Retrieval - Augmented Generation (RAG) with View.io

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 Powered ML Inference

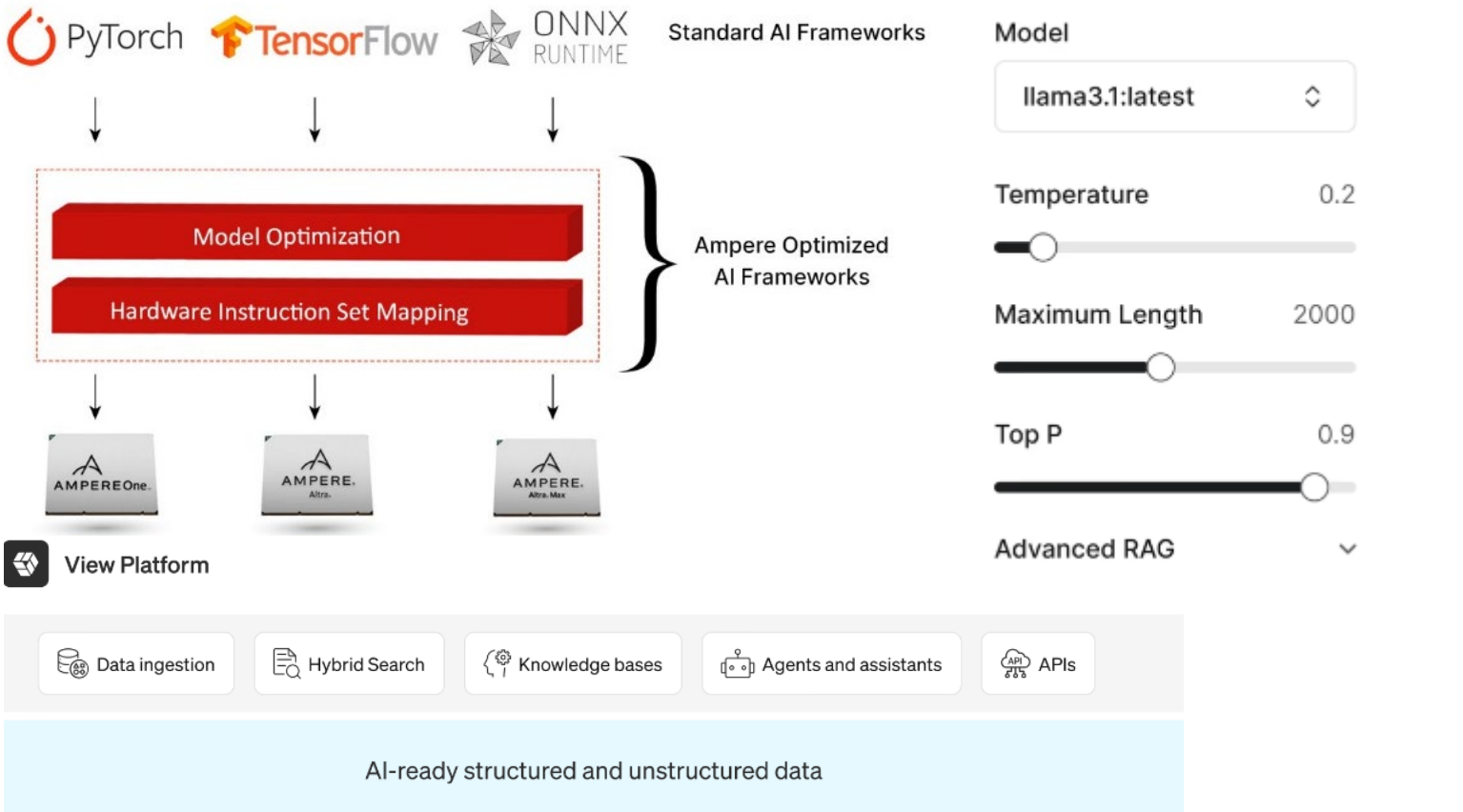
Ampere **Cloud Native Processors** satisfy the performance requirements of widely used AI workloads while **providing the best price-performance and optimizing power draw**. This demo showcases elements of the View.io end-to-end enterprise AI platform. View.io's generative AI solutions optimizes LLM (Large Language Models) performance with View's data management platform and RAG pipeline.

Setup

Deployment of the View.io platform with **Ampere® Optimized AI Frameworks (AIO®)** running on AmpereOne®. Processing of content into AI-ready formats for powerful search and chat capabilities with the use of a number of different LLMs to choose from. View assistant enables real-time conversation with enterprise data with context-aware capabilities.

Key Benefits Demonstrated

- Provides **leading end-user experience** in terms of time to first token and tokens per second (tps).
- Delivers the best **price-performance** in AI inference in both cloud and edge deployment scenarios.
- View.io platform leverages a number of open-source models which can be **securely deployed** to satisfy even the most stringent privacy requirements.



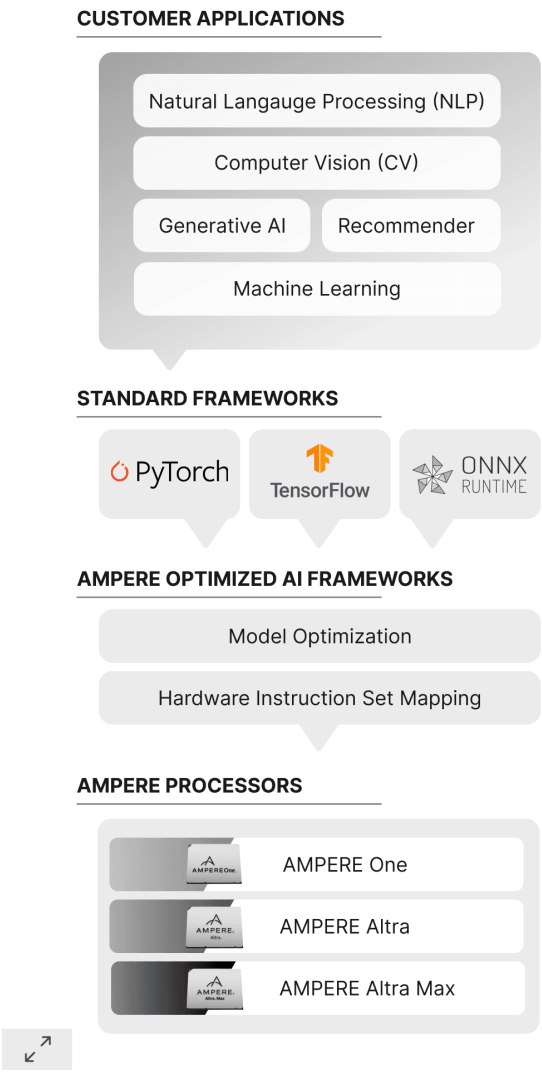
Real-time Object Detection and Classification

This demo performs retrieval-augmented generation (RAG). It's a partner solutions of View.io, specifically their View Assistant. It runs at **real-time performance level**. The performance can be tweaked depending on application requirements, be in terms of time to first token or tokens per second (tps), to meet the desired end-user experience expectations price-performance targets.

Resources

The docker images of Ampere Optimized AI Frameworks (Ampere Optimized TensorFlow, PyTorch, ONNX-RT) can be downloaded and installed free of charge on any edge workstation or server through Ampere AI developers web page. Further details about the View.io platform can be accessed on their website. View.io also offers free trial, which allows you to test RAG performance on Ampere Cloud Native Processors for yourself.

Integration of Ampere Optimized Frameworks with Ampere Cloud Native Processors



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