



Technical Perspective

Bridging AI with Real-Time Systems

By Giorgio Buttazzo

ARTIFICIAL INTELLIGENCE (AI) and machine learning models are making progress at an unprecedented rate and have achieved remarkable performance in several specific tasks such as image classification, object detection, automatic control, strategy games, some types of medical diagnoses, and music composition.

The exceptional performance of machine learning models in perception tasks makes them very attractive for being adopted in a large variety of autonomous systems, which must process sensory data to understand the environment and react in real time to accomplish a given task. Examples of such autonomous systems include self-driving cars, advanced robots operating in unknown environments, and interplanetary space probes. These systems must not only perceive the objects in the scene and their location with a high accuracy, but they also must predict their trajectories and plan proper actions within stringent timing constraints.

Consider, for instance, an autonomous car driving in an urban environment. Its onboard perception system is not only in charge of detecting the road, sidewalks, traffic lights, and road signs, but it is also responsible for identifying and recognizing moving objects, such as pedestrians, bicycles, and other moving vehicles, while predicting their trajectories and planning proper actions to prevent possible impacts with them. In this context, a correct prediction produced too late could cause the system to fail. This example illustrates that guaranteeing a timely response in this type of system is as crucial as producing a correct prediction.

In a complex, highly dynamic scenario like the one considered for a self-driving car, however, not all computational tasks are equally important. For example, objects closer to the vehicle should receive a higher pri-

ority with respect to those located further away. Similarly, objects moving at higher speed should be processed at higher rates with respect to objects that are standing or moving at lower speed.

One problem with the current AI frameworks and hardware accelerators for deep neural networks is that they have been developed for non-critical applications where timing is not an issue. Consequently, when multiple neural models must be executed on the same platform, each model is normally executed non-preemptively (that is, without interruption) or, in the best case, using simple scheduling heuristics that do not take time requirements or task criticality into account.


This means if a highly critical task H is activated just after a low-critical task L has started its execution, H will experience a long delay, since it can only start executing after the completion of L . This phenomenon is referred to as a *priority inversion* and has been studied extensively in the field of real-time systems. However, it represents a serious problem in current AI algorithms, preventing their use in safety-

critical real-time systems, where timing and functional requirements are equally important.

The following paper proposes a new methodology for overcoming the limitations of current AI frameworks to enable the use of deep neural networks in mission-critical systems. The key idea is to split the perception process into multiple tasks associated with different objects and prioritize them to enable more timely response to more critical stimuli.

The system combines range data acquired from a light detection and ranging sensor (LiDAR) with images obtained from a camera. In particular, the 3D objects detected by the LiDAR (based on distances) are projected on the 2D-image plane of the camera. Then, bounding boxes are assigned a priority inversely proportional to their distance from the vehicle, so that closer objects will be attended first. In this way, depending on the overall workload, low-priority objects can be processed by a lower rate.

To overcome the limitation of non-preemptive execution, the deep neural network in charge of processing the cropped images is broken into stages, each consisting of multiple neural layers so the model inference can be preempted between stages. To add flexibility, multiple predictions with different precision are generated at the end of multiple stages to balance accuracy vs. execution time.

The overall system is then able to schedule the various perceptual tasks based on the assigned priority, while avoiding priority inversion during neural inference and enabling a more predictable execution of AI algorithms in mission-critical systems. 

Giorgio Buttazzo is a professor of computer engineering at the Sant'Anna School of Advanced Studies in Pisa, Italy.

The following paper proposes new methodology for overcoming the limitations of AI frameworks to enable the use of deep neural networks in mission-critical systems.