Case Study

Intel® Xeon® Processors

Technology Innovation Institute Democratizes GenAI in Public, Private, and Academic Settings Through its Open-Source Falcon LLMs

TII and Intel collaborate to optimize Falcon 2 11B LLMs for Amazon instances featuring 4th Gen Intel® Xeon® processors, making GenAI deployment faster and more affordable.

Solution Ingredients

- Amazon EC2 C7i
- Custom 4th Gen Intel® Xeon® processors
- Intel® Advanced Matrix Extensions (Intel AMX)
- Intel® Distribution of OpenVINO™ Toolkit

Executive Summary

The Technology Innovation Institute (TII) is a leading global scientific research center and a pillar of Abu Dhabi's Advanced Technology Research Council (ATRC). The organization seeks to help democratize AI through its open-source Falcon large language models (LLMs). TII also offers the enormous REFINEDWEB dataset for fast and customizable training, which is available through Hugging Face. Falcon 2 11B, the first LLM in the Falcon 2 family, is an excellent choice for implementing natural language queries in GenAI use cases. To maximize Falcon 2 11B inference performance, TII chose Amazon EC2 C7i instances featuring custom, cost-effective, and readily available 4th Gen Intel® Xeon® processors that helped speed and ease deployment. The Intel® Distribution of OpenVINO™ Toolkit assisted with additional performance improvements. With optimized instances, LLM inference benchmark results demonstrate significantly improved throughput and reduced latency for INT8 and INT4 quantization.

Challenge

TII leads the Falcon Foundation, which brings together developers, academic experts, and industry leaders to accelerate the democratization of GenAI. The organization wanted to broaden the ecosystem of instances optimized for use with Falcon. Performant and readily available CPU-based instances offered an

Members of the Falcon Foundation led by TII, including developers, academic experts, and industry leaders, collaborate to accelerate the democratization of GenAI.

outstanding option to help reduce the cost of LLM training and make GenAI more accessible to organizations of all sizes. TII also wanted to enhance its open-access large language models to add more user value. TII's Falcon 2 series needed capabilities like multimodal, multilingual, and vision-to-language support. The Falcon 2 series is a unique foundation AI model with vision-to-language capabilities.

Solution

Working together, TII, Amazon, and Intel sought ways to optimize the Falcon solution using Amazon EC2 C7i instances with 4th Generation Intel Xeon processors. The CPUs offered several built-in AI acceleration features, like Intel® Advanced Matrix Extensions (Intel® AMX), which provided a significant performance advantage for LLM training and inference compared with previous CPU generations.

Intel Distribution of OpenVINO Toolkit can optimize AI models deployed on Amazon instances employing Intel's CPU architecture. Intel OpenVINO INT8 and INT4 quantization and weight compression improved Falcon's performance. OpenVINO also reduced model size and computational demand, making 11B usage cost-effective when run on CPU-based instances.

Falcon 2 11B trained on a 5.5 trillion token dataset consisting of web data from RefinedWeb with 11 billion parameters. The LLM offers multilingual support for English, French, Spanish, German, and Portuguese GenAI scenarios. Plus, Falcon 2 11B is the first LLM to offer vision-to-language capabilities for advanced use cases.

> "The deployment of Falcon LLM on AWS c7i instances, enhanced by OpenVINO for inference, marks a significant milestone in our AI journey. We are excited about the future possibilities this technology unlocks, from more sophisticated language models to real-time AI applications that can transform industries."
>
> *— Hakim Hacid, Chief AI Researcher, TII*

Results

Now available to anyone, TII's open Falcon models are among HuggingFace's most downloaded LLMs. Users can derive significant value from Falcon's massive data sets for training and its new features that enable advanced GenAI usage scenarios.

Amazon EC2 C7i instances can now efficiently run Falcon 2 11B on Intel® CPU architectures, providing a performant and cost-effective LLM solution. By embracing Intel Xeon processors with Intel AMX, the open-source developer community and other Falcon users can run more sophisticated AI applications cost-effectively. Plus Intel Xeon processor-based instances remain widely accessible for faster and easier deployment.

Quantization	Batch	Input Prompt Tokens	Output Tokens	1st Token Latency ms/token	2nd Token Latency ms/token	Throughput tokens/s
INT8	1	32	128	148.07	82.51	12.12
INT8	1	64	128	189.98	79.74	12.54
INT8	1	128	128	283.50	80.26	12.46
INT8	1	512	128	1037.25	82.03	12.19
INT8	1	1024	128	1961.91	84.76	11.80
INT8	1	2048	128	4068.90	90.40	11.06

Use Case 1: Quantization technique using INT8 Falcon11B with Intel OpenVINO on Amazon EC2 C7i instances[1]

Quantization	Batch	Input Prompt Tokens	Output Tokens	1st Token Latency ms/token	2nd Token Latency ms/token	Throughput tokens/s
INT4	1	32	128	142.00	59.06	16.93
INT4	1	64	128	195.73	82.58	12.11
INT4	1	128	128	274.67	80.40	12.44
INT4	1	512	128	991.34	82.22	12.16
INT4	1	1024	128	1922.87	85.18	11.74
INT4	1	2048	128	4079.58	90.11	11.10

Use Case 2: Quantization technique using INT4 Falcon11B with Intel OpenVINO on Amazon EC2 C7i instances[1]

Benchmark summary:

The LLM inference benchmark results showcase the impressive performance of INT8 and INT4 quantization using Intel OpenVINO. The results demonstrate significant improvements in latency and throughput, which are essential for real-time applications like ChatQnA RAG and Code Generation.

For more information

- Learn more about TII Falcon.
- Explore Intel Xeon Processors.
- Check out Amazon EC2 C7i instances.

"By working together, TII, Amazon, and Intel sought ways to optimize the Falcon 11B on Amazon EC2 C7i instances with Intel's 4th Generation Xeon processors. The CPUs offer several built-in AI acceleration features, like Intel AMX, which provides a significant performance advantage for LLM inference compared with previous CPU generations. By running the Falcon 11B which was built exclusively on Amazon SageMaker, now running on Amazon's EC2 c7i instances—we have unlocked new possibilities for AI accessibility and democratizing AI further, allowing start-ups and enterprises alike to build innovative ML and GenAI applications addressing a wide range of customer demands."

—Dr. Shivagami Gugan, Chief Technologist, AWS

intel + aws

[1] Testing done by the Technology Innovation Institute on Amazon EC2 C7i instances.

102024/RJMJ/JS/PDF PleaseRecycle 362530-001US