Automatic Speech Recognition (ASR) with Whisper Model

Ampere® Cloud Native Processors with Ampere® Optimized AI Frameworks, deliver best GPU-Free AI inference performance for applications developed in PyTorch, TensorFlow, and ONNX-RT.

Ampere Altra Powered ML Inference

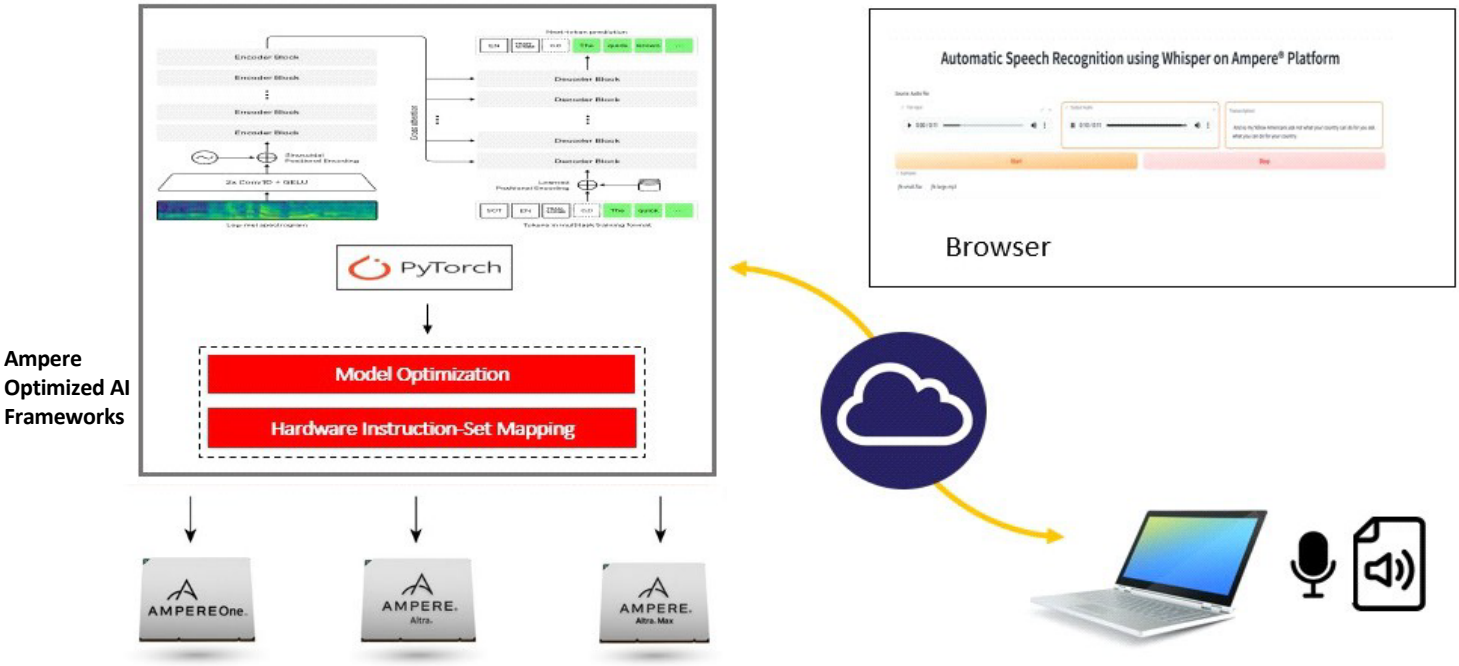
Ampere Cloud Native Processors satisfy the performance requirements of widely used machine learning (ML) workloads while providing the best price-performance and optimizing power draw. This demo performs transcription of audio files into text, using the state-of-the-art Open AI Whisper model. Whisper offers the best-in-class accuracy and capabilities for Automatic Speech Recognition (ASR) use cases.

Setup

The setup includes the deployment of the open-source ASR AI model Whisper, with **Ampere® Optimized PyTorch**. The chosen model, Whisper Medium, is a widely used algorithm for ASR applications where both throughput and latency are critical. Implementation and performance details for the Whisper model by Open AI can be found at https://github.com/openai/whisper.

Key Benefits Demonstrated

- Meets or exceeds the necessary l**ow latency** requirements for real-time ML Automatic Speech Recognition (ASR) applications.
- Provides GPU-Free AI inference performance superior to many competing GPU deployments.
- Delivers the best GPU-Free AI inference **price-performance** in both cloud and edge deployment scenarios, compared to competing CPU-only and GPU deployments.
- The Whisper model can be downloaded from Ampere® AI Model Library (AML) and used as is without any modifications.
- Ampere Altra processor can **easily be scaled** and **dynamically provisioned** based on the performance requirement of the user's applications.



Real-time Automatic Speech Recognition (ASR)

This demo performs ASR inference with a pre-trained Whisper model. It processes audio streams read from audio files. The demo runs at real-time **performance level** (the rate of speech-to-text processing is faster than the rate of the audio stream). The performance can be scaled depending on application requirements by allocating the number of CPU cores to meet the desired price-performance target.

Resources

The Whisper model can be accessed from the Ampere AI Model Library. The docker image of Ampere Optimized PyTorch is available in the downloads section of Ampere AI Solutions web page. Other Ampere® Optimized Frameworks can also be accessed from the same location.

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

Figure 2. Integration of Ampere Optimized Frameworks with Ampere Cloud Native Processors