

Peer Review

Review of: "Optimizing Edge AI: A Comprehensive Survey on Data, Model, and System Strategies"

Xu Kaixin¹¹. Agency for Science, Technology and Research (A*STAR), Singapore, Singapore

The article generally overviews the end-to-end pipeline for Edge-AI from data to end deployment in various real applications, which is good to see as most existing surveys are isolated modules along this whole lifecycle.

However, I feel the article still has significant potential to improve. Below are my comments:

1. I am not very sure about the inclusion of the data preprocessing section, especially the data cleaning part, where the methods introduced are not strongly correlated to the big theme of the article, which is on edge deployment. To me, data cleaning is more native to general AI as a default required procedure. Maybe you could emphasize how it could be done to serve the later edge deployment better.
2. I don't have wide knowledge about other technical parts, but for my expertise scope, e.g., pruning and quantization, a lot of works are missing, and at least some high-impact, highly cited, and essential ones should be discussed for completeness. Just to name a few, for pruning, early papers like (Song Han, NIPS'15) should be introduced. For quantization, there are missing popular papers like Adaround, BRECO, ACIQ, (Jacob, Benoit, Google, CVPR'18), etc. For Inference Serving platforms, many widely used LLM platforms such as ollama, vlm, mlc-llm, are missing.

Declarations

Potential competing interests: No potential competing interests to declare.