Apple flash: Our smart devices will soon be smarter

December 28 2023, by Peter Grad

Credit: Pixabay/CC0 Public Domain

Our smart devices take voice commands from us, check our heartbeats, track our sleep, translate text, send us reminders, capture photos and movies, and let us talk to family and friends continents away.

Now imagine turbocharging those capabilities. Holding in-depth, natural language exchanges on academic or personal queries; running our vital signs through a global database to check on imminent health issues; packing massive databases to provide comprehensive real-time translation among two or more parties speaking different languages; and conversing with GPS software providing details on the best burgers, movies, hotels or people-watching spots trending along your route.

Tapping into the seductive power of large language models and natural language processing, we've witnessed tremendous progress in communications between us and technology that we increasingly rely on in our daily lives.

But there's been a stumbling block when it comes to AI and our portable devices. Researchers at Apple say they are ready to do something about it.

The issue is memory. Large language models need lots of it. With models demanding storage of potentially hundreds of billions of parameters, commonly used smartphones such as Apple's iPhone 15 with a scant 8GB of memory will fall far short of the task.

In a paper uploaded to the pre-print server *arXiv* on Dec. 12, Apple announced it had developed a method that utilizes transfers of data between flash memory and DRAM that will allow a smart device to run a powerful AI system.

The researchers say their process can run AI programs twice the size of a device's DRAM capacity and speed up CPU operations by up to 500%. GPU processes, they say, can be sped up to 25 times current approaches.

"Our method involves constructing an inference cost model that harmonizes with the flash memory behavior, guiding us to optimize in

two critical areas: reducing the volume of data transferred from flash and reading data in larger, more contiguous chunks," the researchers said in their paper titled, "LLM in a flash: Efficient Large Language Model Inference with Limited Memory."

The two techniques they used were:

1. Windowing, which slashes the amount of data that needs to be exchanged between flash memory and RAM. This is accomplished by reusing results from recent calculations, minimizing IO requests and saving energy and time.
2. Row column bundling, which achieves greater efficiency by digesting larger chunks of data at a time from flash memory.

The two processes, say the researchers, "collectively contribute to a significant reduction in the data load and an increase in the efficiency of memory usage."

They added, "This breakthrough is particularly crucial for deploying advanced LLMs in resource-limited environments, thereby expanding their applicability and accessibility."

In another recent breakthrough, Apple announced that it had designed a program called HUGS that can create animated avatars from just a few seconds worth of video captured from a single lens. Current avatar creation programs require multiple camera views. The report, "HUGS: Human Gaussian Splats," was uploaded to *arXiv* Nov. 29.

Their program can create realistic dancing avatars in as little as 30 minutes, far shorter than the two days required for current popular approaches, according to Apple.

More information: Keivan Alizadeh et al, LLM in a flash: Efficient

Large Language Model Inference with Limited Memory, *arXiv* (2023).
DOI: 10.48550/arxiv.2312.11514