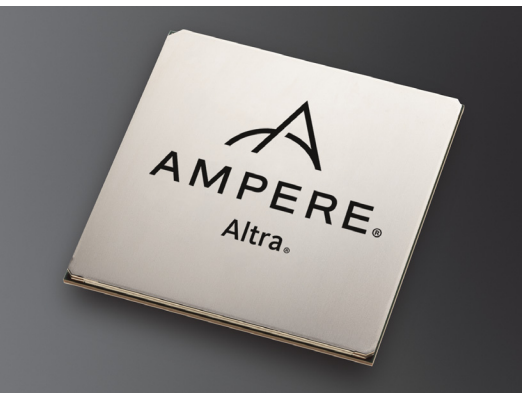
AMPERE®



Text Sentiment Analysis with Natural Language Processing (NLP) — 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

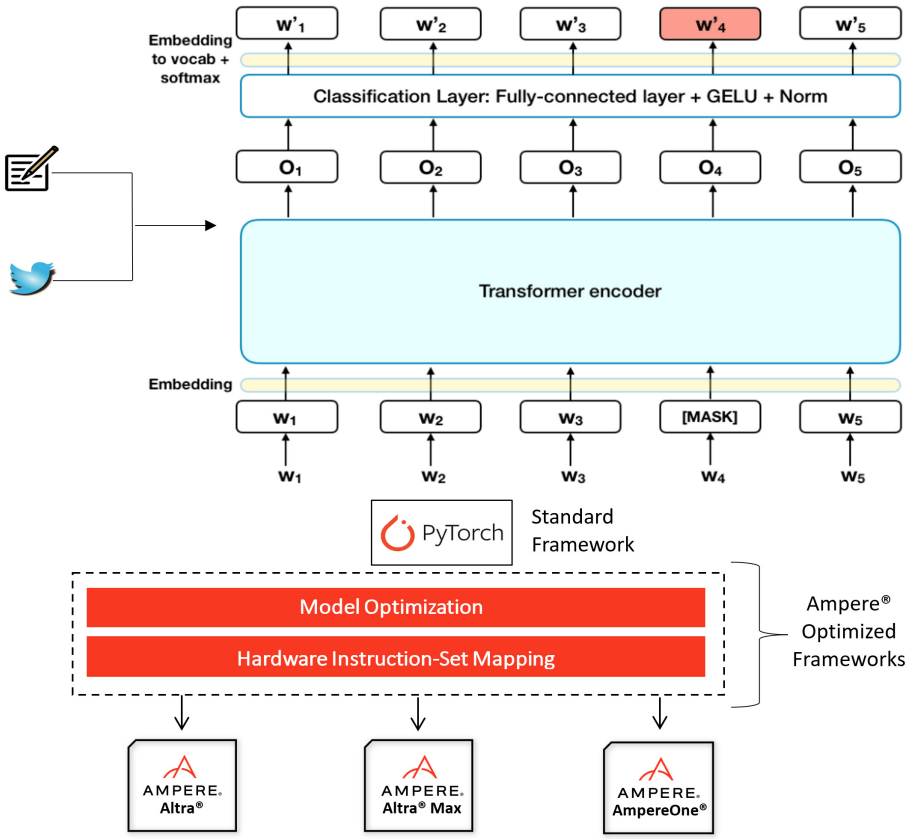
Ampere AI on the Ampere Altra family of Cloud Native processors satisfies the needs of common ML workloads while optimizing the total cost of operations. This specific model was trained for sentiment analysis tasks and was adapted for classification of three classes of sentiments: positive, negative, and neutral.

Setup

Deployment of open-source NLP AI model RoBERTa for sentiment analysis is with Ampere optimized PyTorch running on the Ampere Altra family of Cloud Native processors. We use RoBERTa, a transformer-based model belonging to the BERT family. The model used in this demo, "Twitter-roberta-base" is based on the RoBERTa architecture but was retrained on 124M tweets and fine-tuned for sentiment analysis with the TweetEval benchmark dataset.

Key Benefits Demonstrated

- Provides the necessary **low latency** requirements of real-time ML NLP applications.
- Delivers the best **cost-performance** in CPU-only AI inferencing on cloud deployment.
- NLP can readily take advantage of the **high performance** of Ampere AI.
- Can be **easily scaled** and dynamically provisioned based on the requirements.



Interface

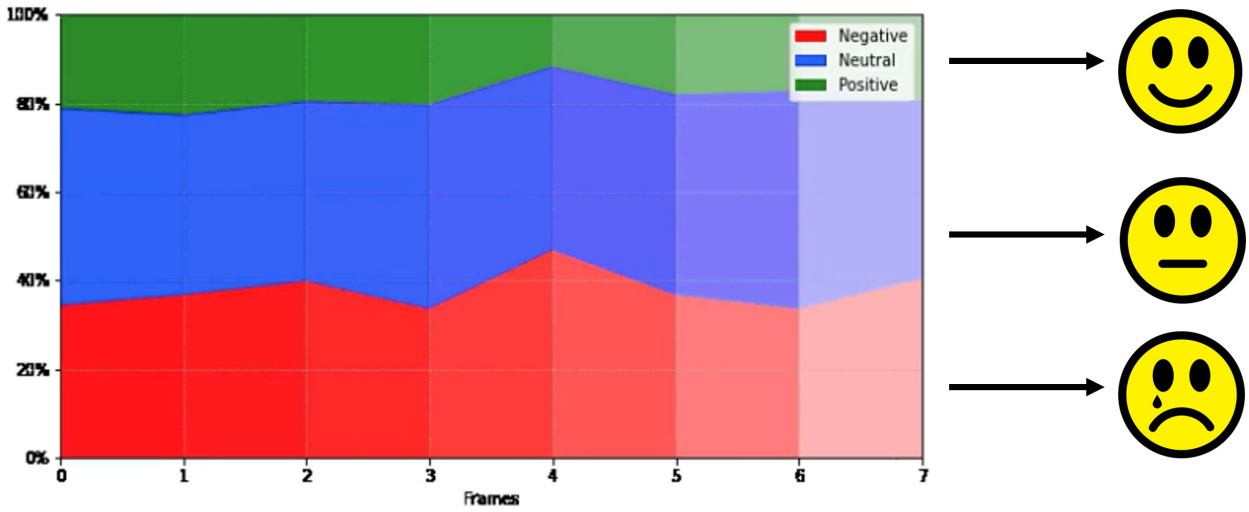
The demo has two interfaces:

1. Interactive Demo where the user can enter a sentence and the web application can immediately receive the corresponding sentiment class probabilities.
2. Non-Interactive Demo where a continuous stream of Twitter feed sentences is fed into the model. The Jupyter Notebook app then plots the percentage of sentences with positive, negative, and neutral sentiments.

Low Latency Demo – Real-Time NLP: Sentiment Analysis

This demo performs sentiment analysis using a pre-trained RoBERTa model. It processes language input from a web application user and from a continuous Twitter feed. The demo runs on an OCI Ampere A1 instance with Ampere Altra.

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 TweetEval benchmark 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 PyTorch listing on OCI marketplace.

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