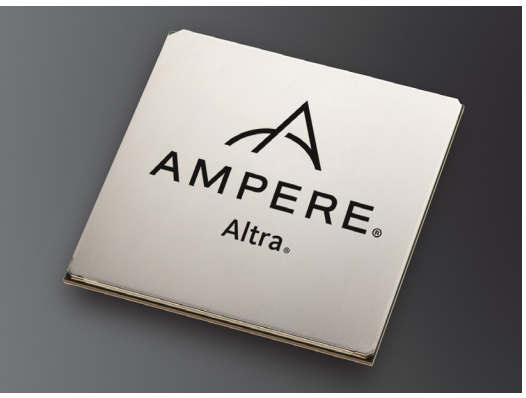
AMPERE®



Privacy Preservation and Face Tracking — 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

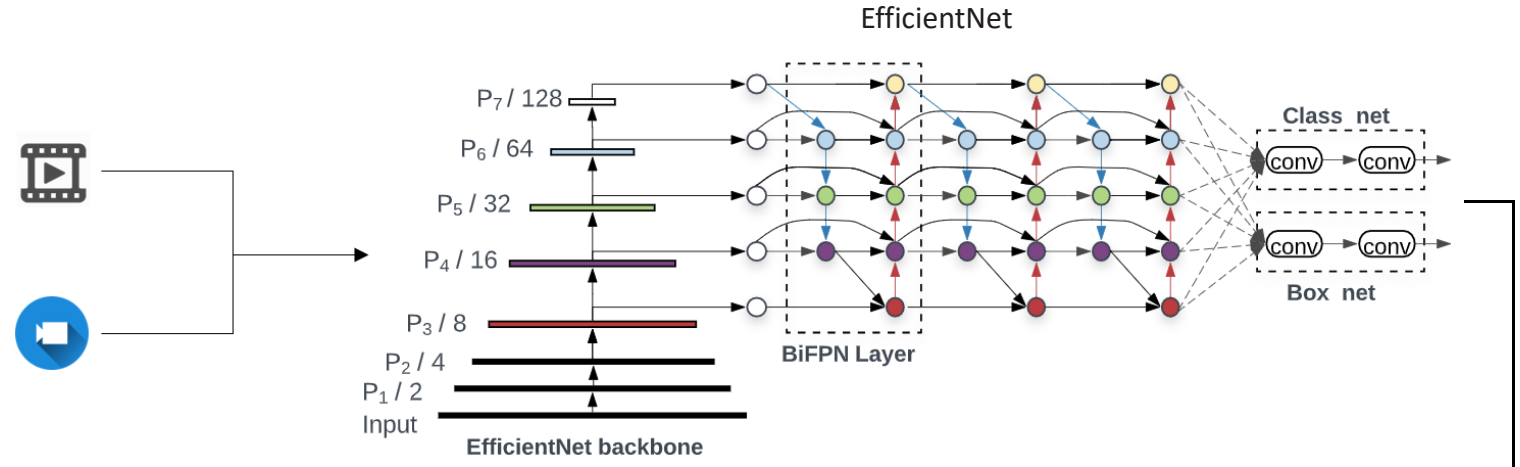
Ampere AI on the Ampere Altra family of cloud-native processors satisfies the needs of common ML workloads while optimizing the total cost of operations. The demo shows real-time face and body masking to conceal the identities of people in the video.

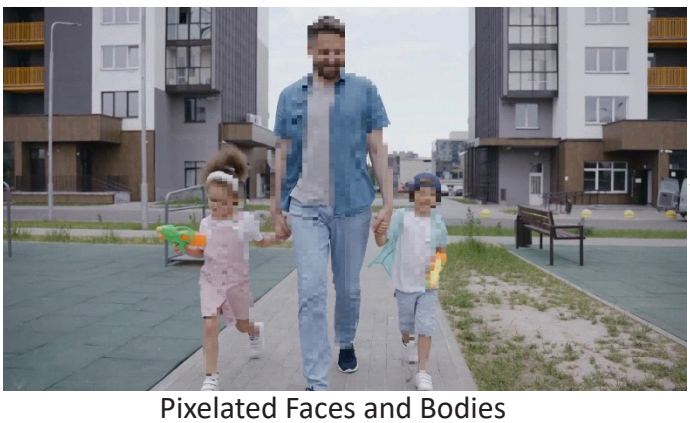
Setup

Deployment of open-source human and body key-point tracking AI models with Ampere optimized TensorFlow, running on the Ampere Altra. The demo uses EfficientDet-Lite2 and MoveNet models.

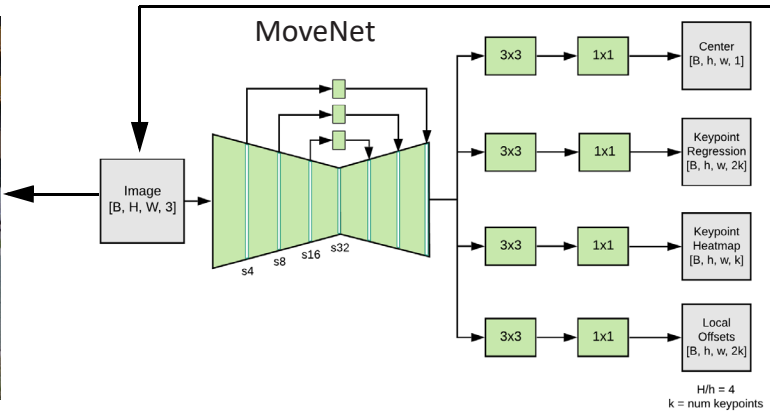
Key Benefits Demonstrated

- 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.)





Pixelated Faces and Bodies



Model Usage & Processing Steps

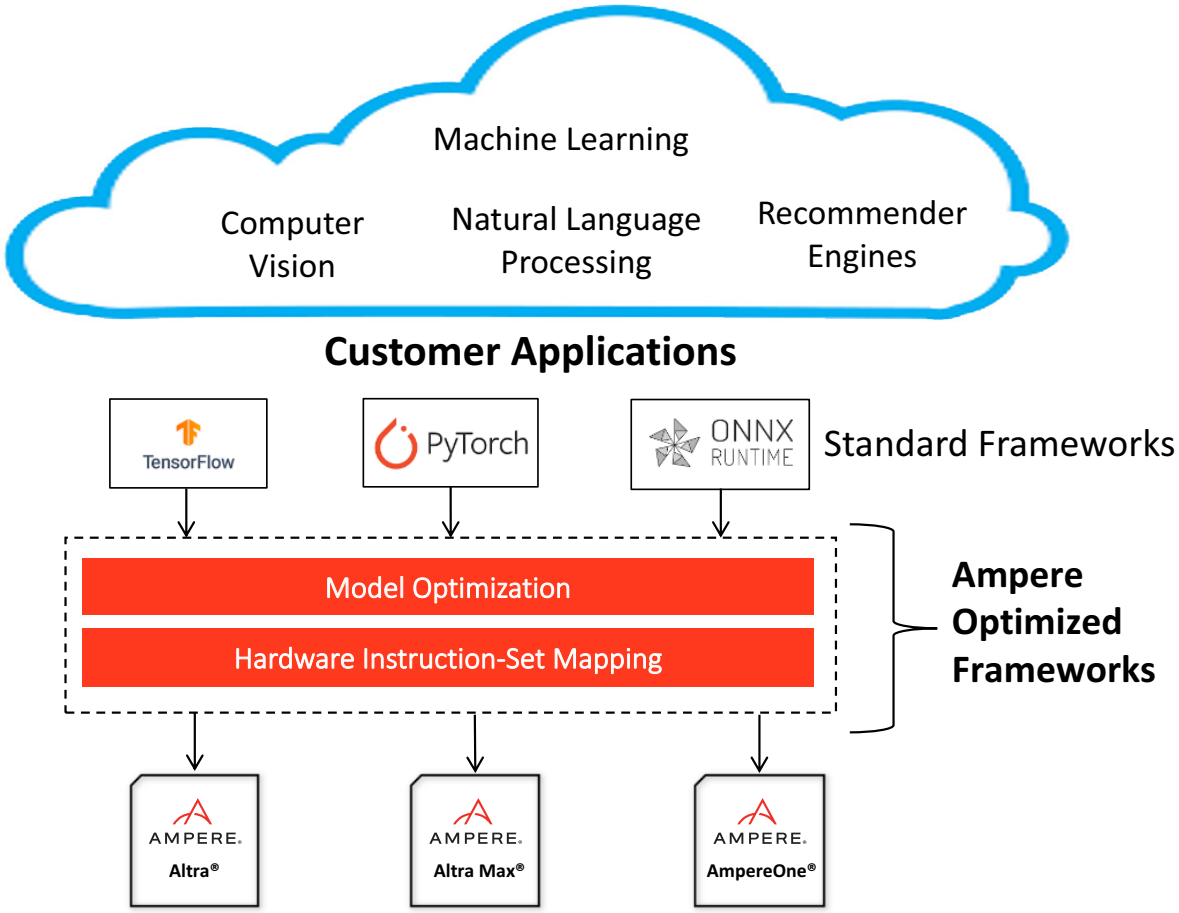
Models integrated:

1. EfficientNet is used to detect a human in the frame. After a human is detected, the human region is passed on to the next model.
2. MoveNet is used to detect the human body's key points (e.g., hands, feet, head, etc.), then either only facial features or the whole body can be masked by pixelation to preserve the individuals' privacy.

Low Latency Demo – Real-time Object Detection and Classification

This demo performs human and body key-point tracking with a pre-trained video file. A video file that contains people is loaded, and the output shows pixelated faces and bodies in real-time. The demo runs on OCI Ampere A1 instance.

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 docker image of Ampere optimized TensorFlow 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 reserves the right to make changes to its products, its datasheets, or related documentation, without notice and warrants its products solely pursuant to its terms and conditions of sale, only to substantially comply with the latest available datasheet.

Ampere, Ampere Computing, the Ampere Computing and 'A' logos, and Altra are registered trademarks of Ampere Computing.

Arm is a registered trademark of Arm Limited (or its subsidiaries) in the US and/or elsewhere. All other trademarks are the property of their respective holders.

Copyright © 2022 Ampere Computing. All Rights Reserved.

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