Numenta and Intel Deliver Cost-Effective, Powerful Inference Performance

Fast, efficient technologies are essential for various natural language processing (NLP) tasks, including text classification, sentiment analysis, and ChatGPT. Transformer models such as BERT and GPT play a crucial role due to their ability to manage intricate text and provide precise outcomes. To meet the throughput and latency demands that sophisticated NLP applications require, customers typically run transformers on GPUs for deep learning inference. However, GPUs can be more costly, and adding extra hardware can mean more maintenance for IT departments. In collaboration with Intel, Numenta combined their proprietary, neuroscience-based solution for dramatically accelerating transformer networks with the new Intel® Advanced Matrix Extensions available in 4th Gen Intel® Xeon® Scalable processors.

"These breakthrough results make CPUs the best option for running transformers. Customers with performance-sensitive AI applications can use the combination of Numenta and 4th Gen Intel® Xeon® Scalable processors to deploy their real-time applications in a lightweight, cost-Effective manner."

Subutai Ahmad, CEO, Numenta

Products and Solutions
4th Gen Intel® Xeon® Scalable processors
Intel® Advanced Matrix Extensions
Intel® Xeon® CPU Max Series

Industry
Software Development

Organization Size
11–50

Country
United States

Learn more
Case Study

1, 2 For more complete information about performance and benchmark results, visit https://www.intel.com/content/www/us/en/customer-spotlight/stories/numenta-transformer-networks-customer-story.html