White Paper

How to Accelerate a Business-to-Business (B2B) Company Search Engines using Intel Optimizations for Deep Learning and Improving Total Cost of Ownership

delphai uses Intel® Neural Compressor and Intel® Extensions for PyTorch together with Intel® AI Engine on Intel® Xeon® Processors to accelerate Natural Language Processing Models for B2B company search engines.

Authors

Albertano Caruso
Field Application Engineer, Intel

Alexander Chaiko
AI Software Development Engineer, Intel

Malek Naski
MLOps Engineer, delphai

Table of contents

Executive Summary

delphai is a Berlin-based AI Natural Language Processing (NLP) software company that provides a company search engine that automatically collects, analyzes, and structures data from companies driving technological and innovative change across industries and provides actionable insights.[1]

delphai aggregates public data in every language from more than 12 million company websites and more than 15,000 global news sources, financial statements, conference websites, investor portfolios, job posts, and patent filings. delphai goes beyond conventional syntactic processing; it also considers the context in which text data is used to interpret it. Various NLP and Natural Language Understanding (NLU) methods are used for data classification, extraction, categorization, linking, clustering, and analysis.

delphai provides an intuitive graphical user interface that clearly conveys content and invites exploration by using semantic search functions and dynamic visualizations that guarantees a high level of information retrieval. Users are able to reduce their company search time to seconds, all while identifying more relevant companies than are found with other incumbent company search platforms.[1]

delphai helps customers like Siemens, ABB, and ING to save time and increase growth opportunities by delivering structured B2B company intelligence. This collaboration with Intel pushes this further by optimizing delphai's technical processes to provide customers with fresh structured updates. The collaboration brings together delphai's innovative solution, built upon sophisticated data and machine learning (ML) pipelines, and Intel's powerful software and hardware optimizations for ML inference.

This whitepaper considers two representative use cases that are crucial to the delphai analytics platform and search engine: multilingual translation and sequence classification.

delphai has an extensive knowledge base of companies worldwide and provides its customers with structured firmographic data about them, hence the sequence-classification task. For instance, this use case in particular classifies companies according to their description by assigning them industry labels.

Using Intel® Neural Compressor and Intel® Extension for PyTorch*, delphai managed to get significant acceleration of critical AI components and improve cost efficiency by moving from expensive GPU Virtual Machines (VM) to cheaper CPU instances in Microsoft Azure* Cloud Service Provider.[2]

Figure 1. Delpai Dashboard Capabilities

The Challenge

As delphai relies on fresh data that are ingested and then processed daily, speed is crucial in order to have the latest information about any company reflected in delphai as soon as possible. Several steps of processing, ML-powered predictions, and transformations are applied to this unstructured data to turn it into useful structured data.

To achieve high-quality reliable predictions, delphai relies mostly on state-of-the-art deep learning models. Many of these are challenging to use in production due to their high inference time. Therefore, it's an unceasing quest to make these models' inference faster.

All delphai workloads and resources are hosted on Microsoft Azure, including ML services which are running on clusters managed by Azure Kubernetes Service (AKS).

Translation inference at delphai was done solely on GPUs as they are typically much faster than CPUs. Yet, it was still not fast enough to keep up with the amount of text that needs to be translated every day. One obvious solution was to have even more replicas of this translation service. However, this comes with great costs as GPU VMs are quite expensive.

The Solution

This led to exploring post-training optimization techniques that would make the models' inference faster without re-training them, since training itself is expensive given the number and complexity of models delphai relies on.

The main tool that could show impressive results is the Intel Neural Compressor, which delphai ended up using in production. Moreover, this also led to replacing Microsoft Azure VMs hosting the ML services by newer ones, specifically ones powered by 3rd Generation Intel® Xeon® Scalable processors.

Part of Intel® AI Analytics Toolkit, Intel Neural Compressor is an open-source python library for model compression that reduces model size and increases the speed of deep learning inference for deployment on CPUs or GPUs by lowering precision of compute operations to INT8.[3]

Additionally, the compressor provides unified interfaces across multiple deep-learning frameworks such as TensorFlow*, PyTorch, MXNet*, and more, in addition to ONNX (Open Neural Network eXchange) runtime for greater interoperability across frameworks. It also supports automatic accuracy-driven tuning strategies to help the user quickly find the best-quantized model.

Intel Neural Compressor leverages Intel® AI Engine acceleration powered by the Intel® AVX-512 instruction set that was first introduced with 1st Gen Intel Xeon Scalable processor.[4] Intel AVX-512 offers vector instructions with two-times wider SIMD registers compared to Intel® AVX2, so that a register fits sixteen 32-bit single-precision floating-point numbers or sixty-four 8-bit integers. Additional improvement of Intel AI Engine have been introduced by VNNI extension which accelerates INT8 convolution operation, adding support for brain floating-point format (BF16) as well as general architectural enhancements in the 2nd and 3rd Gen Intel Xeon Scalable processors.

Another tool used was Intel Extensions for PyTorch. This is because vanilla PyTorch uses Intel® Deep Neural Network Library optimizations by default, while the Intel extension adds more capabilities on top of it.[5] The sum of Intel's neural compressor and PyTorch extension were instrumental in delphai's achieved improvements.

The delphai multilingual translation model that was optimized with Intel Neural Compressor is based on the M2M-100 multilingual encoder-decoder (seq-to-seq) model and uses PyTorch as Framework. The sequence classification one is based on a fine-tuned RoBERTa model that uses ONNX as Framework. The compressor was used for both the translation and sequence classification tasks. For both, post-training dynamic quantization was used while controlling the accuracy drop as these two models are critical to the product.

For GPU workloads (mainly multilingual translation use cases), Intel's neural compressor significantly reduces the number of delphai's GPU-powered VMs overall in favor of more CPU-only VMs[6] that have a better performance-per-cost ratio for delphai's requirements.[7] Figure 2 shows the benefit delphai achieved. It makes the GPU node pool in the AKS cluster get much smaller (After) compared to the original configuration (Before) and to other CPU node pools as well.[6] This translates to improvement of the total cost of ownership (TCO), since the better performance per cost ratio of CPU in comparison of GPU.[7]

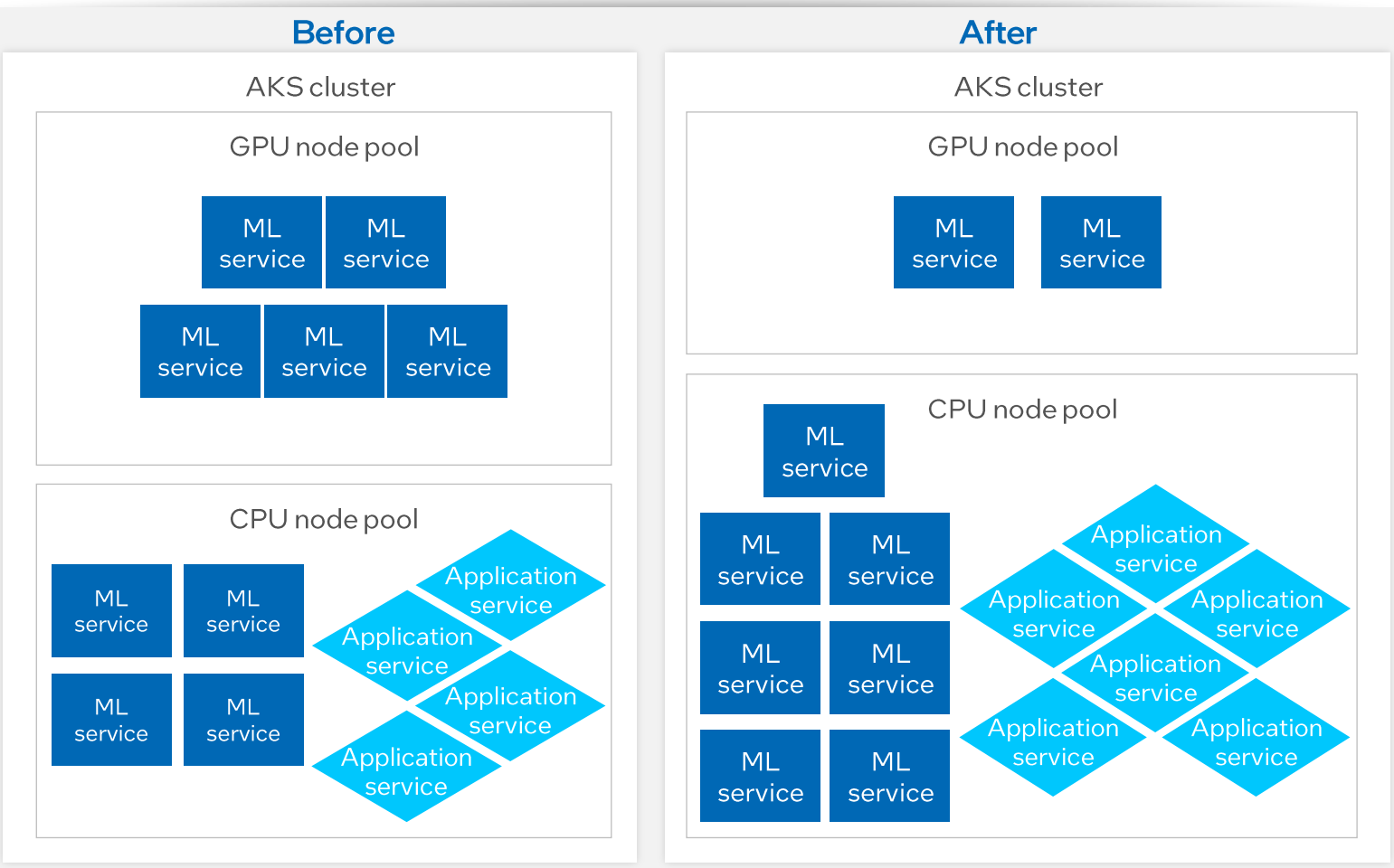


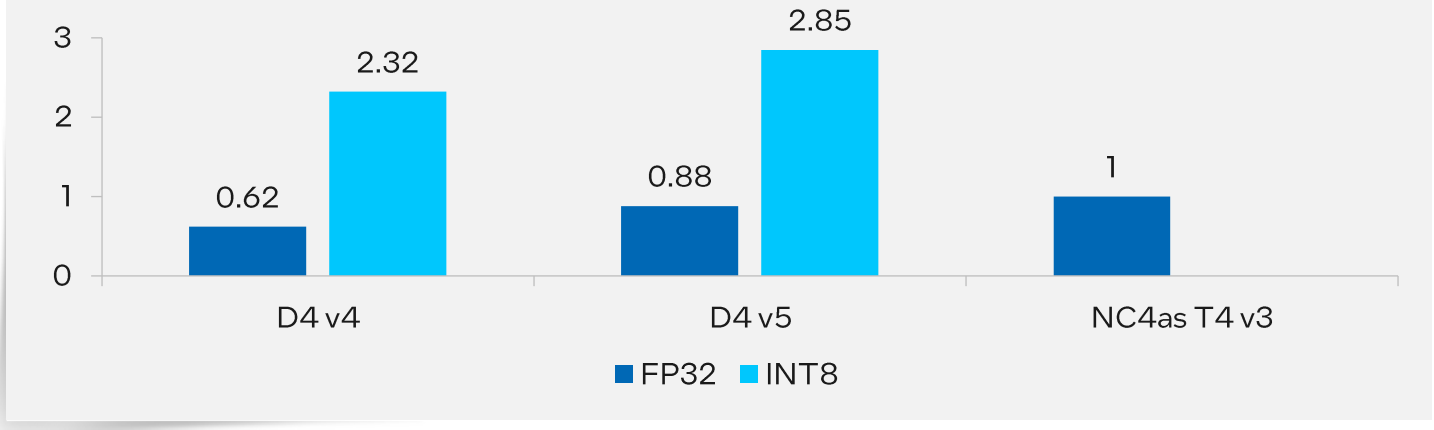
Figure 2. Intel® Neural Compressor reduces the GPU node pool for delphai AKS Cluster[2]

Results

Multilingual translation

For multilingual translation, the overall performance/$ benefit makes a huge difference using Microsoft Azure Dv5 instances.[8] This is due to spot Dv5 instances that are over 6x cheaper than the spot GPU ones. Using 3rd Gen Intel Xeon Scalable processors gives results that are comparable to GPU's inference speed while being much cheaper.[8] This makes it possible to move the translation workloads from running on GPUs to CPUs. In fact, translation services at delphai are currently using a hybrid setup of both GPUs and 3rd Gen Intel Xeon Scalable processors for a better performance. Figure 3 shows the performance per € on D4 v4 and D4 v5, normalized to the GPU NC4as T4 v3 performance. With the Translation Model in FP32 precision the Performance/€ on CPU VM is comparable with the GPU one.[8] Using the Translation Model compressed with Intel Neural Compressor to INT8 precision, the Performance/€ on CPU VM is up to 3 times more efficient compared to GPU.[8] There was not any accuracy drop after using the compressed model resulting in no impact on Translation service.[8]

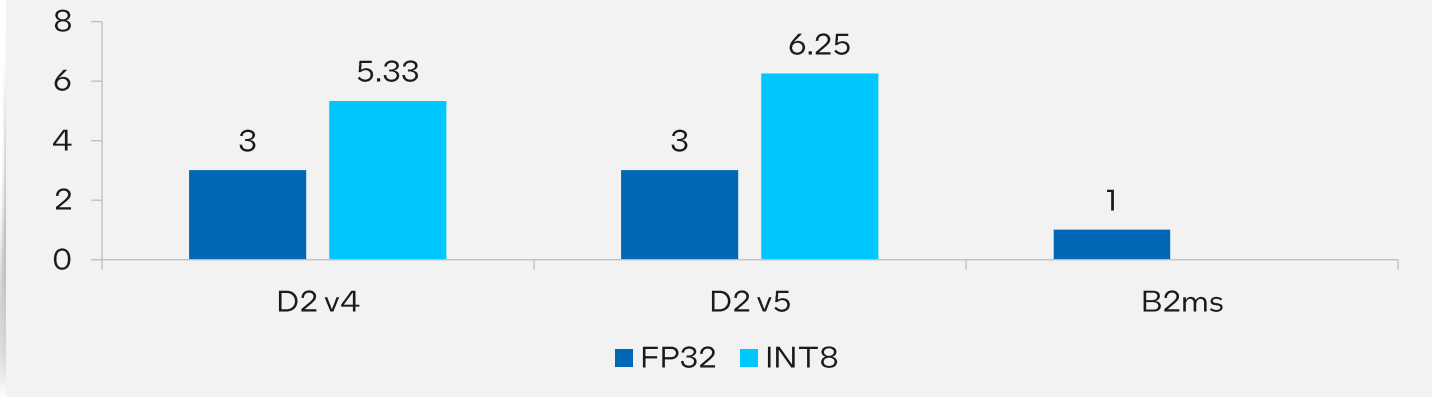
Figure 3. Performance/€ for multilingual translation (baseline is the T4 GPU's performance, higher is better)[8]



Sequence classification

For sequence classification, delphai uses only CPU-based VMs. B2ms VMs were previously used to serve this model and using the latest D2 v4 and D2 v5 VMs provides significant performance improvements.[9] Figure 4 shows the performance per € on D2 v4 and D2 v5, normalized to the CPU B2ms performance. With the Sequence Classification Model in FP32 precision the Performance/€ on the latest VMs is up to 3 times better than B2ms.[10] Using the Sequence Classification Model compressed with Intel Neural Compressor to INT8 precision the Performance/€ on D2 v4 and D2 v5 VM is up to 6 times more efficient compared to B2ms.[10] The accuracy drop here was only 1.4%, which is quite acceptable considering the performance gain.[11]

Figure 4. Performance/€ for sequence classification (baseline is the B2ms performance, higher is better)[10]



Conclusion

This partnership brought Intel and delphai together to combine delphai's AI-powered innovation and Intel's sophisticated AI optimizations to accelerate ML-based processes.[8] Structuring the unstructured global economy can be time consuming and expensive. By using Intel Neural Compressor and Intel Extension for PyTorch on 3rd Gen Intel Xeon Scalable processors with AI Engine, delphai was able to accelerate its NLP models' inference for B2B company search engine.

Learn More

- The delphai Website
- Intel & delphai: Structuring the business world so you don't have to Blog
- Intel® Xeon® Scalable Processors Product Page
- Intel® Neural Compressor Webpage
- Intel® Extensions for PyTorch Webpage

System Configuration Details

All machines used for benchmarking are Microsoft Azure VMs located in the West Europe region.

Instance	vCPU(s)	RAM	Compute	OS
NC4as T4 v3	4	28 GiB	NVIDIA Tesla T4[12]	Ubuntu 18.04.6 LTS
D4 v5	4	16 GiB	Intel® Xeon® Platinum 8370C (Ice Lake)[13]	Ubuntu 18.04.6 LTS
D4 v4	4	16 GiB	Intel® Xeon® Platinum 8370C (Ice Lake) or the Intel® Xeon® Platinum 8272CL (Cascade Lake)[14]	Ubuntu 18.04.6 LTS
D2 v5	2	8 GiB	Intel® Xeon® Platinum 8370C (Ice Lake)[13]	Ubuntu 18.04.6 LTS
D2 v4	2	8 GiB	Intel® Xeon® Platinum 8370C (Ice Lake) or the Intel® Xeon® Platinum 8272CL (Cascade Lake)[14]	Ubuntu 18.04.6 LTS
B2ms	2	8 GiB	Intel® Xeon® Platinum 8370C (Ice Lake), the Intel® Xeon® Platinum 8272CL (Cascade Lake), the Intel® Xeon® 8171M 2.1 GHz (Skylake), the Intel® Xeon® E5-2673 v4 2.3 GHz (Broadwell), or the Intel® Xeon® E5-2673 v3 2.4 GHz (Haswell).[15]	Ubuntu 18.04.6 LTS

Libraries' versions used:

Intel® Extension for PyTorch (IPEX): 1.12.0

Intel® Neural Compressor (INC): 1.12.0

Sources

1. https://www.delphai.com/
2. Moving from NC4as T4 v3 to D4 v5 | IPEX version 1.12.0
3. https://www.intel.com/content/www/us/en/developer/tools/oneapi/neural-compressor.html
4. https://www.intel.com/content/www/us/en/artificial-intelligence/documents/enhance-ai-workloads-built-in-accelerators-pdf.html
5. https://www.intel.com/content/www/us/en/developer/tools/frameworks/overview.html#pytorch
6. Test by Delphai as of 08/08/2022: this took effect after using Intel's optimizations in production
7. Last checked on 21/10/2022 (Azure prices are volatile) | Model: both m2m100 (translation) and fine-tuned RoBERTa (sequence classification)
8. https://www.delphai.com/blog/intel-delphai-structuring-the-business-world/
9. Test by Delphai as of 05/08/2022 | Fine-tuned RoBERTa (sequence classification)
10. Last checked on 21/10/2022 (Azure prices are volatile) | Model: fine-tuned RoBERTa (sequence classification)
11. Test by Delphai as of 05/08/2022 | Fine-tuned RoBERTa (sequence classification) | baseline is delphai's fine-tuned RoBERTa model for industry labels classification that was previously used in production. The benchmark was done on a representative delphai private dataset.
12. https://learn.microsoft.com/en-us/azure/virtual-machines/nct4-v3-series
13. https://learn.microsoft.com/en-us/azure/virtual-machines/dv5-dsv5-series
14. https://learn.microsoft.com/en-us/azure/virtual-machines/dv4-dsv4-series
15. https://learn.microsoft.com/en-us/azure/virtual-machines/sizes-b-series-burstable

Notices & Disclaimers