Novelty Assessment Report

Paper: Efficient Resource-Constrained Training of Vision Transformers via Subspace Optimization
PDF URL: https://openreview.net/pdf?id=0nvQ5kHXf4
Venue: ICLR 2026 Conference Submission
Year: 2026
Report Generated: 2026-01-05

Abstract

In today's world, where AI plays a major role in everyday life, energy consumption and data privacy have become critical concerns. On-device learning offers a promising solution by enabling models to train directly on edge devices, thereby reducing energy usage and minimizing the risk of data leakage. However, the increasing size of modern neural networks poses a serious challenge for on-device training. Although prior work has mainly focused on compact convolutional architectures, we explore a different direction by applying subspace-based training to transformer models. Based on the idea that a model's essential information resides in a fixed subspace, we introduce Weight-Activation Subspace Iteration (WASI), a method designed to overcome the memory bottleneck of backpropagation and improve inference efficiency in transformer-based models by constraining training to this subspace. Our results show that, with accuracy comparable to vanilla training, WASI reduces memory usage by up to $62\times$ and computational cost (FLOPs) by up to $2\times$. Moreover, when tested on a Raspberry Pi 5, WASI delivers approximately $1.5\times$ faster training and inference than vanilla training.

Core Task Landscape

This paper addresses: **Resource-Constrained Training of Vision Transformers via Subspace Optimization**

A total of **21 papers** were analyzed and organized into a taxonomy with **14 categories**.

Taxonomy Overview

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

- **Subspace-Based Training and Optimization Methods**
- **Parameter-Efficient Adaptation of Pretrained Vision Transformers**
- **Model Compression and Pruning for Vision Transformers**
- **Efficient Attention Mechanisms and Architectural Modifications**
- **Deployment and Hardware-Aware Optimization**
- **Interpretability and Mechanistic Analysis of Vision Transformers**
- **Black-Box and Gradient-Free Optimization for Vision Models**
- **Domain-Specific Applications with Resource Constraints**

Complete Taxonomy Tree

- Resource-Constrained Training of Vision Transformers via Subspace Optimization Survey Taxonomy
- Subspace-Based Training and Optimization Methods
  - Direct Subspace Training for Resource Efficiency ★ (2 papers)
  - [0] Efficient Resource-Constrained Training of Vision Transformers via Subspace Optimization (Anon et al., 2026) View paper
  - [6] Learning scalable model soup on a single gpu: An efficient subspace training strategy (Tao Li, 2024) View paper
  - Theoretical Foundations of Subspace Representation Learning (2 papers)
  - [1] Attention-Only Transformers via Unrolled Subspace Denoising (Wang Peng, 2025) View paper
  - [2] White-box transformers via sparse rate reduction (Yu, 2023) View paper
  - Empirical Subspace Analysis and Universal Properties (2 papers)
  - [5] The universal weight subspace hypothesis (Prakhar Kaushik, 2025) View paper
  - [10] Weight Spectra Induced Efficient Model Adaptation (Si, 2025) View paper
- Parameter-Efficient Adaptation of Pretrained Vision Transformers
  - Low-Rank Adaptation Techniques (4 papers)
  - [9] Serial Low-rank Adaptation of Vision Transformer (Cai Ke, 2025) View paper
  - [18] Delta-LLaVA: Base-then-Specialize Alignment for Token-Efficient Vision-Language Models (Mohamad Zamini, 2025) View paper
  - [19] Memory-Efficient Low-Rank Fine-Tuning for AoA Estimation in AI-Enabled RANs (Zhiheng Guo, 2025) View paper
  - [21] EigenLoRAx: Efficient Low Rank Adaptation Using Recycled Principal Subspaces (P Kaushik, n.d.) View paper
  - Structured and Kronecker-Based Adaptation (2 papers)
  - [12] MixPHM: Redundancy-aware parameter-efficient tuning for low-resource visual question answering (Jingjing Jiang, 2023) View paper
  - [15] Generalized Kronecker-based Adapters for Parameter-efficient Fine-tuning of Vision Transformers (Ali Edalati, 2023) View paper
  - General Parameter-Efficient Fine-Tuning Frameworks (1 papers)
  - [4] Parameter-efficient model adaptation for vision transformers (He, 2023) View paper
- Model Compression and Pruning for Vision Transformers
  - Neural Architecture Search and Slimming (1 papers)
  - [3] Vision transformer slimming: Multi-dimension searching in continuous optimization space (Arnav Chavan, 2022) View paper
  - Low-Rank Decomposition for Compression (1 papers)

- [8] Structure-Preserving Network Compression Via Low-Rank Induced Training Through Linear Layers Composition (Zhang Xi-tong, 2024) View paper
- Efficient Attention Mechanisms and Architectural Modifications (1 papers)
  - [7] CS-VLM: Compressed Sensing Attention for Efficient Vision-Language Representation Learning (Kiruluta, 2025) View paper
- Deployment and Hardware-Aware Optimization (1 papers)
  - [14] Mcuformer: Deploying vision tranformers on microcontrollers with limited memory (Liang Yiâ□nan, 2023) View paper
- Interpretability and Mechanistic Analysis of Vision Transformers (1 papers)
  - [11] Steering CLIP's vision transformer with sparse autoencoders (Joseph Sonia, 2025) View paper
- Black-Box and Gradient-Free Optimization for Vision Models (1 papers)
  - [16] Query Efficient Black-Box Visual Prompting with Subspace Learning (Zhaogeng Liu, 2025) View paper
- Domain-Specific Applications with Resource Constraints
  - Self-Supervised Learning for Limited Data Scenarios (1 papers)
  - [13] TransBlast: Self-supervised learning using augmented subspace with transformer for background/foreground separation (Islam I. Osman, 2021) View paper
  - Transfer Learning for Low-Resource Domains (2 papers)
  - [17] Integrating Vision Transformers with Transfer Learning for Enhanced Kidney Cancer Classification in CT Imaging (AM Ali, 2025) View paper
  - [20] Optimizing Transformer-based Models for Low-Resource Languages Using Transfer Learning Techniques (Matthew, n.d.) View paper

Narrative

Core task: resource-constrained training of vision transformers via subspace optimization. The field addresses the challenge of training or adapting large vision transformers when computational resources, memory, or data are limited. The taxonomy reveals several complementary strategies: subspace-based training methods that confine optimization to lower-dimensional parameter manifolds; parameter-efficient adaptation techniques that modify only small portions of pretrained models; model compression and pruning approaches that reduce network size; efficient attention mechanisms that lower computational complexity; deployment-focused methods for hardware-aware optimization; interpretability studies that reveal internal representations; black-box optimization for gradient-free scenarios; and domain-specific applications under resource constraints. Works such as Subspace Optimization Training[0] and Universal Weight Subspace[5] exemplify direct subspace training, while Parameter Efficient Adaptation[4] and Serial Low Rank Adaptation[9] illustrate lightweight fine-tuning strategies. Vision Transformer Slimming[3] and Compressed Sensing Attention[7] represent compression and architectural efficiency, respectively.

A particularly active line of research explores how low-rank or subspace constraints can be imposed during training or adaptation, balancing memory savings against model expressiveness. Some methods like Low Rank Induced Training[8] and Weight Spectra Adaptation[10] focus on inducing structured sparsity or spectral properties, while others such as Model Soup Subspace[6] and Unrolled Subspace Denoising[1] investigate subspace geometry for model merging or denoising. The original paper, Subspace Optimization Training[0], sits squarely within the direct subspace training branch, emphasizing resource efficiency by optimizing in a reduced parameter space from the outset. Compared to nearby works like Model Soup Subspace[6], which explores subspace connectivity among multiple trained models, Subspace Optimization Training[0] targets the training phase itself, aiming to achieve competitive performance without ever instantiating the full parameter set. This contrasts with post-hoc compression methods and highlights an emerging theme: proactive dimensionality reduction as a first-class training strategy rather than a post-processing step.

Related Works in Same Category

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

1. Learning scalable model soup on a single gpu: An efficient subspace training strategy

Authors: Tao Li, Weisen Jiang, Fanghui Liu, Xiaolin Huang, James T. Kwok | **Year/Venue**: 2024 | **URL**: View paper

Abstract

â□¦ Experimental results on various ViT models and data sets show â□¦ optimization target as a hyperplane optimization problem (Sect. 3.1), then introduce our efficient coefficient optimization â□¦

Relationship Analysis

Both papers belong to the Direct Subspace Training for Resource Efficiency category, focusing on training models directly in constrained subspaces to reduce computational and memory costs. The original paper (WASI) addresses resource-constrained training of vision transformers on edge devices by jointly compressing weights and activations through subspace iteration, targeting on-device learning scenarios with extreme memory constraints. The candidate paper (MEHL-Soup) focuses on efficiently learning model soups (weighted combinations of multiple fine-tuned models) through hyperplane optimization and block coordinate gradient descent, primarily addressing the memory burden of combining numerous pre-trained models rather than the training process itself on resource-constrained devices.

Contributions Analysis

Overall novelty summary. The paper introduces Weight-Activation Subspace Iteration (WASI), a method for training vision transformers directly in constrained subspaces to reduce memory and computational costs during on-device learning. Within the taxonomy, it resides in the 'Direct Subspace Training for Resource Efficiency' leaf, which contains only two papers total. This sparse population suggests the research direction—training transformers from scratch in subspaces rather than adapting pretrained models—remains relatively underexplored compared to the broader field of parameter-efficient adaptation, which comprises seven papers across three subcategories.

The taxonomy reveals that neighboring research directions are substantially more populated. The 'Parameter-Efficient Adaptation of Pretrained Vision Transformers' branch contains methods like low-rank adaptation and structured fine-tuning, which modify pretrained models rather than training from scratch. The 'Model Compression and Pruning' category addresses post-training size reduction, while 'Efficient Attention Mechanisms' focuses on architectural modifications. WASI diverges from these by targeting the training phase itself with subspace constraints, positioning it closer to theoretical work on subspace representation learning than to adaptation or compression techniques that assume a trained starting point.

Among the three contributions analyzed, the WASI method itself examined ten candidates with zero refutations, suggesting limited direct prior work on this specific training approach within the search scope. The formalization of the stable parameter subspace hypothesis examined ten candidates and found three refutable matches, indicating some theoretical overlap with existing subspace analysis literature. The activation compression extension examined five candidates with no refutations. These statistics reflect a search of twenty-five total candidates, not an exhaustive survey, meaning the apparent novelty is relative to top-K semantic matches and their citations rather than the entire field.

Given the limited search scope of twenty-five candidates, the work appears to occupy a sparsely populated research direction with modest theoretical overlap. The WASI method shows no direct refutation among examined candidates, while the subspace hypothesis formalization encounters some prior theoretical work. The analysis does not cover exhaustive literature review or adjacent fields outside the semantic search radius, so conclusions about novelty remain provisional and bounded by the candidate set examined.

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

Contribution 1: Weight-Activation Subspace Iteration (WASI) method

Description: WASI is a novel training method that jointly compresses both model weights and activation maps by restricting training to a stable low-dimensional subspace. It applies SVD and subspace iteration to obtain low-rank approximations during each training iteration while controlling information loss through an explained variance threshold.

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. SLTrain: a sparse plus low-rank approach for parameter and memory efficient pretraining

URL: View paper

Brief Assessment

SLTrain[28] focuses on pretraining with sparse plus low-rank weight parameterization, not joint weight-activation compression during fine-tuning. The candidate does not address activation map compression or subspace iteration techniques.

---

2. A memory efficient randomized subspace optimization method for training large language models

URL: View paper

Brief Assessment

Randomized Subspace Optimization[32] focuses on reducing memory for optimizer states and activations through random subspace decomposition for LLM training, but does not employ SVD-based subspace iteration with explained variance thresholds for joint weight-activation compression as in WASI.

---

3. Efficient low-dimensional compression of overparameterized models

URL: View paper

Brief Assessment

Low Dimensional Compression[29] focuses on compressing overparameterized models by reducing intermediate layer widths in deep linear networks for matrix recovery, not on joint weight-activation compression during transformer training with controlled information loss via explained variance thresholds.

---

4. Low-rank momentum factorization for memory efficient training

URL: View paper

Brief Assessment

Low Rank Momentum[30] focuses on maintaining a low-rank SVD factorization of the first-order momentum for memory-efficient training, not on jointly compressing both model weights and activation maps during training as WASI does. The candidate applies factorization to optimizer states (momentum), while WASI applies it to both weights and activations during forward/backward passes.

---

5. Investigating Low-Rank Training in Transformer Language Models: Efficiency and Scaling Analysis

URL: View paper

Brief Assessment

Low Rank Training Analysis[33] focuses on low-rank parametrization of feedforward networks in transformer language models trained from scratch, not joint weight-activation compression during fine-tuning with controlled information loss through explained variance thresholds.

---

6. Lost: Low-rank and sparse pre-training for large language models

URL: View paper

Brief Assessment

Lost Pretraining[31] focuses on low-rank and sparse decomposition for LLM pre-training from scratch, using SVD to initialize both low-rank and sparse components. WASI targets memory-efficient training of vision transformers through joint weight-activation compression during fine-tuning, not pre-training.

---

7. Lora-fa: Memory-efficient low-rank adaptation for large language models fine-tuning

URL: View paper

Brief Assessment

LoRA FA[36] focuses on freezing the projection-down weight A in LoRA adapters to reduce activation memory during fine-tuning, while WASI jointly compresses both model weights and activation maps through SVD-based subspace iteration during training. These are fundamentally different approaches to memory efficiency.

---

8. LoRA-GA: Low-Rank Adaptation with Gradient Approximation

URL: View paper

Brief Assessment

LoRA-GA[34] focuses on initialization strategies for LoRA adapters in parameter-efficient fine-tuning, not on joint weight-activation compression through subspace iteration during training.

---

9. GaLore: Memory-Efficient LLM Training by Gradient Low-Rank Projection

URL: View paper

Brief Assessment

GaLore[27] focuses on gradient low-rank projection for optimizer memory reduction in LLM training, not joint weight-activation compression during forward/backward passes as in WASI.

---

10. Eigen attention: Attention in low-rank space for kv cache compression

URL: View paper

Brief Assessment

Eigen Attention[35] focuses on compressing KV cache in transformer attention mechanisms through low-rank approximation of keys and values, not on joint weight-activation compression during training. The candidate addresses inference memory bottlenecks, while WASI targets training efficiency on edge devices.

Contribution 2: Formalization of stable parameter subspace hypothesis

Description: The authors formalize the hypothesis that the intrinsic subspace of over-parameterized models remains relatively stable during fine-tuning due to small learning rate updates. This theoretical insight motivates their subspace-based compression approach and is empirically verified in their experiments.

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. A kernel-based view of language model fine-tuning

URL: View paper

Brief Assessment

Kernel View Finetuning[37] focuses on kernel-based dynamics during fine-tuning and does not address the stability of parameter subspaces in over-parameterized models during fine-tuning with small learning rate updates.

---

2. SuLoRA: Subspace Low-Rank Adaptation for Parameter-Efficient Fine-Tuning

URL: View paper

Brief Assessment

SuLoRA[46] focuses on parameter space separation for task-specific adaptation in LoRA fine-tuning, not on formalizing stable parameter subspace hypotheses during fine-tuning of over-parameterized models. The candidate addresses parameter interference across tasks rather than subspace stability theory.

---

3. Robust Watermarking for Federated Diffusion Models with Unlearning-Enhanced Redundancy

URL: View paper

Brief Assessment

Watermarking Federated Diffusion[40] focuses on watermarking in federated diffusion models with unlearning mechanisms, not on formalizing stable parameter subspaces during fine-tuning of over-parameterized models.

---

4. Compositional Subspace Representation Fine-tuning for Adaptive Large Language Models

URL: View paper

Brief Assessment

Compositional Subspace Representation[45] focuses on representation-level subspace editing for multi-task adaptation in LLMs, not on formalizing stability of parameter subspaces during fine-tuning of over-parameterized models.

---

5. Distribution-informed neural networks for domain adaptation regression

URL: View paper

Brief Assessment

Distribution Informed Networks[43] focuses on domain adaptation regression using distribution-informed neural networks and does not address stable parameter subspaces during fine-tuning of over-parameterized models. The candidate paper's theoretical framework concerns domain discrepancy measurement and adaptive kernels, not parameter subspace stability during fine-tuning.

---

6. Fine-tuning happens in tiny subspaces: Exploring intrinsic task-specific subspaces of pre-trained language models

URL: View paper

Prior Art Analysis

Intrinsic Task Subspaces[42] demonstrates prior work on stable parameter subspaces during fine-tuning. The candidate paper explicitly states that fine-tuning introduces only minor updates due to small learning rates, causing the intrinsic subspace to remain relatively stable after each training iteration. This directly addresses the same phenomenon that the original paper claims to formalize. The candidate provides empirical evidence showing that parameter subspaces remain stable during fine-tuning trajectories, and that models can be effectively fine-tuned in these low-dimensional subspaces. Both papers build on the same foundational observation from prior work (Radiya-Dixit & Wang, 2020; Li & Zhang, 2021) about parameter space stability during fine-tuning.

Evidence

Evidence 1 - **Rationale**: The candidate paper explicitly hypothesizes and formalizes the existence of stable intrinsic subspaces during fine-tuning, which is the same core theoretical contribution claimed by the original paper. - **Original**: based on the previous studies, we formulate that the essential information of a model parameters resides in a stable subspace throughout fine-tuning (sec. 3.3), which is then verified in sec. 4.2. - **Candidate**: we hypothesize that an intrinsic task-specific subspace exists for each downstream task when fine-tuning a plm. generally, it is intractable to search such an intrinsic task-specific subspace directly. however, if our hypothesis is true, the finetuning trajectory will lie in a low-dimensional subspa...

Evidence 2 - **Rationale**: Both papers provide empirical validation of the stable subspace hypothesis through experiments showing that models can be effectively fine-tuned in low-dimensional subspaces with stable ranks/dimensions. - **Original**: stability of layer ranks.we apply truncated svd to the weight tensors of the linear layers within vit's mlp blocks at each training iteration. we constrain the decomposition by settingε= 0.8 and monitor the layer ranksk i throughout the course of training. as shown in fig. 3a, we observe that the ra... - **Candidate**: we can see that freezing the encoder significantly degrades the model performance as it serves as a naïve baseline (note that it implies fine-tuning in the null space, i.e., v θt = 0, which brings no information to update the model). for intrinsic subspace fine-tuning, we can clearly see that it sho...

---

7. Parameter-Efficient Subspace Optimization for LLM Fine-Tuning

URL: View paper

Prior Art Analysis

Subspace Optimization Finetuning[38] demonstrates that the concept of stable parameter subspaces during fine-tuning was already established in prior literature before the ORIGINAL paper. The candidate explicitly cites and builds upon earlier work by Radiya-Dixit & Wang (2020) and Li & Zhang (2021), who showed that fine-tuned models remain close to pretrained counterparts in parameter space. The candidate paper formalizes this as the 'intrinsic dimensionality' concept and develops it into a practical framework, indicating that the stability of parameter subspaces was not first formalized by the ORIGINAL paper's authors.

Evidence

Evidence 1 - **Rationale**: Both papers acknowledge the motivation for parameter-efficient fine-tuning based on the observation that full parameter updates are unnecessary, suggesting prior understanding of parameter subspace stability. - **Original**: inspired by prior studies on the stability of parameter subspaces during fine-tuning (radiya-dixit & wang, 2020; li & zhang, 2021), we present w asi - **Candidate**: in the realm of fine-tuning, we often have limited labeled data for a downstream task but still wish to adapt the pretrained weights effectively and efficiently. therefore, updating the entire parameter set is memory-intensive. this motivates the study of parameter-efficient fine-tuning (peft) metho...

Evidence 2 - **Rationale**: The candidate paper explicitly formalizes the intrinsic dimensionality concept with mathematical notation, citing prior work [1, 29] and [23] (LoRA), demonstrating that this formalization existed before the ORIGINAL paper. - **Original**: based on the previous studies, we formulate that the essential information of a model parameters resides in a stable subspace throughout fine-tuning (sec. 3.3), which is then verified in sec. 4.2. - **Candidate**: formally, we build on the notion ofintrinsic dimensionalityin llm training [1, 29], recognized in [23] as the origin of lora: there exists a dimension-lifting mapm:r d →r mxn, withd≪mxn, such that the optimal solution Δw ∗ of (1) satisfies Δw∗ ≈ m(ξ∗), ξ ∗ := arg minξ∈rd ℓ(w0 +m(ξ)).

Evidence 3 - **Rationale**: Both papers acknowledge that the concept of low-dimensional intrinsic subspaces in over-parameterized models was established by prior work (Aghajanyan et al., 2020; Li et al., 2018), demonstrating that the foundational hypothesis predates both papers. - **Original**: while prior work has shown that over-parameterized models in fact reside in a low-dimensional intrinsic subspace (aghajanyan et al., 2020; li et al., 2018) - **Candidate**: while (4) may strike to be abstract, many existing benchmarks for llm training can naturally fit in it by considering specific subspaces. here we summarize several notable methods in table 1.

---

8. Twin Learning for Domain Agnostic Time Series Analysis: A Unified Regime-Switch Approach

URL: View paper

Brief Assessment

Twin Learning[41] focuses on time series analysis with regime-switching approaches. The provided context fragments are insufficient to determine whether it addresses stable parameter subspaces in neural network fine-tuning or over-parameterized models.

---

9. Low-Rank Adaptation of Evolutionary Deep Neural Networks for Efficient Learning of Time-Dependent PDEs

URL: View paper

Brief Assessment

Low Rank Evolutionary[39] focuses on evolutionary deep neural networks for time-dependent PDEs with low-rank constraints on parameter velocities, not on fine-tuning stability of over-parameterized models with small learning rate updates.

---

10. Towards Lightweight Adaptation of Massive Neural Network Models

URL: View paper

Prior Art Analysis

Lightweight Adaptation[44] demonstrates that the concept of stable parameter subspaces in over-parameterized models during fine-tuning was already established in prior work. The candidate paper explicitly references that 'parameter updates often lie in low-dimensional subspaces' as a known property of 'overparameterized foundation models', indicating this was not a novel insight. The original paper's formalization builds upon existing observations from Radiya-Dixit & Wang (2020) and Li & Zhang (2021), which the original authors themselves acknowledge, suggesting the hypothesis was already present in the literature rather than being newly formalized.

Evidence

Evidence 1 - **Rationale**: Both papers acknowledge that parameter updates in over-parameterized models lie in low-dimensional subspaces. The original paper presents this as building on prior work (citing Aghajanyan et al., 2020; Li et al., 2018), while Lightweight Adaptation[44] also references this as an established property of overparameterized foundation models, suggesting this was already known in the field. - **Original**: while prior work has shown that over-parameterized models in fact reside in a low-dimensional intrinsic subspace (aghajanyan et al., 2020; li et al., 2018), we further observe that fine-tuning introduces only minor updates at each training step due to the use of a small learning rate. as a result, o... - **Candidate**: parameter updates often lie in low-dimensional subspaces; ... properties of overparameterized foundation models [100].

Evidence 2 - **Rationale**: The original paper explicitly acknowledges that their hypothesis is supported by prior findings from Radiya-Dixit & Wang (2020) and Li & Zhang (2021). This admission indicates that the stability of parameter subspaces during fine-tuning was already documented in the literature, undermining the claim of novel formalization. Lightweight Adaptation[44] discusses this in the context of standard fine-tuning problems for foundation models. - **Original**: This is supported by the findings of radiya-dixit & wang (2020) and li & zhang (2021), who showed that the fine-tuned models are close in parameter space to the pre-trained counterpart. - **Candidate**: standard fine-tuning problem for large foundation models [8].

Contribution 3: Extension of activation compression to 3D tensors with dynamic programming optimization

Description: The authors improve upon prior activation compression methods by introducing a dynamic programming strategy that efficiently determines optimal compression ranks and extending the approach to handle 3D activation tensors, making it applicable to a broader range of transformer architectures.

This contribution was assessed against **5 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. L-GreCo: Layerwise-adaptive Gradient Compression For Efficient Data-parallel Deep Learning

URL: View paper

Brief Assessment

Layerwise Gradient Compression[22] focuses on gradient compression during distributed training using dynamic programming to optimize layer-wise compression parameters. The original paper addresses activation compression for on-device learning with dynamic programming for rank selection across activation tensor modes, which is a different technical problem.

---

2. Adaptive Comcat: Towards Faster and More Precise Low-Rank Decompositon of Attention-Based Vision Models

URL: View paper

Brief Assessment

Adaptive Comcat[26] mentions dynamic programming for rank selection in vision transformers, but the provided context is too fragmentary to establish whether it addresses 3D tensor compression or uses the same dynamic programming strategy as the original paper's ASI method.

---

3. Optimizing low-rank decomposition for efficient attention-based vision models via adaptive neural architecture search

URL: View paper

Brief Assessment

Adaptive NAS Decomposition[23] focuses on low-rank decomposition for attention mechanisms in vision models via neural architecture search, not on activation compression with dynamic programming for optimal rank selection in transformers.

---

4. A comprehensive survey on recent model compression and acceleration approaches for deep neural networks and transformers

URL: View paper

Brief Assessment

Model Compression Survey[25] only briefly mentions iterative low-rank decomposition with dynamic rank selection in passing, without providing technical details about 3D tensor handling or the specific dynamic programming strategy for activation compression in transformers.

---

5. ModelOpt: Research Framework for Zero-Shot Computer Vision Model Optimization With Tree Search and Federated Knowledge Sharing

URL: View paper

Brief Assessment

ModelOpt[24] mentions dynamic programming for rank selection but does not provide details on activation compression methods for 3D tensors or the specific dynamic programming strategy described in the original paper.

Appendix: Text Similarity Detection

No high-similarity text segments were detected across any compared papers.

References

- [0] Efficient Resource-Constrained Training of Vision Transformers via Subspace Optimization View paper
- [1] Attention-Only Transformers via Unrolled Subspace Denoising View paper
- [2] White-box transformers via sparse rate reduction View paper
- [3] Vision transformer slimming: Multi-dimension searching in continuous optimization space View paper
- [4] Parameter-efficient model adaptation for vision transformers View paper
- [5] The universal weight subspace hypothesis View paper
- [6] Learning scalable model soup on a single gpu: An efficient subspace training strategy View paper
- [7] CS-VLM: Compressed Sensing Attention for Efficient Vision-Language Representation Learning View paper
- [8] Structure-Preserving Network Compression Via Low-Rank Induced Training Through Linear Layers Composition View paper
- [9] Serial Low-rank Adaptation of Vision Transformer View paper
- [10] Weight Spectra Induced Efficient Model Adaptation View paper
- [11] Steering CLIP's vision transformer with sparse autoencoders View paper
- [12] MixPHM: Redundancy-aware parameter-efficient tuning for low-resource visual question answering View paper
- [13] TransBlast: Self-supervised learning using augmented subspace with transformer for background/foreground separation View paper
- [14] Mcuformer: Deploying vision tranformers on microcontrollers with limited memory View paper
- [15] Generalized Kronecker-based Adapters for Parameter-efficient Fine-tuning of Vision Transformers View paper
- [16] Query Efficient Black-Box Visual Prompting with Subspace Learning View paper
- [17] Integrating Vision Transformers with Transfer Learning for Enhanced Kidney Cancer Classification in CT Imaging View paper
- [18] Delta-LLaVA: Base-then-Specialize Alignment for Token-Efficient Vision-Language Models View paper
- [19] Memory-Efficient Low-Rank Fine-Tuning for AoA Estimation in AI-Enabled RANs View paper
- [20] Optimizing Transformer-based Models for Low-Resource Languages Using Transfer Learning Techniques View paper
- [21] EigenLoRAx: Efficient Low Rank Adaptation Using Recycled Principal Subspaces View paper
- [22] L-GreCo: Layerwise-adaptive Gradient Compression For Efficient Data-parallel Deep Learning View paper
- [23] Optimizing low-rank decomposition for efficient attention-based vision models via adaptive neural architecture search View paper
- [24] ModelOpt: Research Framework for Zero-Shot Computer Vision Model Optimization With Tree Search and Federated Knowledge Sharing View paper
- [25] A comprehensive survey on recent model compression and acceleration approaches for deep neural networks and transformers View paper
- [26] Adaptive Comcat: Towards Faster and More Precise Low-Rank Decompositon of Attention-Based Vision Models View paper
- [27] GaLore: Memory-Efficient LLM Training by Gradient Low-Rank Projection View paper
- [28] SLTrain: a sparse plus low-rank approach for parameter and memory efficient pretraining View paper
- [29] Efficient low-dimensional compression of overparameterized models View paper
- [30] Low-rank momentum factorization for memory efficient training View paper
- [31] Lost: Low-rank and sparse pre-training for large language models View paper
- [32] A memory efficient randomized subspace optimization method for training large language models View paper
- [33] Investigating Low-Rank Training in Transformer Language Models: Efficiency and Scaling Analysis View paper
- [34] LoRA-GA: Low-Rank Adaptation with Gradient Approximation View paper
- [35] Eigen attention: Attention in low-rank space for kv cache compression View paper
- [36] Lora-fa: Memory-efficient low-rank adaptation for large language models fine-tuning View paper
- [37] A kernel-based view of language model fine-tuning View paper
- [38] Parameter-Efficient Subspace Optimization for LLM Fine-Tuning View paper
- [39] Low-Rank Adaptation of Evolutionary Deep Neural Networks for Efficient Learning of Time-Dependent PDEs View paper
- [40] Robust Watermarking for Federated Diffusion Models with Unlearning-Enhanced Redundancy View paper
- [41] Twin Learning for Domain Agnostic Time Series Analysis: A Unified Regime-Switch Approach View paper
- [42] Fine-tuning happens in tiny subspaces: Exploring intrinsic task-specific subspaces of pre-trained language models View paper
- [43] Distribution-informed neural networks for domain adaptation regression View paper

- [44] Towards Lightweight Adaptation of Massive Neural Network Models View paper
- [45] Compositional Subspace Representation Fine-tuning for Adaptive Large Language Models View paper
- [46] SuLoRA: Subspace Low-Rank Adaptation for Parameter-Efficient Fine-Tuning View paper