Document details - Towards Offline GenAI Fine Tuning Model with LoRA Derivatives for IoT Edge Server

2024 9th International Conference on Informatics and Computing, ICIC 2024

2024

9th International Conference on Informatics and Computing, ICIC 2024; Hybrid, Medan; Indonesia; 24 October 2024 through 25 October 2024; Category numberCFP24G52-ART; Code 208312

Towards Offline GenAI Fine Tuning Model with LoRA Derivatives for IoT Edge Server(Conference Paper)

Yugopuspito, P., Murwantara, I.M., Alim, E.K., Cendana, W., Mitra, A.R.

[a]Universitas Pelita Harapan, Tangerang, Indonesia
[b]Universitas Pelita Harapan, Tangerang, Indonesia
[c]Universitas Pelita Harapan, Tangerang, Indonesia

View additional affiliations

Abstract

The Internet of Things (IoT) has become increasingly pervasive, connecting a vast network of devices to collect and analyze data. However, the reliance on continuous internet connectivity poses challenges in regions with limited or unstable access. This paper investigates the feasibility of deploying offline Generative AI (GenAI) models on IoT edge servers, enabling autonomous data generation in disconnected environments. A significant challenge arises from the resource constraints inherent to edge devices, which often lack the computational power required to run sophisticated AI models. To address this, techniques such as model compression and quantization are considered to reduce the size and computational demands of the models, while maintaining acceptable accuracy. One such technique, Low-Rank Adaptation (LoRA), is examined in this study alongside its various derivatives. The primary contribution of this paper is a comparative analysis of several LoRA derivatives, including Quantized LoRA (QLoRA), Multi-task LoRA (MT-LoRA), and Adaptive LoRA (AdaLoRA), in fine-tuning the LLaMA 3.1 large language model (LLM) for IoT applications. The evaluation focuses on memory optimization and model performance, with experiments conducted using the 4-bit quantized version of LLaMA 3.1 8B. These efforts aim to create realistic simulation environments for testing and evaluating IoT systems under different conditions. © 2024 IEEE.

Author keywords

fine tuning | Generative AI | IoT | LoRA derivative

Indexed keywords

Engineering uncontrolled terms: Computational power | Data generation | Edge server | Fine tuning | Generative AI | Internet connectivity | Low-rank adaptation derivative | Modeling quantizations | Offline | Resource Constraint

Funding details

Funding sponsor	Funding number	Acronym
Lembaga Penelitian dan Pengabdian Kepada Masyarakat		LPPM
Universitas Pelita Harapan	155/LPPM-UPH/VII/2024	UPH

Funding sponsor	Funding number	Acronym
	P-113-FIT/VII/2024	

Funding text

This research is partially funded by Center of Research and Community Development (LPPM), Universitas Pelita Harapan, No. 155/LPPM-UPH/VII/2024 and Faculty of Information Technology No. P-113-FIT/VII/2024 on July 2024.

ISBN: 979-833151760-1
DOI: 10.1109/ICIC64337.2024.10956262
Source Type: Conference Proceeding
Document Type: Conference Paper
Original language: English
Publisher: Institute of Electrical and Electronics Engineers Inc.

Yugopuspito, P.; Universitas Pelita Harapan, Tangerang, Indonesia;

SciVal Topic Prominence

Topic:

Prominence percentile:

About Scopus

What is Scopus

Content coverage

Scopus blog

Scopus API

Privacy matters

Language

日本語版を表示する

查看简体中文版本

查看繁體中文版本

Просмотр версии на русском языке

Customer Service

Help

Tutorials

Contact us