Structured Pruning of Vision Transformers: Latency-Accuracy Trade-offs on ImageNet-1K

Assignee Research

June 7, 2026

Abstract

This report synthesises findings from 15 peer-reviewed papers addressing the following research question: How does structured pruning of vision transformer layers affect the trade-off between inference latency and top-1 accuracy on ImageNet-1k. 6 claims were extracted from source literature; 6 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 7.3/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: ResMLP: Feedforward Networks for Image Classification With Data-Efficient Training. Research question: How does structured pruning of vision transformer layers affect the trade-off between inference latency and top-1 accuracy on ImageNet-1k?.

2 Methodology

Systematic literature search across multiple databases yielded 15 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 7.3/10.

3 Results

15 papers retrieved. 6 claims extracted; 6 independently verified. Quality review score: 7.3/10.

4 Limitations

This report is a machine-generated literature synthesis and does not constitute original research. Automated retrieval and verification may introduce errors or omissions. Review scores reflect automated assessment, not human peer review. Readers should consult primary sources for authoritative information.

5 Extracted Claims

Claim	Verified	Confidence
ResMLP is an architecture built entirely upon multi-layer perceptrons for image classification.	✓	0.35
ResMLP alternates a linear layer where image patches interact independently and identically across channels with a two-l	✓	0.46
When trained with a modern training strategy using heavy data-augmentation and optionally distillation, ResMLP attains g	✓	0.40
ResMLP models can be trained in a self-supervised setup to remove priors from employing a labelled dataset.	✓	0.31
Adapting the ResMLP model to machine translation achieves good results.	✓	0.19
The authors share pre-trained models and code based on the Timm library.	✓	0.26

References

- https://doi.org/10.48550/arxiv.2303.03667
- https://doi.org/10.48550/arxiv.1905.05055
- https://doi.org/10.1109/tpami.2022.3206148