configurations can instead be compared using their Pareto trade-offs:

$$\max_{\theta} \left(A(\theta), T(\theta), -M(\theta), -L(\theta) \right),$$

where A, T, M, and L denote accuracy, throughput, memory consumption, and latency.

An additional MSc-level direction is the use of **surrogate models** to predict hardware performance and reduce the number of expensive measurements required during exploration.

Experimental Evaluation

The project will initially use the SReT-Tiny-Distill model and ImageNet-1K setup from MergeOver to allow direct comparison with the original work.

Experiments should consider substantially different processing platforms, for example:

- an NVIDIA GPU;
- an x86 CPU;
- an ARM-based edge platform.

The automatically discovered configurations will be compared against:

1. the uncompressed SReT baseline;
2. the manually evaluated MergeOver configurations;
3. the configuration selected in the original work.

The evaluation will determine whether DSE can discover configurations that improve the accuracy-performance trade-off for particular deployment scenarios. Results should be presented using hardware-specific Pareto frontiers and an analysis of how the preferred merge strategy changes across platforms and optimisation objectives.

Expected Outcomes

The project is expected to produce:

- a configurable implementation of the MergeOver scheduling mechanism;
- an automated hardware benchmarking pipeline;
- a formalised token-merging design space;
- an implementation and comparison of suitable DSE algorithms;
- Pareto frontiers for different deployment platforms and objectives;
- an analysis of the influence of stage-specific and strategy-specific merge parameters;
- guidelines for selecting token-merging policies for different deployment scenarios.

A further contribution may be a surrogate performance model capable of predicting promising configurations before performing full hardware measurements.

Expected Research Contribution

Whereas MergeOver demonstrates that token merging can successfully be incorporated into recursive hierarchical Vision Transformers, this project moves from **manually designed compression schedules to automatically optimised, deployment-aware schedules**.

The central scientific contribution is to investigate whether token-merging policies should be treated not only as properties of the neural network, but as properties of the combination of **model, workload, optimisation objective, and target hardware**.

The project would therefore turn MergeOver from a fixed post-training compression approach into a configurable hardware-aware optimisation framework in which the amount, location, and strategy of token merging are selected according to the requirements of the target deployment.

Reference

[1] J. Kim, U. Odyurt, and A. Yousefzadeh, "MergeOver: Post-Training Token Merging for Recursive Vision Transformers," *arXiv preprint arXiv:2608.13141*, 2026. [Online]. Available: https://arxiv.org/abs/2608.13141
