Solution Brief

Data Science/AI Algorithm Development
Natural Language Processing

Winning Health NLP Solution Runs Up To 6.30X Faster[1] on 4th Generation Intel® Xeon® Scalable Processors

Optimized Named Entity Recognition algorithm leverages Intel Advanced Matrix Extensions (Intel AMX) and 16-bit quantization to accelerate AI-driven healthcare information integration solution, assisting clinicians to deliver better patient care

intel AI Builders

卫宁健康 WINNING HEALTH

Health Information Systems (HIS) contain a variety of unstructured data about patients coming from many departments and entered by different doctors, nurses, other clinicians, and assistants. This creates a massive repository of unstructured, fragmented, and unconnected data about each patient.

This data is unique to healthcare, with its own language and text characteristics, making it difficult for automated systems to utilize and analyze data holistically for patient care and management. Non-healthcare-focused systems lack an effective and systematic way to structure, integrate, and analyze the data in a meaningful way to enable doctors to gain insights and make more precise clinical decisions.

Winning Health's Natural Language Processing (NLP) platform uses Artificial Intelligence (AI) and machine learning (ML) techniques to analyze and integrate various patient data in the language of healthcare. Their solution assists doctors to arrive at better clinical diagnoses and aid researchers with richer medical research—ultimately delivering better patient care.

The Winning Health Natural Language Processing (NLP) Information System

Winning Health is a leading healthcare software and solutions provider in China. The company develops software systems to enable digital transformation of IT infrastructure in hospitals. Such solutions help their customers integrate business functions, healthcare data, and service delivery on a connected digital platform. With the emergence of Artificial Intelligence (AI) and machine learning (ML), Winning Health created its own AI lab to focus on developing ML-based solutions to assist clinicians and hospital personnel.

The Winning Health NLP platform was designed for hospitals to integrate various sources of medical data, including clinical notes, imaging reports, lab tests, visitation records, and more. The integration unlocks information that is hidden in unstructured and unrelated medical data points. It offers a holistic view of patient data to help doctors make more precise clinical decisions and provide better patient care and research.

"Our Partnership with Intel has been a fantastic experience. The AI Inference team has enabled us to improve the performance of our AI healthcare solutions. We're offering our customers multiple high-performance and cost-effective solutions running on Intel platforms. Most recently, we collaborated with Intel Team in optimizing the NLP workload on the latest Intel 4th Gen Xeon processor with built-in AMX accelerator. We're proud to announce that this optimized solution will be available for our customers once the latest platforms are commercially available in the market."

LIU MINGQIAN, Director of Winning Health AI Lab

The Key NLP Components for AI-Assisted Health Data Insight

The key algorithm modules of the system's workflow comprise a named entity recognition model and a relation extraction model based on medical texts.

Named Entity Recognition

An entity is the basic element of knowledge or a concept. It is an object that can be uniquely identified and distinguished from other entities. Named entity recognition in the Winning Health NLP solution summarizes and differentiates these concepts that appear in medical texts. It then organizes conceptual systems, such as patients, parts, diseases, and symptoms.

The named entity recognition model is composed of a neural network model and a pre-trained language model. These models enable machines to automatically identify entities that appear in medical texts, including identifying the boundaries of entities and determining the type of entities.

Relation Extraction

Semantic relationships describe the association and interaction between entities and concepts. These relationships are one of the core components of knowledge. For example, a diagnostic relationship exists between the patient and the disease, and the disease presents different symptoms.

The relationship extraction model can automatically identify the semantic relationship between different entities in the text. The model can form triples, thereby generating a semantic network, matching the text, structuring the text, and storing it in the form of graph data.

Optimizing Winning Health NLP for 4th Gen Intel Xeon Scalable Processors

It is often misunderstood that AI applications must run on GPUs to achieve acceptable performance. The 4th Generation Intel Xeon Scalable processor family integrates Intel® AMX and other accelerators for AI operations. Additionally, software optimized for Intel architecture and Intel AI acceleration can deliver significant performance boosts for many solutions that allow solution providers to run their software on Intel processors alone, without GPUs, helping to reduce overall costs—including public cloud operations.

Winning Health wanted their NLP solution to run optimally on Intel architecture. For benchmarking, the named entity recognition task was run on a BERT model.

The original Winning Health NLP model was composed in PyTorch. The following optimizations were used to improve inference throughput on 4th Gen Intel Xeon Scalable processors:

- Used the Intel Optimization for PyTorch and further applied the Intel Extension for PyTorch to Winning Health code.
- Enabled use of Intel AMX on 4th Gen Intel Xeon Scalable processors.
- Applied BF16 quantization to the model.

Benchmarking the Optimizations

Engineers benchmarked performance on 3rd Generation Intel Xeon Scalable processor (Intel Xeon 8380 processor) and then on 4th Gen Intel Xeon Scalable processor (Intel Xeon 8480+ processor). From the un-optimized FP32 (baseline) to optimized FP32 model, throughput increased up to 2.64X on Intel Xeon 8480+ processor. Adding BF16 quantization increased throughput performance up to 6.04X from baseline (Figure 1).[1]

Performance improvements resulted from the optimizations for Intel architecture provided in the Intel enhancements for PyTorch, the use of Intel AMX to accelerate matrix calculations, and 16-bit quantization to reduce the magnitude of calculations. These accumulate into as much as a 6.04X throughput boost on the CPU alone.

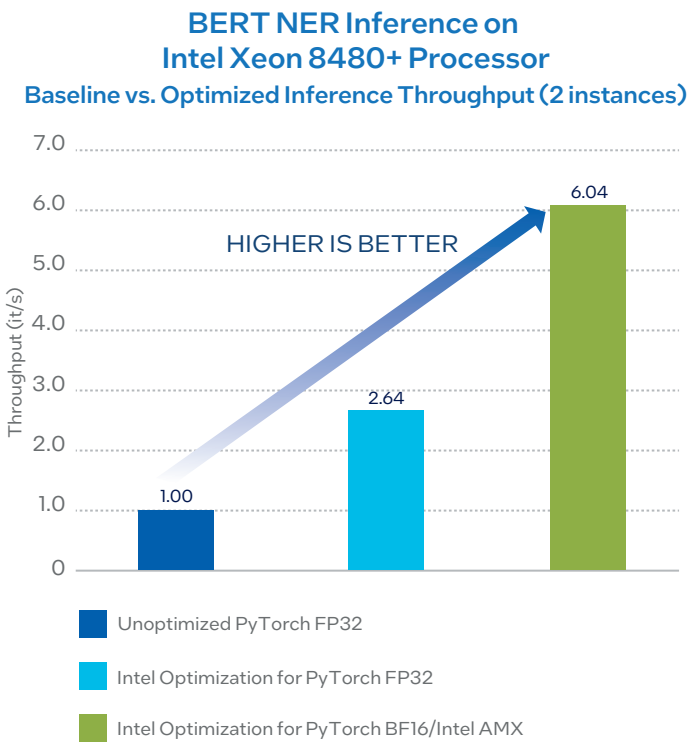

Figure 1. Benchmark results of Winning Health NLP named entity recognition on multiple generations of Intel Xeon Scalable processors.

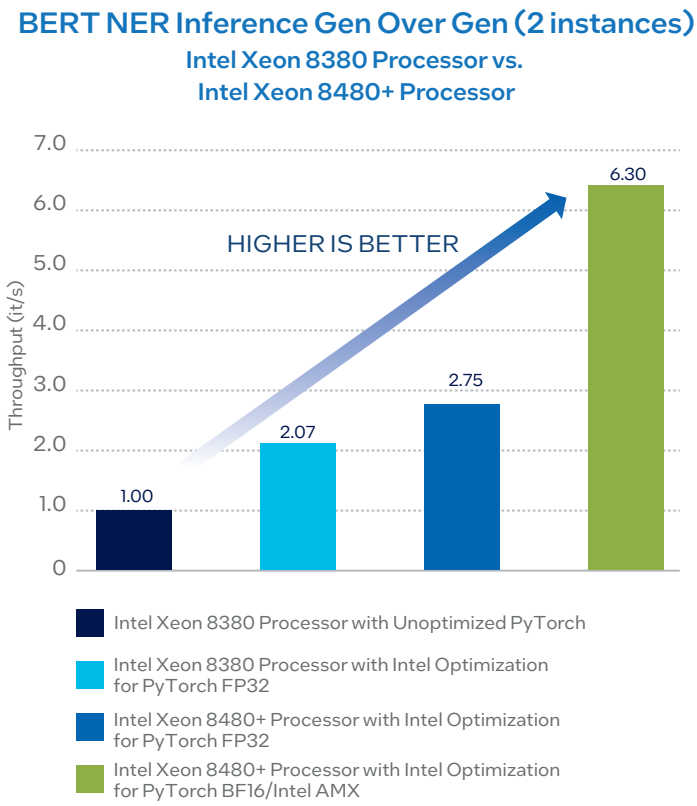

Figure 2. Gen over Gen Comparison—Intel Xeon 8380 Processor vs. Intel Xeon 8480+ Processor.

Optimizations:

OMP_NUM_THREADS=56

KMP_AFFINITY=granularity=fine,compact,1,0

KMP_BLOCKTIME=1

LD_PRELOAD=/home/sdp/.conda/envs/wh/lib/libiomp5.so:/home/sdp/.conda/envs/wh/lib/libtcmalloc.so

Optimizations:

OMP_NUM_THREADS=56 (Intel Xeon 8480+ Processor) / 40 (Intel Xeon 8380 Processor)

KMP_AFFINITY=granularity=fine,compact,1,0

KMP_BLOCKTIME=1

LD_PRELOAD=/home/sdp/.conda/envs/wh/lib/libiomp5.so:/home/sdp/.conda/envs/wh/lib/libtcmalloc.so

The benefits of running on the latest generation of Intel Xeon Scalable processors was shown when engineers benchmarked the performance across two generations of processors—Intel Xeon 8380 processor and Intel Xeon 8480+ processor. With Intel AMX and support of BF16 in the Intel Xeon 8480+ processor, throughput increased by up to 6.30X compared to the previous generation Intel Xeon 8380 processor without Intel AMX and BF16 support (Figure 2).

Conclusion

Since deployment at the end of 2021, the Winning Health NLP solution has enabled hospitals to integrate, link, and analyze 57,000 medical records for patients with gastroenterology- and pancreas-related diseases. The solution has helped researchers extract and identify 32 imaging features and 114 pathological features based on the research subject.

The solution is also currently deployed by a leading hospital in China for pancreatic cancer research and treatment. Pancreatic cancer is regarded as one of the most difficult -to-treat forms of cancer because it is very hard to diagnose in its early stage. Multidisciplinary consultation requires

the participation of imaging, surgery, oncology, and other disciplines. The original image information offers only qualitative analysis and is not traditionally integrated with the electronic medical record system.

With the Winning Health solution, relevant entity recognition and relationship extraction models are used in the evaluation of pancreatic cancer imaging quality control. Using the technology, customers can further analyze the imaging diagnosis and pathological results, improve the diagnostic accuracy of pancreatic cancer, and ultimately improve the prognosis of patients.

By introducing artificial intelligence technology and integrating NLP capabilities into the information system, relevant key information is being extracted from the report to help doctors make accurate diagnosis and conduct clinical research. This scale of a workload on a wide diversity of data is very compute intensive and time consuming.

The NLP platform has collected relevant digital medical record data for 5 years and has completed relevant feature extraction and analysis based on artificial intelligence technology without using GPUs, proving the feasibility and advantage of the overall technical solution on Intel CPUs.

By integrating Winning Health's NLP solution into the hospital IT system, disconnected patient data can be intelligently combined into a more holistic library of information. When optimized for Intel Xeon 8480+ processors using Intel optimizations, Intel AMX, and BF16, the solution delivers improved performance with as much as 6.30X more named entity recognition inference throughput compared to the previous generation Intel Xeon 8380 processor. Accelerating inference can help bring insight to clinicians and researchers faster from a wide variety of data across multiple clinical departments, helping deliver better patient outcomes.

For more information about Winning Health, visit **www.winning.com.cn**

Learn more about the Intel AI Builders program at **builders.intel.com/ai**

intel AI Builders

Winning Health Winning Health offers core IT infrastructure and capabilities to enable digital transformation for the healthcare industry in China. The Winning Health IT infrastructure integrates medical data, technical capabilities and operating processes into a unified system platform that is powered by data and focused on user experiences.

[1] **Baseline/Unoptimized PyTorch on Intel® Xeon® 8480+ Processor (FP32):** Test by Intel as of 10/17/2022. 1-node, 2x Intel® Xeon® Platinum 8480+ CPU @ 2.0GHz Processor, 112 cores, HT On, Turbo On, Total Memory 512 GB (16 slots/ 32 GB/ 4800 MHz [run @ 4800 MHz]), BIOS: 00.01.21, Ucode: 0x2b000041, Ubuntu 22.04.1 LTS, 5.15.0-48-generic, gcc 11.2.0, BERT for NER task inference workload, Framework: Pytorch 1.12.1, Topologies: Bert-Base-Chinese, Dataset: 612 Chinese medical reports in JSON format, Datatype: FP32

Intel® Optimization for PyTorch* on Intel® Xeon® 8480+ Processor (FP32): Test by Intel as of 10/17/2022. 1-node, 2x Intel® Xeon® Platinum 8480+ CPU @ 2.0GHz Processor, 112 cores, HT On, Turbo On, Total Memory 512 GB (16 slots/ 32 GB/ 4800 MHz [run @ 4800 MHz]), BIOS: 00.01.21, Ucode: 0x2b000041, Ubuntu 22.04.1 LTS, 5.15.0-48-generic, gcc 11.2.0, BERT for NER task inference workload, Framework: Pytorch 1.12.1 + Intel Extension for Pytorch 1.12.3, Intel OpenMP, Tcmalloc 2.10, OMP_NUM_THREADS=56, KMP AFFINITY=granularity=fine,compact,1,0, KMP_BLOCKTIME=1, Topologies: Bert-Base-Chinese, Dataset: 612 Chinese medical reports in JSON format, Datatype: FP32

Intel® Optimization for PyTorch* on Intel® Xeon® 8480+ Processor (BF16): Test by Intel as of 10/17/2022. 1-node, 2x Intel® Xeon® Platinum 8480+ CPU @ 2.0GHz Processor, 112 cores, HT On, Turbo On, Total Memory 512 GB (16 slots/ 32 GB/ 4800 MHz [run @ 4800 MHz]), BIOS: 00.01.21, Ucode: 0x2b000041, Ubuntu 22.04.1 LTS, 5.15.0-48-generic, gcc 11.2.0, BERT for NER task inference workload, Framework: Pytorch 1.12.1 + Intel Extension for Pytorch 1.12.3, Intel OpenMP, Tcmalloc 2.10, OMP_NUM_THREADS=56, KMP AFFINITY=granularity=fine,compact,1,0, KMP_BLOCKTIME=1, Topologies: Bert-Base-Chinese, Dataset: 612 Chinese medical reports in JSON format, Datatype: BF16

Baseline/Unoptimized PyTorch on Intel® Xeon® 8380 Processor (FP32): Test by Intel as of 11/8/2022. 1-node, 2x Intel® Xeon® Platinum 8380 CPU @ 2.30GHz Processor, 80 cores, HT On, Turbo On, Total Memory 512 GB (16 slots/ 32 GB/ 3200 MHz [run @ 3200 MHz]), BIOS: SE5C6200.86B.0022.D64.2105220049, Ucode: 0xd000375, Ubuntu 20.04.5 LTS, 5.4.0-131-generic, gcc 9.4.0, BERT for NER task inference workload, Framework: Pytorch 1.12.1, Topologies: Bert-Base-Chinese, Dataset: 612 Chinese medical reports in JSON format, Datatype: FP32

Intel® Optimization for PyTorch* on Intel® Xeon® 8380 Processor (FP32): Test by Intel as of 11/8/2022. 1-node, 2x Intel® Xeon® Platinum 8380 CPU @ 2.30GHz Processor, 80 cores, HT On, Turbo On, Total Memory 512 GB (16 slots/ 32 GB/ 3200 MHz [run @ 3200 MHz]), BIOS: SE5C6200.86B.0022.D64.2105220049, Ucode: 0xd000375, Ubuntu 20.04.5 LTS, 5.4.0-131-generic, gcc 9.4.0, BERT for NER task inference workload, Framework: Pytorch 1.12.1 + Intel Extension for Pytorch 1.12.3, Intel OpenMP, Tcmalloc 2.10, OMP_NUM_THREADS=40, KMP AFFINITY=granularity=fine,compact,1,0, KMP_BLOCKTIME=1, Topologies: Bert-Base-Chinese, Dataset: 612 Chinese medical reports in JSON format, Datatype: FP32

Performance varies by use, configuration, and other factors. Learn more on the Performance Index site.

Performance results are based on testing as of dates shown in configurations and may not reflect all publicly available updates. See configuration disclosure for details.

No product or component can be absolutely secure.

Your costs and results may vary.

Intel technologies may require enabled hardware, software, or service activation.

Intel does not control or audit third-party data. You should consult other sources to evaluate accuracy.

0123/AU/HBD/PDF ♻PleaseRecycle 354287-001US

intel