Novelty Assessment Report

Paper: Vulcan: Crafting Compact Class-Specific Vision Transformers For Edge Intelligence
PDF URL: https://openreview.net/pdf?id=0xE0kNdGIz
Venue: ICLR 2026 Conference Submission
Year: 2026
Report Generated: 2026-01-05

Abstract

Large Vision Transformers (ViTs) must often be compressed before they can be deployed on resource-constrained edge devices. However, many edge devices require only part of the all-classes knowledge of a pre-trained ViT in their corresponding application scenarios. This is overlooked by existing compression methods. Lightweight models produced by these methods retain a substantial amount of class-irrelevant knowledge and suffer suboptimal performance on target classes. To address this, we analyze the knowledge distribution of ViT and reveal a knowledge disentanglement within it: neurons in the feed-forward network (FFN) modules encode class-specific knowledge, while the multi-head attention (MHA) modules capture class-agnostic patterns. Building on this insight, we introduce Vulcan, a pruning-oriented post-training method for deriving compact class-specific models from a pre-trained ViT under given resource budgets. Vulcan follows a novel train-then-prune paradigm, which introduces redundancy into ViTs deliberately by collapsing FFN neurons onto those with the highest class-specific activations and by enforcing low-rankness in MHA weights. This design mitigates the irreversible knowledge loss of direct pruning, so that the post-trained model can be compressed into a compact one with negligible performance loss. Notably, the derived edge ViTs not only achieve significant reductions in size and computation but also even surpass the original ViTs in performance on specific classes. Comprehensive experiments with five base ViTs covering three representative visual tasks on four datasets demonstrate that Vulcan-derived ViTs outperform the base ViTs on class-specific tasks by up to 15.12\% in accuracy, with only 20\%–40\% of their sizes. Compared with state-of-the-art structured pruning methods, Vulcan improves class-specific accuracy by up to 13.92\%. Code is available at Vulcan.

Core Task Landscape

This paper addresses: **Class-Specific Vision Transformer Compression for Edge Deployment**

A total of **50 papers** were analyzed and organized into a taxonomy with **25 categories**.

Taxonomy Overview

The research landscape has been organized into the following main categories:

- **Compression Techniques**
- **Knowledge Distillation and Transfer**
- **Distributed and Collaborative Inference**
- **Hardware-Aware Optimization**
- **Lightweight Architecture Design**
- **Domain-Specific Applications**
- **Training and Adaptation Strategies**
- **Robustness and Adversarial Analysis**
- **Binarization and Extreme Compression**
- **Surveys and Comparative Studies**

Complete Taxonomy Tree

- Class-Specific Vision Transformer Compression for Edge Deployment Survey Taxonomy
- Compression Techniques
  - Pruning-Based Compression
  - Structured Pruning (2 papers)
    - [6] Efficient Vision Transformers for Edge Devices: Pruning and Quantization Approaches (Shashank Pareek, 2024) View paper
    - [16] ViT Hybrid Channel Fit Pruning Algorithm for Co-optimization of Hardware and Software for Edge Device (Fang Liu, 2024) View paper
  - Frequency-Domain Pruning (1 papers)
    - [12] Dct-vit: High-frequency pruned vision transformer with discrete cosine transform (Jong-Ho Lee, 2024) View paper
  - Class-Specific and Task-Adaptive Pruning ★ (2 papers)
    - [0] Vulcan: Crafting Compact Class-Specific Vision Transformers For Edge Intelligence (Anon et al., 2026) View paper
    - [25] NuWa: Deriving Lightweight Task-Specific Vision Transformers for Edge Devices (He Qiang, 2025) View paper
  - Quantization-Based Compression
  - Post-Training Quantization (2 papers)
    - [7] On-Edge Deployment of Vision Transformers for Medical Diagnostics Using the Kvasir-Capsule Dataset (Dara Varam, 2024) View paper
    - [36] Trio-ViT: Post-Training Quantization and Acceleration for Softmax-Free Efficient Vision Transformer (Huihong Shi, 2024) View paper
  - Quantization-Aware Training (2 papers)
    - [11] A 28nm 343.5 fps/W Vision Transformer Accelerator with Integer-Only Quantized Attention Block (Cheng Chen Lin, 2024) View paper

        - [27] Trimming Down Large Spiking Vision Transformers via Heterogeneous Quantization Search (Boxun Xu, 2024) View paper
    - FPGA-Oriented and Hardware-Specific Quantization (1 papers)
        - [46] An FPGA-Oriented Quantization Approach for Vision Transformer with LUT-Friendly Operations (Cheng Xu, 2024) View paper
    - Token Compression and Merging
    - Token Pruning and Selection (2 papers)
        - [21] Token Compression Meets Compact Vision Transformers: a Survey and Comparative Evaluation for Edge AI (Nguyen Phat, 2025) View paper
        - [43] TA-ASF: Attention-sensitive Token Sampling and Fusing for Visual Transformer Models on the Edge (Junquan Chen, 2024) View paper
    - Token Merging and Fusion (2 papers)
        - [4] Efficient Token Compression for Vision Transformer with Spatial Information Preserved (Mao, 2025) View paper
        - [30] Extreme Model Compression for Edge Vision-Language Models: Sparse Temporal Token Fusion and Adaptive Neural Compression (Md Tasnin Tanvir, 2025) View paper
    - Low-Rank Approximation and Hybrid Methods (2 papers)
    - [38] ViT-CAAC: Contribution-Aware Adaptive Compression Framework for Vision Transformers (Yu Zhang, 2025) View paper
    - [45] MLoRQ: Bridging Low-Rank and Quantization for Transformer Compression (Gordon, 2025) View paper
- Knowledge Distillation and Transfer
    - Cross-Architecture Distillation (2 papers)
    - [32] Cross-Architecture Knowledge Distillation (KD) for Retinal Fundus Image Anomaly Detection on NVIDIA Jetson Nano (Berk Yilmaz, 2025) View paper
    - [35] Real-Time Aerial Fire Detection on Resource-Constrained Devices Using Knowledge Distillation (Khan, 2025) View paper
    - Fine-Grained and Manifold Distillation (1 papers)
    - [39] Learning Efficient Vision Transformers via Fine-Grained Manifold Distillation (Hao, 2022) View paper
    - Data-Free and Compression-Integrated Distillation (2 papers)
    - [19] Optimizing Vision Transformers with Data-Free Knowledge Transfer (Habib, 2024) View paper
    - [50] A QoS-Aware Training Framework for ViT Compression, Partition, and Distillation (Changyao Lin, 2024) View paper
- Distributed and Collaborative Inference
    - Edge-Cloud Collaborative Inference (3 papers)
    - [8] SPViT: Accelerate Vision Transformer Inference on Mobile Devices via Adaptive Splitting and Offloading (Sifan Zhao, 2025) View paper
    - [9] StressViT: Splitting and Compressing Vision Transformer Through Edge-Cloud Collaboration (Changyao Lin, 2024) View paper
    - [31] Collaborative Intelligence For Vision Transformers: A Token Sparsity-Driven Edge-Cloud Framework (Monikka Roslianna Busto, 2024) View paper
    - Multi-Device Edge Collaboration (3 papers)
    - [2] Efficient Partitioning Vision Transformer on Edge Devices for Distributed Inference (Xiang Liu, 2025) View paper
    - [15] DeViT: Decomposing Vision Transformers for Collaborative Inference in Edge Devices (Xu Guanyu, 2023) View paper
    - [42] Efficient Inference of parallel partitioned hybrid-Vision Transformers (Oscar Artur Bernd Berg, 2025) View paper
    - Communication-Efficient Parallel Inference (1 papers)
    - [41] Communication-Efficient Multi-Device Inference Acceleration for Transformer Models (Liu, 2025) View paper
- Hardware-Aware Optimization
    - Specialized Hardware Accelerators (3 papers)
    - [10] Vit-slice: End-to-end vision transformer accelerator with bit-slice algorithm (Dong-Jin Shin, 2024) View paper
    - [24] Efficient Edge Vision Transformer Accelerator with Decoupled Chunk Attention and Hybrid Computing-In-Memory (Yi Li, 2025) View paper
    - [26] ViTSen: Bridging Vision Transformers and Edge Computing With Advanced In/Near-Sensor Processing (Sepehr Tabrizchi, 2024) View paper
    - Multi-Stage Optimization Pipelines (2 papers)
    - [22] EdgeFlex-Transformer: Transformer Inference for Edge Devices (Shoaib Mohammad, 2025) View paper
    - [28] Co-optimized Vision Transformer Deployment on Edge Devices: Algorithm-Hardware-Compiler 3D Evolution (Yifan Wu, 2025) View paper
- Lightweight Architecture Design
    - Efficient Attention Mechanisms (2 papers)
    - [5] Design and Implementation of Lightweight Vision Transformer for Low-Power Edge Devices (Kaixin Zheng, 2025) View paper
    - [18] MicroViT: A Vision Transformer with Low Complexity Self Attention for Edge Device (Novendra Setyawan, 2025) View paper
    - Hybrid and Transmission-Friendly Architectures (3 papers)
    - [17] Tformer: A transmission-friendly vit model for iot devices (Lu, 2022) View paper
    - [34] LoRA-Augmented ConvMixed-ViT Architecture for Adaptive Compressive Sensing in Resource-Constrained AIoT Scenarios (Yufeng Zhou, 2025) View paper
    - [47] ConvMixed-ViT Architecture Based on LoRA for Compressive Sense in 6G-AIoT at Resource Constrained Environment (Yufeng Zhou, 2024) View paper
- Domain-Specific Applications
    - Security and Malware Detection (1 papers)
    - [20] ViT4Mal: Lightweight Vision Transformer for Malware Detection on Edge Devices (Akshara Ravi, 2023) View paper
    - Surveillance and Real-Time Vision (1 papers)
    - [23] Dynamic Query Vision Transformers and Hierarchical Latent Compression: Advancing Real Time Surveillance Systems (ASV Rao, 2025) View paper
- Training and Adaptation Strategies
    - On-Device Learning and Fine-Tuning (2 papers)
    - [33] On-device Learning and Inference Optimization for Lightweight Neural Networks and Transformers on Microcontrollers (Dequino, 2025) View paper
    - [44] Block Selective Reprogramming for On-device Training of Vision Transformers (Sreetama Sarkar, 2024) View paper
- Robustness and Adversarial Analysis (1 papers)
    - [48] Attacking Compressed Vision Transformers (Swapnil Parekh, 2022) View paper

- Binarization and Extreme Compression (1 papers)
  - [49] GSB: Group Superposition Binarization for Vision Transformer with Limited Training Samples (Tian Gao, 2023) View paper
- Surveys and Comparative Studies (7 papers)
  - [1] Efficient vision transformer inference on edge devices (Rofes, 2025) View paper
  - [3] Towards efficient vision transformer inference: A first study of transformers on mobile devices (Xudong Wang, 2022) View paper
  - [13] Compressing Vision Transformers for Low-Resource Visual Learning (Eric Youn, 2023) View paper
  - [14] Optimizing Transformer Models for Resource-Constrained Environments: A Study on Compression Techniques for Edge Computing (Yao, 2024) View paper
  - [29] DeepCompress-ViT: Rethinking Model Compression to Enhance Efficiency of Vision Transformers at the Edge (Sabbir Ahmed, 2025) View paper
  - [37] Vision Transformers on the Edge: A Comprehensive Survey of Model Compression and Acceleration Strategies (Lanyu, 2025) View paper
  - [40] Comprehensive Survey of Model Compression and Speed up for Vision Transformers (Feiyang Chen, 2024) View paper

Narrative

Core task: class-specific vision transformer compression for edge deployment. The field addresses the challenge of deploying large Vision Transformers (ViTs) on resource-constrained edge devices by developing methods that reduce model size, computational cost, and memory footprint while preserving accuracy. The taxonomy reveals a rich landscape organized around ten major branches. Compression Techniques encompass pruning-based methods (including class-specific and task-adaptive pruning), quantization, and token reduction strategies such as Efficient Token Compression[4]. Knowledge Distillation and Transfer explores how to transfer learned representations from large teacher models to compact student networks, as seen in works like Manifold Distillation ViT[39]. Distributed and Collaborative Inference investigates partitioning strategies (e.g., Partitioning ViT Edge[2]) and multi-device execution (Multi-Device Transformer Inference[41]). Hardware-Aware Optimization targets specific accelerators and FPGA implementations (FPGA ViT Quantization[46]), while Lightweight Architecture Design focuses on inherently efficient architectures like MicroViT[18] and Lightweight ViT Design[5]. Domain-Specific Applications tailor compression to medical imaging (Medical ViT Deployment[7]) and other specialized tasks, and Training and Adaptation Strategies address parameter-efficient fine-tuning methods such as LoRA ConvMixed-ViT[34].

Several active research directions reveal key trade-offs and open questions. Pruning-based approaches balance granularity (structured versus unstructured) with the need for task or class adaptability, while quantization methods must navigate accuracy-efficiency frontiers across diverse hardware backends. Token compression techniques like those surveyed in Token Compression Survey[21] offer dynamic inference benefits but raise questions about which tokens to retain under varying input conditions. Within this landscape, Vulcan[0] sits in the Class-Specific and Task-Adaptive Pruning cluster, emphasizing tailored compression that adapts pruning decisions to specific classes or tasks. This contrasts with more general pruning frameworks like ViT Hybrid Pruning[16] or broader token-reduction schemes, and aligns closely with NuWa[25], which also explores adaptive strategies. Vulcan's focus on class-specific adaptation addresses a nuanced challenge: ensuring that compression does not disproportionately harm performance on particular categories, a concern particularly relevant for edge deployment where retraining opportunities are limited and diverse workloads are common.

Related Works in Same Category

The following **1 sibling papers** share the same taxonomy leaf node with the original paper:

1. NuWa: Deriving Lightweight Task-Specific Vision Transformers for Edge Devices

Authors: He Qiang, Ziteng Wei, Li Bing, Qiang He, Chen Feifei, et al. (9 authors total) | **Year/Venue**: 2025 | **URL**: View paper

Abstract

Vision Transformers (ViTs) excel in computer vision tasks but lack flexibility for edge devices' diverse needs. A vital issue is that ViTs pre-trained to cover a broad range of tasks are \textit{over-qualified} for edge devices that usually demand only part of a ViT's knowledge for specific tasks. Their task-specific accuracy on these edge devices is suboptimal. We discovered that small ViTs that focus on device-specific tasks can improve model accuracy and in the meantime, accelerate model infe...

Relationship Analysis

Both papers belong to the Class-Specific and Task-Adaptive Pruning category, focusing on deriving compact ViTs tailored to specific classes or tasks for edge deployment. They overlap in addressing the problem of removing class-irrelevant knowledge through structured pruning to create lightweight, task-specific models. However, Vulcan introduces a novel train-then-prune paradigm with class-centric neuron collapse and truncated nuclear norm regularization to deliberately introduce redundancy before pruning, while NuWa follows a conventional prune-then-train approach with dimension-specific pruning strategies (one-shot and adaptive stages) and emphasizes SVD-based compression for attention modules.

Contributions Analysis

Overall novelty summary. The paper introduces Vulcan, a pruning-oriented post-training method for deriving compact class-specific Vision Transformers from pre-trained models. It resides in the Class-Specific and Task-Adaptive Pruning leaf, which contains only two papers (including Vulcan itself and NuWa). This represents a relatively sparse research direction within the broader Pruning-Based Compression branch, suggesting that class-specific adaptation in ViT pruning remains an underexplored area compared to general-purpose structured pruning or token compression methods.

The taxonomy reveals that Vulcan's neighboring research directions include Structured Pruning (two papers), Frequency-Domain Pruning (one paper), and Token Compression methods (four papers across two sub-leaves). While these adjacent areas focus on general-purpose compression or token-level reduction, Vulcan diverges by explicitly targeting class-irrelevant knowledge removal. The broader Compression Techniques branch contains quantization and low-rank methods, but none directly address the class-specific adaptation challenge that Vulcan emphasizes, positioning it at a distinct intersection of pruning and task-aware optimization.

Among the 27 candidates examined through semantic search and citation expansion, none clearly refute Vulcan's three core contributions. The knowledge disentanglement insight (10 candidates examined, 0 refutable) and the Vulcan method itself (10 candidates examined, 0 refutable) appear novel within this limited search scope. The class-centric neuron collapse and truncated nuclear norm regularization (7 candidates examined, 0 refutable) also show no direct prior overlap. However, this analysis is constrained by the search scale and does not constitute an exhaustive literature review.

Based on the top-27 semantic matches and the sparse taxonomy leaf (only one sibling paper), Vulcan appears to occupy a relatively novel position within class-specific ViT compression. The limited number of refutable candidates and the underexplored nature of class-adaptive pruning suggest meaningful originality, though the restricted search scope means potentially relevant work outside these candidates may exist. The knowledge disentanglement insight and train-then-prune paradigm represent the most distinctive contributions within this context.

This paper presents **3 main contributions**, each analyzed against relevant prior work:

Contribution 1: Knowledge disentanglement insight in Vision Transformers

Description: The authors analyze the knowledge distribution within Vision Transformers and discover that feed-forward network (FFN) modules primarily encode class-specific knowledge, while multi-head attention (MHA) modules capture class-agnostic patterns. This insight forms the theoretical foundation for their compression approach.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. Dual Variational Knowledge Attention for Class Incremental Vision Transformer

URL: View paper

Brief Assessment

Dual Variational Knowledge[62] focuses on class incremental learning with attention mechanisms for balancing knowledge across tasks, not on analyzing knowledge distribution between FFN and MHA modules in Vision Transformers.

---

2. BFD: Binarized Frequency-enhanced Distillation for Vision Transformer

URL: View paper

Brief Assessment

BFD[68] focuses on frequency-domain knowledge distillation for binarized Vision Transformers, not on analyzing knowledge distribution across FFN and MHA modules. The candidate addresses high-frequency information preservation during binarization, which is orthogonal to the original paper's analysis of class-specific vs. class-agnostic knowledge disentanglement.

---

3. Pruning self-attentions into convolutional layers in single path

URL: View paper

Brief Assessment

Pruning Self-Attentions[61] focuses on pruning self-attention layers into convolutional layers for efficiency, not on analyzing knowledge distribution between FFN and MHA modules for class-specific vs. class-agnostic patterns.

---

4. Image Recognition with Online Lightweight Vision Transformer: A Survey

URL: View paper

Brief Assessment

Online Lightweight ViT[64] is a survey paper that reviews existing lightweight Vision Transformer techniques. It does not present original research on knowledge distribution within Vision Transformers or analyze FFN vs. MHA module functions, which are the core claims of the original contribution.

---

5. A Survey on Transformer Compression

URL: View paper

Brief Assessment

Transformer Compression Survey[66] provides a broad overview of compression techniques but does not analyze knowledge distribution patterns within Vision Transformers or discuss how FFN and MHA modules encode different types of knowledge.

---

6. Vitkd: Feature-based knowledge distillation for vision transformers

URL: View paper

Brief Assessment

ViTKD[60] focuses on distillation strategies for different layers (shallow vs. deep) but does not analyze knowledge distribution in terms of class-specific vs. class-agnostic patterns in FFN and MHA modules as the original paper does.

---

7. Kformer: Knowledge injection in transformer feed-forward layers

URL: View paper

Brief Assessment

Kformer[65] focuses on knowledge injection in language models through FFN layers for NLP tasks, not on analyzing knowledge distribution patterns in Vision Transformers. The candidate does not address class-specific vs. class-agnostic knowledge disentanglement in visual models.

---

8. KDFAS: Multi-stage Knowledge Distillation Vision Transformer for Face Anti-spoofing

URL: View paper

Brief Assessment

KDFAS[63] focuses on knowledge distillation for face anti-spoofing tasks, not on analyzing knowledge distribution within Vision Transformers. No relevant content available in the provided candidate paper context to assess overlap with the original paper's analysis of FFN and MHA modules.

---

9. RSKD: Enhanced medical image segmentation via multi-layer, rank-sensitive knowledge distillation in Vision Transformer models

URL: View paper

Brief Assessment

RSKD[67] focuses on medical image segmentation using knowledge distillation techniques. No full text context was provided for this candidate paper, making detailed comparison impossible. The query targeted Vision Transformer knowledge distribution analysis, but without accessible content, no refutation evidence can be established.

---

10. Feature-level knowledge distillation for place recognition based on soft-hard labels teaching paradigm

URL: View paper

Brief Assessment

Feature-level KD Place[69] focuses on knowledge distillation for visual place recognition using Vision Transformers, but does not analyze or discuss the internal knowledge distribution within ViT modules (FFN vs. MHA) that forms the core of the original paper's contribution.

Contribution 2: Vulcan method for deriving compact class-specific ViTs

Description: The authors introduce Vulcan, a pruning-oriented post-training method that derives compact class-specific Vision Transformers from pre-trained models. Vulcan follows a novel train-then-prune paradigm that deliberately introduces redundancy before pruning, minimizing irreversible knowledge loss during compression.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. Explainability of vision transformers: A comprehensive review and new perspectives

URL: View paper

Brief Assessment

ViT Explainability Review[59] focuses on explainability methods for vision transformers (attention visualization, pruning for interpretability), not on class-specific model compression or post-training pruning paradigms for edge deployment.

---

2. Lrp-qvit: Mixed-precision vision transformer quantization via layer-wise relevance propagation

URL: View paper

Brief Assessment

LRP-QViT[55] focuses on mixed-precision quantization of Vision Transformers using layer-wise relevance propagation, not on class-specific model derivation or pruning-based compression methods.

---

3. STPM: Spatial-Temporal Token Pruning and Merging for Complex Activity Recognition

URL: View paper

Brief Assessment

STPM[57] focuses on token pruning and merging for video transformers in complex activity recognition, not on deriving class-specific models from pre-trained ViTs through post-training compression methods.

---

4. Mix-QViT: Mixed-precision vision transformer quantization driven by layer importance and quantization sensitivity

URL: View paper

Brief Assessment

Mix-QViT[51] focuses on mixed-precision quantization for Vision Transformers, not on deriving class-specific models through pruning. The candidate addresses bit-width allocation across layers, while the original paper introduces a pruning method for class-specific knowledge extraction.

---

5. Self-distilled vision transformer for domain generalization

URL: View paper

Brief Assessment

Self-distilled ViT[58] addresses domain generalization through self-distillation techniques for Vision Transformers, not class-specific model compression or pruning methods as proposed in Vulcan.

---

6. Rethinking decoders for transformer-based semantic segmentation: A compression perspective

URL: View paper

Brief Assessment

Decoder Compression Transformers[56] focuses on semantic segmentation decoder design through a compression perspective using PCA-inspired attention mechanisms. This is fundamentally different from Vulcan's pruning-oriented post-training method for deriving compact class-specific Vision Transformers through neuron collapse and nuclear norm regularization.

---

7. Parameter-Efficient Fine-Tuning for Individual Tree Crown Detection and Species Classification Using UAV-Acquired Imagery

URL: View paper

Brief Assessment

Tree Crown Detection[52] focuses on parameter-efficient fine-tuning for adapting pre-trained models to tree detection tasks in UAV imagery, not on deriving compact class-specific models through pruning-oriented post-training methods.

---

8. VLTP: Vision-Language Guided Token Pruning for Task-Oriented Segmentation

URL: View paper

Brief Assessment

VLTP[53] focuses on task-oriented segmentation with vision-language guidance and token pruning during inference, not on deriving compact class-specific models through post-training compression as Vulcan does.

---

9. Efficient Partitioning Vision Transformer on Edge Devices for Distributed Inference

URL: View paper

Brief Assessment

Partitioning ViT Edge[2] focuses on splitting ViT models across multiple edge devices for distributed inference, not on deriving compact class-specific models through a train-then-prune paradigm as in the original paper.

---

10. The need for speed: Pruning transformers with one recipe

URL: View paper

Brief Assessment

Pruning Transformers Speed[54] focuses on one-shot pruning for general transformer compression across multiple domains without class-specific specialization. The candidate does not address class-specific model derivation or the train-then-prune paradigm that deliberately introduces redundancy before compression.

Contribution 3: Class-centric neuron collapse and truncated nuclear norm regularization

Description: The authors develop two key technical components: class-centric neuron collapse (CCNC) for FFN modules that collapses neurons onto anchor neurons with highest class-specific activations, and truncated nuclear norm regularization (TNNR) for MHA modules that enforces low-rank structures to enable near-lossless pruning via singular value decomposition.

This contribution was assessed against **7 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. Pela: Learning parameter-efficient models with low-rank approximation

URL: View paper

Brief Assessment

Pela[71] focuses on low-rank approximation for general parameter efficiency in pre-trained models, not on class-specific compression or neuron collapse techniques for vision transformers.

---

2. Dynamic Low-Rank Training with Spectral Regularization: Achieving Robustness in Compressed Representations

URL: View paper

Brief Assessment

Dynamic Low-Rank Training[70] focuses on spectral regularization for adversarial robustness in compressed networks, not on class-specific neuron collapse or nuclear norm regularization for transformer compression. The technical approaches and objectives differ fundamentally.

---

3. Frequency-Aware Token Reduction for Efficient Vision Transformer

URL: View paper

Brief Assessment

Frequency-Aware Token Reduction[74] focuses on token reduction in vision transformers through frequency-domain analysis (high/low-frequency token partitioning), not on neuron collapse regularization or low-rank approximation for transformer compression as described in the original contribution.

---

4. Weight decay induces low-rank attention layers

URL: View paper

Brief Assessment

Weight Decay Low-Rank[72] focuses on how weight decay induces low-rank structures in attention layers during training, not on neuron collapse techniques for FFN modules or pruning-oriented compression methods for class-specific model derivation.

---

5. Projection domain decomposition denoising algorithm based on low rank and similarity-based regularization.

URL: View paper

Brief Assessment

Projection Domain Denoising[75] focuses on denoising algorithms for projection domains using low-rank and similarity-based regularization, which is unrelated to transformer compression or neuron collapse techniques for vision models.

---

6. Symmetry induces structure and constraint of learning

URL: View paper

Brief Assessment

Symmetry Learning Structure[73] focuses on theoretical analysis of mirror symmetries in loss functions and their effects on learning dynamics, not on practical transformer compression techniques like class-centric neuron collapse or truncated nuclear norm regularization for vision transformers.

---

7. Towards an Effective Low-rank Compression of Neural Networks

URL: View paper

Brief Assessment

Low-rank Neural Compression[76] focuses on low-rank compression techniques for neural networks in general, while the original paper specifically addresses class-specific vision transformer compression with novel techniques (CCNC for FFN modules and TNNR for MHA modules) tailored to knowledge disentanglement in ViTs. The candidate does not demonstrate prior work on class-centric neuron collapse or truncated nuclear norm regularization for class-specific model derivation.

Appendix: Text Similarity Detection

Textual similarity detection checked 28 papers and found 1 similarity segment(s) across 1 paper(s).

The following **1 paper(s)** were detected to have high textual similarity with the original paper. These may represent different versions of the same work, duplicate submissions, or papers with substantial textual overlap. Readers are advised to verify these relationships independently.

1. Kformer: Knowledge injection in transformer feed-forward layers

Detected in: Contribution: contribution_1

⚠ **Note**: This paper shows substantial textual similarity with the original paper. It may be a different version, a duplicate submission, or contain significant overlapping content. Please review carefully to determine the nature of the relationship.

References

- [0] Vulcan: Crafting Compact Class-Specific Vision Transformers For Edge Intelligence View paper
- [1] Efficient vision transformer inference on edge devices View paper
- [2] Efficient Partitioning Vision Transformer on Edge Devices for Distributed Inference View paper
- [3] Towards efficient vision transformer inference: A first study of transformers on mobile devices View paper
- [4] Efficient Token Compression for Vision Transformer with Spatial Information Preserved View paper
- [5] Design and Implementation of Lightweight Vision Transformer for Low-Power Edge Devices View paper
- [6] Efficient Vision Transformers for Edge Devices: Pruning and Quantization Approaches View paper
- [7] On-Edge Deployment of Vision Transformers for Medical Diagnostics Using the Kvasir-Capsule Dataset View paper
- [8] SPViT: Accelerate Vision Transformer Inference on Mobile Devices via Adaptive Splitting and Offloading View paper
- [9] StressViT: Splitting and Compressing Vision Transformer Through Edge-Cloud Collaboration View paper
- [10] Vit-slice: End-to-end vision transformer accelerator with bit-slice algorithm View paper
- [11] A 28nm 343.5 fps/W Vision Transformer Accelerator with Integer-Only Quantized Attention Block View paper
- [12] Dct-vit: High-frequency pruned vision transformer with discrete cosine transform View paper
- [13] Compressing Vision Transformers for Low-Resource Visual Learning View paper
- [14] Optimizing Transformer Models for Resource-Constrained Environments: A Study on Compression Techniques for Edge Computing View paper
- [15] DeViT: Decomposing Vision Transformers for Collaborative Inference in Edge Devices View paper
- [16] ViT Hybrid Channel Fit Pruning Algorithm for Co-optimization of Hardware and Software for Edge Device View paper

- [17] Tformer: A transmission-friendly vit model for iot devices View paper
- [18] MicroViT: A Vision Transformer with Low Complexity Self Attention for Edge Device View paper
- [19] Optimizing Vision Transformers with Data-Free Knowledge Transfer View paper
- [20] ViT4Mal: Lightweight Vision Transformer for Malware Detection on Edge Devices View paper
- [21] Token Compression Meets Compact Vision Transformers: a Survey and Comparative Evaluation for Edge AI View paper
- [22] EdgeFlex-Transformer: Transformer Inference for Edge Devices View paper
- [23] Dynamic Query Vision Transformers and Hierarchical Latent Compression: Advancing Real Time Surveillance Systems View paper
- [24] Efficient Edge Vision Transformer Accelerator with Decoupled Chunk Attention and Hybrid Computing-In-Memory View paper
- [25] NuWa: Deriving Lightweight Task-Specific Vision Transformers for Edge Devices View paper
- [26] ViTSen: Bridging Vision Transformers and Edge Computing With Advanced In/Near-Sensor Processing View paper
- [27] Trimming Down Large Spiking Vision Transformers via Heterogeneous Quantization Search View paper
- [28] Co-optimized Vision Transformer Deployment on Edge Devices: Algorithm-Hardware-Compiler 3D Evolution View paper
- [29] DeepCompress-ViT: Rethinking Model Compression to Enhance Efficiency of Vision Transformers at the Edge View paper
- [30] Extreme Model Compression for Edge Vision-Language Models: Sparse Temporal Token Fusion and Adaptive Neural Compression View paper
- [31] Collaborative Intelligence For Vision Transformers: A Token Sparsity-Driven Edge-Cloud Framework View paper
- [32] Cross-Architecture Knowledge Distillation (KD) for Retinal Fundus Image Anomaly Detection on NVIDIA Jetson Nano View paper
- [33] On-device Learning and Inference Optimization for Lightweight Neural Networks and Transformers on Microcontrollers View paper
- [34] LoRA-Augmented ConvMixed-ViT Architecture for Adaptive Compressive Sensing in Resource-Constrained AIoT Scenarios View paper
- [35] Real-Time Aerial Fire Detection on Resource-Constrained Devices Using Knowledge Distillation View paper
- [36] Trio-ViT: Post-Training Quantization and Acceleration for Softmax-Free Efficient Vision Transformer View paper
- [37] Vision Transformers on the Edge: A Comprehensive Survey of Model Compression and Acceleration Strategies View paper
- [38] ViT-CAAC: Contribution-Aware Adaptive Compression Framework for Vision Transformers View paper
- [39] Learning Efficient Vision Transformers via Fine-Grained Manifold Distillation View paper
- [40] Comprehensive Survey of Model Compression and Speed up for Vision Transformers View paper
- [41] Communication-Efficient Multi-Device Inference Acceleration for Transformer Models View paper
- [42] Efficient Inference of parallel partitioned hybrid-Vision Transformers View paper
- [43] TA-ASF: Attention-sensitive Token Sampling and Fusing for Visual Transformer Models on the Edge View paper
- [44] Block Selective Reprogramming for On-device Training of Vision Transformers View paper
- [45] MLoRQ: Bridging Low-Rank and Quantization for Transformer Compression View paper
- [46] An FPGA-Oriented Quantization Approach for Vision Transformer with LUT-Friendly Operations View paper
- [47] ConvMixed-ViT Architecture Based on LoRA for Compressive Sense in 6G-AIoT at Resource Constrained Environment View paper
- [48] Attacking Compressed Vision Transformers View paper
- [49] GSB: Group Superposition Binarization for Vision Transformer with Limited Training Samples View paper
- [50] A QoS-Aware Training Framework for ViT Compression, Partition, and Distillation View paper
- [51] Mix-QViT: Mixed-precision vision transformer quantization driven by layer importance and quantization sensitivity View paper
- [52] Parameter-Efficient Fine-Tuning for Individual Tree Crown Detection and Species Classification Using UAV-Acquired Imagery View paper
- [53] VLTP: Vision-Language Guided Token Pruning for Task-Oriented Segmentation View paper
- [54] The need for speed: Pruning transformers with one recipe View paper
- [55] Lrp-qvit: Mixed-precision vision transformer quantization via layer-wise relevance propagation View paper
- [56] Rethinking decoders for transformer-based semantic segmentation: A compression perspective View paper
- [57] STPM: Spatial-Temporal Token Pruning and Merging for Complex Activity Recognition View paper
- [58] Self-distilled vision transformer for domain generalization View paper
- [59] Explainability of vision transformers: A comprehensive review and new perspectives View paper
- [60] Vitkd: Feature-based knowledge distillation for vision transformers View paper
- [61] Pruning self-attentions into convolutional layers in single path View paper
- [62] Dual Variational Knowledge Attention for Class Incremental Vision Transformer View paper
- [63] KDFAS: Multi-stage Knowledge Distillation Vision Transformer for Face Anti-spoofing View paper
- [64] Image Recognition with Online Lightweight Vision Transformer: A Survey View paper
- [65] Kformer: Knowledge injection in transformer feed-forward layers View paper
- [66] A Survey on Transformer Compression View paper
- [67] RSKD: Enhanced medical image segmentation via multi-layer, rank-sensitive knowledge distillation in Vision Transformer models View paper
- [68] BFD: Binarized Frequency-enhanced Distillation for Vision Transformer View paper
- [69] Feature-level knowledge distillation for place recognition based on soft-hard labels teaching paradigm View paper
- [70] Dynamic Low-Rank Training with Spectral Regularization: Achieving Robustness in Compressed Representations View paper
- [71] Pela: Learning parameter-efficient models with low-rank approximation View paper
- [72] Weight decay induces low-rank attention layers View paper
- [73] Symmetry induces structure and constraint of learning View paper
- [74] Frequency-Aware Token Reduction for Efficient Vision Transformer View paper
- [75] Projection domain decomposition denoising algorithm based on low rank and similarity-based regularization. View paper
- [76] Towards an Effective Low-rank Compression of Neural Networks View paper