#16 Mobile Foundation Model as Firmware

PC conflicts
Dongqi Cai
Jinliang Yuan

Review #16A
Review #16B
Review #16C
Comments

Artifacts Available&Functional&Reusable&Replicated

Submission (60kB) Mar 6, 2024, 4:16:54 PM UTC · 5a26e390

Abstract

In the current AI era, mobile devices such as smartphones are tasked with executing a myriad of deep neural networks (DNNs) locally. It presents a complex landscape, as these models are highly fragmented in terms of architecture, operators, and implementations. Such fragmentation poses significant challenges to the co-optimization of hardware, systems, and algorithms for efficient and scalable mobile AI.

Inspired by the recent groundbreaking progress in large foundation models, this work introduces a novel paradigm for mobile AI, where mobile OS and hardware jointly manage a foundation model that is capable of serving a wide array of mobile AI tasks. This foundation model functions akin to firmware, unmodifiable by apps or the OS, exposed as a system service to Apps. They can invoke this foundation model through a small, offline fine-tuned ``adapter'' for various downstream tasks. We propose a tangible design of this vision called M4, and prototype it from publicly available pre-trained models. To assess its capability, we also build a comprehensive benchmark consisting of 38 mobile AI tasks and 50 datasets, spanning 5 multimodal inputs. Extensive experiments demonstrate \sys's remarkable results: it achieves comparable accuracy in 85% of tasks, offers enhanced scalability regarding storage and memory, and has much simpler operations. In broader terms, this work paves a new way towards efficient and scalable mobile AI in the post-LLM era.

Authors

J. Yuan, C. Yang, D. Cai, S. Wang, X. Yuan, Z. Zhang, X. Li, D. Zhang, H. Mei, X. Jia, S. Wang, M. Xu [details]

Attachments and options

(Conditionally) accepted version of your MobiCom paper (4.1MB)

Which artifact badges do you intend to receive for your submission?

Artifacts Available
Artifacts Evaluated -- Functional
Artifacts Evaluated -- Reusable
Artifacts Evaluated -- Results
Replicated

Hardware Dependencies

desktop computer with a GPU server and Edge AI Developer Kit

Software Dependencies

Python, docker and bash scripts.

Paper ID for Accepted Paper: 148

	RevExp	ArtAva	ArtEvaFun	ArtEvaReu	ResVal	DisArt
Review #16A	2	2	2	2	2	
Review #16B	4	2	3	2	2	1
Review #16C	3	2	2	3	4	

This artifact should be awarded the "ACM MobiCom 2024 Distinguished Artifact" award.