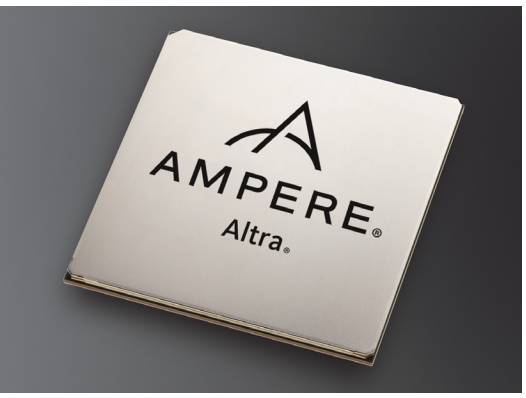
AMPERE®



Object Detection with YOLOv5 — Running Machine Learning Workloads on Ampere® Altra®

Ampere Altra, with high performance Ampere® AI inference engine, offers the best-in-class consistent machine learning (ML) inference performance on standard frameworks including PyTorch, TensorFlow, and ONNX.

Ampere AI Powering ML Inference Workloads

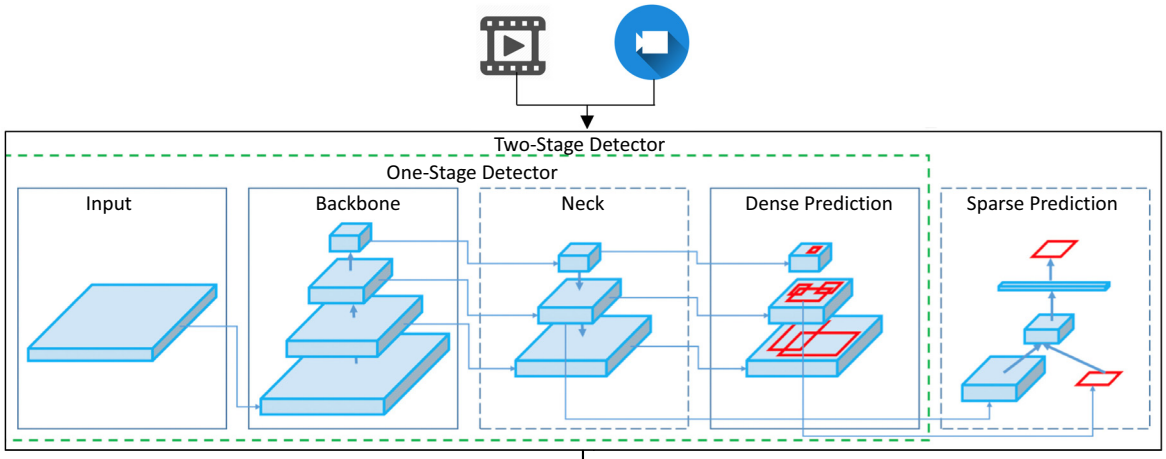
Ampere AI on the Ampere Altra family of Cloud Native processors satisfies the needs of common ML workloads while optimizing the total cost of operations. This demo demonstrates a video analytics use case that detect common objects such as vehicles, pedestrians, and traffic signs.

Setup

Deployment of open-source **computer vision** object detection AI model YOLOv5 is done using **Ampere optimized PyTorch**, running on Ampere Altra. The chosen model, YOLOv5, is the go-to algorithm for real-time applications where both throughput and latency are critical.

Key Benefits Demonstrated

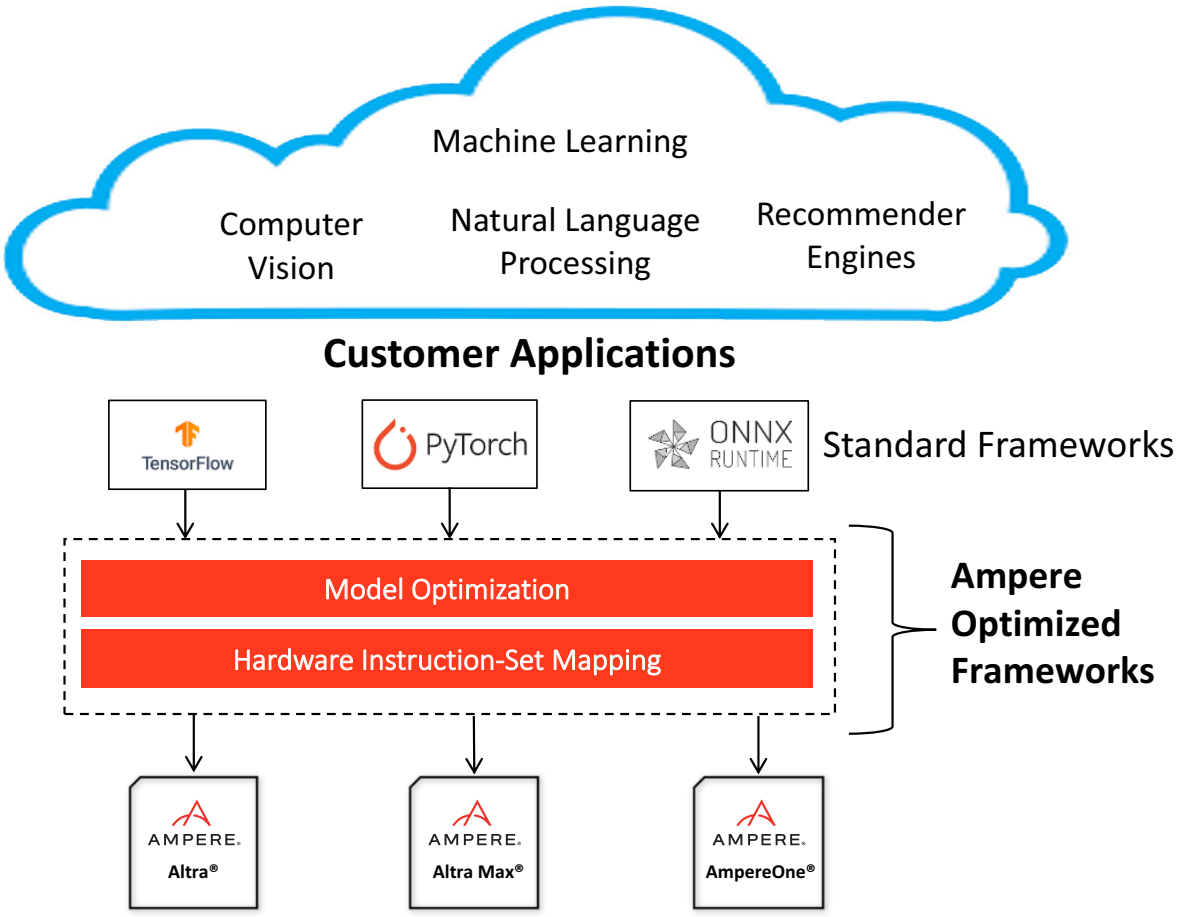
- Provides the necessary **low latency** requirements of real-time ML object detection applications.
- Delivers the best **cost-performance** in CPU-only AI inferencing on cloud deployment.
- Object detection use cases can readily take advantage of the **high performance** of Ampere AI.
- Can be **easily scaled** and dynamically provisioned based on the needs (e.g., target FPS, number of channels, etc.)



Low Latency Demo – Real-time Object Detection and Classification

This demo performs object detection and classification inference with a pre-trained YOLOv5 model. It processes images and video files from a web app and video frames from a camera source. The demo runs on OCI Ampere A1 instance with Ampere Altra at real time performance level (30 fps), or close to it (depending on model size).

The same workload also runs on x86 and Graviton3 based instances for comparison purposes. It demonstrates that the Ampere Altra family of cloud-native processors consistently outperforms competing x86 and ARM64 (e.g., AWS Graviton) platforms.



Resources

The demo is available in the form of a docker image. The YOLOv5 model can be accessed from here. The docker image of Ampere optimized PyTorch is available in the downloads section on the Ampere AI Solutions website and can be accessed for free. Other Ampere optimized frameworks can be accessed in the same manner. You can try out Ampere A1 instance via Oracle's free tier. Access additional information on Ampere A1 Compute and look up the Ampere optimized TensorFlow listing on OCI marketplace.

Ampere Computing® / 4655 Great America Parkway, Suite 601 / Santa Clara, CA 95054 / www.amperecomputing.com