PL
Unified Orchestration for ENTERPRISE AI with ORACLE
O-RAN & DAS Transforming Indoor Connectivity with SOLiD
Accelerating AUTONOMOUS NETWORKS with ERICSSON
RF COVERAGE FOR NEXT-GEN IoT
6G & WIRELESS TRENDS
PERVASIVE MOBILE CONNECTIVITY
5G SA & AI-NATIVE 6G MOBILE MESSAGING
MOBILE & WIRELESS
MOBILE AI • INTERNET BACKUP • ROAMING INTELLIGENCE

www.pipelinepub.com
Volume 22, Issue 4

Mobile AI

By: Mark Cummings, Ph.D.

What is the outlook for AI on mobile devices? Powerful LLMs (models) are getting very large. While progress has been made in running large models on Edge devices, there are serious limitations facing large models on mobile devices. Technical progress has been made in moving the power of large models to mobile devices. The challenge that faces us is learning how to convert that power into effective applications. A co-opetition process is the best way to quickly increase the capability to build effective mobile AI applications.

Background

The frontier AI companies have been on a trajectory of 10X (ten times) larger models each generation. Size is generally measured by the number of parameters. We have gone from Millions to Billions and now Trillions of parameters. Recently, this series of generations has created dramatic capabilities. We are likely to see the beginnings of this next 10X generation in mid to late 2027.

Advances in hardware and software have made running large LLMs (the brains of GenAI) in Edge devices possible, but with latency issues. So, mobile platforms still have trouble running large models locally.

For the next few years, we are likely to see mobile devices locally run relatively small LLMs.

New technology has appeared that makes these small LLMs very powerful.

There is a very broad mobile application space. We are very early in the process of learning how to be effective in applying GenAI. One way to move up the learning curve quickly is to share experience in a co-opetition setting.

Technology Overview

The limiting hardware factors for running LLMs on mobile devices are memory size, memory bandwidth, power consumption, and sometimes heat dissipation. By their nature, mobile devices tend to be limited in these areas. The limitations come from size, weight, and cost constraints.

Latency is another limitation. "Edge AI: Changing GenAI Balance Between Edge and Data Center" described software innovation. It showed how very large LLMs could be run in Edge devices. Since then, the Inferencer tool has been enhanced to add multi-system cooperative operation to its previous SSD streaming capability. Further increasing the ability of Edge devices to run large LLMs.

Generally, mobile devices have more restricted hardware platforms than do Edge devices. It doesn't appear that the Inferencer solution will have the capability to meet the latency requirements of most mobile applications. However, it does make Edge devices another option for data centers for mobile remote access.

There is a stream of hardware innovations that are attempting to address these limitations. Quadric has introduced a hardware product that increases the efficiency of mobile AI infrastructure. It is in the supply chain, but it may still be some time before it is fully utilized. However, the large LLMs are growing so big, so fast. They are staying beyond the capabilities of these kinds of innovations.

Longer-term hardware innovations initially targeted at data centers may trickle down to mobile. An example of such an innovation developed for higher performance / lower power in data centers is Neurophos. These appear to be several years away from widespread data center adoption, then several years more before they apply the technology to mobile.

Application Approaches

The result of mobile device technical limitations is that for the next few years, there will be four ways that AI is implemented in mobile devices:

1.) accessed through a browser on the mobile device

2.) accessed through a cooperative local application communicating with a remote AI

3.) running a small local model cooperating with a large remote AI

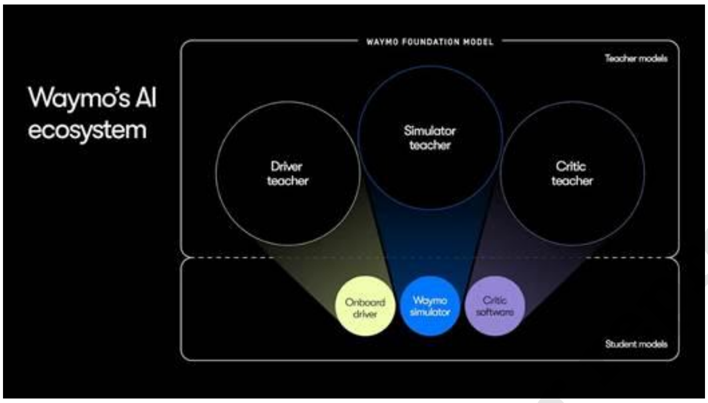
4.) running a small model locally

From the point of view of privacy, security, low latency, and reliable access, running an LLM locally has a big advantage. The question is whether the result has a fully adequate quality and functionality. In AI, quality is a measure of how useful the information the AI outputs is. This includes how accurate it is. Accuracy includes the hallucination problem and other measures of completeness, precision, fidelity, etc. Functionality has to do with the range of outputs. Does the range of output fully meet the application requirements?

For some relatively simple applications, small LLMs may fully meet the requirements for functionality and quality. In the past, there have been concerns about the ability of small models to meet the requirements of more complex applications.

Waymo has come up with a very interesting solution for how to meet the requirements of complex applications in the mobile space with small LLMs. Waymo has a way that a large AI can 'create' a very powerful AI focused on a single specific domain. This is done through a process called 'Distillation'.

In distillation, a very large model is trained on a very large training corpus. Then, the large model's inference results are analyzed. Instead of looking for the correct answers, the analysis attempts to determine the causes of the differences in the very low probability results that are never presented to an end user. It is trying to determine how the model determines the difference between a pretty wrong answer and a terribly wrong answer. The resulting knowledge can be characterized as a description of the general way that the large model makes decisions. This general way is transferred to the small model, which is then trained. The result is a model with many fewer parameters than the large model, but that performs in a fashion close to that of the large model. This small, powerful model is then deployed in mobile applications.



In Waymo's implementation of this technology, an extremely large model is trained. This extremely large model is used to create the Teacher models shown in the illustration below. The teacher models, in turn, are used to train their respective small models. Each focuses on a specific domain.

Because of Waymo's business focus, the three small models are a Driver that is deployed in an autonomous car. A Simulation model that is used to test the Driver. And a Critic model that is used to assess the results of the test in the Simulation model. The Critic can also be deployed in the car to assess the performance of the Driver in actual field operation.

Applications

For mobile chat applications, ways 1.) through 4.) above will be used to access AI's. For the enhancement of other simple mobile applications, way 2), accessing through a cooperative local application communicating with a remote AI, will be used.

For mobile intelligent agent applications, all four ways may be used. The more demanding applications will move to the Waymo distillation approach. Examples of the range of potential mobile applications include those briefly described below.

The technology can be used in other segments of the transportation area, such as drones, railroads, pipelines, etc.

There is a broad range of medical applications, including portable and transportable medical equipment, autonomous crash carts in medical facilities, wearable medical equipment, and medical equipment that moves around in the body (electronics packaged in a swallowable capsule, things that move around in the bloodstream, etc.).

There is a range of mobile applications in field service, construction, and operations. Operations applications can be in electrical, communication, and other kinds of networks. Operations in Automated factories also hold promise.

Recent progress in robotics is making humanoid robots seem feasible. These robots would be a particularly fertile application area.

There is interesting work being done with AI applications for the neurodiverse community. One AI application reads human expressions and tells the neurodiverse individual, who has trouble reading them, what the other person is showing on their face. Currently, this is implemented in conjunction with Zoom. Clearly, it would be valuable if implemented in a mobile platform that a neurodiverse person could have with them all the time. One mobile packaging could be in smart glasses.

As can be seen from these examples, the range of applications can be quite broad.

Co-opetition to Move Up the Learning Curve

One thing that is clear is that we are very low on the learning curve. As a society, we are just beginning to learn how best to utilize the power that is being created by GenAI. One way of moving up the learning curve faster is through cooperation. That is, those building the tools and those using the tools come together and share what they have learned. Some fear that this will help their competitors. In past generations of technology, a system of cooperating to build a shared knowledge base, then using that base in a proprietary fashion to compete with each other, has proven effective. That approach has come to be called co-opetition. A contraction of cooperation and competition. An AI group is exploring AI co-opetition. More information about that group can be found at https://www.bacesecurity.org/form/aiwg.

Conclusion

There is a broad range of valuable mobile AI applications. Powerful LLMs are getting very large. There is technical progress being made in moving that power to mobile devices. The challenge that faces us is learning how to convert that power into effective applications. A co-opetition process is the best way to quickly increase the capability to build effective mobile AI applications.