White Paper

Healthcare
Biopharmaceutical Research

Accelerating Natural Language Processing Inference Models using Processor Optimized Capabilities

AbbVie Uses Intel® Xeon® Processors and the Intel® Distribution of OpenVINO™ Toolkit to Accelerate Natural Language Processing Models for Biopharmaceutical Research

Authors

G Anthony Reina
Chief AI Architect for Health & Life Sciences - Intel Corporation

Mehmed Sariyildiz
Data Scientist for AI and NLP - AbbVie

Brian Martin
Head of AI in R&D Information Research – AbbVie

Andrew Lamkin
Platform Solutions Architect - Intel Corporation

Jason Lee
Software Engineer - Intel Corporation

Table of Contents

Executive Summary

AbbVie is a research-based biopharmaceutical company that serves more than 30 million patients in 175 countries. With its global scale, AbbVie partnered with Intel to optimize processes for its more than 47,000 employees. This whitepaper highlights two use cases that are important to AbbVie's research. The first is Abbelfish Machine Translation, AbbVie's language translation service based on the Transformer NLP model, that leverages second-generation Intel® Xeon® Scalable processors and the Intel® Optimization for TensorFlow with Intel oneAPI Deep Neural Network Library (oneDNN). AbbVie was able to achieve a 1.9x improvement in throughput for Abbelfish language translation using Intel Optimization for TensorFlow 1.15 with oneAPI Deep Neural Network Library when compared to TensorFlow 1.15 without oneDNN.[1] The second use case is AbbVie Search, which is a BERT-based NLP model. AbbVie Search scans research documents based on scientific questions and returns relevant results that enable the discovery of new treatments for patients pharmaceuticals and manufacturing methods. Using the Intel Distribution of OpenVINO toolkit, AbbVie Search was accelerated by 5.3x over unoptimized TensorFlow 1.15 on the same second-generation Intel Xeon processor hardware.[1] AbbVie's NLP AI deployments demonstrate how CPUs can be highly effective for edge AI inference in a large organization without the need for additional hardware acceleration.

Enabling Data-Driven Research

AbbVie is a global research-based biopharmaceutical company that serves more than 30 million patients in 175 countries (AbbVie, 2020). Biopharmaceutical research generates a great number of scientific articles and clinical trial reports—in fact, AbbVie researchers alone have produced more than 1,100.

To keep up with the scale of the data, AbbVie utilizes the latest AbbVie takes advantage of the latest NLP algorithms to allow researchers to quickly return relevant data for its biomedical research and development. Given the recent breakthroughs in deep-learning-based Transformer NLP models, such as BERT, the Bidirectional Encoder Representations from Transformers (Devlin et al., 2018), AbbVie partnered with Intel to optimize Transformer NLP models in their deployments across the company.

The Challenge

With large amounts of biopharmaceutical data being produced in different languages worldwide, it is crucial to ensure that the right data is quickly accessible at the right time. AbbVie found that commercially available language translation services did not always provide the most accurate translations for its highly domain-specific biomedical text. To solve this problem, AbbVie decided to create its own translation and search tools.

[1]See backup for configuration details. For more complete information about performance and benchmark results, visit www.intel.com/benchmarks.

The Solution

Abbelfish Machine Translation

Abbelfish Machine Translation is a language translation service that AbbVie originally created for its 2,600 researchers in Germany. AbbVie has customized their model to provide better translation for biomedical terminology that might not be found in standard English/German translation models.

Over one million texts are currently translated each year with the Abbelfish service using up to 10 concurrent translations per minute. The service is hosted on Intel Xeon processor-based servers. AbbVie parallelized the inference requests by serving a copy of the model on every core. They divide the input text into 300-word sections chunks and serve each section chunk to a different CPU core separately. This allows them to stream the translation service with low latency. In addition to input batching, the Abbelfish model is trained on paragraphs instead of sentences, which allows the model to generate more accurate translations for the overall document.

Abbelfish will soon expand its translation capabilities to include Spanish, Italian, French, Portuguese, Russian, Chinese, and Japanese languages. For this updated model, a multidirectional, multilingual, Transformer-based topology has been created to work with AbbVie's scientific language translations. The model includes 24 layers and over 500 million parameters, which took over four months to train. Abbelfish's custom topology provides significantly more accurate translations than commercially available translation models (BLEU 38.41 versus 34.92 for Japanese ↔ English and BLEU 41.36 versus 36.91 for English ↔ German).

AbbVie Search

AbbVie Search is a question/answer-based search tool for biomedical research. It is based on the BioBERT transformer model (Lee et al., 2020). BioBERT starts with a base BERT model that has been fine-tuned on the SQuAD 1.1 dataset (Figure 1). It further refines the question/answer model by fine-tuning on AbbVie's internal datasets and the BioASQ 6b and 7b datasets, which are an open-source collection of semantic indexing and question/answer texts based on biomedical research articles (Tsatsaronis et al., 2015).

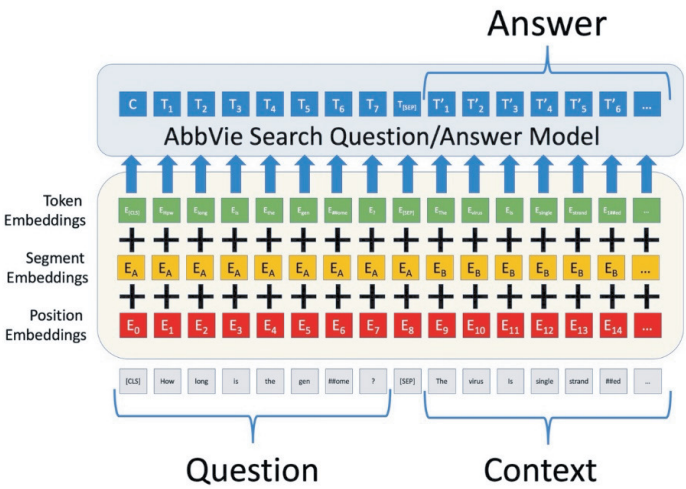

Figure 1. AbbVie Search. A transformer-based NLP question/answer model (based on BioBERT) that was trained on a biomedical corpus. This allows AbbVie to develop models that are specific to biopharmaceutical research. The question and context sentences (for example, scientific article) are transformed into an embedding space and fed into the AbbVie Search model. The model outputs the position in the context where the answer can be found.

Figure 2 shows an example of how the question/answer model works. When AbbVie researchers submit a question to AbbVie Search, both this question and a collection of research articles/clinical notes (also known as context) are fed into the transformer-based model. The model predicts the most likely character position in the context where the answer can be found (highlighted section of the context). One benefit of this approach is that the model is only able to create answers based on the context. Therefore, the answers AbbVie Search returns are immediately verifiable by a quick scan of the original scientific article.

Context

Symptom severity scores were quantified using the following five measures: (i) individual symptom score for 20 symptoms, (ii) the upper respiratory symptom score, calculated as the sum of severity scores for earache, runny nose, sore throat, and sneezing, (iii) the lower respiratory symptom score, calculated as the sum of severity scores for cough, difficulty breathing, hoarseness, and chest discomfort, (iv) the gastrointestinal symptom score, calculated as the sum of severity scores for diarrhea, vomiting, anorexia, nausea, and (Table 1) . There was season-to-season variability in the leading causes of ... The findings of our study, conducted over a five-year period at five geographically dispersed sites in the USA, demonstrate that human coronavirus (HCoV) is an important cause of influenza-like illness (ILI) ranged from 4% to 22%. [8] [9] [10] [11] [14] Additionally, we found HCoV-OC43 to be the most common species among adults, as has been reported elsewhere. [8], [9], [11], [12], [14] HCoV-OC43 and HCoV-229E were the most common strains in alternate seasons, reflecting a season-to-season variability of HCoV strain circulation that has been reported in other multiyear studies.

Question

What is the most common species of Human Coronavirus among adults?

Answer

HCoV-OC43

Figure 2. AbbVie Search Question and Answer Example. The user inputs a question along with a collection of contexts of scientific articles/clinical notes. The model highlights where in the context the answer can be found.

[1]See backup for configuration details. For more complete information about performance and benchmark results, visit www.intel.com/benchmarks.

Results

Abbelfish Machine Translation

The 2020 version of the Abbelfish model is an 8-language translation Transformer model with 24 layers and 500 million parameters. A large portion of the Abbelfish workload is done as a batch process. To meet the needs of its researchers, AbbVie set a target of translating one sentence every two seconds. It is important to note that although faster inference times are always desirable, AbbVie recognized that they don't add significant value to the product given the added expense of dedicated hardware accelerators, such as GPUs. Their strategy allows AbbVie to offload its translation workloads onto the Intel Xeon processor-based servers to free up other hardware resources.

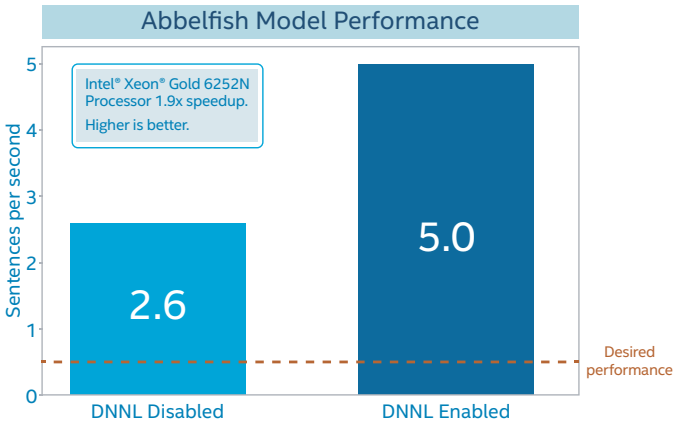

Figure 3. AbbVie's Abbelfish translated over five sentences per second using Intel Optimization for TensorFlow with oneAPI Deep Neural Network Library (oneDNN).[1]

The Abbelfish translation model was benchmarked on a 2nd gen Intel Xeon Scalable processor-based server (Intel Xeon Gold 6252N Processor, 2.30 GHz, 2 sockets, 24 cores per socket). The model and data pipeline included the Tensor2Tensor library, which only works with version 1 of TensorFlow. In addition to using multiple CPU threads during inference, AbbVie was also able to create parallel inference streams by pinning each stream to half of the CPU socket using the numactl command.

Figure 3 compares the model's performance between un-optimized TensorFlow 1.15 (without) and the Intel Optimization for TensorFlow (with). In both cases, there were four parallel inference streams with a batch size of 256 sentences. The effective throughput of all four parallel inference streams was 2.6 sentences per second without DNNL and 5.0 sentences per second with. This is a 1.9x speedup over the unoptimized TensorFlow 1.15 (without DNNL) and 10x the performance requirements needed by AbbVie.[1] Note that even without optimizations, the raw CPU performance was sufficient to meet the AbbVie requirement. More importantly, the performance of just a single stream using was over 2x the AbbVie requirement, and only used 12 of the 48 cores on the server. This means that AbbVie could easily process Abbelfish translation workloads and still have 36 CPU cores (75 percent of the remaining compute) for other simultaneous workloads (Figure 4). This greatly simplifies AbbVie's deployment because they can seamlessly adjust the Intel Xeon processor-based server to their desired balance of AI versus non-AI workload performance.

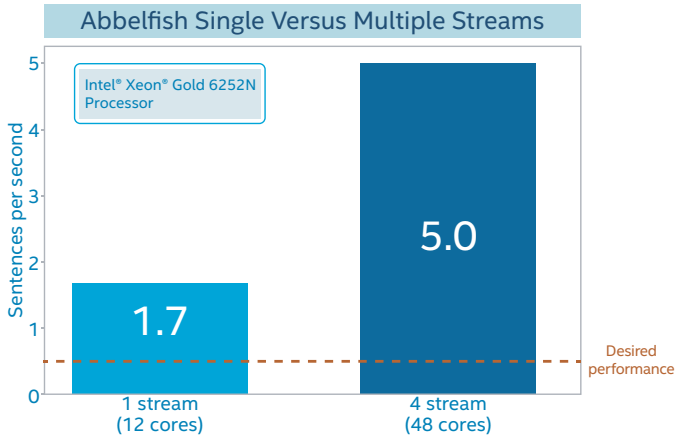

Figure 4. Abbelfish performance using Intel Optimization for TensorFlow 1.15 with oneDNN.[1]

AbbVie Search

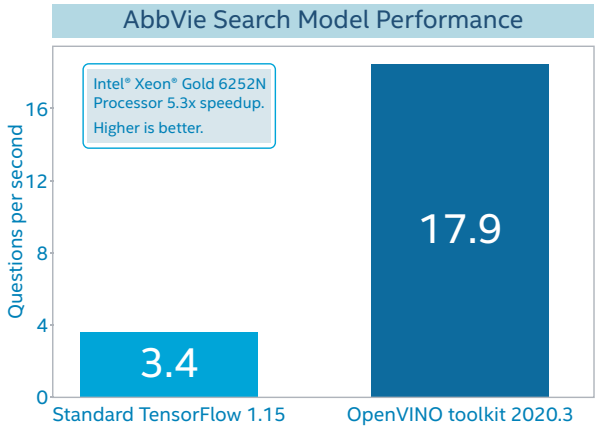

Figure 5. Comparison of AbbVie Search inference between unoptimized TensorFlow 1.15 (oneDNN disabled) and OpenVINO toolkit 2020.3.[1]

Figure 5 shows the performance of the AbbVie Search question/answer model. The same Intel Xeon processor -based server was used for this benchmark. The Intel Distribution of OpenVINO toolkit provided a 5.3x speedup in the number of questions answered per second compared with the unoptimized version of TensorFlow 1.15 (without oneDNN).[1] This allows AbbVie researchers to answer more than 17 questions per second from a given scientific article or clinical report. In the future, AbbVie Search can be scaled across the company by leveraging the OpenVINO Model Server, a gRPC-based microservice that is compatible with existing TensorFlow Serving applications.

[1]See backup for configuration details. For more complete information about performance and benchmark results, visit www.intel.com/benchmarks.

Conclusion

AbbVie's use of the oneAPI Deep Neural Network Library (oneDNN) and the Intel Distribution of OpenVINO toolkit have allowed them to speed up their NLP services while leaving CPU resources for their additional non-AI workloads. This greatly simplifies AbbVie's deployment and reduces the need for additional hardware accelerators. Abbelfish and AbbVie Search leverage these AI-optimized libraries to scale NLP inference across their worldwide research using Intel Xeon processor-based servers.

For more information about Intel Xeon Scalable Processors, visit: intel.com/xeonscalable

For more information about the Intel Distribution of OpenVINO toolkit, visit: intel.com/openvino.

Download the Intel Distribution of OpenVINO Toolkit

References

- AbbVie, "A Closer Look", https://www.abbvie.com/content/dam/abbvie-dotcom/uploads/PDFs/AbbVie_Corporate_Presentation.pdf, accessed online 20 April 2020
- Devlin J, Change MW, Lee K, Toutanova K. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding", https://arxiv.org/pdf/1810.04805.pdf
- Lee J, Yoon W, Kim S, Kim D, Kim S, So, Jaewoo Kang. "BioBERT: a pre-trained biomedical language representation model for biomedical text mining", Bioinformatics, Volume 36, Issue 4, 15 February 2020, pages 1234–1240
- Tsatsaronis G, Balikas G, Malakasiotis P, Partalas I, Zschunke M, Alvers M, Weissenborn D, Krithara A, Petridis S, Polychronopoulos D, Almirantis Y, Pavlopoulos J, Baskiotis N, Gallinari P, Artieres T, Nogonga Ngomo AC, Heino N, Gaussier E, Barrio-Alvers L, Schroeder M, Androutsoploulos I, Paliouras G. "An overview of the BIOASQ large-scale biomedical semantic indexing and question answering competition". BMC Bioinformatics, 16, 138 (2015). https://doi.org/10.1186/s12859-015-0564-6

Backup: System Configuration Details

All tests were performed by Intel in June 2020.

Intel® Xeon® Gold 6252N Processor @ 2.30 GHz, two sockets, 24 cores per socket, 394 GB DDR4 RAM, Intel Hyper-Threading Technology enabled, Intel® Turbo Boost enabled, NUMA enabled, BIOS version 4.1.12, Microcode 0x500002c, Ubuntu 18.04.4 LTS, Linux Kernel 4.15.0-101-generic, Spectre/Meltdown mitigated, Software: Intel® Optimization for TensorFlow* version 1.15 with DNNL and Intel® Distribution of OpenVINO™ toolkit.