

Peer Review

Review of: "GenAI at the Edge: Comprehensive Survey on Empowering Edge Devices"

Lars Wulfert¹

1. Embedded Software & AI, Fraunhofer Institute for Microelectronic Circuits and Systems, Duisburg, Germany

Summary

This paper provides an overview of existing methods for implementing GenAI on edge devices. It addresses the key challenges of deploying resource-intensive models in environments with limited computing resources. The focus is on software optimization strategies such as model compression and neural architecture search (NAS), as well as hardware solutions such as specialized accelerators. In addition, frameworks designed to improve inference latency, memory efficiency, and overall energy consumption are explored.

Novelties

- First survey on GenAI at the edge

Strengths

- Provides helpful background information on the discussed methods at the beginning of the subsections
- Well-written paper and clear summaries of the referenced papers within the survey
- Great figure illustrating an overview of GenAI at the edge
- Informative table about hardware accelerators for GenAI

Weaknesses

- Background section about GenAI/transformers, hardware accelerators, edge devices, etc., is missing
- An overview of the systematic literature review process is missing

- Section about future directions and open issues is missing
- A table summarizing models deployed on edge devices (including model sizes and hardware used) would be very useful

Minor Issues

- **Section: Introduction**
 - Significant challenges are mentioned but not explicitly identified
 - “By reviewing state-of-the-art techniques from top-tier conferences and journals, this work offers a roadmap for researchers seeking to apply GenAI in edge.” => Which conferences/journals are meant? More details about the review process are needed
 - Figure 1 is not referenced in the text
- **Section: Software Optimization**
 - **Sub Section: Model Compression**
 - “...to address these challenges...” => Which challenges are meant? Clarify
 - “This challenge...” => Which challenge?
 - “Firstly...Secondly...Thirdly...” => Citations are missing in these sentences.
 - **Sub Section: Quantization**
 - “However, significant challenges remain...” => Specify the challenges
 - **Sub Section: Pruning**
 - “...components.Unstructured...” => Missing space after the period
 - **Sub Section: Neural Architecture Search**
 - “FL-NAS[60] have proposed an approach which leverages LLM to find high-performance DNNs for resource-constrained systems.” => Interesting approach, but is it relevant to the focus of the paper? Was it developed for running NAS directly on edge devices?
 - What NAS methods were used in [61] and [62]? What are the differences between them?
 - **Sub Section: Open-Source GenAI models**
 - Paper: “One of the key contributions to the advancement in GenAI is open source innovations, specifically for edge scenarios in which the resources are limited.” => Why is open source specifically important for edge scenarios? Elaborate.
- **Section: Hardware Optimization**
 - **Sub Section: Hardware Accelerators**
 - The abbreviation *GenAI* is introduced multiple times unnecessarily (repetition)

- There is a page with three-quarters blank space before this section begins
- Table description appears on a different page – should be placed directly with the table
- **Section:** Attention Optimization
 - The connection to edge devices is unclear. This connection is not sufficiently explained in the text
- **Overall Result**
 - Although the paper is well-written and the structure of the survey is solid, key components (e.g., future directions and open issues) are missing. Due to these omissions, the final rating is **2.5/5**. However, since a 2.5 cannot be officially granted, the final result is **3/5**.
It is strongly recommended that the paper undergoes a thorough revision before being considered for official publication.

Declarations

Potential competing interests: No potential competing interests to declare.