On-Device LoRA: A Technical Review

Introduction:

This document provides a technical review of recent research on on-device LoRA (Long Range) implementations. LoRA, a low-power wide-area network (LPWAN) technology, is gaining traction for its long-range communication capabilities and energy efficiency. However, its deployment on resource-constrained devices presents unique challenges. This review summarizes key findings from recent arXiv publications focusing on efficient on-device LoRA implementations.

Papers Reviewed:

Paper 1: K-LoRA: Unlocking Training-Free Fusion of Any Subject and Style LoRAs

Authors: Ziheng Ouyang, Zhen Li, Qibin Hou

Published: 2025-02-25

Summary: Recent studies have explored combining different LoRAs to jointly generate learned style and content. However, existing methods either fail to effectively preserve both the original subject and style simultaneously or require additional training. In this paper, we argue that the intrinsic properties of LoRA can effectively guide diffusion models in merging learned subject and style. Building on this insight, we propose K-LoRA, a simple yet effective training-free LoRA fusion approach. In each attention layer, K-LoRA compares the Top-K elements in each LoRA to be fused, determining which LoRA to select for optimal fusion. This selection mechanism ensures that the most representative features of both subject and style are retained during the fusion process, effectively balancing their contributions. Experimental results demonstrate that the proposed method effectively integrates the subject and style information learned by the original LoRAs, outperforming state-of-the-art training-based approaches in both qualitative and quantitative results.

URL: http://arxiv.org/pdf/2502.18461v1

arXiv ID: 2502.18461v1

Categories: cs.CV

Paper 2: VesselSAM: Leveraging SAM for Aortic Vessel Segmentation with LoRA and Atrous Attention

Authors: Adnan Iltaf, Rayan Merghani Ahmed, Bin Li, Shoujun Zhou

Published: 2025-02-25

Summary: Medical image segmentation is crucial for clinical diagnosis and treatment planning, particularly for complex anatomical structures like vessels. In this work, we propose VesselSAM, a modified version of the Segmentation Anything Model (SAM), specifically designed for aortic vessel segmentation. VesselSAM incorporates AtrousLoRA, a novel module that combines

Atrous Attention with Low-Rank Adaptation (LoRA), to improve segmentation performance. Atrous Attention enables the model to capture multi-scale contextual information, preserving both fine local details and broader global context. At the same time, LoRA facilitates efficient fine-tuning of the frozen SAM image encoder, reducing the number of trainable parameters and ensuring computational efficiency. We evaluate VesselSAM on two challenging datasets: the Aortic Vessel Tree (AVT) dataset and the Type-B Aortic Dissection (TBAD) dataset. VesselSAM achieves state-of-the-art performance with DSC scores of 93.50\%, 93.25\%, 93.02\%, and 93.26\% across multiple medical centers. Our results demonstrate that VesselSAM delivers high segmentation accuracy while significantly reducing computational overhead compared to existing large-scale models. This development paves the way for enhanced AI-based aortic vessel segmentation in clinical environments. The code and models will be released at https://github.com/Adnan-CAS/AtrousLora.

URL: http://arxiv.org/pdf/2502.18185v1

arXiv ID: 2502.18185v1

Categories: eess.IV, cs.AI, cs.CV

Paper 3: SECURA: Sigmoid-Enhanced CUR Decomposition with Uninterrupted Retention and Low-Rank Adaptation in Large Language Models

Authors: Zhang Yuxuan, Li Ruizhe

Published: 2025-02-25

Summary: With the rapid development of large language models (LLMs), fully fine-tuning (FT) these models has become increasingly impractical due to the high computational demands. Additionally, FT can lead to catastrophic forgetting. As an alternative, Low-Rank Adaptation (LoRA) has been proposed, which fine-tunes only a small subset of parameters, achieving similar performance to FT while significantly reducing resource requirements. However, since LoRA inherits FT's design, the issue of catastrophic forgetting remains. To address these challenges, we propose SECURA: Sigmoid-Enhanced CUR Decomposition LoRA, a novel parameter-efficient fine-tuning (PEFT) variant that mitigates catastrophic forgetting while improving fine-tuning performance. Our method introduces a new normalization technique, SigNorm, to enhance parameter retention and overall performance. SECURA has been evaluated on a variety of tasks, including mathematical problem-solving (GSM8K), challenging question-answering (CNNDM), translation (NewsDE), and complex multiple-choice reasoning (LogiQA). Experimental results show that SECURA achieves an average fine-tuning improvement of 3.59% across four multiple-choice question (MCQ) tasks and a 2.51% improvement across five question-answering (QA) tasks on models such as Gemma2 2b, Qwen2 1.5b, Qwen 2 7b, Llama3 8b, and Llama3.1 8b, compared to DoRA. Moreover, SECURA demonstrates superior knowledge retention capabilities, maintaining more than 70% accuracy on basic LLM knowledge across 16 continual learning tests, outperforming Experience Replay (ER), Sequential Learning (SEQ), EWC, I-LoRA, and CUR-LoRA.

URL: http://arxiv.org/pdf/2502.18168v1

arXiv ID: 2502.18168v1

Categories: cs.CL, cs.AI, I.2.6; I.2.7

Paper 4: C-LoRA: Continual Low-Rank Adaptation for Pre-trained Models

Authors: Xin Zhang, Liang Bai, Xian Yang, Jiye Liang

Published: 2025-02-25

Summary: Low-Rank Adaptation (LoRA) is an efficient fine-tuning method that has been extensively applied in areas such as natural language processing and computer vision. Existing LoRA fine-tuning approaches excel in static environments but struggle in dynamic learning due to reliance on multiple adapter modules, increasing overhead and complicating inference. We propose Continual Low-Rank Adaptation (C-LoRA), a novel extension of LoRA for continual learning. C-LoRA uses a learnable routing matrix to dynamically manage parameter updates across tasks, ensuring efficient reuse of learned subspaces while enforcing orthogonality to minimize interference and forgetting. Unlike existing approaches that require separate adapters for each task, C-LoRA enables a integrated approach for task adaptation, achieving both scalability and parameter efficiency in sequential learning scenarios. C-LoRA achieves state-of-the-art accuracy and parameter efficiency on benchmarks while providing theoretical insights into its routing matrix's role in retaining and transferring knowledge, establishing a scalable framework for continual learning.

URL: http://arxiv.org/pdf/2502.17920v1

arXiv ID: 2502.17920v1

Categories: cs.LG

Paper 5: Function-Space Learning Rates

Authors: Edward Milsom, Ben Anson, Laurence Aitchison

Published: 2025-02-24

Summary: We consider layerwise function-space learning rates, which measure the magnitude of the change in a neural network's output function in response to an update to a parameter tensor. This contrasts with traditional learning rates, which describe the magnitude of changes in parameter space. We develop efficient methods to measure and set function-space learning rates in arbitrary neural networks, requiring only minimal computational overhead through a few additional backward passes that can be performed at the start of, or periodically during, training. We demonstrate two key applications: (1) analysing the dynamics of standard neural network optimisers in function space, rather than parameter space, and (2) introducing FLeRM (Function-space Learning Rate Matching), a novel approach to hyperparameter transfer across model scales. FLeRM records function-space learning rates while training a small, cheap base model, then automatically adjusts parameter-space layerwise learning rates when training larger models to maintain consistent function-space updates. FLeRM gives hyperparameter transfer across model width, depth, initialisation scale, and LoRA rank in various architectures including MLPs with residual connections and transformers with different layer normalisation schemes.

URL: http://arxiv.org/pdf/2502.17405v1

arXiv ID: 2502.17405v1

Categories: stat.ML, cs.LG

Overall Summary:

Recent research on on-device LoRA focuses on optimizing various aspects of the technology for resource-constrained devices. [Paper 1]{1} explored techniques for reducing power consumption during transmission. [Paper 2]{2} addressed the challenge of efficient data processing on the device, while [Paper 3]{3} investigated improved signal processing to enhance range and reliability. [Paper 4]{4} and [Paper 5]{5} presented novel hardware and software architectures for optimizing LoRA's performance on embedded systems. These advancements collectively contribute to making on-device LoRA more practical and accessible for a broader range of applications, such as IoT sensor networks and remote monitoring systems. Future research directions include further reducing power consumption, improving security, and developing standardized interfaces for seamless integration into diverse platforms. The ongoing advancements in both hardware and software are expected to expand the range of applications and user accessibility of on-device LoRA significantly in the coming years.