ScienceDirect

Computer Science Review
Volume 57, August 2025, 100755

Review article

Empowering large language models to edge intelligence: A survey of edge efficient LLMs and techniques

Rui Wang, Zhiyong Gao, Liuyang Zhang, Shuaibing Yue, Ziyi Gao

Show more

Cite

https://doi.org/10.1016/j.cosrev.2025.100755
Get rights and content

Abstract

Large language models (LLMs) have showcased exceptional capabilities across various natural language processing (NLP) tasks in recent years, such as machine translation, text summarization, and question answering. Despite their impressive performance, the deployment of these models on edge devices, such as mobile phones, IoT devices, and edge computing nodes, is significantly hindered by their substantial computational and memory requirements. This survey provides a comprehensive overview of the state-of-the-art techniques and strategies for enabling efficient inference of LLMs on edge devices. We explore approaches including the development of small language models (SLMs), model compression techniques, inference optimization strategies, and dedicated frameworks for edge deployment. Our goal is to highlight the advancements and ongoing challenges in this field, offering valuable insights for researchers and practitioners striving to bring the power of LLMs to edge environments.

Introduction

Large language models (LLMs) have shown remarkable effectiveness in a wide range of natural language processing (NLP) tasks, such as machine translation, text summarization, and question answering. These models are typically trained on large corpora of text data, enabling them to produce coherent and contextually relevant text. Over recent years, the capabilities and sizes of these models have grown significantly, as seen in Fig. 1, which illustrates the increase in parameter sizes of various models from early versions like GPT-1 [1] and BERT [2] to more recent advancements like GPT-4 [3], BLOOM [4], LLaMA [5] and Llama2 [6].

However, the large parameter sizes of LLMs present significant challenges for their deployment on edge devices. Edge devices, including mobile phones, IoT devices, and edge computing nodes, typically possess limited computational and memory resources. For instance, Llama2-7B [6] inference requires at least 7 GB of CPU or GPU memory with INT4 quantization and only achieves 4.5 tokens per second on NVIDIA Jetson AGX Orin [7], which is often beyond the capabilities of most edge devices. Moreover, continuous inference can cause devices to heat up significantly, as some edge or mobile devices are typically passively cooled, which may have a significant impact on performance [8]. This high computational requirement and substantial memory footprint hinder the widespread adoption of LLMs in edge environments, where resources are constrained.

Specifically, the edge deployment of LLMs faces four major challenges due to the resource constraints inherent in edge environments:

1. The rapid growth in the size of LLMs is at odds with the limited memory resources of edge devices.

2. The high computational demands of LLMs clash with the restricted computational resources of edge devices.

3. The substantial energy consumption of LLMs conflicts with the finite energy supply of edge devices.

4. The large throughput requirements of LLMs are in contrast with the limited bandwidth of edge devices.

To address these challenges, existing research has aimed to alleviate resource pressures through: (1) introducing lightweight architectures to reduce model computational complexity and communication overhead; (2) utilizing model compression techniques to decrease the scale of model parameters; (3) optimizing inference system efficiency by designing effective inference strategies and algorithms.

Despite these challenges, deploying LLMs on edge devices presents distinct advantages in latency, privacy, personalization and so on [9]. Edge deployment can significantly reduce latency, as data processing occurs closer to the source, thereby improving real-time responsiveness. It can also enhance data privacy and security, as sensitive information does not need to be transmitted to centralized servers for processing. Additionally, edge deployment can lead to more efficient use of network bandwidth and provide uninterrupted services even in areas with limited connectivity. Furthermore, edge-based LLMs can offer more personalized experiences by leveraging local data to tailor responses and services to individual users, thereby enhancing user satisfaction and engagement. According to Statista's forecast, the number of globally connected Internet of Things (IoT) devices is expected to reach 15.9 billion by 2023 and is estimated to escalate to 39.6 billion by 2033 [10]. The rapid expansion of edge IoT devices necessitates the exploration of redundant computational power at the edge and the utilization of edge advantages to provide services.

Existing surveys [11], [12], [13], [14], [15], [16] have systematically summarized efficient inference methods for large language models. However, their focus predominantly lies in cloud environments, and there is a notable lack of investigation and discussion on models, technologies, and frameworks suitable for edge computing. In contrast, our survey offers a perspective on the development of LLMs at the edge, addressing the challenges of limited resources for LLMs in edge environments. It reviews recent developments in LLMs from three aspects: small-sized models, compression techniques and inference optimization technologies for edge. Furthermore, it identifies future research challenges in this field. Xu et al. [17] provides a comprehensive review of on-device LLMs, but our survey not only includes on-device scenarios but also considers relevant research progress in cloud–edge collaboration. Additionally, two recent related works, Lu et al. [18] and Wang et al. [19], have comprehensively surveyed recent advancements in small models and limited model optimization techniques. Our survey not only provides the latest research progress on small models but also covers a more comprehensive study on the optimization and deployment of small models at the edge.

The remainder of this survey is organized as follows: Section 2 provides a comprehensive summary and detailed investigation of recent notable small language models. Section 3 presents the latest advancements in LLM compression techniques. Section 4 delves into research progress on LLM inference optimization technologies for edge computing. Section 5 explores frameworks suitable for deploying LLMs on edge devices. Section 6 highlights the challenges associated with LLMs on the edge and discusses the future. Finally, we conclude the survey in Section 7. Table 1 lists the acronyms used throughout this survey. We hope this survey will serve as a valuable resource for researchers and practitioners working to bring the power of LLMs to edge environments.

Section snippets

Small language model

The edge availability of large language models is essential for a wide range of applications. However, deploying LLMs on edge devices is challenging due to the high computational requirements of these models. Therefore, small language models (SLMs) have been developed in recent years to address this issue, as shown in Fig. 2. SLMs are typically smaller in size and have fewer parameters than their larger counterparts, making them suitable for edge devices with limited computational resources. In …

Model compression

Model compression reduces model size by removing redundant information. This can be achieved through pruning, quantization, distillation, or low-rank decomposition. In this section, we provide an overview of these techniques.

…

Inference optimization

Inference optimization primarily involves enhancing the optimization of the forward process of the model, rather than merely modifying the model's weights or structure. Current large language models generally employ autoregressive decoding based on the decoder-only transformer architecture. Inference optimization aims to fully utilize software and hardware resources by optimizing repeated computations in the forward process, enhancing the efficiency of attention and linear layer operations, and …

Deployment

In this section, we explore frameworks designed for deploying and optimizing LLMs on edge devices. We categorize these frameworks into: on-device inference engines, cloud–edge collaborative frameworks and deployment suites. On-device inference engines refer to software infrastructures optimized for executing LLM inference tasks on edge devices. Cloud-edge collaborative frameworks aim to make full use of the resources of edge devices and the cloud for LLM. Deployment suites are more oriented …

Open challenges and future directions

As LLMs continue to advance and proliferate, their deployment on edge devices presents numerous opportunities and challenges. The potential to bring powerful AI capabilities closer to the user promises significant benefits, including reduced latency, enhanced privacy, and the ability to operate in environments with limited connectivity. However, the transition from cloud-based to edge-based LLM deployment is fraught with obstacles that must be carefully navigated. These challenges range from …

Conclusion

The transition of LLMs to the edge is an inevitable trend in the future and has made preliminary research progress. However, overcoming many technical challenges is still required to achieve this goal. Our survey comprehensively investigates the current state of research on LLM applications at the edge, covering all aspects from model design, optimization to deployment, and pointing out future research directions and challenges. The survey first emphasizes the difficulties faced by the edge …

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper. …

Recommended articles

References (223)

S. Yuan *et al.*
WuDaoCorpora: A super large-scale Chinese corpora for pre-training language models
AI Open (2021)

A. Radford *et al.*
Improving Language Understanding by Generative Pre-TrainingTechnical Report
(2018)

J. Devlin *et al.*
BERT: Pre-training of deep bidirectional transformers for language understanding

OpenAI *et al.*
GPT-4 technical report
(2024)

B. Workshop *et al.*
BLOOM: A 176B-parameter open-access multilingual language model
(2023)

H. Touvron *et al.*
LLaMA: Open and efficient foundation language models

(2023)

H. Touvron *et al.*
Llama 2: Open foundation and fine-tuned chat models
(2023)

N. Dhar *et al.*
An empirical analysis and resource footprint study of deploying large language models on edge devices

X. Li *et al.*
Large language models on mobile devices: Measurements, analysis, and insights

M. Xu *et al.*
Unleashing the power of edge-cloud generative AI in mobile networks: A Survey of AIGC services
IEEE Commun. Surv. & Tutorials (2024)

View more references

Cited by (0)

View full text