Researchers create an algorithm that maximizes IoT sensor inference accuracy using edge computing

October 25 2022

Credit: IMDEA Networks Institute

We are in a fascinating era where even low-resource devices, such as Internet of Things (IoT) sensors, can use deep learning algorithms to tackle complex problems such as image classification or natural language processing (the branch of artificial intelligence that deals with giving

computers the ability to understand spoken and written language in the same way as humans).

However, deep learning in IoT sensors may not be able to guarantee quality of service (QoS) requirements such as inference accuracy and latency. With the exponential growth of data collected by billions of IoT devices, the need has arisen to shift to a distributed model in which some of the computing occurs at the edge of the network (edge computing), closer to where the data is created, rather than sending it to the cloud for processing and storage.

IMDEA Networks researchers Andrea Fresa (Ph.D. Student) and Jaya Prakash Champati (Research Assistant Professor) have conducted a study in which they have presented the algorithm AMR^2, which makes use of edge computing infrastructure (processing, analyzing, and storing data closer to where it is generated to enable faster, near real-time analysis and responses) to increase IoT sensor inference accuracy while observing latency constraints and have shown that the problem is solved. The paper "An Offloading Algorithm for Maximizing Inference Accuracy on Edge Device in an Edge Intelligence System" has been published this week at the MSWiM conference.

To understand what inference is, we must first explain that machine learning works in two main phases. The first refers to training when the developer feeds their model with a set of curated data so that it can "learn" everything it needs to know about the type of data it is going to analyze. The next phase is inference: the model can make predictions based on real data to produce actionable results.

In their publication, the researchers have concluded that the inference accuracy increased by up to 40% when comparing the AMR^2 algorithm with basic scheduling techniques. They have also found that an efficient scheduling algorithm is essential to support machine learning algorithms

at the network edge properly.

"The results of our study could be extremely useful for Machine Learning (ML) applications that need fast and accurate inference on end devices. Think about a service like Google Photos, for instance, that categorizes image elements. We can guarantee the execution delay using the AMR^2 algorithm, which can be very fruitful for a developer who can use it in the design to ensure that the delays are not visible to the user," explains Andrea Fresa.

The main obstacle they have encountered in conducting this study is to demonstrate the theoretical performance of the AMR^2 algorithm and validate it using an experimental testbed consisting of a Raspberry Pi and a server connected through a LAN. "To demonstrate the performance limits of AMR^2, we employed fundamental ideas from linear programming and tools from operations research," highlights Fresa.

However, with this work, IMDEA Networks researchers have laid the foundations for future research that will help make it possible to run machine learning (ML) applications at the edge of the network quickly and accurately.

More information: Andrea Fresa et al, An offloading algorithm for maximizing inference accuracy on edge device in an edge intelligence system, *MSWiM proceedings* (2022). dspace.networks.imdea.org/handle/20.500.12761/1613

Conference: mswimconf.com/2022/

Provided by IMDEA Networks Institute