Object Detection with YOLOv8

Running Machine Learning on Ampere® Altra® Max

Ampere Altra family of processors with high performance Ampere® AI inference engine, deliver best-in-class AI inference performance on standard frameworks, including PyTorch, TensorFlow, and ONNX-RT.

Ampere® Altra® Max Powered ML Inference

Ampere Altra family of **cloud-native processors** meets the needs of widely used ML workloads across transportation, industrial, telecommunication, and others while **optimizing the total cost of** sources, webcam or video files detecting still and moving objects like pedestrians, laptop, chair, cup, etc., using the popular YOLOv8 model.

Setup

Deployment of open-source **computer vision** object detection AI model, YOLOv8 with **Ampere® Optimized PyTorch**, running on Ampere Altra Max. The chosen model, YOLOv8, is a widely used algorithm for real-time applications where both throughput & latency are critical. Implementation and performance details for the YOLOv8 model developed and released by Ultralytics can be found here: https://github.com/ultralytics/ultralytics/tree/main/ultralytics/yolo/v8

Key Benefits Demonstrated

- Meets or exceeds the necessary **low latency** requirements for real-time ML object detection applications.
- Delivers the best **price-performance** in CPU-only AI inferencing in both cloud and edge deployment scenarios.
- The YOLOv8 model can be downloaded from Ampere® AI Model Library (AML) and used as is without any modifications.
- Ampere Altra Max processor with 128 cores can **easily be scaled** and **dynamically provisioned** based on the performance requirements of the user's application such as target frame rate, number of video

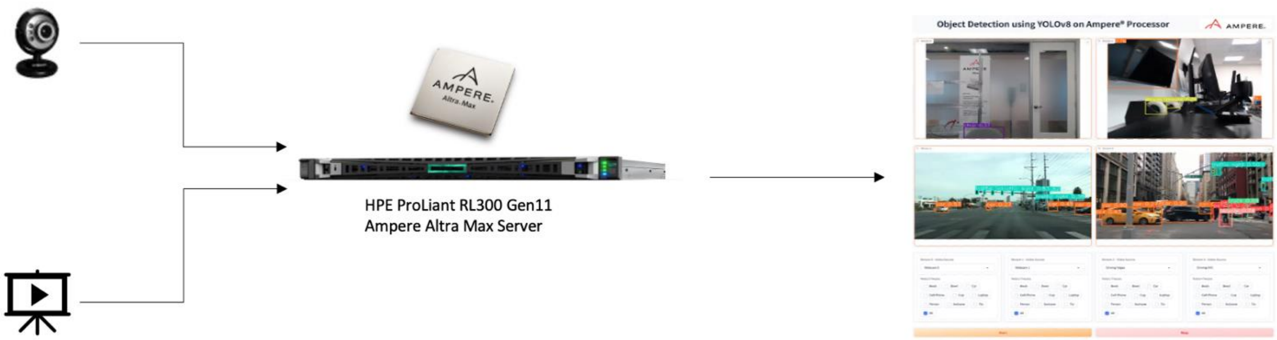

Figure 1: Ampere Altra Max YOLOv8 demo runs on local Ampere Altra Max server.

Low Latency Demo – Real-time Object Detection and Classification

This demo performs object detection and classification inference with a pre-trained YOLOv8 model. It processes images and videos from an incoming real-time video streaming from a camera or video files. The demo runs on a **local Ampere Altra Max server** at real time **performance level**. The performance can be scaled up or down depending on application requirements by assigning processor cores and adding or removing CPU instances to meet a desired price-performance target.

The same workload also runs on x86 for comparison purposes. We demonstrate that the **Ampere Altra family of cloud-native processors consistently outperforms x86 platforms.**

Resources

YOLOv8 demo supports AI inference on Ampere Altra family processors and with NVIDIA GPU. The YOLOv8 model can be accessed from Ampere AI Model Library. The docker image of Ampere® Optimized PyTorch is available in the downloads section of Ampere's AI Solutions web page. Other Ampere® Optimized Frameworks can also be accessed from the same location.

Ampere Optimized TensorFlow, PyTorch, ONNX-RT can also be downloaded and installed free of charge on any edge workstation or server through Ampere's AI Solutions web page.

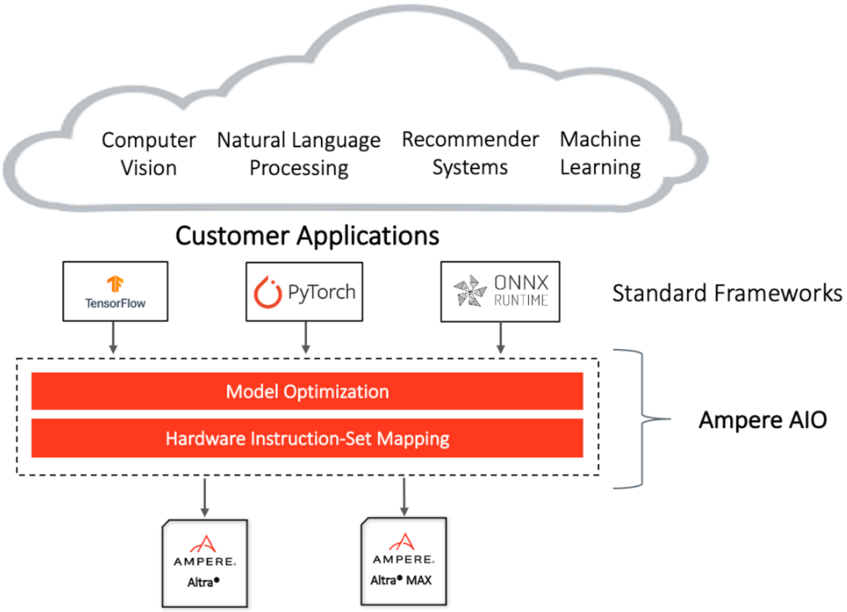

Figure 2: Ampere Altra instances are available from a variety of cloud providers, including Oracle Cloud Infrastructure, Google Cloud, Microsoft Azure, Tencent, Equinix, etc.

AMPERE® AI